Australian Government
Department of Defence
Defence Science and
Technology Organisation

Selecting Measures to Evaluate Complex Sociotechnical Systems: An Empirical Comparison of a Task-based and Constraint-based Method

David J. Crone

Joint and Operations Analysis Division
Defence Science and Technology Organisation

DSTO-RR-0395

ABSTRACT

Researchers need better measures for evaluating human performance in complex socio-technical systems. A constraint-based and a task-based method were compared for their effectiveness in identifying measures that would be sensitive to a system modification. Working in an advanced simulator, aircrews conducted tactical missions with or without a modification to a specific aircraft system. Across two experiments there was no significant difference between the methods in the sensitivity or suitability of the measures that they suggested. Nonetheless, observations made during the program of research suggested that the constraint-based method for identifying measures is a viable alternative to the task-based method.

RELEASE LIMITATION

Approved for public release

Published by

Joint and Operations Analysis Division
DSTO Defence Science and Technology Organisation
Fairbairn Business Park Department of Defence
Canberra ACT 2600 Australia

Telephone: 1300 DEFENCE
Fax: (02) 6128 6332

AR-015-654
July 2013

APPROVED FOR PUBLIC RELEASE

Selecting Measures to Evaluate Complex Sociotechnical Systems: An Empirical Comparison of a Task-based and Constraint-based Method

Executive Summary

Evaluating whether a modification to a complex socio-technical system will be successful involves the important step of selecting measures that are sensitive to that modification. If the wrong measure is selected, then the success of the evaluation is jeopardised. The evaluation of systems that users have had no previous experience of ("future systems") is becoming more common and is placing pressure on existing methods for selecting measures. This thesis compares two methods for selecting measures ("measure-selection methods") that are used to evaluate complex socio-technical systems. The first method is centred on the analysis of system tasks ("task-based") and the second method is centred on the analysis of system constraints ("constraint-based").

Under a task-based method, measures (for example, "participant reaction time") are identified from the properties of a task. The task-based method is used within the Human Engineering Process (HEP) and represents the current best practice. The HEP is widely used in the evaluation of complex systems in both laboratory and in operational settings. Task-based methods, in general, have been criticised on the grounds that the measures used for the evaluations are selected using guidelines, are not theoretically grounded, and that the approach is not appropriate for future systems.

The constraint-based method is used within the context of Cognitive Work Analysis. Under a constraint-based method measures are identified from constraints acting on the "work" that is being performed. For example, landing an aircraft on a landing site may be constrained by a window of opportunity of a few minutes. A constraint-based measure is "arrive no later or earlier than 15 seconds of planned arrival time". Although other constraint-based methods for selecting measures have been used successfully in laboratory settings and are theoretically grounded, they have not been extensively tested in applied situations and have been restricted to a subset of system constraints. The analysis of constraints, in general, is believed to be uniquely suited for evaluating future systems.

Although some research has evaluated task-based and constraint-based methods for system evaluation, no research has compared them specifically on the sensitivity of the measures that they suggest for current and future systems in an operational setting. The main question asked in this thesis is: is there a difference between a constraint-based and a task-based method for evaluating complex socio-technical systems (both current and future) in operational contexts? More specifically: are the measures suggested by the two methods sensitive to a system modification for a current and future system? And are the methods suitable for use in operational, not laboratory, settings?

Three experiments (exploratory, Experiment 1, Experiment 2) were conducted in an advanced simulator that emulated operational conditions. The experiments required one aircrew team (one Pilot and one Aircraft Captain) per experiment to conduct a number of tactical missions with or without a modification to an aircraft system—specifically, a modification to a Radar Warning Receiver (RWR). The measures that each of the two methods predicted would be sensitive to the RWR modification were collected and tested.

The results of the exploratory experiment suggest that the advanced simulator met a number of important design requirements, including providing a simulation experience that emulated operational conditions.

The results of Experiment 1 suggest that there is no significant difference between the measure-selection methods in terms of the sensitivity of the measures that they suggest for a current system ($X2$ (1, $N = 67$) = 3.22, $p = 0.07$). However, it was found that some measures produced by Work Domain Analysis, a constraint-based method, were sensitive to the modification of the current RWR system. The results also suggested that no measures produced by the task-based and Control Task Analysis, a constraint-based method, were sensitive. The results of Experiment 1 also suggested that there was no significant difference between the two measure-selection methods in terms of their suitability for use in the operational setting used in the experiment ($X2$ (1, $N = 67$) = 1.36, $p = 0.24$). This was because both methods suggested measures that were affected by factors common to operational conditions.

The results of Experiment 2 suggest that there is no significant difference between the measure-selection methods in terms of the sensitivity of the measures that they suggest for a future system ($X2$ (1, $N = 67$) = 0.02, $p = 0.88$). However, it was found that the measure-selection methods produced a similarly low number of sensitive measures. This was different from Experiment 1. The results of Experiment 2 also suggested that there was no significant difference between the two measure-selection methods in terms of the suitability for use in the operational setting used in the experiment ($X2$ (1, $N = 67$) = 0.01, $p = 0.75$). Like Experiment 1 this was because both methods suggested measures that were affected by the same factors seen in experiment 1. However, unlike Experiment 1 measures produced by each method were affected differently by those conditions.

The results were discussed in the broader context of the contribution to existing theory, system evaluation, limitations of the study and opportunities for future research. It was suggested that using a theoretically grounded approach is more likely to produce sensitive measures than if guidelines are used, and that the constraint-based perspective provides a framework to categorise different measures and this was seen to be useful for system evaluation. Factors that limit the results from this program of work include, using a process of reliability assurance for the methods rather than a formal test of reliability, the strategy used for testing the methods, the number of participants used in the experiments, and that only two of the five CWA phases were incorporated into the constraint-based method.

The conclusion reached from the experiments was that there was not a statistical significant difference between the measure-selection methods in terms of sensitivity and suitability. However, observations made during the experiments suggest that the original constraint-based method developed in this program of research, with its foundation in theory, and a focus on evaluating complex systems in operational settings is a viable alternative to the task-based approach.

This page is intentionally blank

Author

David J. Crone
Joint and Operations Analysis Division

David J. Crone is a Senior Research Scientist with Joint and Operations Analysis Division (JOAD). Prior to his move to JOAD he was a Senior Human Factors Engineer with Air Division and has been Director Program Office with DCDS Stategy and Programs. David joined DSTO in 1998 after working at British Aerospace and at the Institute of Aviation Medicine in the United Kingdom. He was awarded a Bachelor of Science degree with Honours in Psychology from the University of Plymouth and a Master of Science degree in Advanced Systems Engineering from Salford University. He has recently been awarded a PhD in Psychology from the University of Queensland. His current research focusses on developing models for representing future complex socio-technical systems.

This page is intentionally blank

Contents

GLOSSARY

PREFACE

ACKNOWLEDGEMENTS

1. INTRODUCTION 1
1.1 Problem statement 1
1.2 Aim and Scope 2
1.3 Thesis overview 4

2. APPROACHES TO EVALUATING COMPLEX SOCIO-TECHNICAL SYSTEMS 5
2.1 Task-based perspective and Human Engineering Process 7
2.1.1 Task-based perspective 7
2.1.2 Human Engineering Process 9
2.2 Constraint-based perspective and CWA 13
2.2.1 Constraint-based perspective 14
2.2.2 CWA framework 14
2.3 Analytic products, complexity, and system life cycle 22
2.3.1 Task-based analytic products for this program of research 22
2.3.2 Constraint-based analytic products for this program of research 25
2.3.3 Summary 30
2.4 Methods used for selecting measures 30
2.4.1 Task-based measure-selection-method 31
2.4.2 Constraint-based measure-selection method 35
2.4.3 Testing which measure-selection method is more effective 40
2.5 Conclusion 41

3. TEST SYSTEM USED FOR THE RESEARCH 42
3.1 Test case and test environment 42
3.1.1 Black Hawk Helicopters, RWRs and Airmobile Operations 43
3.1.2 The simulation environment 46
3.1.3 The Black Hawk helicopter simulator 47
3.1.4 The simulated world 50
3.1.5 Summary 51

4. OVERALL RESEARCH AIMS, RELIABILITY, VALIDITY AND DESIGN 51
4.1 Research aims, questions and structure 52
4.1.1 Stage 1 53
4.1.2 Stage 2 53
4.1.3 Stage 3 54
4.1.4 Stage 4 54
4.2 Reliability and validity 54
4.2.1 Reliability and validity of analytic products (Stage2) 54

4.2.2 Reliability and validity of measure-selection methods (Stages 3 and 4) ... 57
4.3 Experimental design and hypotheses ... 60
4.4 Data collection methods ... 63
4.4.1 Modified Critical Decision Method ... 64
4.4.2 Direct observation method ... 67
4.4.3 Automated quantitative data collection methods ... 67
4.5 Conclusion ... 68

5. TASK-BASED AND CONSTRAINT-BASED ANALYTIC PRODUCTS ... 69
5.1 Method used to develop analytic products ... 70
5.1.1 Participants ... 70
5.1.2 Procedure ... 72
5.2 Outputs and discussion ... 74
5.2.1 Constraint-based analytic products ... 74
5.2.2 Task-based analytic products ... 93
5.3 Conclusion ... 99

6. SELECTING MEASURES FOR EVALUATING THE TEST CASE SYSTEM ... 100
6.1 Process used to develop measure-selection methods ... 101
6.2 Description of measure-selection methods ... 101
6.2.1 Constraint-based measure-selection method flowcharts ... 101
6.2.2 Task-based measure-selection method flowchart ... 110
6.2.3 Subject Matter Expert process for assessing the Task-based method ... 113
6.3 Selecting and refining sets of measures for testing ... 113
6.3.1 Measures for the RWR system ... 114
6.4 Conclusion ... 119

7. EXPERIMENT 1: COMPARING METHODS USING A CURRENT SYSTEM ... 119
7.1 Background, aims and hypotheses ... 120
7.2 Method ... 122
7.2.1 Participants ... 122
7.2.2 Apparatus and materials ... 123
7.2.3 Design ... 125
7.3 Procedure ... 128
7.3.1 Training phase ... 129
7.3.2 Experimental sessions ... 129
7.3.3 Final wrap-up ... 130
7.4 Results and discussion ... 130
7.4.1 Assessing sensitivity of measures ... 131
7.4.2 Assessing suitability of methods ... 145
7.5 Are analytic products valid? ... 150
7.6 Conclusion ... 151

8. EXPERIMENT 2: COMPARING METHODS WITH A FUTURE SYSTEM ... 153
8.1 Background, aims and hypotheses ... 154

8.2 Method 155
8.2.1 Participants 155
8.2.2 Apparatus and materials 156
8.2.3 Design 156
8.3 Procedure 157
8.4 Results and discussion 157
8.4.1 Assessing sensitivity of variables that methods suggested 157
8.4.2 Assessing suitability of methods 169
8.5 Are the analytic products valid? 174
8.6 Conclusion for Experiment 2 175
8.7 Comparing results of Experiment 1 and Experiment 2 176
8.7.1 Comparing measure-selection methods on sensitivity 176
8.7.2 Comparing measure-selection methods on suitability 179
8.8 Conclusion 181

9. GENERAL DISCUSSION AND CONCLUSION 181
9.1 Summary of research 182
9.2 Significant and original outcomes of research 186
9.2.1 Comparison of methods for system evaluation 186
9.2.2 Development of WDA-CTA measures framework 186
9.2.3 Development of constraint-based method 187
9.2.4 Contributions to system evaluation 191
9.3 Theoretical implication: Let theory guide measure selection 192
9.4 Limitations 193
9.5 Further research 195
9.6 Future vision for community of practice of using a constraint-based perspective 196
9.7 Conclusions 198

APPENDIX A: EXPLORATORY EXPERIMENT 203
Introduction 203
Method 204
Participants 204
Apparatus and materials 204
Design 204
Procedure 207
Results and discussion 209
Conclusion 213

APPENDIX B: EXAMPLE OF THE DATABASE 214

APPENDIX C: FUNCTION FLOW DIAGRAMS 216

APPENDIX D: INPUT AND OUTPUTS 219
Inputs and Outputs for the WDA method 219
Inputs and Outputs for the CTA method 221

APPENDIX E: MISSION SCENARIOS FOR EXPERIMENT 1 AND 2 223
Test results for experiment 1 225
Test results for experiment 2 225

Glossary

Abstraction-Decomposition Space (ADS) - An analytic product produced from the information gathered in a Work Domain Analysis. The ADS is a table that represents the work domain in two dimensions, abstraction and decomposition. The abstraction dimension shows the objects, functions, values and priorities and purpose of the work domain. The decomposition dimension shows the domain in terms of system, subsystem, unit and component. Means-end relationships are not shown.

Abstraction Hierarchy (AH) - An analytic product produced from the information gathered in a Work Domain Analysis. The AH is a figure that represents the objects, functions, values and priorities and purpose, of the work domain. The objects, functions, values and priorities and purpose, of the work domain are shown in terms of levels of abstraction and decomposition and their means-ends relationship. The means-end relationships are shown by lines.

Activity Analysis (AA) - An analysis of the constraints that are derived from the control tasks that have to be performed by the system. Usually used in the context of Control Task Analysis.

Analytic product - Any output or product (e.g. Abstraction Hierarchy) produced by following a prescribed data collection method or form of analysis (e.g. Work Domain Analysis), that represents data in one form or other.

Apparent validity - Refers to the extent that an analytic product identifies a measurable property (of the domain) that appears in data (discussion and/ or observation and/or transcription) from an actor.

Approach - An approach is proposition that describes the relationship between a perspective, a method (or methods) and an analytical product (or products), that are used to solve a problem.

Cognitive Work Analysis (CWA) - A framework for work analysis that is based on the concept of behaviour-shaping constraints and that contains models of the work domain, control tasks, strategies, social-organisational factors, and worker competencies in a single, integrated framework.

Complex socio-technical system - An entity that incorporates humans and machine components in a purposeful way to produce an outcome and that meets some, if not all, of the system complexity characteristics as listed by Vicente (1999).

Complex system - A system that meets most of the defining criteria for complexity that Vicente (1999) lists.

Concurrent validity - Refers to the extent to which a method produces a result that is consistent with the results of other methods.

Constraint-based method - A method that is centred on the analysis of system constraints

Constraint-based perspective - A theoretical position that emphasises that system behaviour is governed by factors that remove the degrees of freedom.

Construct validity for a method - Refers to whether the underlying theoretical perspective that the method uses is accepted by the community that uses the method.

Construct validity for an analytic product - Refers to whether the analytic product accurately reflects the theoretical construct from which it is derived.

Content validity for a method - Refers to the extent that a method appears to do what it purports to do.

Control Task Analysis (CTA) - An analysis of what needs to be done by the system. CTA is typically thought of an activity analysis. The CTA identifies what needs to be done, independently of how or by whom.

Current system - A system that has a number of functions of which the operators have previous experience.

Data collection methods - Any method that is used to collect data for various applications. Common data collection methods are semi-structured interview and observation.

External validity for an analytic product - Refers to whether the elements and relationships shown in the analytic product correspond to the elements and the relationships in the world that it is representing.

Future system - A system that has a number of functions of which the operators have no previous experience.

High-level dependent variable - A dependent variable formed out of an aggregation of low-level dependent variables that is used for hypothesis testing in this thesis.

Internal validity for an analytic product - Refers to whether the elements shown in the analytic product are coherent, i.e. whether a logical relationship exists between the elements represented.

Low-level dependent variable - A property of a task or system that has been suggested by the measure-selection methods as being sensitive to the system modification.

Measure - A property of a task, human or system that can be parameterised.

Measure of effectiveness (MOE) - A measure of how well a system meets a criterion. MOEs are usually mission or purpose oriented. In the case of a complex system, an example of a MOE is whether an aircraft can deliver a ten tonne load.

Measure of performance (MOP) - A measure that reflects a property of the system. MOPs are usually object or system oriented. In this thesis, a MOP is used to mean a measure that reflects the result of a test and is related to hardware, software and human characteristics such as probability of detection, false alarm rates and human reaction times.

Measure-selection method – A method that is used to select measures.

Method - A replicable series of steps that if followed will result in a goal being achieved.

Operational setting - An environment that represents the conditions in which the system will operate in when in service.

Perspective - The theoretical foundation of a method.

Predictive validity of a method - Predictive validity refers to whether the output of the method meets criteria set by users of the method. In this thesis the predictive validity of the measure-selection methods is shown if measures suggested by the methods are statistically *sensitive* to the system modification and if the methods are *suitable* for use in the operational setting.

Process - A process is a description that shows how methods and analytic products are integrated.

Radar Warning Receiver (RWR) - A system that is used to alert the aircrew to the presence of a threat radar system.

Reliability assurance - A process used in this thesis that aims to ensure that the decisions made, and data used for the production of an analytic product and measure-selection method is recorded and inspectable.

Reliability tests - Formal processes that aim to assess empirically whether different analysts, or the same analyst over time, produce the same analytic product or measure from a measure-selection method.

Sensitivity of a measure - Sensitivity of a measure is shown if the measure reflects a change in response when a relevant variable is manipulated. In this thesis a measure is sensitive if it shows a statistically significant difference in response to a system modification.

Social Organisation Analysis (SOA) - A Cognitive Work Analysis phase that is aimed at identifying the constraints on a system emerging from the interaction between individuals, and between individuals and the organisation.

Socio-technical system - An entity that incorporates humans and machine components in a purposeful way to produce an outcome.

Strategy Analysis (SA) - A Cognitive Work Analysis phase that is aimed at identifying the constraints that affect how the activities identified in the Control Task Analysis can be done.

Suitability of a method - Suitability of a method is shown if the measures that the methods suggests are not affected by common pragmatic limitations (the resources available, data collection methods used, number of data gathering opportunities and theory).

System Life Cycle (SLC) - Encompasses all the activities (of which evaluation is one) that move a product from conception to retirement.

Task-based method or Task-based measure-selection method - A method that aims to identify sensitive measures based on a task-based analysis of the system and the use of guidelines to select measures.

Task-based perspective - A theoretical or pragmatic position that states that the analysis of goals and tasks will provide a description of the system that can be used for system design and evaluation.

Temporal-coordination - Control Task Analysis (TC-CTA) - An analytic product that represents the control tasks and the temporal constraints between them.

The Human Engineering Process (HEP) - Is designed to facilitate the design and evaluation of complex socio-technical systems within the overarching Systems Engineering management framework. The HEP is a standard process that has been used to select and integrate the various task-based data collection methods and analytic products for military systems.

Work Domain Analysis (WDA) - A Cognitive Work Analysis phase that is aimed at identifying the functional structure of the work domain, including constraints from the physical, functional and purposeful nature of the environment.

Worker Competency Analysis (WCA) - A Cognitive Work Analysis phase that is aimed at identifying the human competencies that are required by users of the system.

Preface

The purpose of this report is to inform DSTO scientists and engineers and the wider Defence community about the results of a doctoral program of research. The research was undertaken while the author was a member of Air Operation Division (AOD) and was supervised by Professor Penelope Sanderson from the University of Queensland and Dr Neelam Naikar from Air Operations Division.

This report is essentially a copy of the doctoral thesis (reference: David Crone (2011) *Selecting measures to evaluate complex socio-technical systems: an empirical comparison of a task-based and constraint-based method* PhD Thesis, School of Psychology, The University of Queensland) with some minor formatting and stylistic changes.

Acknowledgements

This work would not have been possible without the help and support of many people. In particular I would like to thank Dr. Simon Parker who provided the initial funds and resources and, above all, encouragement for me to undertake this research. I would like to thank Professor Penelope Sanderson (University of Queensland) and Dr. Neelam Naikar (DSTO) for agreeing to be my supervisors, for their insightful comments, encouragement, drive and enthusiasm. Without them this work would never have been more than just an idea.

To the members of the Air Operation Simulation Centre, Jason Ashwell, Jodie Doman, Tim Fagan, John Fulton, Maria Grabovac, Dennis Hourigan, Steve Kent, Ian Kerton, Cameron Lewis, Bradley MacPherson, Sylvain Manso, Andrew Robbie, Graeme Simpkin, Mike Spataro, Simon Tartaggia, John Yildiz and the late Reg Worl, thank you for your dedication to produce the systems that were specified. Thank you for working weekends, your patience with me and your humour – you are a great bunch of people.

I am indebted to the aircrews from the Australian Army for volunteering to be participants in the experiments. Finally, I wish to thank the Chiefs of Division, Research Leaders, and Heads of Group from the Defence Science Technology Organisation that have supported me.

1. Introduction

1.1 Problem statement

Analysts within Government organisations play an important role in providing advice to customers about the procurement of complex socio-technical systems. For example, scientists at the Defence Science and Technology Organisation (DSTO) provide advice to the Australian Army about the procurement of sensor systems for helicopters. This advice is usually given after the system in question is evaluated against various requirements. For example, a requirement could be that the aircrew transport their cargo to the target area in a safe and timely fashion and return to base. If the use of the system results in the requirement being met the advice may be that the system should be procured.

One of the most important steps in evaluating a system is selecting the measures of performance and measures of effectiveness that should be used in determining whether a requirement has been met or exceeded. A measure of performance is a measure that reflects the result of a test and is related to hardware, software and human characteristics such as probability of detection, false alarm rates and human reaction time [1]. A measure of effectiveness is a measure that is used to reflect how well a human or system meets a criterion. If a measure is selected that is not sensitive, then the evaluation program and the advice given cannot be valid. Analysts must choose the method that will deliver sensitive measures that can be used in programs designed to evaluate improvements to in-service systems and to future systems.

This thesis compares two methods for selecting measures for evaluating complex socio-technical systems. The first method is centred on the analysis of system tasks ("task-based") and the second method is centred on the analysis of system constraints ("constraint-based").

Under a task-based method, measures are identified from the properties of a task. For example, a task may be "detect a threat" and one measure is "participant reaction time". The task-based measure-selection method is used within the Human Engineering Process and represents the current best practice. This process is widely used in the evaluation of complex systems in both laboratory and operational settings. The task-based measure-selection method has been criticised, however, on the grounds that the measures used for the evaluations are selected using guidelines, are not theoretically grounded, and that the approach is not appropriate for future systems.

Under a constraint-based method, measures are identified from constraints acting on the "work" that is being performed. For example, landing an aircraft on a landing site may be constrained by a window of opportunity of a few minutes. A constraint-based measure is "arrive no later or earlier than 15 seconds of planned arrival time". The constraint-based measure-selection method used in this thesis was developed by the author and is placed within the theoretical approach known as Cognitive Work Analysis. Although other constraint-based methods for selecting measures have been used successfully in laboratory

[1] Note all the technical terms that appear in this thesis are defined in the glossary.

settings and are theoretically grounded, they have not been extensively tested in operational situations and have been restricted to a subset of system constraints. The analysis of constraints, in general, is believed to be uniquely suited for evaluating future systems.

Although some research has evaluated task-based measure-selection methods and constraint-based measure-selection methods for system evaluation, no research has compared them for use during the evaluation of current and future systems in operational settings, including advanced simulations. The main question asked in this thesis is: Is there a difference between a constraint-based and a task-based method for evaluating complex socio-technical systems (both current and future) in operational settings? This question will be further refined into specific hypotheses as material for the thesis is presented.

1.2 Aim and Scope

One way of thinking about the research question is in terms of the predictive validity of the methods – do the methods correctly identify sensitive measures and are the methods suitable for use in operational settings. The aim of this thesis is to compare the predictive validity of the task-based and constraint-based methods for evaluating a current complex socio-technical system and a future complex socio-technical system in an operational setting. Predictive validity is shown if the methods meet two criteria. First, the methods should correctly suggest measures that are sensitive, i.e. show statistically significant differences to changes in human-system performance when there is a system modification. Second, the methods should be suitable for use in an advanced simulator (as judged against four criteria that will be described later).

Figure 1-1 shows the structure of the research. Four research stages must be completed, which are shown as the rectangles on the main left diagonal. The theoretical work underpinning each stage is shown in circles along the base of the figure. The rounded rectangles linking the stages highlight the evaluation needed to ensure that the output of each stage is reliable and valid, so providing a solid basis to progress to the next stage.

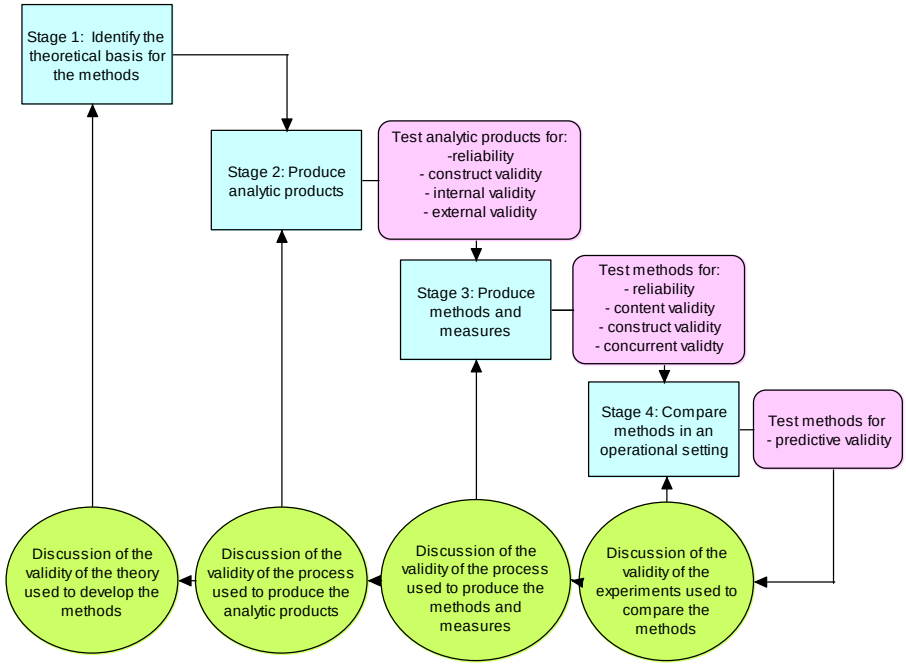


Figure 1-1 Structure of the research performed in the thesis, described as four stages.

Stage 1: The theoretical basis for the measure-selection methods should be identified, which for this thesis are the task-based and constraint-based approaches.

Stage 2: Analytic products should be produced from which measures will eventually be derived, and those analytic products should be valid and produced through a reliable process. In the present case, task-based and constraint-based analytic products will be developed that are consistent with the description of a Radar Warning Receiver during the Preliminary Definition Phase of the System Life Cycle.

Stage 3: Once the analytic products are developed, valid and reliable methods (from the task-based and constraint-based perspectives) should be developed for selecting measures using the analytic products. The methods should then be used to select the measures that will test the effect of an RWR modification in both a current and future system.

Stage 4: Once the measures have been selected, they should be used in the simulation environment where they will be evaluated for sensitivity to the system modification. Statistical sensitivity will indicate that the method for producing the measure is valid. They will also be evaluated for suitability for use in operational settings.

In the chapters that follow, each of these stages and its associated methods is considered in more detail. Variants of the figure will locate each part of the thesis in the context of the stages shown. A little more detail is provided below about each of the four stages.

1.3 Thesis overview

Figure 1-2 provides an overview of the structure thesis. Chapter 1, the present chapter, provides an overview of the thesis. In Chapter 2 the literature on the use of task-based and constraint-based methods for selecting measures to evaluate systems is presented. I will show that a world's best practice task-based method for selecting measures exists but that a constraint-based method is required. I will also show that task-based evaluations of complex systems have been criticised on the grounds that the measures used for the evaluations are selected using guidelines and are therefore not theoretically grounded, and that the approach is not appropriate for future systems. The information in Chapter 2 will show that constraint-based methods have been successfully used in laboratory settings for selecting measures that are theoretically grounded. However, unlike task-based methods, they have not been tested in operational settings. Constraint-based methods purport to be applicable for evaluating future systems. Finally, I will show that it is important to compare the two methods for evaluating complex future technical systems.

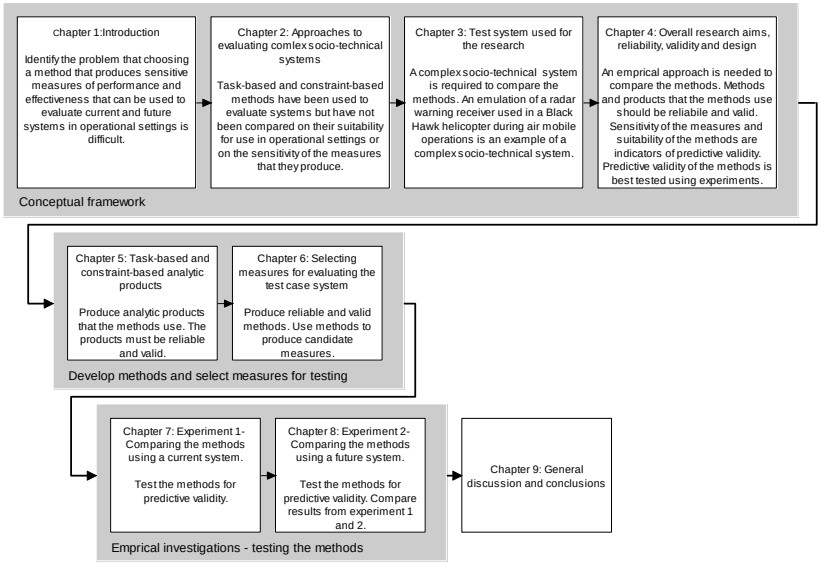


Figure 1-2 Thesis overview

Chapter 3 introduces the technical system on which I compare the two methods of selecting measures for evaluation. The military platform is the Black Hawk helicopter and the operational context is airmobile missions. The specific technical system is a Radar Warning Receiver (RWR). In this chapter I make an assessment on whether the technical system is a complex-socio-technical system. I also describe the Black Hawk simulation environment in which the evaluations take place.

Chapter 4 details the most appropriate research method that should be used to compare the task-based and constraint-based methods. In this chapter I describe the important concepts of reliability and validity. I show that it is important to assess the reliability and validity of the methods and products developed throughout the research program and I outline how this is best achieved. I also show that the predictive validity of the task-based and constraint-based methods is best tested through experiments.

In Chapter 5 the development of task-based and constraint-based analytic products is described. The constraint-based analytic products that are produced are Abstraction Hierarchy (AH), Abstraction-Decomposition Space (ADS) and Control Task Analysis (CTA). The task-based analytic products produced are Mission Narrative (MN), Function Flow Diagram (FFD), ad hoc Function Allocation (FA) and Time Line Analysis (TL). The reliability and validity of the analytic products will be discussed.

In Chapter 6 the development of the actual methods for selecting the measures from the analytic products is described. The reliability and validity of the methods is discussed. In addition, sets of potentially sensitive measures are produced and defined for both the current systems and the future system.

In Chapter 7 the first simulator-based experiment is described: Experiment 1. This experiment investigates whether the measures suggested by the two methods are statistically sensitive to a system modification for a current system and whether the methods are suitable for use in operational settings.

In Chapter 8 the second simulator-based experiment is described: Experiment 2. This experiment is designed to investigate whether the measures suggested by the two methods are statistically sensitive to a system modification for a future system and whether the methods are suitable for use in operational settings. At the end of this chapter results from a comparison of Experiment 1 and 2 are presented.

In Chapter 9 the results obtained from Experiment 1 and 2 are discussed and general conclusions are drawn. In this chapter a summary of the results is given, significant and original outcomes of this research are described, theoretical implications are stated, limitations of the current research are identified, further avenues for research are presented and overall conclusions are stated.

2. Approaches to Evaluating Complex Socio-Technical Systems

Analysts within Government organisations (e.g. scientists at Defence Science and Technology Organisation) have an important role in providing advice to customers (e.g. the Army) about the procurement of complex socio-technical systems (e.g. sensor systems for helicopters). This advice is usually given after the system in question is evaluated against various requirements (e.g. a requirement may be that the aircrew should achieve their mission in a timely fashion). If the system meets or exceeds these requirements the advice may be to suggest that the system should be procured.

One of the most important steps in evaluating a system is selecting the measures that should be used in determining whether a requirement has been met or exceeded. If a measure is selected that is not sensitive then the evaluation program and the advice given cannot be valid.

The problem that analysts face is choosing the method that will deliver sensitive measures that can be used in evaluation programs designed to assess improvements to current systems and to future systems. The choice of the method is complicated because it relies on the analyst understanding several factors including the theoretical perspective that underpins the method and the effect of limitations on the evaluation exercise.

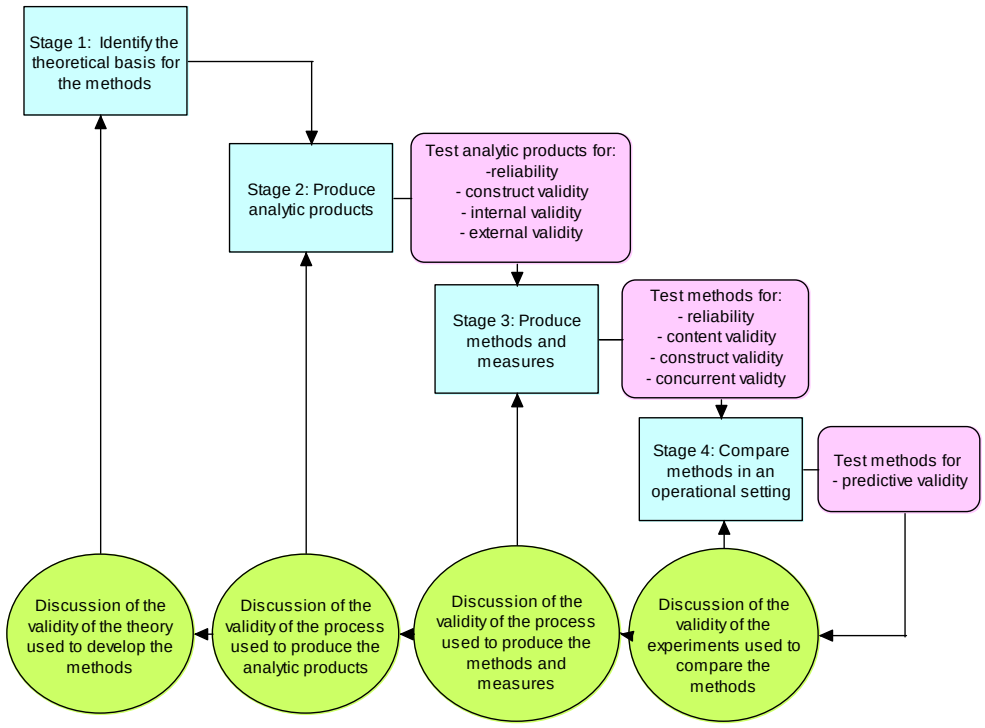


Figure 2-1 Framework of the research conducted in this thesis. The material presented in this chapter relates to Stage 1, in which the theoretical basis for evaluation methods is established

As shown in Figure 2-1 this chapter provides a review of the literature concerning the evaluation of complex-socio technical systems and in particular focuses on the methods used by analysts to select measures of performance and measures of effectiveness.

In the following sections each of the factors that influence the choice of the method will be discussed.

- In the first and second sections (Section **2.1**, Section **2.2**) the distinction between two dominant perspectives used to evaluate system performance is made. The first section (Section **2.1**) describes the task-based perspective and the associated Human Engineering process for system evaluation. The second section (Section **2.2**) describes the constraint-based perspective and its application in Cognitive Work Analysis. The perspective chosen by the analyst is important because it defines two main aspects necessary for selecting measures. These aspects are, first, the type of analytic products produced and, second, the data collection methods used in the evaluation program.

- In the third section (Section **2.3**) the limitations or constraints that shape the evaluation program are examined. These constraints include the phase of the System Life Cycle (SLC) in which the evaluation occurs and the complexity of the system. In this section I will show that system evaluation activities occur in all SLC phases. I will also show that the SLC is used by analysts to guide them on which analytic products are most appropriate to use.
- In the fourth section (Section **2.4**) the task-based and constraint-based methods ("measure-selection methods") used to select measures of performance and measures of effectiveness will be described. In this section I will show that the measure-selection methods incorporate the analytic products and data collection methods that were identified using the SLC. I will also show that there are several potential implications of theory (Kantowitz (1992) that are particularly important for measure-selection methods and the measures that they may suggest. These are:
    i. Theory allows for interpolation of results when data cannot be collected, or where limited data points can be gathered.
    ii. Predictions based on theory can be used in the design of a system, before the system is built.
    iii. Theory can be used to aid measurement and system design.
    iv. Theory can be used as a means for representing normative human behaviour or system performance.

From the review of the literature I will show that an important omission from current research is an empirical comparison between the effectiveness of the task-based and constraint-based methods for evaluating the effect of new systems. In addition, I will show that such a comparison should assess whether the two methods have predictive validity; that is, whether the methods produce "sensitive" measures of performance and effectiveness and whether they are "suitable" for evaluating "current" and "future" complex systems in operational settings.

2.1 Task-based perspective and Human Engineering Process

In this section I describe the task-based perspective, and then I describe the task analysis data collection methods and analytic products and how they are integrated into the Human Engineering process for evaluating a system.

2.1.1 Task-based perspective

A task-based perspective to system design and evaluation emphasises the task as the unit of analysis and describes the behaviour of the system in terms of what the system can do. In this section I describe the main characteristics of the task-based perspective. Briefly, these characteristics are that it has no "theoretical" foundation but has developed from a pragmatic need to study human "tasks", is designed to capture the human physical and cognitive characteristics of current tasks, is associated with a number of methods (of which the first was used for the design of training programs), is now the "organising element"

for system design, has only relatively recently been used to predict future tasks and finally, a key requirement for its use is an accurate representation of the environment in which the task takes place.

The origins of the task-based perspective for the analysis of human work may be traced back to the practical work of Taylor (1911) and his time-study procedure rather than to theory. The work that Taylor performed was motivated by the need to develop standards for the time needed to complete physical tasks. Later on he extended his work to include personnel selection, work methods, labour standards and an individual's motivation to perform work. His work became less relevant as tasks became more cognitively complex because his method could not account for an individual's ability to analyse information and make decisions.

The understanding that human performance (the ability of a human to perform a task) could be influenced by psychological factors rather than purely physical aspects was first identified by the Hawthorne Studies (from 1927 to 1932) and later Hertzberg (Hertzberg, 1966; cited in Crystal and Ellington, 2004). The results of the Hawthorne studies found that attention given to workers by managers was more important than environmental effects (e.g. lighting). Other factors such as "motivation" and "hygiene" were identified by Hertzberg as also being important to workers. Hertzberg analysed the tasks that people were employed to do and was able to show that non-physical factors such as job satisfaction and the psychological states of workers contributed to their performance.

As the complexity of tasks increased, the requirement grew to have a reliable means to collect and record data, and to produce analytic products for reporting. The requirement to have a reliable means to record data in turn led to a requirement to understand tasks at a deeper level. Chapanis (1959) (cited in Crystal and Ellington, 2004) developed linear flow diagrams to aid the analysis of complex tasks that involved control, planning and problem-solving. The diagrams were used by researchers for developing formal models of human performance. The development of these models in turn emphasised the importance of having a detailed understanding of the task.

More recently comes Hierarchical task Analysis (HTA: Annett et al, 1971) which is probably the best known "method" that adopts a task-based perspective. HTA was influenced by work on planning and problem solving by Miller, Galanter and Pribram (1970). Originally the method was used to identify where training was needed. To do this, high level system goals associated with operating the system were decomposed into smaller ones. Modern applications of HTA focus on tasks (although there is some debate surrounding whether it is valid to do so; Diaper, 2004a). A "typical" task analysis of an aircraft system may result in the goal/ task "Fly aircraft" decomposed into "Manipulate aircraft surfaces using flight controls" and "Control thrust using engine controls".

The importance of representing human work as tasks in a variety of domains is strongly supported by Meister (1999a). In his review Meister emphasises not only that the task is the organising element in system design and evaluation, but also that a task description should be prescriptive: "A task describes the steps by which the human shall interact with the machine" (p.46). He also expands on the definition of a task by noting that a task is

"man-made" and "must contain a setting, a description of the environment, and conditions in which that action takes place" (Meister, 1999; p.47). A task is therefore a human construct that may be defined at various levels of granularity and it is event dependent.

Meister also notes that task representativeness is crucially important if an evaluation of a system is to be effective. If the task is not representative of the action that the system performs (or will perform) then there will be little confidence (or usefulness) in the results of the evaluation activity.

In summary, an analysis of the literature revealed that the task-based perspective of human work was born not out of psychological theory, but out of pragmatism. The initial analysis of system goals has led to the analysis of system tasks. The task-based perspective has developed to a point where the task is the fundamental unit of analysis for system evaluation. In addition, a task is defined in specific terms that relate to the properties of the human and machine agents that perform the task and also the environment in which it takes place. As I will show in the next section the task-based perspective has led to a number of task-based data collection methods and analytic products and these have been integrated into a standardised process for the design and evaluation of human-machine systems.

2.1.2 Human Engineering Process

The previous section indicated that the task-based perspective has been adopted by many analysts and has become the unit of analysis for system design. In this section I review the Human Engineering (HE) literature and show that the HE analysts have produced many different task-based data collection methods and analytic products for many applications. I also show that a process that incorporates important HE aspects has been developed (the Human Engineering Process; HEP) and is widely used by analysts to guide them on the use of task-based data collection methods and development of its analytic products.

Human Engineering may be defined as:

> *"The application of knowledge about human capabilities and limitations to system or equipment design and development to achieve efficient, effective, and safe system performance at minimum cost and manpower, skill, and training demands. Human engineering assures that the system or equipment design, required human tasks, and work environment are compatible with the sensory, perceptual, mental, and physical attributes of the personnel who will operate, maintain, control and support it." (EIA, 2002, p.16)*

The definition of Human Engineering highlights four important aspects for system design and evaluation. First, HE is a human-centred approach to system design. Second, HE is associated with a process that should result in a system that is designed to meet the characteristics of the human operator (HE "assures that the system…"). Third, human work is described in terms of tasks. Fourth, HE is very broad.

The literature reviewed in Table 2-1 encompasses all of these aspects. The table summarises the literature and reveals that there are many types of task-based data collection methods and analytic products that have been used or have been recommended to be used for various military and civil applications. As can be seen from the table the

analytic products and data collection methods are generalisable to a wide range of systems. For example, the task-based data collection method of observation and the analytic product Hierarchical Task Analysis is cited by Kirwan and Ainsworth (1992), Diaper and Stanton (2004a) and Stanton et al (2005) as being applicable to various military and civilian systems.

The HEP is designed to facilitate the design and evaluation of complex socio-technical systems within the overarching Systems Engineering management framework. The HEP is a standard process that is used to select and integrate the various task-based data collection methods and analytic products for military systems.

Table 2-1 Task-based data collection methods and analytic products commonly used in system design and evaluation

Authors	Task-analysis data collection methods	Analytic products	Area of application (including case studies and general examples)
Kirwan and Ainsworth (1992)	Activity sampling Critical Incident Technique Observational techniques Questionnaires Structured interviews Verbal protocol analysis Task description Simulation Behaviour assessment Task requirement evaluation	Charting and network techniques Decomposition methods Hierarchical Task Analysis Link analysis Operational sequence diagrams Timeline analysis	Staffing levels for nuclear power plant Assessment of communications requirements for a drilling platform. Panel design for a nuclear power plant Workload assessment for a control room Workload assessment for a command system Safety analysis for a nuclear power plant Training analysis for maintenance personnel Analysis of human performance and errors associated with an inspection task Analysis of potential human error associated with the operation of a solid storage plant Analysis of the operation of a nuclear chemical plant
Beevis (1999)	Data collection methods were not explicitly stated but following resources were identified: Interview data from subject matter experts Document analysis Analysis of similar systems	Narrative mission descriptions Graphic mission profiles Function flow diagrams Sequence And Timing diagrams Structural Analysis and Design Technique Information flow and processing analysis State-transition diagrams Petri nets Behaviour graphs	General analysis of military aircraft (fixed wing and rotary wing), navy vessels (ship, submarine, fast patrol boat) and army vehicles (general vehicle and main battle tank)

Authors	Task-analysis data collection methods	Analytic products	Area of application (including case studies and general examples)
		Ad hoc function allocation Fitts' list Review of potential operator capabilities Function allocation evaluation matrix Requirements allocation sheets Timelines Flow process charts Operational Sequence Diagrams Information/ action or Action/ information tabulations Critical task analysis Decision tables	
Diaper and Stanton (2004a)	Interviewing experts Direct observation Document analysis	Hierarchical Task Analysis Computer software design products	Human-Computer interface design for various systems
Stanton et al (2005)	Interviews Questionnaires Observation	Many including: HTA Verbal Protocol Analysis Task Decomposition The Sub-Goal Template Method Tabular Task Analysis	Crew situation assessment Team assessment Interface analysis Human error identification

The HEP is typically made up of a number of integrated phases (Pearce, 1990, cited in Beevis 1999; DEF STAN, 2008).

Mission analysis. Mission analysis defines the overall requirements of the system. The analyses define what the system must do and also in which environment it will operate.

Function Analysis. Function analysis methods are designed to analyse the system in terms of the function (high level activities) that it should perform rather than on the subsystems (technologies) that could be used to meet the requirements of the system.

Function Allocation. Function allocation methods are designed to allocate the system functions rationally to the subsystem (including human) that is most suited.

Task Analysis. Task analysis methods are used to identify the tasks that are necessary for the operator to perform.

Table 2-2 shows that each of the HEP phases identifies different task-based analytic products that should be produced (it is important to note that although one of the HE evaluation phases is labelled task analysis, all the phases produce task-based analytic products). In general, as part of the evaluation activity typically seen during system procurement, all the HEP phases are performed in the following sequence: mission analysis, function analysis, function allocation, task analysis. As I will show later the level, or depth, at which each of the HEP phases is performed is also guided by the SLC phase. Examples of the different types of information contained in the analytic products and how that information is integrated between phases is presented below.

Narrative mission descriptions are used during the Mission Analysis phase of the typical HEP. They are used to describe the typical or probable events of a mission in detail (Beevis, 1999, pp 39). Major mission phases, major system functions, the timescale of activities, and the external events that trigger the activities are all elements of the mission that should be included. Beevis notes that inputs to the narrative mission should include the description of the system's missions, required capability, operational environment and systems dynamics and system boundaries. He also notes that the use of subject matter experts (SMEs) with experience of similar missions is essential to develop the narrative. The output of the descriptions should provide sufficient detail to identify the upper–level functions provided by the system (Beevis, 1999).

Function flow diagrams are used during the Function Analysis phase of the HE process. Function flow diagrams show the sequence of the functions that are required to perform the mission. The sequence of the functions reflects the order that the functions are performed. AND/OR logic is used to indicate functions that are performed in parallel or in series. Information from the narrative mission analysis and similar systems with similar operational requirements are required to construct the diagrams (Beevis, 1999).

Table 2-2 Task-based analytic products by HEP phase

Human Engineering evaluation phase	Typical task-based analytic products
Mission analysis	Narrative mission descriptions Graphic mission profiles
Function analysis	Function flow diagrams Sequence And Timing diagrams Structural Analysis and Design Technique Information flow and processing analysis State-transition diagrams Petri nets Behaviour graphs
Function allocation	Ad hoc function allocation Fitts' list Review of potential operator capabilities Function allocation evaluation matrix Requirements allocation sheets

Human Engineering evaluation phase	Typical task-based analytic products
Task analysis	Timeline analysis Flow process charts Operational Sequence Diagrams Information/ action or Action/ information tabulations Critical task analysis Decision tables

Ad Hoc Function Allocation is used during the Function Allocation phase of the HEP. Ad Hoc Function Allocation is a process by which the functions described in the Function Flow diagrams are allocated to hardware, software and human parts of the system. As Beevis notes, the allocation could be based on knowledge of predecessor systems or similar systems.

Timeline analysis is used during the Task Analysis phase of the HEP. Timeline analysis is used to show the temporal relationship between system tasks as a basis for workload and resource estimation. The tasks are derived from the function flow and allocation diagrams. A critical task is one that has a high workload or that is critical to system safety or mission success (STANAG 3994 AI, quoted in Beevis 1999). In order to produce the timeline analysis, the sequence and performance criteria for the operator's tasks and details of the human-machine interface design should be included. STANAG 3994 indicates that following information should be included for each task: information required, perceptual load, decision required, action taken, communications required, interface constraints, workspace constraints, and environmental constraints.

A typical example of the application of the HEP phases is given by Campbell and Herdman (2003). In their evaluation of head-down displays for a fighter aircraft the authors completed a number of the HEP evaluation activities. These included: completing a mission analysis in which descriptions of the equipment suite and agreed capabilities were identified, identifying the functions required to attain the mission objectives and collecting narrative and graphical data to describe the fighter operations, allocating the system functions to either the pilot or machine based on pilot abilities, and producing operational sequence diagrams. Once these steps had been taken a simulator was designed to reflect the operational conditions of the fighter aircraft. An empirical study was then conducted to assess the benefit of the new head-down displays.

In summary, many task-based data collection methods and analytic products have been developed for use in system design and evaluation. The HEP is a standardised approach that is designed to facilitate the design and evaluation of military socio-technical systems. The HEP provides a guide to system designers on which of the task-based data collection methods and analytic products are most appropriate to use.

2.2 Constraint-based perspective and CWA

In this section I describe the constraint-based perspective and the Cognitive Work Analysis framework for design and evaluation of socio-technical systems.

2.2.1 Constraint-based perspective

A constraint-based perspective to system design and evaluation describes the behaviour of the system in terms of what constrains it. To date, Cognitive Work Analysis (CWA) is the only approach to the design and evaluation of socio-technical systems that adopts a constraint-based perspective (Vicente, 1999). CWA is a conceptual framework that was developed by Rasmussen et al. (1994). It is designed to facilitate the identification and analysis of work. The framework's theoretical roots are in the systems perspective and in "an ecological perspective to human factors" (Flach, Hancock, Caird and Vicente, 1995; cited in Vicente, 1999, p.48).

Proponents of the systems perspective consider that a system is a collection of components organised in a logical manner. The proponents also consider that the output (or value) of the system is more than the sum of its parts. Therefore, it is essential that that the system is treated as a whole rather than as a collection of individual components. The ecological perspective emphasises the role that the environment plays in constraining the behaviour of actors in that environment. The constraint-based perspective considers human-machine output to be a product of the interaction between different classes of constraint.

2.2.2 CWA framework

In this section I describe the CWA framework. Following the definition of CWA and a brief discussion of its purported benefits over task-based perspective, I show that the data collection methods used are the same as the task-based data collection methods, but the analytic products are different. I then describe the five CWA phases and the analytic products associated with the phases in some detail.

Cognitive Work Analysis (Vicente, 1999) adopts the constraint-based perspective to the analysis of human work and is defined as:

> *"A framework for Work Analysis…It is based on the concept of behaviour-shaping constraints and contains models of the work domain, control tasks, strategies, social-organisational factors, and worker competencies in a single, integrated framework." Vicente (1999, p. 5)*

This definition clearly identifies a number of differences and similarities with the HEP. The most obvious point of similarity between the HEP and CWA is that both purport to be complete representations of the system. The implication is that once the HEP or CWA is finished, a complete description of the system is gained – no more analysis is required. The most obvious point of difference is that CWA identifies constraints rather than tasks as the unit of analysis. Unlike the HEP, CWA is not a standardised approach to system design and evaluation, but like the HEP it emphasises the goal of making human work safe, productive and healthy. It is also similar to the HEP in that it identifies a number of phases, each of which has distinct data collection methods and analytic products. The CWA definition also identifies that analysis should be conducted to identify the constraints associated with all of the phases, and like the HEP all the phases are integrated.

CWA has developed in response to limitations with traditional system design and evaluation methods that have emerged from an over reliance on the task-based, procedural approach to system design and evaluation (Vicente, 1999). Three of the limitations are as follows. The first limitation is that traditional system design does not

cater for unanticipated events, which are often the very events that cause the greatest negative impact. The second limitation is that contemporary systems do not take full advantage of technical possibilities because system designers tend to develop new systems on the basis of what the old technology has offered (an evolutionary approach) rather than what the new technology can produce (a revolutionary approach). The third limitation is that the human interface of past systems has been designed solely from a cognitive, or human information processing viewpoint, rather than from an ecological one. Vicente argues that if human-machine interfaces are designed from an ecological point of view, operators should be able to recover from unanticipated events (system errors) in a safer and more productive way.

CWA has been used successfully during the design and evaluation of complex socio-technical systems for civil and military applications such as aircraft, ships, command and control systems, hospital systems (See Bisantz and Burns, 2009, and Jenkins et al, 2009a, for overviews).

A review of the literature concerning the use of CWA reveals that many data collection methods used to analyse the constraints on human work are common to the task-based perspective (compare Table 2-1 with Table 2-3) but that the information the methods produce is presented via analytic products unique to CWA. The review also reveals that CWA has been used in a wide range of system evaluation activities. Table 2-3 summarises some of the literature concerning the production of constraint-based analytic products across several areas of application.

Table 2-3 Constraint-based data collection methods and analytic products commonly used in system design and evaluation

Authors	Constraint-based data collection methods	Analytic products	Area of application
Naikar et al (2005, 2006)	Observation, focussed field observations Interviews with domain experts using walkthroughs, talk throughs, table top analysis Critical decision method Document analysis	Abstraction - decomposition space (ADS)	Crewing concepts for Airborne Early Warning and Control aircraft F/A-18 aircraft training simulator requirements
Bizantz et al (2003, 2001)	Semi-structured interview Document analysis	Abstraction - decomposition space (ADS)	Information requirements for Navy ship personnel
Jenkins et al (2008)	Semi-structured interview using walkthroughs	Abstraction - decomposition space (ADS)	Command and control system for the army.
Sanderson and Naikar (2000)	Semi-structured interview Document analysis	Temporal Coordination-Control Task Analysis (TC-CTA)	Crewing concepts for Airborne Early Warning and Control aircraft

Authors	Constraint-based data collection methods	Analytic products	Area of application
Rasmussen, (1974)	Observation Interview	Decision ladder	Developing representations of cognitive processes to be used in the development of human-machine interface
Naikar et al (2005, 2006)	Same methods as used for the ADS	Contextual Activity Template	Crewing concepts for Airborne Early Warning and Control aircraft
Rasmussen (1980, 1981)	Document analysis Interviews	Information Flow maps	Developing representations of cognitive processes to be used in the development of human-machine system

CWA contains five analytical phases. Each phase captures a particular type of constraint and each phase uses unique analytic products to represent those constraints. Table 2-4 shows the relationship between the CWA phase and analytic product. The five CWA phases are work domain analysis (WDA), control task analysis (CTA), strategies analysis (SA), social organisation and cooperation analysis (SOA) and worker competencies analysis (WCA). Vicente (1999) proposes that the analysis of system constraints should proceed in the following order, WDA, CTA, SA, SOA, and WCA. Each of the five phases is described below.

Table 2-4 constraint-based (CWA) analytic products

Cognitive Work Analysis phase	Typical constraint-based analytic products
Work Domain Analysis	Abstraction hierarchy (AH) Abstraction decomposition space (ADS)
Control Task Analysis	Activity Analysis in work domain terms (AA/WD) Temporal coordination control task analysis (TC-CTA) Decision ladder (DL) Strategies Analysis Information flow maps (IFM) Charting techniques
Strategies Analysis	Information flow maps (IFM) Decision ladder (DL) Charting techniques
Social Organisation and Cooperation Analysis	Abstraction hierarchy (AH) Abstraction decomposition space (ADS) Activity Analysis in work domain terms (AA/WD) Temporal coordination control task analysis (TC-CTA) Decision ladder (DL) Information flow maps (IFM) Social network analysis (many types) Link analysis (many types)
Worker Competencies Analysis	Skills, rules and Knowledge (SRK)

The first CWA phase, Work Domain Analysis (WDA), is an analysis of the constraints on human behaviour coming from the physical, functional and purposeful nature of the work

environment. The WDA is most commonly represented as one or more abstraction hierarchies.

WDA represents the domain in which activity takes place, not the activity itself (Vicente, 1999). With a WDA, the analyst aims to capture both the physical elements of the system and the reasons why the physical elements are present. The WDA therefore captures the constraints on operator behaviour that are imposed by the environment (domain) in which the operator works. The Abstraction Hierarchy and the Abstraction Decomposition Space are ways to present the WDA. An AH has been adopted in this thesis and is shown, in general terms, in Figure 2-2 and important aspects, as signified by the boxes and connections between boxes, are described in detail below

The AH captures three important aspects of a Work Domain Analysis. First, the work domain of interest is concurrently represented at several levels of abstraction and at several levels of decomposition. Figure 2-3 highlights the abstraction dimension, which runs vertically. Using labels borrowed from Xiao et al (2008) this dimension can be described in terms of the physical objects ("Physical Objects") and their engineering function ("Physical functions"); these two levels make up the physical domain. The abstraction dimension also shows why these objects and functions are useful within a particular work domain. This "purposive" aspect is not an engineering property, nor is it what the subsystem, unit or component was engineered to do, but instead it is a reflection of the human purposes of the system; what the subsystem or component could be used to do, or is used for, in that particular domain.

The "purposive" aspect of the WDA is captured by the "Domain functions" level (or why a particular function is in the work domain), the "Domain values and priorities" level (what are the priorities or what aspects are of value in the work domain); and finally the "Domain purpose" level (the reason that the whole work domain exists).

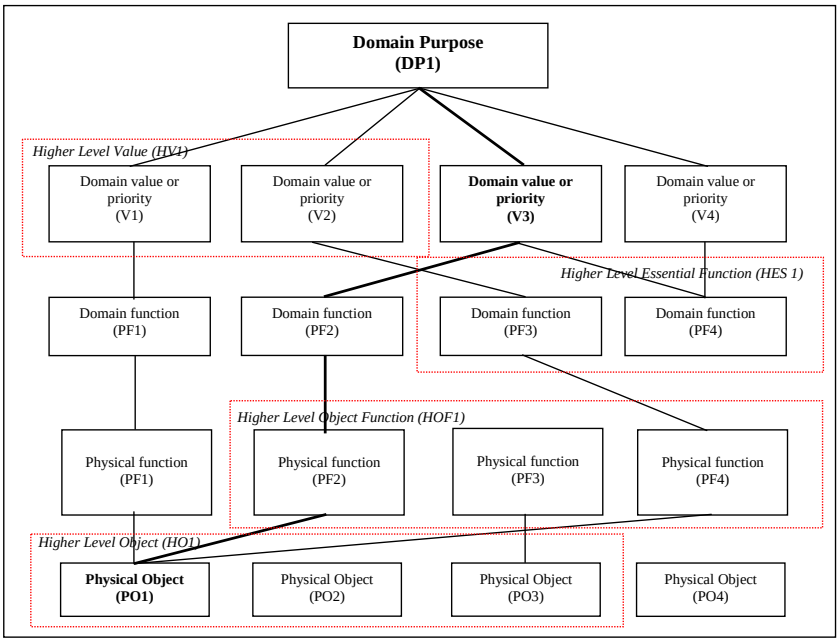


Figure 2-2 AH representation. The labels used in the boxes come from Xiao et al (2008).

The abstraction dimension makes the AH a unique analytic product because it lets a work domain be viewed in different ways that are meaningful for different questions. For example, if the analyst were concerned with deriving the absolute mass of a system then the abstraction layer of most use would be the "Physical objects" level. The analysts would be able to assign a mass for each object and simply sum the total. However, if the analyst were interested in comparing one system against another in terms of safety, then the most useful domain representation would be the "Domain values and priorities" (assuming safety is a domain value).

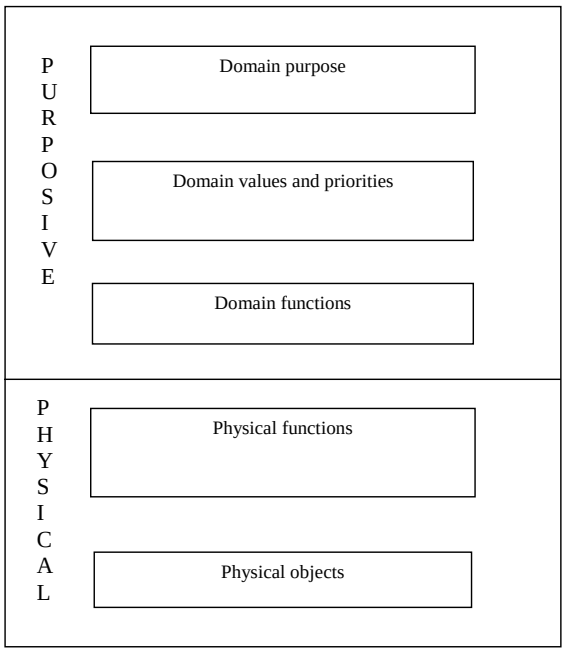


Figure 2-3 Five abstraction layers of an AH

In contrast to the abstraction dimension, the system decomposition dimension represents the system at different levels of granularity (from a system level, through sub-system and unit level, to an individual component level). There are many ways to show decomposition relationships. Within Figure 2-4 nested boxes represent the various decomposition levels. Objects enclosing other objects are at a higher level of decomposition. Hence, "Higher level physical object (HPO1)", a system level object, may be decomposed into "Physical object, PO1" and "Physical object, PO2".

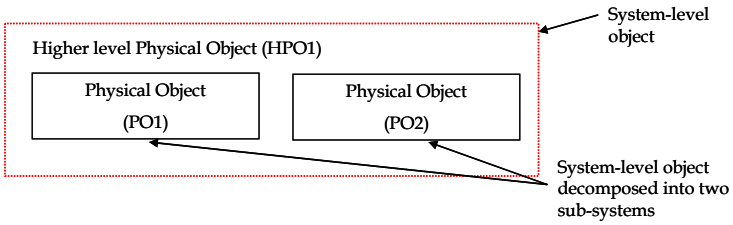


Figure 2-4 Abstraction Hierarchy decomposition dimension

The second important aspect of the WDA is that each object, function and purpose may be described in terms of their properties. These properties may be physical in the case of objects (for example, "mass" or "volume") or they may be functional in the case of

functions. For example, the Domain function, Tactical Operations, may have two properties: timeliness and surprise.

The third important aspect of the WDA is the means-end (or how-why) relationship between objects, functions, and purposes across different layers of the abstraction dimension, which are shown as links. This means that as well as describing the work domain in terms of objects, functions, values and purposes, the WDA may be parsed in a meaningful way that illustrates the relationship between objects, function and purposes at different levels. For example, Figure 2-5 shows that if one selects a property and asks the question "Why is the property seen in the function PF2 important?" and follows the link to the function above, the answer is given (in this case DF2). Similarly, if one selects a value, say V3, and asks the question: "How is this value achieved?" and follows the link down the answer is given: "By achieving DF2". The structural why - how relationship (a means-ends relationship) is important because it allows one to develop an understanding of the whole work domain: the objects, their properties (with respect to the work domain) and functions.

Data collection methods that have proved to be useful when constructing an ADS are observation, focussed field observations and interviewing domain experts using walkthroughs, talkthroughs, tabletop analyses, critical decision method and analysis of documents (Naikar et al., 2006). Other authors also examined documents and used semi-structured interviews (e.g. Bisantz et al, 2003, 2001) and semi-structured interviews and walkthroughs (e.g. Jenkins et al, 2008).

The second CWA phase, control task analysis (CTA) is an analysis of the constraints that are derived from the actions that have to be performed by the system, and is typically thought of as an activity analysis. The CTA is most commonly represented as a decision-ladder (Rasmussen, 1974), Temporal Coordination–Control Task Analysis (TC-CTA; Sanderson & Naikar, 2000) and more recently the Contextual Activity Template (Naikar, et al 2006). Although these are different products they all share a common feature. The data collection method used to generate each representation focuses on activity. Note, however, that the term "activity" should not be confused with "task" as used in the HEP. The data collection methods used in CTA are used to examine the inputs and outputs (the constraints) to an activity, which is thought of as a "black-box". The black-box "describes what needs to be done, not how or who" (Vicente, 1999, p. 183). This is in contrast to a task that is the description of the "steps by which a human shall interact with the machine" (Meister, 1999a, p.46).

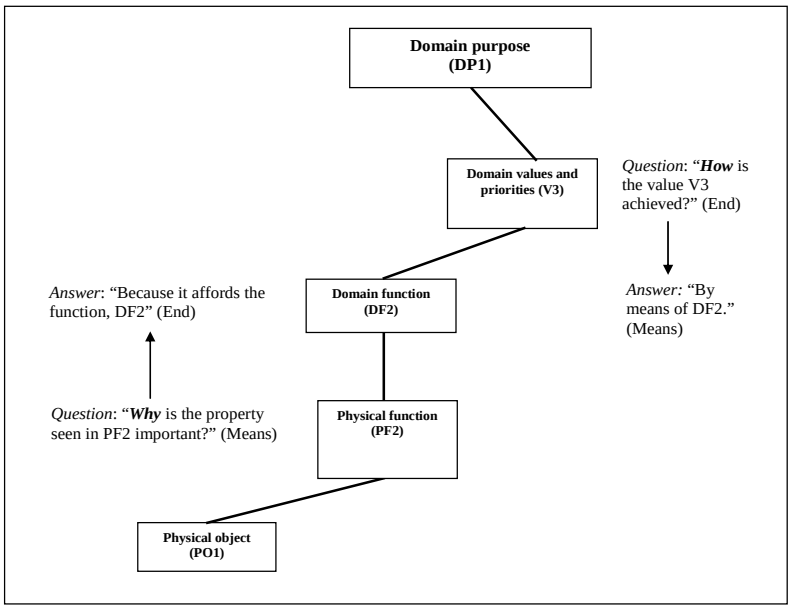


Figure 2-5 Abstraction Hierarchy means-end relationship

Figure 2-6 presents a generic TC-CTA. In addition to the standard TC-CTA features of representing functions on the y-axis and mission phases and other significant constraints on temporal ordering (such as landing and take-off) and the appropriateness of action on the x- axis, the TC-CTA includes the systems that are used during the control task. In the TC-CTA control tasks are represented as if they are movable beads on rods of different lengths, occurring at any point within a particular time span defined by the span of the rod. Each task has a number of task-specific properties associated with it and all the tasks have task-generic properties that describe relationships between tasks. Hence, a task-specific property is a property that describes a task independently of other tasks, whereas a task-generic property describes some relationship such as the sequence that the tasks occur in. For example, control task 1 may include a number of task-specific properties (for example, "duration" and "time") and may also have generic properties associated with it (for example, it may be the highest priority task of all tasks and may be the first task in the sequence of all tasks). Typical TC-CTA data collection methods include semi-structured interviews as well as document analysis. For an example see Sanderson and Naikar (2000).

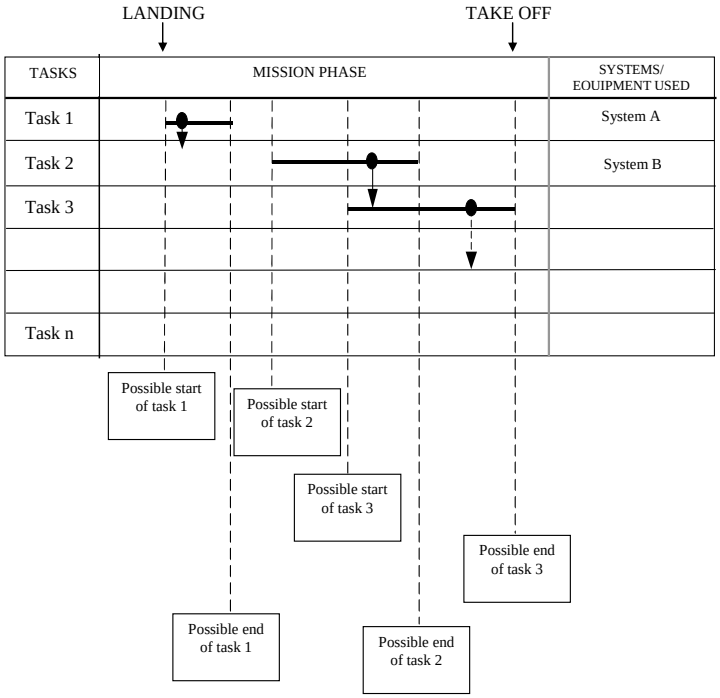


Figure 2-6 Temporal coordination CTA

The third CWA phase, SA, is an analysis of the constraints that are derived from how the activities identified in CTA can be done (Vicente, 1999). The analytic product used is information flow maps (Rasmussen, 1980, 1981). Constraint-analysis data collection methods include document analysis and interviewing domain experts (Vicente, 1999).

The fourth CWA phase, SOA, is an analysis of the constraints that are derived from the interaction of individuals, and individuals and the organisation. The constraints are primarily concerned with the identification of who does what. SOA constraints are usually represented as annotations to ADS, CTA and information flow maps (Vicente, 1999). There are no constraint-based data collection methods unique to this phase.

The fifth CWA phase, worker competencies analysis (WCA) is an analysis of the human competences that are required by the users of the system. Rasmussen's (1983) skills, rules, knowledge (SRK) taxonomy is often used to represent these constraints. There are no constraint-based data collection methods unique to this phase.

In summary, many constraint-analysis data collection methods and analytic products are used for system design and evaluation. The data collection methods used are common to both the task-based and constraint-based perspectives. CWA is the only constraint-based approach that describes how a constraint-based analysis of a system may be performed and provides guidance to system analysts on which of the analytic products is most appropriate to use for each CWA phase.

2.3 Analytic products, complexity, and system life cycle

The goal of this section is to identify the analytic products that are most suitable for use in this program of research and in particular the analytic products that should form the basis of the measure-selection methods. The previous section has shown that there are many different task-based and constraint-based data collection methods and analytic products that have been used by analysts during various system evaluation activities. This section now shows that analysts can be guided on what task-based analytic products to produce by using the SLC phase and the category of complexity of the system. However, I also show that this guidance does not exist for the constraint-based products, but that by reviewing the literature it is possible to identify which analytic products are most suitable. Before that the SLC will be briefly described.

The "System Life Cycle" (SLC) is often characterised as "cradle to grave" and encompasses all the activities (of which evaluation is one) that move a product from conception to retirement. There is no one "correct" SLC. SLCs are developed and used almost on a project-by-project or country-by-country basis. Beevis (1999) identifies four general SLC phases:

i. Preliminary systems studies phase
ii. Concepts formulation and validation phase
iii. Design and development phase
iv. Procurement and use

The preliminary systems studies phase is concerned with taking the needs statement (this identifies a capability shortfall) and identifying the requirements to satisfy the needs. Once a specific need has been identified, high-level concepts are generated during the concepts formulation and validation phase. The aim of this phase is to provide a number of possible concepts that meet the specific need, evaluate these concepts against a number of criteria and recommend one concept to be taken further in the design process. Technological research may be conducted if there is insufficient information to support a decision. During detailed design and development the preferred concept is developed in more detail, prototypes are developed and evaluated. Finally, the system is procured and used in the field.

2.3.1 Task-based analytic products for this program of research

Table 2-5 (Beevis, 1999) shows the relationship between the SLC and system complexity and indicates which analytic products are recommended during each phase of the Human Engineering Process (HEP) phases. Beevis produced the table to summarise the recommendations made by NATO panel members. The table is designed to guide contractors and government organisations on the most common and beneficial task-based products to use in each of the SLC phases during system evaluation. Beevis' recommendations were based on an extensive survey of manufacturers and government organisations across the world. Beevis (1999) is the primary reference used in this thesis for the HEP approach because the work reflects the most current available survey of the

analytic products used during complex system procurement and provides a record of world best practice.

The table lists all the HEP phases and products along the x-axis and system complexity by System Life Cycle (SLC) phase along the y-axis. The table shows the SLC phases as: P- Preliminary phase, C- Concept phase, D- Design phase, U- Use, and A- average over all phases. Each analytic product is given a level of recommendation as follows: H-high recommendation, M-medium recommendation, L-low recommendation, N-not recommended.

As can be seen from the groupings of table columns, Beevis (1999) categorises a system in terms of complexity: simple, medium-complexity, high-complexity single-operator and high-complexity multi-operator. Charting the four categories of complexity on one axis and the task-based analytic products on the other NATO panel members were able to recommend suitable analytic products for use with systems of different complexity. For example, the panel recommended that for an evaluation of a medium complexity system, during the Preliminary systems study phase, which is the category of our technical system case is categorised (see Chapter 4 for a detailed description of the Radar Warning Receiver used on a Black Hawk helicopter), the following analytic products are most appropriate:

- Narrative Mission Descriptions during the Mission Analysis phase.
- Function Flow Diagrams during the Function Analysis phase.
- Ad Hoc Function Allocation during the Function Allocation phase.
- Timelines during the Task Analysis phase.

In the next section the constraint-based approach is considered and the analytic products that the constraint-based measure-selection method should use are identified.

Table 2-5 Applicability of task-based analytic products to the system life cycle phase (Beevis, 1999)

Human Engineering Analysis Products by System Development Phase[2]	System Complexity																			
	Simple System					Medium-Complexity System					High-Complexity Single Operator					High Complexity Multiple-Operator				
	A	P	C	D	U	A	P	C	D	U	A	P	C	D	U	A	P	C	D	U
Mission analysis																				
Narrative mission descriptions	H	H	H	M	L	H	H	H	L	L	H	H	H	L	L	M	H	H	M	M
Graphic mission profiles	N	N	N	N	N	N	L	N	N	N	H	H	H	M	M	H	H	H	M	M
Function analysis																				
Function flow diagrams	L	L	L	L	N	M	L	M	M	N	H	M	H	H	L	H	H	H	H	M
Sequence And Timing diagrams	N	N	N	N	N	L	N	L	N	N	M	L	M	M	N	H	M	H	H	M
Structural Analysis and Design Technique	N	N	N	N	N	L	L	M	L	L	M	L	M	M	L	H	M	H	H	M
Information flow and processing analysis	N	N	N	N	N	L	L	M	L	L	L	L	L	L	L	L	L	M	L	L
State-transition diagrams	N	N	N	N	N	L	L	L	L	N	M	L	M	M	L	M	N	M	L	N
Petri nets	N	N	N	N	N	N	N	N	N	N	M	L	M	M	N	M	N	M	M	N
Behaviour graphs	N	N	L	N	N	M	L	M	M	L	H	L	H	H	L	H	L	H	H	L
Function allocation																				
Ad hoc function allocation	N	N	L	N	N	L	L	M	N	N	L	L	M	N	N	L	L	M	L	N
Fitts' list	N	N	N	N	N	L	N	L	L	N	L	L	L	N	N	N	L	N	N	N
Review of potential operator capabilities	N	N	L	N	N	L	L	M	N	N	M	L	H	L	N	M	M	M	L	L
Function allocation evaluation matrix	N	N	N	N	N	L	N	L	L	N	M	L	M	M	L	M	L	H	M	N
Requirements allocation sheets	N	N	N	N	N	L	L	L	N	N	M	L	M	M	L	M	L	M	M	N
Task Analysis																				
Timelines	L	N	L	L	N	M	L	M	M	N	M	N	H	M	L	H	M	H	M	L
Flow process charts	L	L	L	N	N	N	N	N	N	N	N	N	L	N	N	L	N	L	L	L
Operational Sequence Diagrams	H	L	H	H	H	H	L	H	H	H	H	L	H	H	H	H	L	H	H	M
Information/ action or Action/ information tabulations	H	H	H	H	L	M	L	H	H	L	M	L	M	M	L	L	L	M	L	L
Critical task analysis	L	N	L	L	N	L	N	L	L	L	H	L	H	H	M	H	L	H	H	M
Decision tables	N	N	N	N	N	L	N	L	L	N	M	L	M	M	N	M	L	M	M	N

[2] (A=average over all phases, P=Preliminary phase, C=Concept phase, D=Design phase, U=Use; H=high recommendation, M=medium recommendation, L=low recommendation, N=not recommended).

2.3.2 Constraint-based analytic products for this program of research

There is no state of the art survey that shows the relationship between the SLC, system complexity and which analytic products are recommended for each of the CWA phases. Therefore, it is necessary to review the literature to determine which analytic products are commonly produced during the SLC phases.

Determining the "complexity" of systems reported in the CWA literature is also problematic because unlike Beevis (1999) there is no agreed set of criteria. Vicente (1999) however, does identify 11 characteristics that may be used to assess the complexity of a system. By applying Vicente's characteristics it seems that all the systems in the reviewed literature may be described in Beevis' terms as Medium-complexity, High-complexity single operator or High-complexity multiple operator.

Table 2-6 shows a summary of the CWA literature reviewed. The table indicates whether an analytic product is suggested for a particular SLC phase (+) or whether empirical evidence indicates that it is suitable (++). The column at the far right is a product of the other columns. Therefore, the table represents the results of an assessment, based on the research available, of what the most suitable analytic product is. Blank cells indicate that there is no information to support the use of that analytic product, not that it cannot be used. Cells that are merged indicate that the research reviewed did not identify the specific SLC phase and that the analytic product is likely to be suited to either SLC phase. The table clearly indicates that although researchers have claimed that all the CWA phases are applicable to all the SLC phases, empirical research has focussed on three of the five CWA phases and the two early SLC phases: Preliminary system studies and Concept Development.

The empirical research has shown that the following analytic products are most appropriate for this program of research:

- Abstraction Hierarchy and Abstraction-Decomposition Space for the Work Domain Analysis;
- Activity Analysis in Work Domain terms, Temporal Coordination-Control Task Analysis and Decision Ladder for the Control Task Analysis;
- Information Flow Maps, Decision Ladder and undefined charting analytic products for the Strategies Analysis.

Table 2-6 Applicability of CWA products to system life cycle phase

Constraint-based Products by System Life Cycle[3]	Area of investigation												Mapping of CWA products and the SLC Phases			
	Research concerning complex systems				Research applying CWA analytic products to system development and design				Research reporting on the use of CWA for evaluating systems in general (SLC Phases deduced by the author)							
	P	C	D	U	P	C	D	U	P	C	D	U	P	C	D	U
Work domain Analysis (WDA)																
Abstraction hierarchy (AH)	+	+	+	+					++				++	+	+	+
Abstraction-Decomposition space (ADS)	+	+	+	+					++				++		+	+
Control Task Analysis (CTA)																
Activity Analysis in work domain terms (AA/WD	+	+	+	+					++				++	+	+	+
Temporal coordination control task analysis (TC-CTA)	+	+	+	+	+	+			++				++			
Decision ladder (DL)	+	+	+	+					++				++		+	+
Strategies Analysis (SA)																
Information flow maps (IFM)		+	+	+					++				++		+	+
Decision ladder (DL)									++				++			
Charting techniques (CT)									++				++			
Social Organization Analysis (SOA)																
Abstraction hierarchy		+	+	+	+								+	+	+	+
Abstraction-Decomposition space		+	+	+	+								+	+	+	+
Activity Analysis in work domain terms		+	+	+	+								+	+	+	+
Temporal coordination control task analysis		+	+	+	+								+	+	+	+
Decision ladder		+	+	+	+								+	+	+	+
Information flow maps		+	+	+	+								+	+	+	+
Social network analysis (many types)									+				+			
Link analysis (many types)									+				+			
Worker Competencies Analysis (WCA)																
Skills, rules and Knowledge		+	+	+	+	+							+	+	+	+

[3] (++= empirical evidence, += proposed).

In the remainder of this section I provide a summary of the literature cited in the table. Looking at the table the literature falls into three main categories. The first category is research concerning complex systems that identifies the relationship between System Life Cycle phase and CWA. The second category is research that reports on the application of CWA analytic products to system design and evaluation. The third category reports on a variety of research using CWA for evaluating a system in general.

2.3.2.1 Research on complex systems

The research reviewed in this section is concerned with the evaluation of systems that are broadly equivalent, in complexity terms, to our test case system (a RWR used on a Black Hawk helicopter). The research reviewed here also identifies the CWA-based analytic products that are suitable for different SLC phases.

Research on the application of CWA to the concept development phase (a SLC phase) of the design of a next generation (future) US Navy surface combatant ship was carried out by Bisantz et al (2003, 2001). The focus of their work was to support the design of watchstander tasks, functions, and support systems in the bridge and combat command centre, located onboard ship. The authors report that they were able to support the design by producing and using an abstraction hierarchy (AH), decision ladders (a product of CTA) and other non-CWA products.

Other research has focussed on the on the application of WDA evaluate design proposals for complex systems. For example, Naikar and Sanderson (2001, 2000a, 2000b) describe how WDA was used, in part, to assess a number of manufacturers' tenders for an Airborne Early Warning and Control (AEW&C) aircraft for the Australian Air Force. The WDA (using the ADS product) produced by the authors guided the Australian Operations and Technical Tender Evaluation Working Groups to evaluate tenders at different levels of abstraction. Using the WDA and CTA in this was seen to complement the normal tender evaluation process and was deemed to be useful by the Operations and Technical Tender Evaluation Working Group.

WDA (using the ADS product) has also been used to produce a training simulator for combat aircraft (Naikar and Sanderson, 1999). This research extends the use of WDA to provide functional specifications for training systems in general.

2.3.2.2 Research applying CWA analytic products to system development and design

The research reported in this section shows the relationship between the CWA analytic products and the SLC. Sanderson et al (1999) note that CWA supported all system development and evaluation phases (these may be broadly mapped onto the SLC phases used in this thesis. The relationship between the use of the CWA analytic products and the system development and evaluation phases is shown in Table 2-7.

Table 2-7 Suggested use of the CWA analytic products during SLC activities (from Sanderson et al, 1999)

SLC Activities	CWA Phase				
	WDA	CTA	SA	SOA	WCA
Requirements Specifications	Support	Support			

SLC Activities	CWA Phase				
	WDA	CTA	SA	SOA	WCA
Design	Support				
Simulation	Support	Support	Support	Support	Support
Evaluation of designs	Support				
Implementation	Support				
Test	Support	Support	Support	Support	Support
Operator selection					Support
Operator training	Support	Support	Support	Support	Support
Routine use	Support	Support	Support	Support	Support
Non-routine use	Support	Support	Support	Support	Support
Maintenance	Support	Support	Support	Support	Support
Research (HF studies)	Support	Support	Support	Support	Support
Upgrades	Support	Support			
System retirement	Support	Support			

Table 2-7 is a summary of Sanderson et al (1999) and shows that the CWA analytic products may offer some support to all the SLC activities. For example, WDA analytic products may support the SLC requirements gathering activity. In their paper, Sanderson et al specify the actual analytic product that is deemed to be useful for each activity. These are listed here:

- ADS (a WDA product), supported "requirement development",
- Activity Analysis in Work Domain Terms (AA/WD) (a CTA product) supported "system specification development",
- ADS (a WDA product)supported "design",
- all analytic products from CWA supported "simulation",
- ADS (a WDA product)supported "evaluation of designs",
- ADS (a WDA product)also supported "implementation",
- analytic products from all CWA phases supported "test",
- a WCA product (not specified by the authors) supported "operator selection",
- all analytic products from all CWA phases supported "operator training",
- all analytic products from all CWA phases supported "routine, non-routine, and maintenance activity",
- all analytic products from all CWA phases supported "research (Human Factors)",
- ADS (a WDA product)and the AA/WD (a CTA product) supported "upgrades", and
- ADS (a WDA product) and AA/WD (a CTA product) supported "system retirement".

Other authors have expressed concerned that ISO13407 (ISO, 1999) on Human Centred Design Process (ISO) provides little guidance to cognitive engineers on the methods and analytic products that are appropriate to meet the ISO13407. Hori et al (2001) built on work by Sanderson et al (1999) and studied three approaches for system design; ISO13407 on Human Centred Design Process (ISO), a representation of the System Life Cycle (following that presented by Sanderson et al, 1999), and Cognitive Work Analysis. The authors used a

case study approach and specifically evaluated whether the CWA phases could be used to provide the information required by ISO. Based on their analysis the authors proposed that the CWA phases be used to develop system models (products) and that the models would provide information for engineers to comply with ISO requirements. The advice offered by the authors was that the WDA, CTA and WCA could support the ISO process, "understand and specify the context of use"; that WDA, SOCA and WCA could support the ISO process, "specify the user and organisational requirements"; that WDA and CTA could support the ISO process, "produce design solutions"; and, that all of the CWA phases could offer limited support to the final ISO process, "evaluate design against requirement". The actual products used were: for the WDA, Abstraction-decomposition space (ADS); for the CTA, decision-ladder; and for the SOCA, decision ladder. No unique product was used for the WCA.

2.3.2.3 Research reporting use of CWA for evaluating systems in general

CWA has been used to evaluate a variety of systems including medical, and air traffic control systems. In early work concerning the impact of technology Benda and Sanderson (1999, 1998) proposed the use of WDA and CTA to describe and predict the impact of technology on work practices. In their work the authors used WDA and CTA to analyse the impact of a new automated anaesthesia record keeping system when it was introduced into the operating rooms of a university medical centre. The WDA product they used was an abstraction hierarchy (AH) and the CTA product was an Activity Analysis in Work Domain Terms (AA/WD) – a precursor to the temporal coordination CTA. The authors mapped several examples of how the introduction of the anaesthesia system affected the work practices of the operators (taken from a published review of the introduction of the system) onto their WDA and CTA. Their analysis showed that WDA and CTA could be used to show how technology limited behaviour (represented in the CTA) and how the many possible effects on the work domain (represented in the WDA as relationships between objects, functions, values and priorities and purposes) could be visualised. The authors noted that further work including empirical testing of the approach was needed. WDA (represented as an AH) and CTA (represented as decision-ladders) and SA (represented as two charting techniques) have been used during an analysis of cardiac care nurses performing teletriage (Burns et al, 2009).

Several of the CWA phases have been used to analyse (a precursor to evaluation) a simulation of an air traffic control environment. Kilgore et al (2009) conducted a restricted CWA of a PC-based microworld simulation of an air traffic control environment - TRACON. The motivation for conducting the analysis was pedagogical. In their analysis they analysed the system using all the CWA phases and used the following product: WDA (represented as an AH); CTA (represented as a decision ladder); SA (represented as Information Flow maps, IFM); SOA (represented as IFM); and WCA (represented as Skills, Rules and Knowledge, SRK).

Some authors have suggested that information gained by using CWA products should be supplemented with other products from other disciplines. For example, Pfautz and Pfautz (2009) argue that there has been little guidance on how to conduct a SOA and how that information should be represented. The authors contend that guidance is needed if the result of an analysis of a system is to be successful. The authors argue that the CWA products should be supplemented with products from disciplines such as organisational

psychology, social science, management science and cultural anthropology to provide that guidance. Products from the other disciplines considered to be useful were social network analysis methods and link analysis. In conclusion, the review of the literature has identified several CWA analytical products that should be used as a basis for the constraint-based measure-selection method.

2.3.3 Summary

The goal of this section was to identify the analytic products that should form the basis of the measure-selection methods. The task-based perspective, being the dominant approach to complex system evaluation, has benefited from an extensive history of use that has resulted in a useful mapping of analytic products to system complexity and SLC phase. The constraint-based perspective has a much shorter history of complex system evaluation. By considering research, however, it is possible to map appropriate analytic products onto the SLC phases. The mapping reveals that empirical research has focussed on medium to high complexity systems and has found that WDA and CTA are best used for system evaluation during the Preliminary and Concept Development phases of the SLC.

It was found that the analytic products that should form the basis of the task-based measure-selection method are:

- Narrative Mission Descriptions during the Mission Analysis phase.
- Function Flow Diagrams during the Function Analysis phase.
- Ad Hoc Function Allocation during the Function Allocation phase.
- Timelines during the Task Analysis phase.

It was also found that the analytic products that should form the basis of the constraint-based measure-selection method are:

- Abstraction Hierarchy and Abstraction-decomposition Space for the Work Domain Analysis;
- Activity Analysis in Work Domain terms, Temporal Coordination-Control Task Analysis and Decision Ladder for the Control Task Analysis;
- Information Flow Maps, Decision Ladder and undefined charting analytic products for the Strategies Analysis.

Given that suitable analytic products have been identified, the question of how to use them to select appropriate measures to be used in the system evaluation activity remains. That question is addressed in the next section.

2.4 Methods used for selecting measures

In this section I will identify the measure-selection methods that use the analytic products described previously. I will also describe the criteria that are used to judge whether the

method is appropriate for use during system evaluation. First, the task-based measure-selection method that has been typically used during the HEP will be reviewed. I will show, amongst other things that the task-based measure-selection method is based on the application of guidelines for the selection of measures and although some recommendations have been made to improve the method for selecting measures the HEP does not require those improvements. Second, the constraint-based measure-selection method will be reviewed. I will show that a method does not currently exist but one can be produced that conforms to the constraint-based perspective.

2.4.1 Task-based measure-selection-method

The goal of this section is to present a task-based measure-selection method that can be compared to a constraint-based measure-selection method. There is no formal task-based measure-selection method (as mandated by a formal standard) to be used during system evaluation. The informal method that has been observed in the literature involves the analysis of system tasks and the application of general guidelines to select the most sensitive measures out of alternative ones (examples of the guidelines include identify measures that are "reliable" and "not intrusive"). The development of the measure-selection method has been influenced by two main factors. The first factor is the potential impact of analyst experience on the choice of the most appropriate measure. Analysts have developed criteria to achieve objectivity in selecting measures. The second factor is the generalisability of laboratory-based results (that are based on measures suited to those conditions) to operational settings. Analysts have identified factors that influence the choice of measures that support better generalisability to operational settings. These developments will be briefly reviewed.

2.4.1.1 Criteria of acceptability

The selection of measures based on the experience of the analyst has been cited as an area of concern during system evaluation. For example, Charlton and O'Brien (1996) point out that measures used for system evaluation are "often selected solely on the basis of the expertise and experience of the individual tester" (p21), and although this may not result in the wrong measure being selected it does "open the door to selection of measures out of familiarity and convenience rather than careful consideration of the system to be tested" (p21). The authors argue that in order to remove the potential impact of the analyst's experience with certain measures, measures should be selected on the basis of the functions and tasks of the system being evaluated. Once done, measures should be selected using generic criteria or guidelines to select between potential measures.

Meister (1999) uses a number of criteria to select amongst a set of prospective measures once an analysis of the system tasks have been completed. The criteria including the following:

- reliability – will the same result be found if the measure is repeated during the same conditions?
- validity – does the measure really measure what it is meant to measure?
- detail – does the measure reflect performance at the level of granularity that is meaningful?

- sensitivity – does the measure reflect the change in conditions?
- diagnosticity – does the measure discriminate between operator capabilities?
- intrusiveness – does the measure affect the task being performed?
- requirements – what system resources does the measure require?
- acceptance – will test personnel tolerate the measure?
- objective – will the measure be moderated by the researcher and quantitative – is the measure able to be recorded numerically?
- cost - what is the cost of using the measure?

Other authors have included other criteria. For example, Gawron (2000) provides a handbook of human performance measures that can be used for system evaluation. In the handbook the reader is directed to a number of criteria that should be used when a measure is considered for use. These criteria are similar to those of Meister and include: reliability, validity and quantitative, and also include relevance and comprehensiveness. Relevance refers to whether a measure is relevant for the research question being asked. Comprehensiveness refers to the fact that it is better to measure many dimensions of performance during one experiment than to repeat the experiment measuring separate dimensions on each occasion.

The use of guidelines is not the only method of removing the potential impact of analyst experience. Once measures are selected using guidelines some authors advocate that they should be tested before they are used in the evaluation exercise. For example, in a two step process, Muckler and Seven (1992) use criteria to select measures but they also test the measure before it is used.

The first step in Muckler and Seven's (1992) process identifies what needs to be measured. It entails analysing the task functions and dimensions that are to be measured through task analysis and various analytic products. A parallel activity is to understand exactly what the analyst needs to know and what needs to be measured in order that the analyst gets the information that he or she is seeking.

Once the measurement dimensions and information needs have been met, several methods of measuring the dimensions are then proposed by the analyst. The methods are then compared and evaluated against each of the following eight criteria: validity; reliability; precision (detail); non reactivity (lack of intrusiveness); resources; relative simplicity (the most relevant or critical measures that are easily interpreted and defined should be chosen); generalisability (a goal may be to develop a set of generalisable measures); and multiple criteria (measures could be chosen on the basis of multiple criteria rather than just one).

The second step in Muckler and Seven's process looks at how the selected measure will be evaluated. The goal is to define, select and test the final set of measures. Clearly a number of different sets of measures may be developed that could be applied to the system in question. The sets of measures would then be traded off against each other on the four following areas:

- Information –the sets of measures should provide useful and needed information
- Instrumentation and data processing requirements – such requirements should be considered to ensure that the data can actually be collected and processed
- Cost effectiveness – the cost-effectiveness of each set of measures should be considered
- Credibility of the measure sets –the set of measures should be suitable for their purpose from the perspectives of all involved.

From the review of the literature given in this section it is clear that the informal task-based measure-selection method is based on the application of guidelines. These guidelines are designed to make the selection of the measures objective.

2.4.1.2 Measures for generalisation to operational settings

Some authors have voiced concerns about whether laboratory-based measures are generalisable to operational settings. Some of these (e.g. Kantowitz, 1992) argue that generalisability would be better if theory were used to select the measures. Others (e.g. Hennessy, 1990) argue that subjective data should be used rather than objective data in operational environments.

Kantowitz (1992) argues that if human factors research is to be useful in operational settings then the results obtained from laboratory studies (e.g. simulator studies) should be more generalisable than "pure" research. This is done by selecting measures using theory. In his view there are two ways in which measures are traditionally selected. The first is to use SMEs to pick variables they regard as being important. The second is to use theory to select the variables studied. Kantowitz argues that there are five generic benefits of using a theory to solve human factors problems in operational settings. These follow below and will be further discussed in Chapter 9.

- Theory allows for interpolation of results where data cannot be collected, or where limited data points can be gathered. A theory helps analysts predict how a variable will affect results.
- Predictions based on theory can be used in the design of a system, before the system is built.
- By using theory, analysts may recognise similarities across a range of practical problems so that time is not wasted "re-inventing the wheel".
- Theory can be used as a means for representing normal human behaviour or system performance. In this way, cost-benefit relationships may be ascertained.
- Theory can be used to aid measurement and system design.

Even though authors like Kantowitz have advocated the use of theory for evaluating a system, including selecting measures, the approach has not been taken up by the wider practitioner community. A review of the literature reveals that even though military standards and guidelines stress the importance of integrating Human Engineering (HE) methods with systems methods during the design and evaluation of complex systems, the treatment of the area of performance measure selection is limited. The advice given,

generally, is to select measures of performance on a task/function basis (i.e. understand what the crew activities are) and select ways to measure how well the crew performs that activity. For example, Advisory Publication 61/116/12, Crew Performance Measurement (1996), which applies HE to design, development, test and evaluation of aircrew systems, states "*There is no general theory to guide Crew Performance Measurement and to relate behavioural processes within the individual to performance of the task, and to link task to total system performance*" (p.13). The document lists a range of crew performance measures (such as reaction time), but only gives general guidance on their selection and utility.

The usefulness of applying laboratory–based experimental protocols in operational settings has been questioned by Hennessy (1990). Hennessy comments on the validity of the laboratory based, or classic empirical (experimental), approach for system evaluation. He notes that often this approach is used in operational settings (e.g. field or simulator) as a means to understand human performance. Here, measures are controlled, and "objective" data are collected and analysed in a way that, as Hennessy notes, "is logical and understood by all". However, there are some limitations to this approach, which according to Hennessy limits its usefulness.

The first of the limitations that Hennessy (1990) notes concerns pressures from the decision-maker (often the sponsor of the research program) that force data collectors to adopt a classic empirical approach and the relationship between the decision-maker and data collector. Sometimes there is a difference in what the decision-maker expects of the method and what the method will actually produce. For example, the decision-maker expects quantitative data to be collected yet constrains the experimental design through time and resource limitation. The data collector will tend to yield to the decision-maker's requests and therefore designs, gathers and analyses data in a way that meets the decision-maker's requirements.

The second of the limitations that Hennessey (1990) notes concerns the accuracy of the results gained from the traditional empirical approach. Hennessy argues that several problems emerge when the laboratory approach is operational to field-testing: data are missing; part of the data variability is due to unknown factors (not identified and controlled in the design); statistically low-level variables are identified but found to be insignificant; objective measures are difficult to interpret in terms of the test questions; subjective ratings and observational data tend to pick up differences in effectiveness; and conclusions tend to be based on subjective measures, but arguments are found by researchers to show that the objective measures show the same conclusions.

The third of the limitations that Hennessey (1990) notes, which is related to the second, is that objective data collection tends to be regarded as more valid and reliable than subjective data collection even though other authors (e.g. Muckler and Seven, 1992) have identified subjectivity in all stages of experimental design and analysis. In addition, the objective data are usually the output of the system, such as time of button selection, or some other low-level item that can be automatically sampled. Hennessy argues that in this case the data point collected (e.g. button selection) may be at the end of a human process that is invariably not captured. He also points out that there is a major problem of aggregating low level data (e.g. button selection) into something meaningful.

The limitations of objective methods point to the need to use alternative methods. Hennessy argues that more subjective, observational, methods should be adopted. He argues that although both objective and subjective data collection have inherent errors associated with them, the cumulative error associated with subjective data is smaller than that associated with objective data. In addition, Hennessy sees several advantages that subjective data collection has over automated data collection. He argues as follows: gathering subjective data is more economical in terms of time and cost; subjective data are seen to directly measure performance of interest; the context can be taken into account; small differences in conditions can be picked up; the results are quickly available; concurrent tasks can be recorded by observer; and cognitive tasks can be measured. Although Hennessy argues strongly for the use of observational data collection, he notes that automated data gathering should be used in conjunction with observational data, wherever possible. The timing of significant events is still of benefit, for example.

Although he does not specify the exact mechanism, Hennessy argues that human performance measurement can be facilitated through a three-stage process of developing a performance measure hierarchy, obtaining performance measure weightings and videotaping the subject trial. A performance measure hierarchy would not only show that lower level measures can combine to form higher level measures, but also show how (the rules by which) the measures combine. Once the hierarchy has been developed the weights of the components in the performance hierarchy can then be allocated by using subjective methods such as policy capture techniques. Here, weights are assigned to factors that are seen to be of importance in the decisions made by experts. Videotaping (including sound) system usage is advocated as it provides a means to re-visit the test to extract new data, or validate old data. Hennessy provides a Human Performance Measurement Model that encompasses the development of the performance measurement hierarchy, assignment of weights and the use of videotaping. He also provides an example of its use. However, the model was never empirically tested.

In conclusion, a formal task-based method selection method does not exist. The informal method involves analysing the human-machine tasks and choosing measures that meet a number of generic guidelines. Potential improvements to this method involve using theory to guide the selection of measures, pretesting the measures on a number of dimensions and making better use of subjective measures. To date these improvements have not been formally tested or mandated in existing formal standards. In terms of the implications for this thesis the lack of a formal task-based measure-selection method means that the method adopted to compare against the constraint-based method must be at least representative of what system evaluators would use and be world's best practice. The task-based measure-selection method that was developed is shown in Chapter 6.

2.4.2 Constraint-based measure-selection method

The goal of this section is to present a constraint-based measure-selection method that can be compared to the task-based measure-selection method. In this section I will show that there is little research from the constraint-based community directly aimed at the problem of choosing measures. I will also show that a constraint-based measure selection method does not exist that is suitable for operational settings but that a theoretical framework can be identified and can be used to develop such a method.

Some authors have argued that measures that can be used in system evaluation may be attributed to different types of constraint. For example, Vicente (1999) proposed that each CWA phase be used to determine different types of measures. Vicente proposes that WDA helps to define variables that identify the state of the work domain. For example, one measure could be how close the work domain is to exceeding a safety boundary. He notes that CTA could be used to identify measures associated with what subjects do. For example, measures could be task completion time, number of errors, number of control actions, number of worker verbalisations and non-verbal behaviours. SA can be used to identify measures that describe how subjects do what they do. For example, measures could emerge from eye movement and verbal protocol data. SOA can be used to identify measures of team or group communication and cooperation. For example, measures could relate to the direction and frequency of communication between multiple actors. WCA can be used to identify measures associated with the subject's level of expertise. For example, measures could include mental workload and situation awareness.

Other authors have stressed the importance of developing novel human performance measures that are system specific. For example, building on Vicente (1999), Xinyao, Lau, Vicente and Carter (2002) use a laboratory task to illustrate how the Abstraction-Decomposition Space (ADS) representation of a WDA can produce "novel" human performance measures that not only conform to measurement criteria (objectivity, quantitative and sensitivity) but also are theoretically grounded. Xinyao et al note that in process control environments operators have difficulty dealing with unforeseen system errors. The difficulty is due to the operator "knowing where they need to go (error correction), but not how to get there". The ADS describes the same system but in different terms (abstraction) and levels of detail (decomposition). By overlaying onto the ADS the pathways or trajectories that operators take to "reach the error reduction solution", Xinyao et al. were able to trace the operators' heuristic decisions for error correction. By analysing the variance of the trajectory and comparing the variance of the trajectory between subjects, the authors could demonstrate the relationship between both subjective and objective results - a relationship that they had not detected with traditional performance measures. Hence, the ADS provided a framework to develop theoretically grounded and novel human performance measures.

The theoretical position by Hajdukiewicz and Vicente (2004) is particularly interesting because although it does not focus on the issues surrounding selecting sensitive measures per se, it does clearly recognise that the functions and purposes of the WDA could be defined in quantitative terms and so be useful for defining the state of the work domain.

Some authors have used constraint-based products to identify measures that can be used to judge whether a system meets its design aims. For example, Naikar and Sanderson (1999) in their paper on the use of WDA to produce a training simulator propose that a WDA can be used to identify "measures of performance"[4] that can be used to test whether the purpose(s) of the simulator has been realized. The measures that they identified were primarily associated with the priorities and values level of the AH. The simulator would be judged as meeting the purposes if the measures of performance, identified at the

[4] Quotation marks are used here to reflect Naikar and Sanderson's use of the term measure of performance. Their use is at odds with the definition used in this thesis.

priorities and values level, were achieved. In addition, the authors note that data from the other (lower levels) levels could be used to influence the measures of performance.

More recently Jenkins et al (2009b) argue that WDA can also be used to derive measures of performance to assess the impact of technological change. In their work the authors use expert judgement to identify the effects that a technological change (at the lower levels of the AH) would have on upper levels of the AH. The measure that the authors identified as being indicative of system performance was concordance. Effectively this means that the performance of a system was assessed as "better" than another system if the relationship between two objects, functions, values and priorities or domain values was judged as being stronger (by a stakeholder or stakeholders). Similarly, the performance was assessed as "worse" if the strength of the relationship was reduced or absent. Using a case study method the authors were able to show that expert judgement could be mapped onto the WDA they had produced. The resulting product could then be used to assess the potential impact of new technology. In their paper the authors identify that their work was subject to further research.

Building on Vicente's theoretical position Crone et al (2003) proposed that a method for selecting measures of performance and measures of effectiveness for evaluating complex socio-technical systems could be developed using constraint-based analytic products. The specific method that was developed is described later in Chapter 6. However, in the next few paragraphs the theoretical framework that forms the core to the method is described. The theoretical framework is described in two stages: first, the products of a WDA and CTA are expressed in terms of measures of performance and measures of effectiveness and second, the WDA and CTA products are integrated.

Figure 2-7 shows the result of describing the WDA in MOP and MOE terms. The left of the figure shows the labels used for each of the five levels used in the AH. This part of the figure is consistent with the description of an AH presented in Section 2.2.2 in that each layer of the AH may be used to describe the same work domain but in different terms. Additionally, implicit in this part of the figure (but not shown) is the idea that by decomposing objects, functions and purposes into properties a high level of system detail may be achieved and that each of these objects, functions and purposes may be parameterised. The right side of the figure is the result of considering each of the levels in terms of whether they represent measures of performance or measures of effectiveness - remember MOPs are related to the physical properties of a system and MOEs are related to criteria for assessing whether the system has been effective for the purpose for which it was designed.

Figure 2-7 shows that properties of physical objects may be mapped onto system performance measures; properties of physical functions may be mapped onto system function measures; properties of the domain functions may be mapped onto mission function measures; properties of the domain priorities and values may be mapped onto mission priority measures and finally, properties domain purpose may be mapped onto mission effectiveness measures.

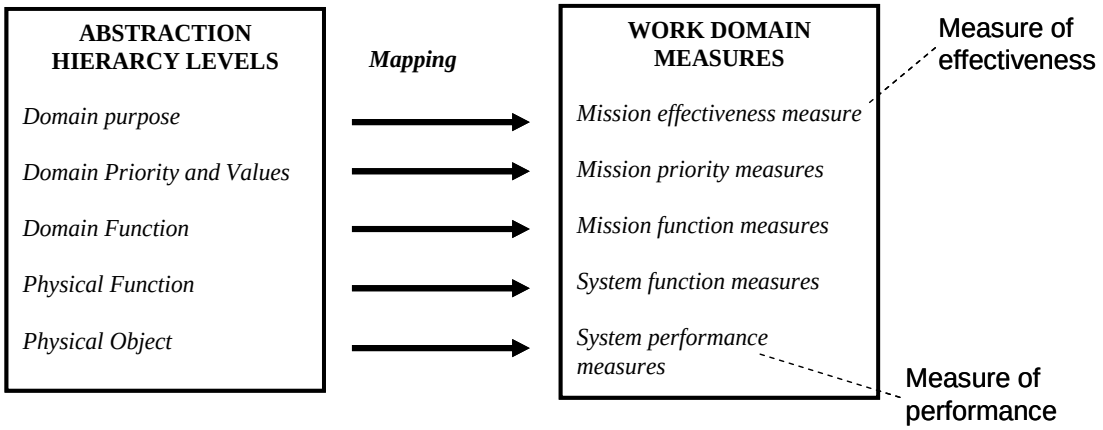


Figure 2-7 Mapping work domain abstraction levels to domain measures

The material presented in Section 2.2.2 has also indicated that the temporal coordination CTA may be used to assess the impact of system performance on task-specific properties and task-generic properties. As with the AH, here the task properties may be mapped onto different classes of measures; Task-generic properties may be mapped onto measures applicable to all tasks, and Task-specific properties may be mapped onto measures specific to a task (Figure 2-8).

The theoretical construct that the above mapping represents is concerned with the nature of task properties. Vicente (1999) considers that control tasks represent "the goals that need to be achieved, independently of how they are to be achieved or by whom" (p183). Typically a control task is seen to be a "black box" and is defined in terms of the inputs to the task, the outputs achieved after the task has been completed, and the constraints that govern the task. Vicente also attributes measures to control tasks that describe what subjects do. Hence, it is clear that there are measures associated specifically with tasks (what subjects do) and measures associated with the constraints on the task. In our case, task specific properties (what subjects do) can be mapped onto measures specific to a task, and task generic properties (constraints on the task) can be mapped onto measures applicable to all tasks. This distinction will be used to explain the different types of measures used in the experiments in Chapters 7 and 8.

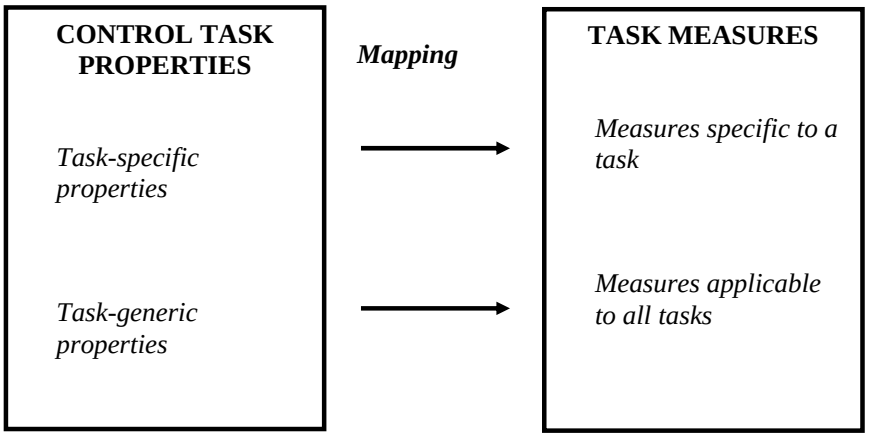


Figure 2-8 Mapping control task properties to task measures

The integration between WDA and CTA is important for developing a measure-selection method. The general relationship between WDA and CTA is illustrated in Figure 2-9. This figure indicates that the system properties, represented in the WDA, afford various tasks, as seen in the temporal coordination CTA. This means that a change in a work domain property will affect a control task property. In addition, the figure shows that the control tasks themselves will act on the work domain of interest. In other words, the performance of an activity is dependent on the environment in which it takes place.

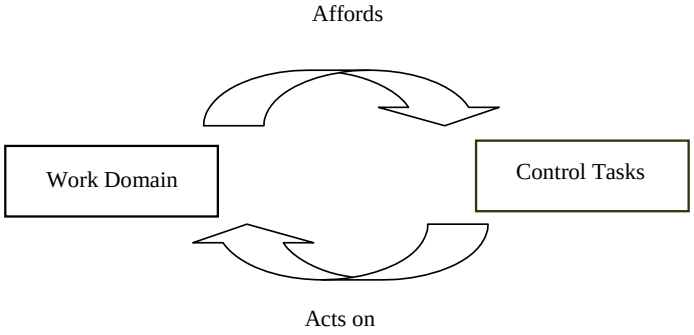


Figure 2-9 Relationship between the work domain and control tasks (modified from Vicente, 1999)

Given that various measures of performance and effectiveness can be mapped onto an AH (Figure 2-7) and temporal coordination CTA (Figure 2-8) and given that a theoretical relationship exists between Work Domain and Control Tasks (Figure 2-9) it is possible to construct an integrated framework of classes of measures. Figure 2-10 is that framework.

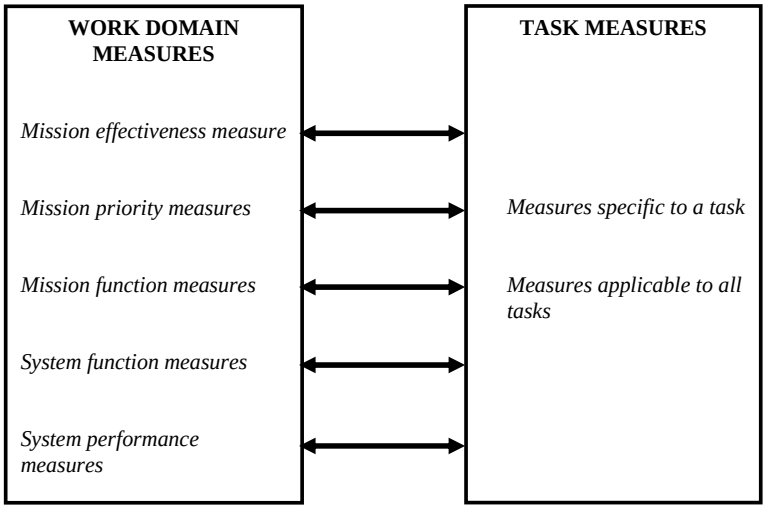


Figure 2-10 Integrated framework of classes of measures

The framework shows that the various classes of measures are related. From the figure one would expect to see that a change in property of the system (seen in the work domain) would cause a change a property of a task and vice versa. For example, if an analyst is interested in evaluating the system independently of the activity he or she would use the left side of the framework (work domain point of view). However, if the same analyst is interested in the relationship between the work domain and activity (between system performance and human or system behaviour) the right side of the figure can be

incorporated. In this example, the analyst will move from system performance measures (work domain side) to measures specific to a task and measures applicable to all tasks (activity domain side). Using this framework the analyst can explore the effect of changing a system from different levels of abstraction and from different task specificity depending on the purpose of the evaluation.

The framework is important for a number of reasons. First, it provides a means to integrate the work domain and the activity domain in a way that is theoretically valid. Second, the framework provides a number of propositions that are testable. For example, one proposition is that system performance measures are "related" to measures specific to a task. Third, using the framework, together with the AH and temporal coordination CTA, an analyst should be able to select measures of performance that are important for system evaluation. In other words the framework provides a basis for a method to ground the selection of measures a priori; the method is not based purely on guidelines or experience. Fourth, the framework is generalisable to any complex socio-technical system. Fifth, it provides a structure that can be used to develop and codify a method for selecting measures. Finally, the framework provides mechanisms to distinguish between measures of performance and measures of effectiveness.

In conclusion, researchers in CWA have only recently considered the problem of measure selection. Vicente suggest that each CWA phase may be used to derive measures that are specific to different types of constraints. Other research has focussed on measures applicable to microworlds. Only recently has the problem of measure selection been considered in complex operational settings. However, analysis of that research reveals that although some measure-selection methods have been suggested they require empirical testing by their authors. The framework proposed by Crone et al (2003) is important for several reasons. One of the more important reasons is that it provides a basis for a measure-selection method that can be developed and empirically tested. The constraint-based method developed from the framework is given in Chapter 6 and is based on the WDA and CTA phases of CWA.

2.4.3 Testing which measure-selection method is more effective

The goal of this section is to describe, in general terms, how the task-based and constraint-based measure-selection methods will be compared. Figure 2-11 illustrates the relationship between measure-selection methods and the population of actual measures. The circle indicates the limit of the comparison. It can be seen that the comparison will involve measures from WDA and CTA and some measures from the task-based methods. Measures from the other CWA phases (i.e. WCA, SA and SOA) and some task-based measures (e.g. workload measures and situational awareness measures) are not considered. The choice of what task-based measures to include was derived from the work by Vicente (1999) as discussed in Section 2.5.2. The implication for the results of this program of work of excluding some task-based measures is discussed in Chapter 9.

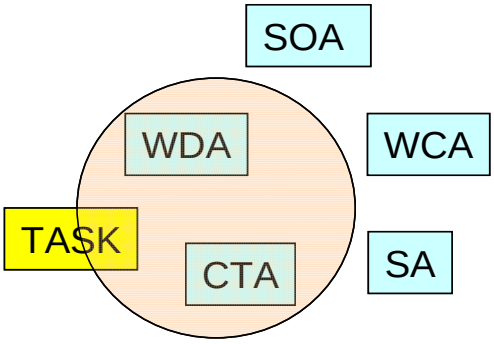


Figure 2-11 The research compares the measures from the constraint-based WDA and CTA with task-based measures. The circle indicates the boundary of the comparison.

2.5 Conclusion

This chapter has described two main approaches for complex system evaluation: a task-based approach and a constraint-based approach. In the task-based approach the task-based perspective to work analysis has been adopted in the Human Engineering Process (HEP). The HEP provides general guidance to the analyst on what data collection methods and analytic products are suitable for evaluating the system. The choice of analytic products is constrained by the System Life Cycle (SLC) and the complexity of the system. The measure-selection method that the analyst adopts consists of the application of a set of guidelines.

In the constraint-based approach the constraint-based perspective has been adopted into the CWA framework. Like the HEP CWA provides guidance on what data collection methods and analytic products are suitable for evaluating the system. Unlike the task-based perspective there is no established measure-selection method for selecting measures for system evaluation in operational settings. A framework that can be used to develop a method has been proposed by the author.

Now that more future systems are being procured, the task-based approach for system evaluation has come under increasing pressure. Although task-based and constraint-based approaches have been compared on various dimensions, there has been no empirical research comparing them on their ability to produce sensitive measures of performance and effectiveness. The choice of these measures is a critical step in system evaluation.

There are several significant implications of the research reviewed above when comparing task-based and constraint-based measure-selection methods. First, the comparison must include an evaluation of a complex socio-technical system, because I am interested in the helping analysts in their role in informing procurement decisions. Second, the two measure-selection methods should be compared while evaluating modifications to both a current system and a future system, again because this is typical of the work that the analysts do and because one of the claims made against the task-based approach is that it does not cater well for novel human-system behaviour seen in future systems. Third, the evaluation should be of an operational system that is used in the field. This is because such environments are typically being used in recent evaluation of systems. Fourth, the research

must reflect the evaluation processes seen in actual procurement. This can be achieved by selecting a SLC phase which in turn would identify the data collection methods and the analytic product suitable for the measure-selection methods. Fifth, the two measure-selection methods should be compared on whether they produce measures that are sensitive to a system modification, because this is a logical requirement for measures. Sixth, sensitivity should be defined in statistical terms. In this way the results of the comparison should be unambiguous. Finally, the two methods should be compared on their suitability for use for system evaluation in operational settings. This is because there is some evidence to suggest that the measures selected using the task-based measure-selection method are not suitable for operational settings. In the next chapter the test system used to compare the two methods is described.

3. Test System Used for the Research

Chapter 2 showed that an important part of evaluating a system is selecting sensitive measures of performance and effectiveness. The chapter described two measure-selection methods. The task-based measure-selection method for system evaluation is based on current international best practice for human engineering, and is widely used in both laboratory and operational settings. However, the task-based measure-selection method has been criticised on the grounds that it is based on the application of guidelines that may not be appropriate for future, first-of-a-kind, systems. A constraint-based measure-selection method has not been developed for use in operational settings. However, the constraint-based approach, embodied in CWA, has a strong theoretical basis and proponents claim that it should be uniquely useful for evaluating future systems in operational settings.

As shown in Figure 1-1, in this thesis I will develop a method for selecting measures to be used in system evaluation. Stages 2, 3, and 4 of the method are conducted in the context of one particular socio-technical system. Before describing the method for selecting measures any further, I describe the test environment.

In the remainder of Chapter 3 the work system used for the study is presented. The purpose and functioning of the system are described and an assessment is made of it as a "complex socio-technical system". The simulation environment for the system is also described. In the next chapter, Chapter 4, the aim and research questions are stated and the method to answer the questions is presented.

3.1 Test case and test environment

Chapter 2 showed that for a valid comparison of the measure-selection methods, a test case is needed that is classified as a complex socio-technical system, that is suitable for evaluation, and that is representative of a real system (to ensure that any behaviour observed is valid). A test case that meets all these requirements is a Radar Warning Receiver (RWR) operating in a Black Hawk Helicopter during Airmobile Operations

(Patrol Insertion). In addition, because a goal of the thesis is to replicate current evaluation processes for assessing systems it is important to consider the phase of the System Life Cycle (SLC) in which the evaluation takes place. The SLC phase is important because it will govern the types of analytic products, data collection methods and resultant measures that are produced (as seen in the previous chapter). By applying Beevis' (1999) system of classification, the RWR was found to be a medium complexity system. The SLC phase for the system evaluation that was chosen is the Preliminary Definition Phase. The Preliminary Definition Phase was chosen because it is the first stage at which system evaluation is conducted using simulators, i.e. this is the first time that the evaluation is conducted in an operational setting. Essentially, for the purpose of testing the measure-selection methods, the system evaluation question I will pose is the following: Will the modification to an RWR result in better mission effectiveness when compared to a RWR that has not been modified? The research question then is whether measures derived from task-based approaches are more or less sensitive to any changes to mission effectiveness with the RWR than are measures derived from constraint-based approaches. In the following sections the main characteristics of the RWR and how they are emulated in the simulation environment are described.

3.1.1 Black Hawk Helicopters, RWRs and Airmobile Operations

The Radar Warning Receiver (RWR) is a system that is fitted to the Black Hawk helicopter and used in Airmobile Operations (Patrol Insertion). RWRs are used in Black Hawk helicopters and other military aircraft to detect radar emissions from potentially hostile weapon systems. The hostile systems use radar to detect an aircraft and then launch a weapon against that aircraft.

One of the properties of an RWR is radar detection sensitivity. Detection sensitivity may be defined as the ability of the RWR to detect radar emissions. One of the ways in which detection sensitivity is characterised is in terms of distance. An RWR system with high radar detection sensitivity will detect a weapon system at a greater distance than an RWR system with low radar detection sensitivity. Once the RWR detects the emission from a radar system, the aircrew is informed via auditory and visual displays.

The present situation is a helicopter fitted with a low sensitivity RWR – the unmodified system. This unmodified system is compared to the helicopter fitted with a high sensitivity RWR - the modified system. Two types of RWRs are also used, a "current" one and "future" one. The "current" RWR (as modelled in the simulator) is one that the aircrew participants have experienced and it provides bearing information to the threat only. The future RWR has several properties that are new to the aircrew, including providing range and bearing information to the threat. More details are given in Chapter 6 and 7 respectively.

The RWR system used in the experiments was designed to emulate operational RWR systems where national security limitations permitted. Some characteristics of the system (for example, attenuation of radar returns by different weather conditions) were not included. The choice of what to emulate and what not was based on SME input and the analytic products (see Chapter 5). One Australian Army Captain and one Senior Electronic Warfare Scientist took part in determining what aspects of the RWR should be emulated in

the experiments[5]. Note, that the involvement of SME input in defining characteristics to be simulated is a potential limitation of the research that will be discussed in the Limitations section in Chapter 9. Hence, for the purposes of this program of research the RWR would only detect a threat radar system when three conditions were met; the radar had to be emitting, there must be line-of-sight between RWR and the emitting radar, and the radar must be in detection range of the RWR.

In addition to these conditions the "current" RWR was also designed to emulate the error in resolving threats that typifies older type RWRs. This error is manifested as the RWR not being able to separate and display a second threat within a sector that is defined as 15 degrees either side of the first threat. The error results in only one threat being displayed when in fact there are two threats in a similar location.

The range over which the RWR could detect a radar system is a function of whether the RWR is modified. For example, an unmodified RWR can first detect a SA6 radar (a threat system) in Search mode if the RWR is within 70km of the SA6. If the RWR is further away then it would not detect the radar. If on the other hand a modified RWR is operating, one would expect it to detect the same radar anywhere in the simulated world[6,7].

An Airmobile Operation may be defined as "An operation in which combat forces and their equipment move about the battlefield in helicopters under the control of a ground force commander to engage in ground combat" (Australian Army, 2001). Airmobile Operations are the most common kind of operations that Australian Black Hawk crews are trained to perform and are used during the testing of Black Hawk systems.

Essentially an airmobile operation involves a helicopter moving people and equipment in a tactical environment as opposed to a non-tactical environment. In the case of Patrol Insertion, the aim of the Airmobile Operation is to insert troops into a combat area. Clearly, the success of the mission will depend in part on the performance of the RWR system. Hence, the ability to identify the potential performance benefits of modifying the RWR system is important.

Chapter 2 established that the problem faced by government analysts is choosing a method for selecting measures of performance and effectiveness to aid procurement decisions for complex socio-technical systems. Before the test case can be used to help solve this problem it is important to establish whether the test case is representative of complex socio-technical systems in general. To do this the test case will be compared to each of the characteristics that Vicente (1999) produced to assess the complexity of socio-technical systems.

---

[5]More information about the SMEs is provided in Chapter 5. The Australian Army Captain and the Senior Electronic Warfare Scientist are designated ARMY SME1 and EW SME in section 5.2.1.1.

[6] It is noted that the decision to include only some properties of the RWR limits the results of this program of work. This is further discussed in Chapter 9.

[7] The maximum distance that the modified and unmodified RWR can detect radar emissions from the threat systems when they are operating in different modes is available from the author on request.

Table 3-1 provides a summary of the assessment. The table shows that the test system meets 10 of the 12 complexity characteristics and may be deemed to be representative of a complex socio-technical system.

Table 3-1 RWR complexity assessment

Complexity characteristic	Assessment
Large problem space. Complex systems have a high number of variables that interact with each other. This interaction tends to be complex and not easily quantified.	In the case of the test system a large problem space exists. There are many variables associated with the operation of RWR systems. Although it is possible to identify many variables, some variables interact in a way that is not expected. The test case system is classified as complex under this characteristic.
Social. The number of people involved with the system determines whether the system is complex or simple. The larger the number of people, the more complex the system because factors such as communication play a major part in how well a system performs.	In the case of the RWR, two pilots are typically involved; the Aircraft Captain is primarily concerned with using the system, interpreting the information from it, and communicating his decisions to the Flying Pilot and passengers. The Flying Pilot is responsible for carrying out the orders of the Aircraft Captain. The outcome of this interaction will be a function of the communication between the two pilots. The test case system is classified as complex under this characteristic.
Heterogeneous perspective. The degree of variation in factors such as the background, values and views of the workers/ users of the system will alter the degree of complexity of the system because outcomes, for example, reaching consensus, may be more difficult to achieve.	Given that the aircrew are trained in operation of the RWR and follow standard operating procedures for the most part, the test case system is classified as simple.
Distributed. The social coordination may be hampered by both geographical location and cultural aspects of the personnel involved in the design of the system.	The aircrew are contained within one platform and are from the same cultural background. Both are trained in RWR operation and it is the Aircraft Captain that interprets the information presented. The test case system is classified as simple under this characteristic.
Dynamic. Systems that are complex often show a high degree of dynamic behaviour with long response times. This means that for an outcome to be successful, personnel will have to anticipate the results of their actions well before the result of the action is seen.	The RWR integrates data from many different sources, so its output will reflect the rapidly changing tactical environment. In addition, decisions made early in the mission by the aircrew could affect the success of the mission. The test case system is classified as complex under this characteristic.
Hazard. An error made in operation of a complex system could result in a catastrophic result.	Errors made in the operation of RWR systems can lead to catastrophic results and to death. The test case system is classified as complex under this characteristic.

Complexity characteristic	Assessment
Coupling. Complex systems exhibit a high degree of interaction between sub-systems thus making it difficult to predict the overall behaviour of the system.	Determining the success of a mission based on the performance of the RWR subsystem and other subsystems is difficult to predict. Hence, the test case system in conjunction with the mission is classified as complex under this characteristic.
Automation. A high degree of automation is characteristic of complex systems. In complex systems the worker/ operator who monitors the system is expected to intervene quickly and decisively to overcome abnormal situations that tend to be infrequent. The intervention is generally cognitive rather than psychomotor.	The RWR is highly automated but prone to errors, the aircrew are expected to monitor the display and decide whether the display provides a correct or incorrect picture. The test case system is classified as complex under this characteristic.
Uncertainty. Complex systems tend to present an incomplete picture of what is actually happening this may be due to failure of sensors, random drift or a real change in the system. Hence, workers will need to act as problem solvers with possibly impoverished data.	Once again, the very nature of the environment in which the RWR operates ensures that there is always a degree of uncertainty associated with information derived from RWR systems. The test case system is classified as complex under this characteristic.
Mediated interaction. Users of the system may get their view of the system (or world) from a device and may need to bring to bear significant cognitive resources to make sense of what is being viewed.	Aircrew using RWRs get a highly mediated view of the system operation and environment. The test case system is classified as complex under this characteristic.
Disturbances. Complex systems tend to exhibit behaviour that may occur infrequently and be unanticipated by the designers. Workers are then expected to understand what the behaviour means and act to bring the system back to operating within normal conditions.	Failure of one part of the RWR system may result in the degradation of the system gracefully or an unanticipated event may occur that causes the failure of the system in an unanticipated manner. The test case system is classified as complex under this characteristic.

3.1.2 The simulation environment

The simulation environment is designed to provide the infrastructure that supports the comparison of the task-based and the constraint-based measure-selection methods. It consists of the Black Hawk simulator and the simulated world that the Black Hawk operates in. The infrastructure must support the theoretical underpinnings of the two methods, and meet the requirements for the experiment.

The two theoretical underpinnings that the simulation environment should support are, first, the representativeness of the physical environment and, second, the type of behaviour that may be observed in that environment. From a theoretical point of view both the constraint-based (represented by CWA) and task-based perspectives (represented by HE) emphasize the need to evaluate systems in an environment that is representative of the field setting. However, they differ from one another in what is considered to be representative of the field setting. In the case of the two phases of CWA in which I am interested, WDA and CTA, representativeness is considered in terms of the ecology of the environment and the activity that is required. In the case of HE, representativeness is considered in terms of the task being performed and the physical environment in which the task is performed.

The above difference between constraint-based and task-based perspectives has important implications for the design of the simulation environment. From a constraint-based perspective, the difference means that a simulation object must not just look like "the real thing" but key properties and key relationships between properties in the world must be faithfully modelled in the simulator. Given that CWA is not prescriptive–we do not know in advance what the key relationships are–an important step is to identify all the properties of an object, and then through testing include the properties that are needed for the broad range of activities identified in the CTA. From the HE perspective, the simulation design is centred on identifying specific tasks and then identifying the environment necessary to support that task. Hence, the range of situations to model is significantly reduced, making the design process easier. By contrast, in the case of the CWA perspective and CTA in particular an activity may be "Fly aircraft". This would mean modelling a wide range of the helicopter flight profiles, with the commensurate wide variation in the visual database. In the case of HE perspective, the task "Fly helicopter" could mean modelling the flight profile and visual database at the altitude typical of normative or expected behaviour.

The simulation environment must provide support for the type of system behaviour that is specified by CWA and by HEP. In the case of CWA specific system behaviour is never specified. Instead the constraints on the behaviour are specified. This means that during an evaluation the simulation environment must be able to support a wide range of possible behaviour. In the case of the HEP where specific system behaviour is specified the simulation environment should be designed to support that behaviour.

The difference between the requirements for conducting a system evaluation by CWA and HEP means that the simulation environment must support "novel" behaviour (i.e. behaviour that is not defined in standard operating procedures but behaviour that is "defensible" by the aircrew), as well as standard behaviour. This has implications for the design of the experimental task. The task that the aircrew is asked to perform must be both familiar to them and the behaviour of aircraft and threat systems should be doctrinally accurate. The experimental task should also be designed to allow known behaviour (SOPs) to be performed and also to allow novel behaviour to be expressed. This is particularly important in the design of the mission scenarios.

In the following paragraphs the simulation environment and equipment used to meet the theoretical and experimental requirements of the research are described (See Appendix A for a full description of the development and testing).

3.1.3 The Black Hawk helicopter simulator

The Black Hawk simulation environment used is housed within the Air Operations Simulation Centre (AOSC) at the Defence Science and Technology Organisation (DSTO) in Melbourne. The simulation environment consists of the Crewed Universal Battle Environment (CUBE), the Black Hawk helicopter cockpit and the simulation control centre.

The CUBE consists of four projectors, four mirrors and four rear projection screens that are used to provide the forward, left, right and upper view of the simulated external environment. The floor of the CUBE is not used as a display surface. The sixth 'side' of the CUBE is open to allow various aircraft cockpits to be slid in and out of the structure. The resolution of each of the four display screens is 1280x1024. It should be noted that edge smoothing between the display screens is not possible given the fact that the displays are perpendicular to each other.

Figure 3-1 shows a view of the CUBE from the outside-in. The figure shows the position of the aircraft cockpit relative to the display screens and the projectors.

Figure 3-1 Relative position of the main structures of the CUBE (Copyright Commonwealth of Australia, 2008)

Figure 3-2 Black Hawk cockpit positioned in the CUBE. (Copyright Commonwealth of Australia, 2008)

Figure 3-2 shows the Black Hawk cockpit positioned in the CUBE. The Black Hawk helicopter cockpit is mounted on wheels and is an accurate spatial representation of the cockpit area of operational Black Hawk helicopters. The aircraft cockpit is a fixed-based, rather than motion-based, simulation system. As can be seen from the figure the cockpit is made up of a fibre-glass structure that accurately limits the external field of view of

participants. It has fully adjustable seats, helicopter controls (collective, cyclic and pedal)[8], a main instrument panel, and a centre equipment panel. The cockpit structure also provides support for mounting computers and equipment that are used during the simulation.

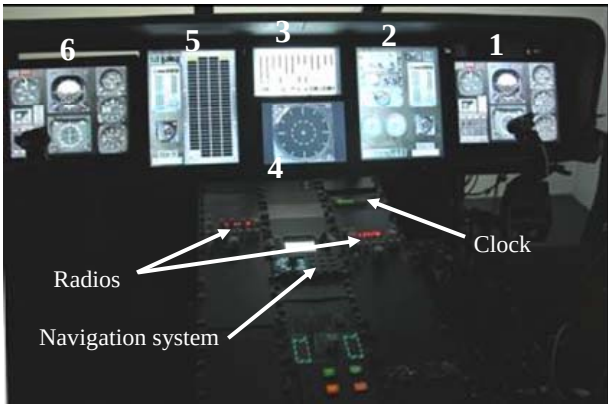


Figure 3-3 Main and centre instrument panels for the Black Hawk simulator. The RWR is located in a position that can be seen by both the Aircraft Captain and the Pilot. (Copyright Commonwealth of Australia, 2008)

The main instrument panel, see Figure 3-3, consists of six liquid crystal display panels that can be configured to display Black Hawk flight instruments.

- Panel 1 is a repeat of panel 6 and was configured to display the vertical situation indicator, horizontal situation indicator, radar altimeter, barometric altimeter, instantaneous vertical velocity indicator, airspeed indicator, stabilator position, stabilator position placard, internal communication system mode selector and the Vertical Situation Indicator/ Horizontal Situation Indicator mode selector panel. Panel 2 was configured to display the radio call placard, radar altimeter dimming control, Pilot display unit, clock, Navigation switches and indicator light, blade de-ice control panel, blade de-ice test panel, liquid water content indicator, fuel flow indictor, internal communication system switch identification placard.
- Panel 3 displayed the central display unit.
- Panel 4 displayed the radar warning receiver.
- Panel 5 displayed the caution advisory panel, the pilot's display unit, the radio call placard, radar altimeter dimming control, Pilot display unit, clock, navigation switches and indicator light.

The central equipment panel was used to house a clock (with stopwatch), radio controls, and the navigation system. In addition to the Black Hawk related equipment, two small eye-tracking cameras were positioned on the main instrument panel and orientated toward the participants. The eye-tracking cameras were used to record eye and head movements. A further camera was positioned behind the crew to record crew interaction.

[8] The helicopter controls consist of a cyclic that controls aircraft pitch and roll, a collective which controls power and torque and a pair of pedals that control yaw.

The Black Hawk flight model, originally developed and validated by DSTO scientists, was used to represent the flight dynamics of operational Black Hawk helicopters during the experiment.

The AOSC coordinates simulation experiments and consists of a number of networked Windows, Linux, and SGI Onyx PCs and displays. Figure 3-4 shows the general layout of the AOSC. The computers provide the visuals for the external environment displayed in the CUBE, record human and system responses, and provide the means for the researcher to communicate with the participants.

Figure 3-4 AOSC control centre (Copyright Commonwealth of Australia, 2008)

3.1.4 The simulated world

The simulated world consisted of the terrain data base and the Scenario Toolkit And Generation Environment (STAGE) software (STAGE is a commercial product from Presagis™, Montreal, Quebec).

The terrain data base used was the east coat of Australia. This was chosen because it provided a variety of terrain conditions and features that would be readily recognisable to the participants and because it provided a variety of conditions that did not overly limit their behaviour.

The threats appearing on the RWR display were computer-generated entities created using STAGE. The threats were of five different types of entities: SA8, SA6, Ship, ZSU23-4 and Unknown. Four of the entities could change between three radar modes: Search, Acquisition and Track. The ZSU23-4 entity could change between the radar modes Search and Track. The Unknown entity always is in Search mode.

A mode is a state of the radar system. Each mode is defined in terms of a number of properties. For example, radar in "Search" "observes" a much greater volume of sky than radar in "Track" mode. Threats changed mode based upon the range at which the Black Hawk was detected. For example, an SA8 in "search mode" will not change into "acquisition mode" until it can detect the Black Hawk and it can determine the Black

Hawk is within range of its acquisition radar. A change in radar mode was indicated to the aircrew by a change on the RWR display.

3.1.5 Summary

This chapter has presented the test system and test environment that will be used to compare the task-based and constraint-based measure-selection methods. The test system emulates a RWR and the test environment provides an environment in which the RWR can be used.

RWRs are complex socio-technical systems that are used in military aircraft (including Black Hawk helicopters) to detect radar emissions from potentially hostile weapon systems. In this thesis the RWR will be modified to increase the range that it can detect threats. Two types of RWRs are used to compare the measure-selection methods. The current RWR provides information on the bearing from the helicopter of the threat, the type of threat and the radar mode that the threat is in. The future RWR provides the range that the threat is away from the aircraft as well as all of the information that the current RWR does.

The test environment consists of a Black Hawk helicopter simulator and simulated world. The Black Hawk simulator is a fixed based physical mock-up of a Black Hawk cockpit. The cockpit contains all the flight controls and essential instruments that are needed to conduct simulated missions. The simulated world consists of a visual database that represents parts of Australia and threat systems. Using the Black Hawk simulator aircrew can "fly" simulated missions and use the RWR to detect and avoid threats. Throughout a mission data can be collected automatically on various human, aircraft, threat and environment behaviours and states.

In the next chapter, the foundation of the thesis is presented. In that chapter the research aims are stated and the program of research that will be used to compare the measure-selection methods is described.

4. Overall Research Aims, Reliability, Validity and Design

In Chapter 4 I present the main aims of the research, together with some of the conceptual questions relating to the reliability and validity of analytic products, measure-selection methods, and the measures themselves that need to be resolved. Only once these questions are resolved is it possible to derive the task-based and constraint-based measures of the RWR and to test those measures for their sensitivity in the Black Hawk simulator environment.

In Section 4.2 the research aim and research questions are stated and the research method to answer the questions is presented. In Section 4.3 the important foundational concepts of reliability and validity are discussed. This section will show that it is important to test the

analytic products that underlie the measure-selection methods that use these products. This section will also describe how reliability of the measure-selection methods was assured and describe the tests for validity that are necessary to test the measure-selection methods. It will be shown that comparing the task-based and constraint-based methods for the sensitivity of the measures that they produce and for their suitability for use in operational settings constitutes a test of the method's predictive validity. The design of the experiments (including a statement of the research hypotheses) used to test the methods is then described in Section 4.4. Finally, Section 4.5 describes the data collection methods used during the experiments.

4.1 Research aims, questions and structure

In this section I briefly describe the aim of the research in this thesis and I present a model of the framework for the research that will be used throughout the thesis. Then I state the research questions and describe the structure of the research program that is needed to answer the questions.

The aim of the program of research reported in this thesis is to compare the task-based and the constraint-based measure-selection methods. The literature review in Chapter 2 revealed that the task-based measure-selection method is used widely in the evaluation of complex systems in both laboratory and operational settings. The task-based method is based on the application of a number of guidelines. It has been reported that although the task-based method has been successfully used for selecting sensitive measures for use in laboratory settings and for current systems, it may not be successful for selecting measures for use in operational settings and for future systems.

The literature review also revealed that there have been constraint-based methods for selecting measures for use in the laboratory, but these were not designed for selecting measures for testing operational systems in operational settings. The literature review revealed that, in general, researchers believe that the analysis of constraints is uniquely suited for evaluating future systems. A constraint-based measure-selection method that can be used to select measures for operational systems in operational settings was produced and will be presented in Chapter 6.

The lack of empirical research comparing task-based and constraint-based measure-selection methods has led to the following research question: Is there a difference between a constraint-based and a task-based measure-selection method for evaluating complex socio-technical systems (both current and future) in operational settings? More specifically: Are the measures suggested by the two methods sensitive to a system modification for a current and future system? And are the methods suitable for use in operational, not laboratory, settings?

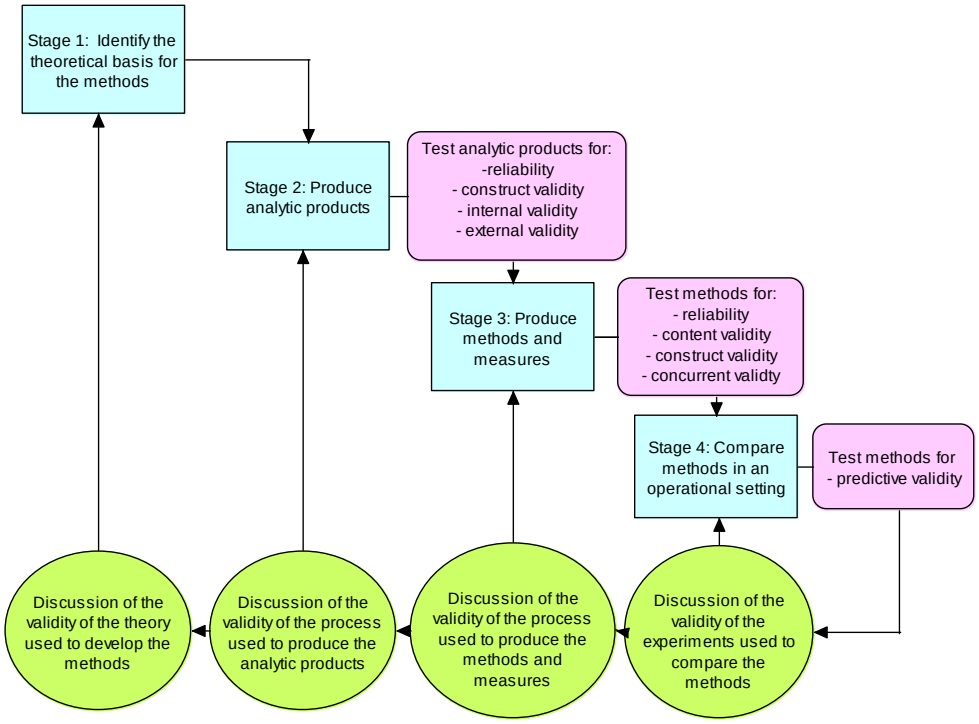


Figure 4-1 Structure of the research performed in the thesis, described as four stages

Figure 4-1 shows the structure of the research in this thesis, as introduced in Chapter 1. In the following sections, some more information is provided about the challenges that emerge at each stage, and in the transition between stages. This information provides the basis for the work on reliability and validity that needs to be achieved for the thesis to achieve the research goals.

4.1.1 Stage 1

Figure 4-1 shows that the research in Stage 1 of the method (reported in Chapter 2) describes the theoretical basis for the measure-selection methods. It was concluded in Chapter 2 that the constraint-based perspective has a strong theoretical foundation that could be a sound basis for measure-selection methods. It was also concluded that the task-based measure-selection method does not have such a strong theoretical basis.

4.1.2 Stage 2

Stage 2 focuses on developing reliable and valid analytic products that represent the test system. The method adopted to develop the analytic products will be described in more detail in Chapter 5, but it involves the following steps:

1. Produce the analytic products, identified in Chapter 2, using appropriate data collection methods.
2. Use expert independent practitioners to judge whether the analytic products are reliable and valid for the test system.

4.1.3 Stage 3

Stage 3 focuses on developing measure-selection methods and selecting measures from the analytic products with the methods, and it will be described in more detail in Chapter 6. It involves the following activities:

1. Produce measure-selection methods, identified in Chapter 2, using a process that is auditable
2. Assess whether the methods are reliable and valid.
3. Select measures for testing.

Where a measure-selection method already exists, such as the task-based method, expert independent practitioners are tasked to assess whether the measure-selection method produced by the author is representative of the method they (as practitioners) would use. The validity of the output of the methods (the measures) are also assessed. In the case of the constraint-based method, where there is no established measure-selection method, a method is developed and assessed for reliability and validity. The task-based and the constraint-based methods are then used to produce sets of measures (a set of task-based and a set of constraint-based measures) that may be sensitive to the system modification.

4.1.4 Stage 4

Stage 4 focuses on testing the measures produced by following the measure-selection methods identified in Stage 3. Following an exploratory experiment (described in Appendix A) that confirmed a number of technological requirements and tested the experimental protocol, two main experiments were performed. Experiment 1 tested the degree to which the measures suggested by the methods were sensitive to a modification to a current system. Experiment 2 tested the degree to which the measures were sensitive to a modification to a future system.

4.2 Reliability and validity

The concepts of reliability and validity are central to the research method used in this program of work. Without ensuring the reliability and validity of the measure-selection methods and the analytic products that are used by the methods the results gained from the comparison of the methods cannot be valid. As can be seen from Figure 4-1, stages 2, 3 and 4 of the research method represent different forms of reliability and validity. In this section I discuss reliability and validity as they relate to each of those stages.

4.2.1 Reliability and validity of analytic products (Stage2)

Figure 4-2 shows forms of reliability and validity assurance associated with the analytic products. The figure shows that Stage 2 (“Produce analytic products”) produces a requirement (rounded rectangle) to establish that the analytic products are reliable and have various forms of validity.

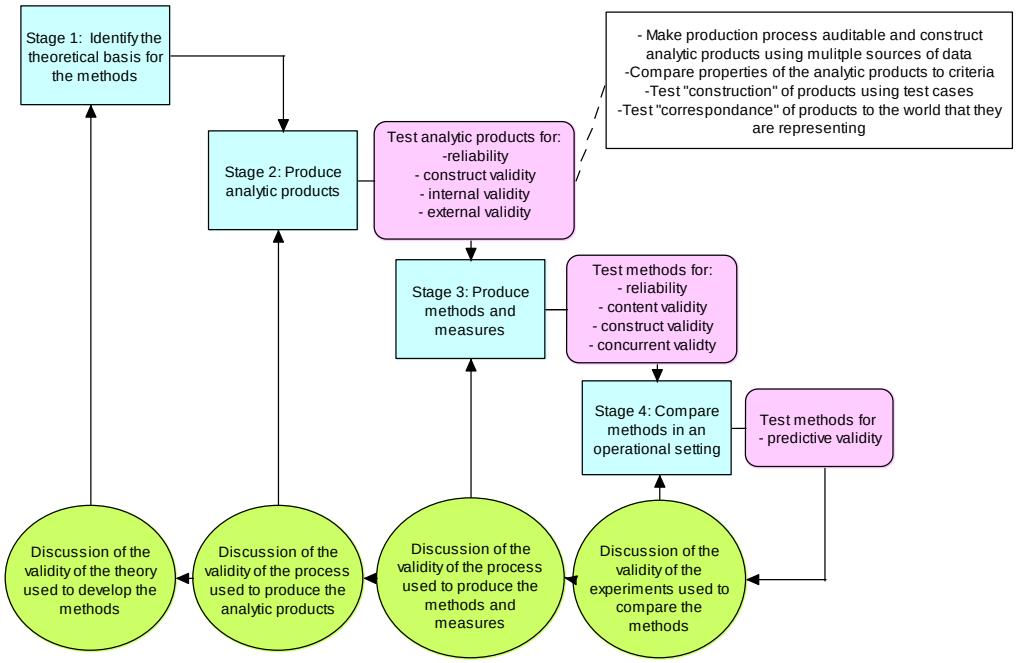


Figure 4-2 Testing the reliability and validity of analytic products

4.2.1.1 Reliability of analytic products

Analytic products are deemed reliable if the same analytic product is produced on a number of different occasions using the same information by the same analyst or if the same analytic product is produced by different analysts using the same source material. When considering reliability the analytic product need not be a true representation of the domain, nor is it important how the analytic product may be used.

The form of reliability assurance for analytic products that is used in this program of work involves producing the analytic products using multiple forms of data (for example, SME interview data and technical documents) and creating an open, inspectable database representing the analytic products so that it can be reviewed by other researchers in the future (Yin, 2009). In principle, for future use, the various forms of raw data on which such a database is produced would also be made available. This method is best suited to assessing the analytic products for complex systems, where a high degree of domain knowledge is required, where it is not always possible to train another researcher in that domain to perform traditional reliability checks and where several iterations of a development-test cycle for analytic products are not possible. Further rationale and details are given in Section 5.1.

4.2.1.2 Validity of analytic products

Analytic products are deemed valid if they represent what they are meant to represent. In general, there are three main dimensions of validity along which analytic products should be assessed. The first, construct validity, refers to whether the analytic product accurately reflects the theoretical construct from which it is derived. The second form of validity, internal validity, refers to whether the elements and the relationships between the elements in the analytic product are coherent, i.e. whether the elements in the analytical

product are logically related. Finally, the third type of validity, external validity, refers to whether the elements and relationships shown in the analytic product correspond to the elements and relationships in the world that it is representing. As with reliability, further development of the analytic products is informed through use. In this thesis the term apparent validity (not shown in Figure 4.2) is used to indicate whether an actor in a domain refers to a property that is identified in the analytic product and is defined as a measurable property (of the domain) that appears in data (from the actor) via discussion and/ or observation and/or inference from the transcription data.

The standard to aim for with all form of validity testing is to use multiple individuals to assess the products against a number of criteria. However, the domain explored for this thesis does not offer multiple available experts and it is unrealistic to train people for this specific purpose. An alternate strategy is necessary which includes using fewer numbers of highly experienced SMEs to assess the analytic products.

For the constraint-based abstraction hierarchy (AH) the test for construct validity involves experienced SMEs assessing whether the analytic product complies with several criteria. These criteria (taken from Vicente, 1999) are:

- The AH should represent the ecological properties of a work domain and those properties should be related to each other in a structural means-ends hierarchy. If the analytic product were constructed using other constructs, such as action means-ends or goal-directed behaviour, then the analytic product would not be valid.
- The AH should be event independent, i.e. the objects, functions and purposes represented in an AH should not represent a specific event – there should only be one AH for a work domain.
- The language used to describe objects, function and purposes on the same level of abstraction should be the same, and different language should be used to describe different abstraction levels.
- The objects, functions and purposes represented in the AH should be described as nouns rather than verbs. This is because the AH represents the domain in which activity takes place, not the activity itself (Vicente, 1999).

For the constraint-based temporal coordination-CTA (TC-CTA) that represents the activity (not tasks) that the system should support. A good test of construct validity is to see whether the TC-CTA describes the generic activity needed to achieve known goals rather than describing specific actions (tasks) needed to achieve a goal – the latter would be representative of a task-based analytic product. Other criteria (Vicente, 1999) that inform construct validity are, when applied to the TC-CTA:

- The TC-CTA should be described using verbs instead of nouns (to distinguish it from a AH).
- The TC-CTA should describe what needs to be done, not how or by whom[9].

[9] Although the TC-CTA shown in this thesis does identify the actors that perform the control task this is only a temporary addition of material to highlight the fact that this analysis deals only with

- The TC-CTA should be device independent.
- The goal identified in the TC-CTA should be achievable in different ways on different occasions (requisite variability).
- The input and output requirement should be identified for each activity.
- The constraints that act on the activities must be identified.

For the task-based analytic products, construct validity may be tested by assessing whether the analytic products represent goal-directed behaviour.

For the analytic products produced from the constraint-based AH and TC-CTA, internal validity can be tested by having individuals assess whether the analytic products are well formed and coherent. In the case of the AH, a suitable test is to assess whether the analytic product conforms to a structural means-ends relationship for the specific test case from the domain of interest. In the case of TC-CTA, the test involves ensuring that the activities represented in the analytic product occurred during the correct phase of the mission. Another test is to assess whether and the relationships between activities are logical. For example, the operation of certain aircraft system (for example, a RWR) should not occur before the aircraft is in flight because the RWR system is not designed to be operated on the ground.

For the constraint-based analytic products a good test for external validity is to assess the product against a number of different test cases, i.e. test that the analytic product corresponds to real cases. Burns et al (2001) applied a similar approach to their work.

For the task-based analytic products internal validity can be achieved by having independent SMEs assess the products for completeness; essentially assessing whether all the elements in a specific mission that should be included are included, and assessing that the analytic products contain the correct elements. For example, one would expect that the narrative mission analysis (an analytic product) should contain major mission phases that reflect the mission (each mission phase is an element), that major system functions reflect the mission phase functions (also an element), and that that the time scale of the tasks and the external events that trigger tasks are included.

For the task-based analytic products a good test for external validity is to assess whether the elements represented have been generated from a variety of mission types and when tested, are representative to other similar mission types. For example, a mission narrative may state that the altitude that the aircraft flies at during a mission is 35,000ft or 16,000ft. If it observed that the aircraft actually flies at 25,000ft it can be concluded that the mission narrative does not have external validity.

4.2.2 Reliability and validity of measure-selection methods (Stages 3 and 4)

The previous section has shown that having reliable and valid analytic products is important for system evaluation. Equally important is having reliable and valid methods

---

the Aircraft Captain and does not include other Social Organisation Analysis (SOA) material. No SOA-based measures, for example, number of interactions between aircrew, are extracted.

that use those analytic products. Although the concept of reliability is the same for measure-selection methods as for analytic products, there are subtle differences in definitions of validity and in the way that the tests for them are implemented. In addition, the concept of "predictive validity" is included here, something that was not considered before. Hence, this section will consider the concepts of reliability and validity specifically in the context of producing the measure-selection methods. Figure 4-3 shows the reliability and validity tests associated with the measure-selection methods.

4.2.2.1 Reliability of measure-selections methods

In general, a reliable method is one that can be followed by another analyst to produce an outcome that is the same as the outcome produced by the original analyst. In this thesis, I am examining whether the same measures would be produced by different analysts using the same method.

There is agreement between authors on the importance of assessing a method's reliability. There is also agreement that reliability is shown if results of following a method are the same if the method is followed a second time. Yin (2009) presents the case that reliability (for a case study) of a method is achieved if a second analyst follows the procedures identified by an earlier analyst and, using the same data, produces the same results and conclusions. Other authors, such as Annett (2002) echo this point of view and argue that reliability of a method (used in ergonomics) "is essentially about repeatability of results either by another observer or the same observer at different times or under different conditions" (p. 229). Hoffman (1998) also indicates that reliability of a method is shown if independent researchers generate the same results when using the same method.

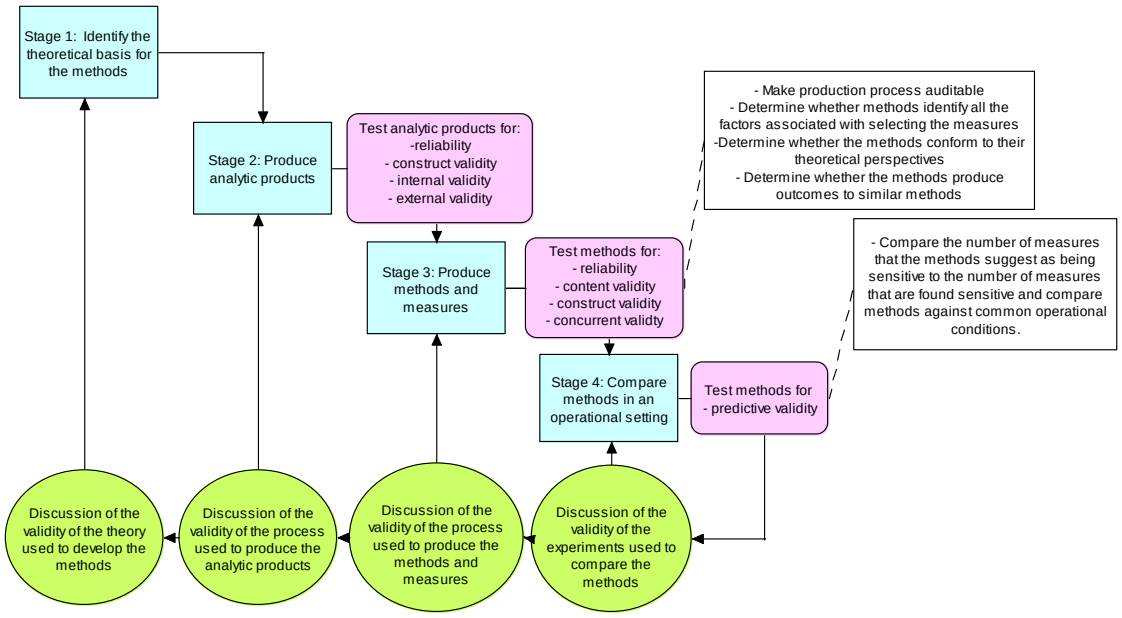


Figure 4-3 Testing the reliability and validity of measure-selection methods

Intra-tester reliability is shown if the same researcher at different times or under different conditions produces the same results; whereas inter-tester reliability is shown if two different researchers produce the same results.

The reliability tests outlined above are the ideal, and may need to be modified because of the constraints of the present research program. For an inter-tester reliability test to be performed the second tester would have to have a high level of domain familiarity, specialist RWR knowledge and either have a constraint-based (CWA) or task-based background or both. It was not practicable to find or train another analyst to act as an independent tester. In addition, given that the method was being performed just after having been developed, it was impractical to attempt intra-tester reliability. Therefore, the method of reliability assurance adopted by the author included following the method through and documenting all issues found. The method was specified in sufficient detail that another reasonably knowledgeable person could probably follow it and produce similar results. All the measures that were suggested as being sensitive to the system modification were recorded

4.2.2.2 Validity of measure-selection methods

Again, there are various forms of validity testing relating to methods in general. Annett (2002) argues that there is a clear relationship between a method and the model underlying the method. Annett argues that construct validity of a method refers to whether the underlying theoretical model that the method uses is accepted. For example, he notes Swain and Weston's Human Reliability Analysis (HRA; Swain and Weston, 1988) and Baber and Stanton's Task Analysis for Error Identification (TAFEI; Baber and Stanton, 1994) methods are based on accepted thinking on human error and thus have construct validity.

Concurrent validity refers to the extent to which a method produces a result that is consistent with other methods that have been shown to produce the same result. Predictive validity refers to whether the output of the method meets the aim of the researcher; specifically, whether the output correlates to a criterion. Hence, a method for selecting sensitive measures should produce statistically sensitive measures (if the criterion is that the measure should be statistically significant). Finally, content validity refers to the extent that a method appears to do what it purports to do (Diaper and Stanton, 2004a).

To test the construct validity of the constraint-based and task-based measure-selection methods it is necessary to consider whether the theoretical basis for the method is accepted by the community of practice (Human Factors practitioners) and also show that the methods clearly reflect their theoretical basis.

For the constraint-based measure-selection method, the test adopted in this thesis, for construct validity is for the author to assess whether the method is based on a constraint-based perspective of system design and whether the relationships between the different classes of constraints (representing the different types of measures) are clearly identified.

For the task-based measure-selection method assessing whether the task-based measure-selection method has construct validity is more difficult. I could say that the method has construct validity because the impact of the system and task are clearly identified or I could say there is no construct validity if I consider the test (construct validity) is at the level at which measures are selected. In this thesis because I am interested in the method for producing measures and because it has been shown earlier that there is no theory

governing the selection of measures it is concluded that the task-based measure-selection method cannot have construct validity.

To test the content validity of the measure-selection methods it is necessary to use independent experts to assess whether the methods account for all the factors (that affect measure-selection), and whether the methods if followed, would result in measures being produced.

Assessing the constraint-based measure-selection method for concurrent validity is not appropriate because the one tested is a first of a kind. In the case of the task-based measure-selection method where there is no empirical evidence for its effectiveness or the effectiveness for other similar methods an assessment of its concurrent validity is also not appropriate.

Testing the predictive validity of the measure-selection methods may be done empirically by comparing the number of measures that the methods suggest as being sensitive to the number of measures that are found to be sensitive. This is best achieved using an experiment. The experimental design used to investigate the predictive validity of the methods is described in the following section.

4.3 Experimental design and hypotheses

In this section the experimental design is presented. This includes a description of the independent variables, dependent variables, controlled variables, configuration of experiments, and presentation of the research hypotheses. A brief outline of the experimental task and an example of the process of data collection are also provided.

The literature review given in Chapter 2 has shown that for a valid comparison of the measure-selection methods several factors should be considered. These factors were the measure-selection method (task- or constraint-based), system type (current or future) and system modification state (unmodified and modified). The review also indicated that two important points of comparison between the measure-selection methods were, first, whether the methods produced measures that were sensitive to a modification and, second, whether the methods are suitable for use in an operational setting. The need to compare the measure-selection methods in terms of the sensitivity of the measures that they suggest and the suitability of the method for use in operational settings gives rise to a 2X2X2 experimental design which is shown in Table 4-1. The dependent variables are described in the following paragraphs. There are two classes of dependent variables, low-level and high-level.

The low-level dependent variables are the measures from the measure-selection methods and are defined in Chapter 6 and are shown in Table 4-1. The high-level dependent variables are aggregations of the low-level dependent variables. There are five high-level dependent variables. These are now described.

The first high-level dependent variable is used to compare the measure-selection methods in terms of the sensitivity of the measures that they suggest and is, Number of measures

that are statistically significant. This is defined as the number of measures that are statistically sensitive to the system modification.

Table 4-1 Experimental design. Table shows the low-level dependent variables.

		Independent Variable 1 (Measure-selection method)			
		Constraint -based		**Task -based**	
		Independent Variable 3 (System modification)		Independent Variable 3 (System modification)	
		Unmodified	**Modified**	**Unmodified**	**Modified**
Independent Variable 2 (System type)	**Current RWR** (Expt 1)	Constraint-based measure 1 ... Constraint-based measure 42		Task-based measure 1 ... Task-based measure 25	
	Future RWR (Expt 2)	Constraint-based measure 1 ... Constraint-based measure 42		Task-based measure 1 ... Task-based measure 25	

The remaining four high-level dependent variables are used to compare the measure-selection methods in terms of their suitability for use in operational settings. The first high-level dependent variable is, *Percentage of measures for which the collection of data is limited by simulation resources*. This is defined as the number of measures that could not be tested for statistical sensitivity because system models (for example, a property of the RWR system, a property of the world, and a property of the helicopter and mission) cannot be developed, divided by the total number of measures.

The second high-level dependent variable used to compare the measure-selection methods in terms of their suitability for use in operational settings is, *Percentage of measures for which the collection of data is limited by the data collection method used*. This is defined as the number of measures that could not be tested for statistical sensitivity because the data collection method used to collect the data for the measure was not adequate (because of problems associated with gathering data points manually or automatically), divided by the total number of measures.

The third high-level dependent variable used to compare the measure-selection methods in terms of their suitability for use in operational settings is, *Percentage of measures for which data is limited by the number of data gathering opportunities*. This is defined as the number of measures that could not be tested for statistical sensitivity because the number of data collection opportunities was not sufficient to meet statistical protocols (for example, if the number of data points was less than 10 or if data were not present in all categories - if categorical data were collected), divided by the total number of measures.

The fourth high-level dependent variable used to compare the measure-selection methods in terms of their suitability for use in operational settings is, *Percentage of measures for which data could not be collected because of theory*. This is defined as the number of measures that could not be tested for statistical sensitivity because theory restricted the collection of data, divided by the total number of measures. In other words, if it can be shown that data could not be collected because the underlying theoretical perspective excluded that data, then this measure would be used.

Table 4-1 shows how the variables are configured for each of the experiments. The table shows that Experiment 1 is concerned with comparing the measure-selection methods using a current system and is reported in Chapter 7. Experiment 2 is concerned with the comparison of the measure-selection methods using a future system and is reported in Chapter 8.

In Experiment 1, there are there are five hypotheses related to the issues of measure sensitivity and five hypotheses related to the issues of method suitability. The hypotheses are the same for the current and future systems. They are as follows.

Measure sensitivity:

H1 None of the task-based and none of the constraint-based low-level dependent variables are sensitive to the system modification.

H2 All the task-based and all the constraint-based low-level dependent variables are sensitive to the system modification.

H3 Significantly more of the constraint-based low-level dependent variables than the task-based low-level dependent variables will be sensitive to the system modification.

H4 Significantly more of the task-based low-level dependent variables than the constraint-based low-level dependent variables will be sensitive to the system modification.

H5 Some of the task-based and constraint-based low-level dependent variables will be sensitive to the system modification.

Method suitability:

H6 The task-based and constraint-based low-level dependent variables will be affected by all of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H7 The task-based and constraint-based low-level dependent variables will not be affected by all of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H8 The constraint-based low-level dependent variables will not be affected by some of the following as the task-based low-level dependent variables will be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H9 The task-based low-level dependent variables will not be affected by some of the following as the constraint-based low-level dependent variables will be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H10 The task-based low-level dependent variables and the constraint-based low-level dependent variables will not be affected by some of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

In order to collect data to test the hypothesis the aircrew are required to complete a number of airmobile operations in which they must complete their mission while surviving a number of threats. A mission consists of a briefing, flight, and debriefing phase. All three mission phases are necessary because some measures suggested by task-based and constraint-based measure-selection methods can be assessed only during specific phases whereas others could be assessed only by comparing data across phases. For example, it is postulated that the modification to the RWR will affect whether the Aircraft Captain correctly observes standard operating procedures (SOP) in response to a threat. For an assessment to be made about whether the Aircraft Captain correctly implemented the SOP I have to know what that SOP is and then compare the observed action of the Aircraft Captain to the SOP. In the experiments the SOP is defined by the Aircraft Captain during the mission briefing phase; the actual behaviour of the Aircraft Captain is then observed during the mission flight phase, and the reasons for the Aircraft Captain' behaviour are explored during the mission debriefing phase. Hence, the opportunity to collect data about the correct implementation (or not) of the SOP occurs once in the briefing phase, once in the flight phase and once in the debriefing phase. From the observations made during the three phases data can be used in statistical tests. During the experiments, qualitative data from both observation and interview data collection methods augment the quantitative data gained from the simulator.

4.4 Data collection methods

Section 3.2 has described the four stages of the process needed to compare the task-based measure-selection method with the constraint-based measure-selection method. Two of those stages (Stage 3, produce analytic products, and Stage 4, compare methods in an operational setting) have specific requirements for the types of data that are required. In Chapter 2 those requirements were identified and the available literature concerning the use of data collection methods for the development of analytic products and experiments was reviewed.

The review in Chapter 2 identified that of all the data collection methods, semi-structured interview methods (including Critical Decision Method), have been adopted or recommended by Kirwan and Ainsworth (1992), Beevis (1999), Diaper and Stanton (2004a), Stanton et al (2005), Naikar et al (2005), Bisantz et al, (2003, 2001), and Jenkins et al (2008) as most applicable to the development of the task-based and constraint-based analytic products and to their use during the experiments.

The review also identified that observational methods have been used, or recommended, by Kirwan and Ainsworth (1992), Diaper and Stanton (2004a), and Stanton et al (2005) and that such methods seem to be particularly suited to the experimental assessment. Automated data collection methods, while not specifically identified in the literature review, are particularly suited to the experiments. Figure 4-4 shows the data collection methods that were used in this program of research. As can be seen from the figure a modified Critical Decision Method is used in the production of the analytic products (Stage 2 of the research program) and the experiments comparing the measure-selection methods (Stage 4 of the research program) and automated data collection and direct

observation are used in the experiments (Stage 4). In the following sections each of the data collection methods will be briefly described.

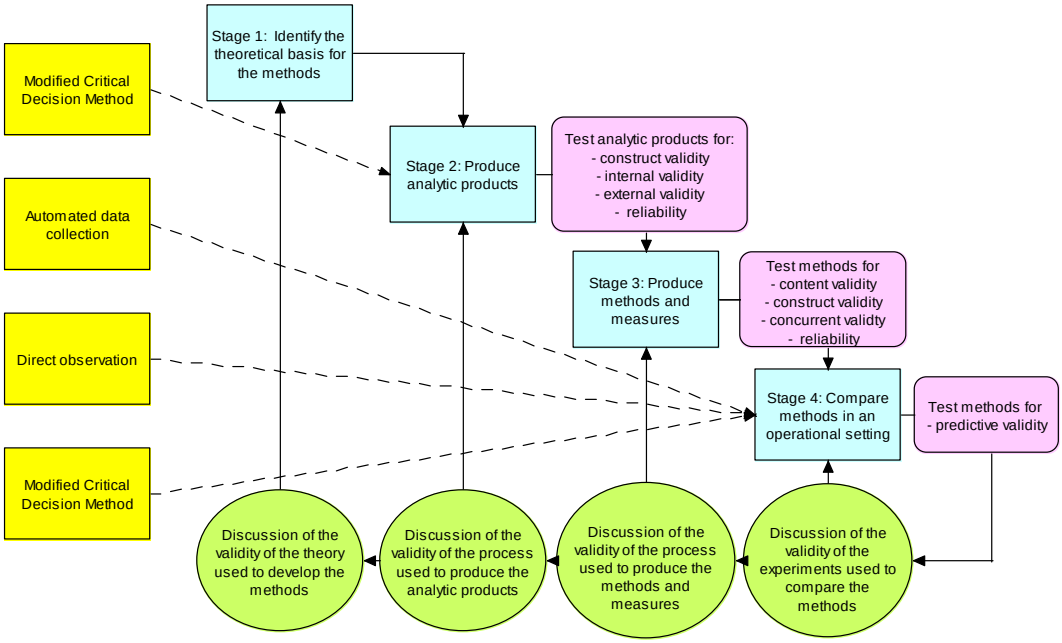


Figure 4-4 Structure of the research program described in this thesis with the data collection methods highlighted

4.4.1 Modified Critical Decision Method

The modified Critical Decision Method (modified CDM) is part of the family of interview methods. In this program of work it was used to gather data from SMEs. These data were then used to produce the analytic products (an abstraction hierarchy for the WDA, activity analysis for the CTA and various Human Engineering analytic products). It was also used with experiment participants during the exploratory experiment, Experiment 1 and Experiment 2 to provide data to augment other data gathering activities. In the next sections interview methods and the CDM in particular are introduced and finally the modified CDM used in the thesis.

4.4.1.1 Background to CDM

Kirwan and Ainsworth (1992), Stanton (2004) and Stanton et al (2005) provide good overviews of the use of interview methods including the Critical Decision Method (CDM; Klein et al, 1989). Stanton et al (2005) note that there are three kinds of interview: structured; semi-structured and unstructured. Semi-structured interviews (of which the CDM is an example) elicit information in response to a number of questions. However, other probe questions, in addition to those planned, may be used. The interviewer can also divert the focus of the interview to gather data on unexpected issues. In general, the semi-structured interview is best suited for collecting data from these experiments.

The CDM is a knowledge elicitation method that "is designed to model tasks in naturalistic environments characterised by high time pressure, high information content

and changing conditions" (Klein et al, 1989). The purpose of the CDM is to understand the basis for decision making and situation assessment during critical incidents. Differences between experts and novices can be identified. Tasks are modelled and then used to help understand the role of technology in supporting both the cognitive and behavioural aspects of human performance, designing training programs, and supporting the development of technologies that will automate the critical task functions that have an aspect of human reasoning. Klein et al. argue that experts make decisions based on their "skill at recognising situations as typical and familiar" and developed the Recognition-Primed Decision (RPD) model of decision-making (see Klein at al, 1989). The authors developed a series of cognitive probes designed to elicit information to populate each part of that model.

Although CDM has been used widely (for example Hoffman et al, 1998) there are some practical and conceptual limitations with the method. Naikar et al (2005) report on the theoretical concepts and methodology associated with the development of Work Domain Analysis and analytic products and also identify the benefits of interview methods, including CDM. However, Naikar et al sound a note of caution about using semi-structured interviews (and by implication CDM). In particular they draw attention to the high cost (time, effort) of the method. Although this does not exclude the method it does emphasize the need for appropriate planning, where the time and resources needed to collect the data is considered carefully.

Bisantz et al (2003, 2001) have also used CDM in their work and have modified it to collect data that can be used for the development of WDA and CTA analytic products. The modification involved adding specific prompts to capture specific WDA and CTA aspects. The information gained with this modification could then be mapped directly on to the WDA.

4.4.1.2 My use of CDM

Given the applicability of CDM, my approach was to augment the existing CDM prompts with ones that specifically provided data for all the constraint and task-based analytic products. The modified CDM was designed to meet the objectives of this research program. Specifically, it was designed to support the development of the constraint-based and task-based analytic products and to provide data that would help assess the behaviours of the aircrew (participants) during the experiment. These data included how the aircrew intended to conduct their mission, the tactics that they planned to use when encountering threat systems, and characteristics of the environment that were considered important. Additionally, the modified CDM was designed to elicit information from aircrew about events that took place during the flight phase of the experiment. Table 4-2 shows the prompts used during the production of the analytic products and experiments.

Table 4-2 Modified CDM prompts mapped onto the analytic products. The prompts were also used during the experiments to provide data for assessing aircrew behaviour.

Prompts (with CDM types in parenthesis)	Application to constraint and task-based analytic products
Do you have a general philosophy for selecting different tactics that you follow when flying a mission?	Constraint-based - WDA (Domain Values and priorities) Task-based – Function flow
What are your expectations going into this event?	Constraint-based – WDA (all levels) and CTA (all aspects) Task-based – All analytic products
What are your priorities at this point?	Constraint-based - WDA (Domain Values and priorities). Task-based – Mission narrative
What were your specific goals at this time (Goals)?	Constraint-based - WDA (Domain purpose) Task-based – Critical tasks
If you were to consider yourselves as part of the Black Hawk system, what functions are you performing at this point? What are you doing at this point?	Constraint-based - WDA (Domain functions, Physical function), CTA (Activities). Task-based – Functions and time line of critical tasks, function allocation
How are your functions different from each other?	Constraint-based - CTA (Activities) Task-based – Function allocation
What tools/ systems are you using at this point? (e.g. map, instruments, etc)	Constraint-based - WDA (Physical objects), CTA (Systems). Task-based – Function flow, function allocation
What mental strategies are you employing at this point? (e.g. Trying to make more time, radar modes and tactics, fuel remaining and routes, etc.)	Constraint-based – WDA and CTA (Physical objects and Activities described in ecological property terms)
What could have gone wrong at this point? (e.g. system failures)	Constraint-based - WDA (Domain Values and priorities), Task-based – Function flows, critical tasks
What might a less experienced pilot/NFP have done at this point?	Constraint-based – CTA (Activities) Task-based – Function flows
What were you seeing, hearing (Cues)?	Constraint-based - WDA (Physical objects) Task-based – Function allocation
What information did you use in making this decision and how was it obtained (Knowledge)?	Constraint-based - WDA (Physical Objects, Physical Functions, what-why-how relationships)
What other courses of action were considered by or available to you (Options)?	Constraint-based – CTA (Activates) Task-based – Function Flow
How was this option selected/other options rejected? What rule was being followed (Basis)?	Constraint-based - WDA (Domain values and priorities)

Prompts (with CDM types in parenthesis)	Application to constraint and task-based analytic products
	Task-based – Function Flow
What specific training or experience was necessary or helpful in making this decision (Experience)?	Task-based – Critical tasks
If the decision was not the best, what training, knowledge, or information could have helped (Aiding)?	Constraint-based – WDA(Physical objects, physical function) Task-based – Critical tasks
How much time pressure was involved in making this decision (Time pressure)?	Constraint-based - WDA (Domain values and priorities)
Imagine that you were asked to describe the situation to a relief officer at this point: how would you summarize the situation (Situation assessment)?	Task-based – Mission narrative

4.4.2 Direct observation method

The observation data collection method was used during the experiments (Stage 4 of the research design, Figure 4-4). Stanton and Young (1999), Kirwan and Ainsworth (1992), Diaper and Stanton (2004a) and Wilson and Corlett (1995) all provide good explanations of observation methods. The advantages of observation include the following factors: the data collected provide a "real-life" insight to the activities performed, a wide range of data can be recorded, the method has been widely used, objective information may be provided, and the interaction between the system, participants and the environment can be studied. The method seems well suited to this program of work.

Broadly speaking there are two main classes of observation methods: direct observation and indirect observation (Wilson and Corlett, 1995). Direct observation methods generally involve observers collecting information directly from the subjects or from recordings of the subject's behaviour. Indirect observation methods involve the observer using reports made by the subjects or using data collected by other individuals rather than the observer (Wilson and Corlett, 1995). Because I wish to augment the quantitative data gathered during the experiments, direct observation was seen to be the most appropriate. Chapter 6 and 7 explain how the observations were made in the context of the experiments.

4.4.3 Automated quantitative data collection methods

Figure 4-4 shows that quantitative data collection methods were used during the experiments (Stage 4 of the research design). The quantitative data that were collected may be divided up into two main categories: data relating to participant behaviour and system data. Both categories of data were synchronised and collected every 16 milliseconds during the mission flight phase. The participant data included all video, audio and cockpit button presses. Video and audio were continuously recorded. The type of button pressed and the time when it was pressed were recorded.

The system data that was collected included the following:

- Threat and aircraft interaction information displayed to the crew (the time when a threat signal is displayed to the crew via the RWR system (both visually and aurally); the threat's behaviour (mode) on a continuous basis – time and position stamped; whether the aircraft had been damaged by a weapon at what level of damage; and the chaff usage (how much and when).
- Aircraft data (6 degrees of freedom of the helicopter - x, y, z, heading, pitch, roll); position and velocity; latitude and longitude of helicopter; airspeed – true and indicated; pedal, cyclic, collective position; altitude – radar and barometric; fuel levels and fuel flow; chaff and radio button states and cockpit warning information.
- Threat data (6 degrees of freedom of lead aircraft - x, y, z, heading, pitch, roll); latitude and longitude of threats; threat mode information – frequency, scan info, tracking; Missile information and whether line of sight to the helicopter existed.
- Navigation data (buttons pressed, current waypoint information – heading to waypoint; waypoint number; distance to waypoint and position of waypoints).

A description of how the raw data is used in the measures is given in Section 6.4.1.

4.5 Conclusion

To recap, the method proposed in this thesis for selecting measures has four main stages:

- Stage 1: Identify the theoretical basis for the methods (see Chapter 2).
- Stage 2: Produce analytic products
- Stage 3: Produce methods and measures
- Stage 4: Compare methods in an operational setting

This chapter has outlined the research methods that will be used in Stages 2, 3, and 4 to compare the task-based and constraint–based measure-selection methods. The concepts of reliability and validity were discussed and appropriate methods identified to help us determine how adequately each stage is completed. The basic design of the experiments that will compare the measures suggested by the measure-selection methods was described, plus some of the methods that would be used in the experiments.

In the chapter that follows, Stage 2 of the research method is described in detail. In Stage 2, the task-based and constraint-based analytic products are generated and described.

5. Task-based and Constraint-based Analytic Products

Chapters 1 and 4 presented the four-stage research process that is needed to compare the task-based and the constraint-based methods. Chapter 2 identified the task-based and constraint-based approaches to developing methods for selecting measures to use in system evaluation, which is Stage 1 in the process. Stage 2 in the research process is to select and produce reliable and valid analytic products, and it is discussed in this chapter. Figure 5-1 indicates the location of Stage 2 in the logical flow of the thesis.

Developing reliable and valid analytic products is a critical step because the measure-selection methods draw upon the information that is represented in the products. If reliable and valid analytic products cannot be produced, then reliable and valid methods based on the analytic products cannot be produced. This chapter will present the analytic products developed to represent the test case.

As indicated in Chapter 2 there are many task-based and constraint-based analytic products that can be used and it is important to select the correct ones for the test case. As concluded in Section 2.4, the most suitable task-based analytic products for evaluating a medium complexity system during the Preliminary phase of the system life cycle are narrative mission descriptions, function flow diagrams, ad hoc function allocation and timelines. The most suitable constraint-based analytic products were found to be the abstraction hierarchy (AH) and the temporal coordination-control task analysis (TC-CTA). These analytic products are described in this chapter.

Chapter 4 indicated that to establish the reliability of the analytic products it was important to make the development process auditable. In addition, independent SMEs should be used to test the products' internal, external and construct validity. In this Chapter I will show that the reliability of the analytic products was assured by producing analytical products that are inspectable and making available the data used to construct them. I will also show that the validity of the analytic products was determined by independent Human Factors practitioners.

In the following section (Section 5.2) the method that was used to develop and test the reliability and validity of the analytic products is described. The analytic products that were produced are then presented and discussed in Section 5.3. Conclusions are drawn in Section 5.4.

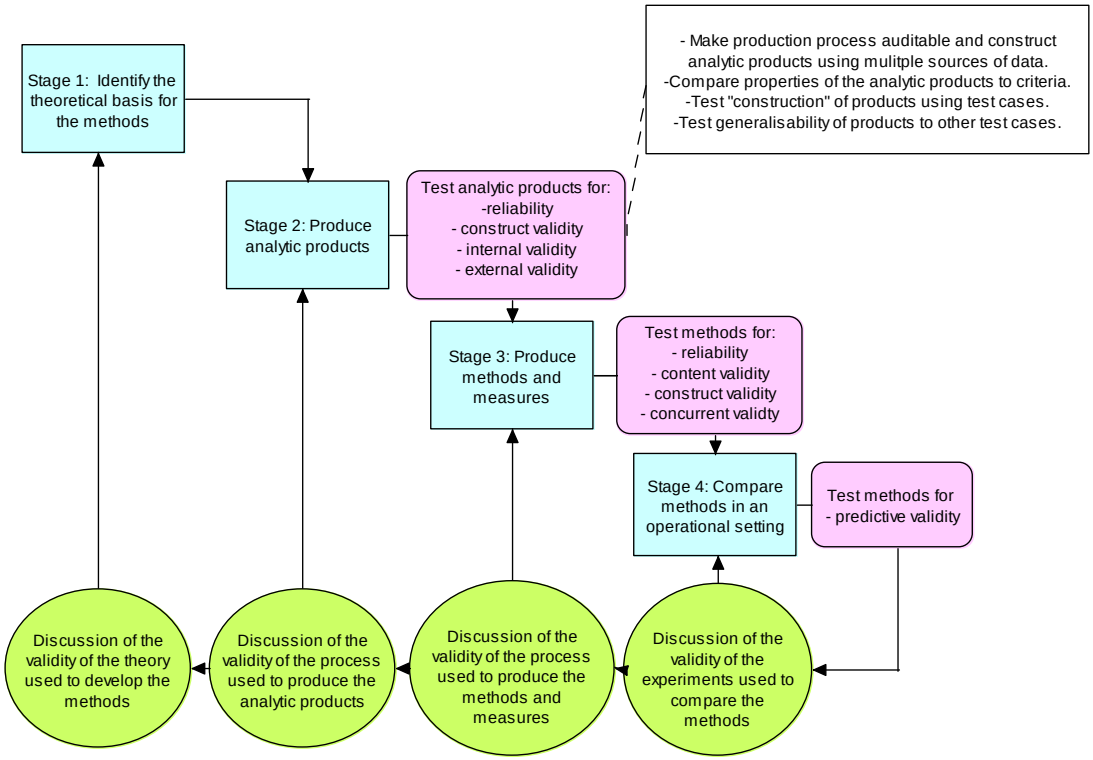


Figure 5-1 Structure of the research program. The analytic products are developed and tested for validity and assured for reliability.

5.1 Method used to develop analytic products

In this section the method used to develop the analytic products is described.

5.1.1 Participants

One set of participants took part during initial development of the analytic products and another set took part for reliability and validity testing.

5.1.1.1 Initial analytic product development

Two serving Australian Army Captains and one Senior Electronic Warfare Scientist took part in the development of the analytic products (ARMY SME1, ARMY SME2 and EW SME respectively).

ARMY SME1 had 1100 flight hours, 800 of which were flying Black Hawk, had been a Troop Commander of a Black Hawk squadron, was qualified as a 'C' Category Pilot and Night Vision Goggles (NVGs) Captain. He had overseas operational experience using the RWR equipment used in the simulation.

ARMY SME2 had 500 hours flying the Black Hawk and had varied experience of Airmobile Operations and was qualified to fly at night with NVGs. EW SME was a Senior Electronic Warfare Engineer at the Defence Science Technology Organisation with more than ten years experience with Electronic Warfare equipment.

5.1.1.2 Reliability and validity assurance

Three individuals took part in the assessment of the analytic products. CWA SME is a recognised national and international expert in Cognitive Work Analysis and has published widely in a variety of learned journals. CWA SME has over 20 years experience in producing CWA products, tools and techniques and has worked for the Department of Defence on a number of CWA related tasks. CWA SME was familiar with the use of RWRs in Black Hawk missions.

HE SME1 is a national expert in Human Engineering and has over 15 years experience as a researcher and practitioner. HE SME1 has worked for both government and industry organisations. A substantial amount of this individual's time has been spent working for the Department of Defence and in particular with the Army, Navy and Air Force.

HE SME2 is an international expert in Human Engineering and has over 30 years experience as a researcher and practitioner. HE SME2 has provided consultancy to North Atlantic Treaty Organisation (NATO), UK Ministry of Defence (MoD) and as consultant to BAE SYSTEMS produced a set of Human Factors Integration Guidelines. HE SME2 was familiar with RWRs and had some experience of their use in helicopters.

5.1.1.3 Information used

The information for the constraint-based analytic products (the WDA and CTA), and for the task-based analytic products was obtained from analysing Army helicopter aviation operational documents, including aircrew flight and standard operating procedures manuals, aircrew checklist and other documents as indicated in Table 5-1.

Table 5-1 Information sources (left) that assisted in the construction of the analytic products (right). Acronyms are given in the Glossary.

Information Source	Analytic products to which information from the sources at left was applied
Australian Army, Land Warfare Doctrine, Part Three, Volume 3, Pamphlet No 3, Airmobile Operations, 2001	AH and ADS (Domain Purpose, Domain Values and Priorities, Physical Functions, Physical Objects), TC-CTA, Task-based (All).
Flight Manual Black Hawk (S-70A-9), Royal Australian Air Force, Australian Air Publication, 7210.015-1, 1997	AH and ADS (Domain Values and Priorities, Physical Functions, Physical Objects), Task-based (All).
Flight Crew Checklist Black Hawk, Royal Australian Air Force, Defence Instruction (Air Force), AAP 7210.015-1CL, 1995	AH and ADS (Physical Functions, Physical Objects), Task-based (All).
Standard Operating Procedures. Part 3 – Flying Operations. 5th Aviation Regiment, 1997	AH and ADS (Domain Functions, Physical Functions, Physical Objects), TC-CTA, Task-based (All).
Interviews with subject matter experts/ Air Mobile Operations scenarios.	AH and ADS, TC-CTA, Task-based (All).

5.1.2 Procedure

The constraint-based and task-based analytic products were developed using the following three steps. First, ARMY SME1, ARMY SME2 and EW SME were interviewed; second, the data from the interviews were collected and supplemented using a number of additional information sources and the analytic products produced; and third, the analytic products were tested for validity and a process of reliability assurance followed.

ARMY SME1 was interviewed using a semi-structured interview technique that was based on a modified Critical Decision Method (CDM) interview technique (Klein, Calderwood & MacGregor, 1989). Prior to the interview ARMY SME1 received instructions to develop two representative airmobile missions. During the interview ARMY SME1 was asked to recount each of the missions in detail and questioned by the interviewer using the prompts shown in Table 4-2. Interviews were taped and the tapes transcribed. A Microsoft Access™ database was developed and used to record and organise the information. During a second interview session ARMY SME1 was again interviewed. The aim of this interview was to clarify information that was not clear from the first interview. The second interview was taped and the tapes transcribed. The database was updated with the new information.

The aim of interviewing ARMY SME2 was to corroborate the information gained from ARMY SME1. ARMY SME2 was interviewed using the CDM-based semi-structured interview technique. Once again, the interview was taped, the tape transcribed and the data entered into the database as was required to meet the reliability requirements.

EW SME was also interviewed with the CDM-based semi-structured interview technique, and information concerning the operation of electronic warfare systems, and radar theory in general, was gathered. The interview was taped and the tape transcribed.

The constraint-based and task-based analytic products were then produced using the interview data and the information from a number of documents (Table 5-1). The analytic products were produced using Winflow™ and Microsoft Office Software.

The steps used to test the construct, internal and external validity of the constraint-based analytic products are as follows.

1. I checked the AH for content omissions. Random parts of the transcribed interview data were selected and the AH was consulted to assess whether the objects, functions, values and priorities and purposes that were identified in the interview data were represented.
2. The CWA SME then assessed the AH against the following questions: Does the AH represent the ecological properties of the work domain? Does the AH represent the properties in a structured means-ends hierarchy? Is the AH event independent? Is the language used to describe objects, function and purposes at the same level of abstraction the same? Is different language used to describe different abstraction levels? Are the objects, functions and purposes represented in the AH described as nouns rather than verbs.
3. The CWA SME then randomly selected parts of the transcribed interview data and assessed whether the objects, functions, values and priorities and purposes that were identified in the interview data were represented in the AH. The size of the selected parts of the transcribed interview data were not defined a priori. Individual words, sentences and paragraphs were used depending on the part of the analytic product that was being assessed.
4. The CWA SME assessed the TC-CTA against the following questions: Is generic activity needed to achieve known goals rather than specific actions (tasks) needed to achieve a goal described? Does the TC-CTA describe what needs to be done, not how or who? Is the TC-CTA device independent? Can the goal identified in the TC-CTA be accomplished in different ways on different occasions? Are the input and output requirements identified for each activity? Are the constraints that act on the activities identified?
5. I used data from the exploratory experiment to check the AH and TC-CTA for content omissions.

HE SME1 and HE SME2 were contracted (paid a commercial fee) to assess whether the task-based analytic products were valid representations of what Human Factors practitioners would produce and whether they were complete, i.e. whether the products contained all the information that should have been included, and whether the information was correct.

The process that HE SME1 and HE SME2 took to assess the analytic products follows:

1. I made available to HE SME1 and HE SME2 the analytic products that were developed, the raw data that was used to construct them, and a list of references that were consulted during the production of the analytic products. The author then asked them to judge whether the analytic products were representative of what Human Factors practitioners would produce.
2. HE SME1 and HE SME2 assessed whether the list of references were complete, i.e. they wanted to establish whether the references used were the ones that Human Factors practitioners would use to guide them in producing the analytic products. They considered each item and assessed whether the list as a whole represented a "good" cross-section of the material available.
3. HE SME1 and HE SME2 assessed whether the definition of the RWR as a medium complexity system was correct by using the classification in Beevis (1999) as a guide.
4. HE SME1 and HE SME2 assessed whether the analytic products were the ones that were most appropriate for a Preliminary Definition Phase of the System Life cycle again by using the classification in Beevis (1999) as a guide.
5. HE SME1 and HE SME2 viewed the raw data that was used to produce the analytic products and considered whether any more data were required.
6. HE SME1 and HE SME2 assessed the content of each analytic product against the raw data to see if the analytic product "correctly" represented the data. The assessment was made by identifying the information elements in each analytic product and comparing that to what the references identified as being required.

5.2 Outputs and discussion

In the following section the constraint-based analytic products will be presented first and then the task-based ones. Within each group the concepts of reliability and validity, as applied to the different types of analytic products, will be discussed with reference to the SMEs assessments.

5.2.1 Constraint-based analytic products

Three CWA analytic products were produced.

- Black Hawk Airmobile (Patrol Insertion) AH (Figure 5-2).
- Black Hawk Airmobile (Patrol Insertion) Electronic Warfare AH (Figure 5-3).
- Black Hawk Airmobile (Patrol Insertion) Electronic Warfare temporal coordination CTA (Figure 5-4).

The first, the Black Hawk Airmobile (Patrol Insertion) AH represents all the components that make up the Black Hawk Airmobile (Patrol Insertion). This was the general model.

The second model, the Black Hawk Airmobile (Patrol Insertion) Electronic Warfare AH, is the model that specifically relates to the RWR system and may be seen to contain a subset

of the information in the general model. This is the model that will be used in the measure-selection method to select measures for evaluation and is the specific model.

The third model is the Black Hawk Airmobile (Patrol Insertion) temporal coordination CTA. This analytic product contains the activity "Manage EW systems" which is the activity central to the process for selecting measures.

5.2.1.1 Black Hawk Airmobile (Patrol Insertion) WDA – The general model

The AH of the Black Hawk Airmobile (Patrol Insertion) is shown in Figure 5-2. As can be seen from the figure, colour and shading have been used to help categorise physical objects, physical functions and domain values and priorities. For example, there are 13 aircraft systems, coloured orange, seen at the Physical Objects level. Lines are also coloured to help distinguish them from each other and show the relationships between objects, functions, priorities and values and the purpose.

The AH indicates that a complex relationship exists between physical objects, their intended functions, and how the intended functions are beneficial in this particular domain. All the relationships between objects, functions, priorities and values and the purpose that were identified via the interview process are recorded in the database that was mentioned above (a screen shot of the data base is given in Appendix B). The Abstraction-Decomposition Space (ADS) is additional to the AH, and it is also shown as Table 5.2. Through the use of the AH and ADS the work domain is represented at several levels of abstraction and at several levels of decomposition. In addition, means-ends (or how-why) relationships are indicated by the links between objects, functions, values and priorities, and the purpose, represented in the AH. In addition, there are a number of properties associated with each of the objects, functions, values and priorities, and the purpose of the AH.

In Figure 5-2 and Table 5.2 it can seen that the ADS shows the same information in the AH but presents this information across abstraction and decomposition axes. The terms "System", "Subsystem", "Unit" and "Component" are awkward because they can readily be used only for physical objects. However, when used in the context of human ideas, or the result of action, for example "Plan", they lose meaning[10]. Hence, although the ADS is shown for all levels, the functions contained within the Domain Purpose and Domain Functions are not decomposed.

The first row in Figure 5-2 and Table 5.2 shows the Domain Purpose of the Black Hawk during Air Mobile Missions which is as follows:

> *"Transport personnel and their equipment from one point to another in a safe and efficient manner within a specified time".*

This statement sums up and provides a rationale for including the objects, functions, and values seen in the remainder of the work domain. All the objects, functions and values are necessary for achieving this purpose. The Domain Purpose is at the "System" level and is not decomposed.

[10] Again it is important to stress that "Plan" is used here in the context of the result of action. Not action itself. It is not used as a verb. Hence, we do not decompose "Plan" into actions like: "Gather information together", "Perform risk analysis", etc.

The Domain Values and Priorities considered to be important are presented in the second row of Figure 5-2 and Table 5.2. These items may be used to judge whether the Domain Purpose is met. For example, "Power limits of aircraft" is one of several values that may be grouped into a set of measures of "Operational limits of the aircraft" which is relevant from a "Safety" point of view. Therefore, one important value that should be adhered to during "Transport personnel and their equipment from one point to another in a safe and efficient manner within a specified time" is "Operation within power limits of aircraft". Failing to observe this value will be breach of the safety constraints.

The Domain Functions considered to be essential within this domain are presented in the third row of the table. "Navigation", "Plan", "Coordination", "Communication", "Flight", "Systems and Resources Management" are all units of "Transportation" and "Tactical Operation" is the only unit of "Tactical Operation". If one takes "Plan" as an example and refers to Figure 5-2 one can see that without the Domain Function "Plan" the individual components of "Optimum tactics" (a value) may not be achieved (which has implications in terms of the mission), which in turn would mean that the Domain Purpose may not be achieved.

The fourth row shows the Physical Functions of the objects that are available in the work domain. These functions are only related to the objects to which they are linked and are independent of a particular work domain. For example, "Physical protection" (is a part of "Aircraft and Cargo Protection", which is a unit of "Black Hawk Resource Functions") is directly linked to the object "Aircraft structure". One could take the object "Aircraft structure" and its function "Physical protection" and apply it to another work domain. However, it is important in this Black Hawk domain because it supports the Domain Function "Systems and Resource Management".

Figure 5-2 and Table 5.2 show all the physical objects [11] that are important in the work domain. These physical objects have a meaning that is specific to the work domain. For example, "Tree" is important not because it is wood and has leaves, but because it absorbs sound. The absorption of sound is just one property of the tree and it is exploited by the aircrew that use it to "Mask" the approach of the aircraft (see Figure 5-2).

The assessment by CWA SME revealed that the Black Hawk Airmobile (Patrol Insertion) AH was in general valid. However, it is noted that strictly speaking if the requirements for construct validity are to be met, the analytic products should represent ecological properties of the world in terms that convey those properties. Although the importance of representing the ecological properties is recognised, it is also important to recognise that it is important to operationalise the ecological properties of the work domain in terms that are useful for system evaluation. Hence, the work domain could have been described in ecological terms, such as time, space, latent energy, etc. but those terms are not useful for differentiating meaningful measures for system evaluation. As an example, "potential

[11] A number of components are described as the system and Human-Machine Interface (HMI) together, e.g. "Electronic Warfare Self Protection (EWSP) system and HMI". The alternative way to describe them is the system and separate displays and controls, e.g. the EWSP system and the HMI system. The former has been adopted because the EWSP system (including its displays and controls) is one replaceable unit.

energy" is an important ecological property associated with manoeuvring helicopters (the higher the altitude the greater the potential energy and the greater the speed that may be achieved in a dive), however, is it far easier to communicate the idea of "operate aircraft within airspeed limits" than it is to communicate the idea of "operate within potential energy limits" especially as a very complex relationship exits between altitude, mass, speed and potential energy. Hence, one can argue that the WDA analytic product has construct validity because it represents important ecological properties even though these properties are conveyed in terms that are not generally used.

This page is intentionally blank.

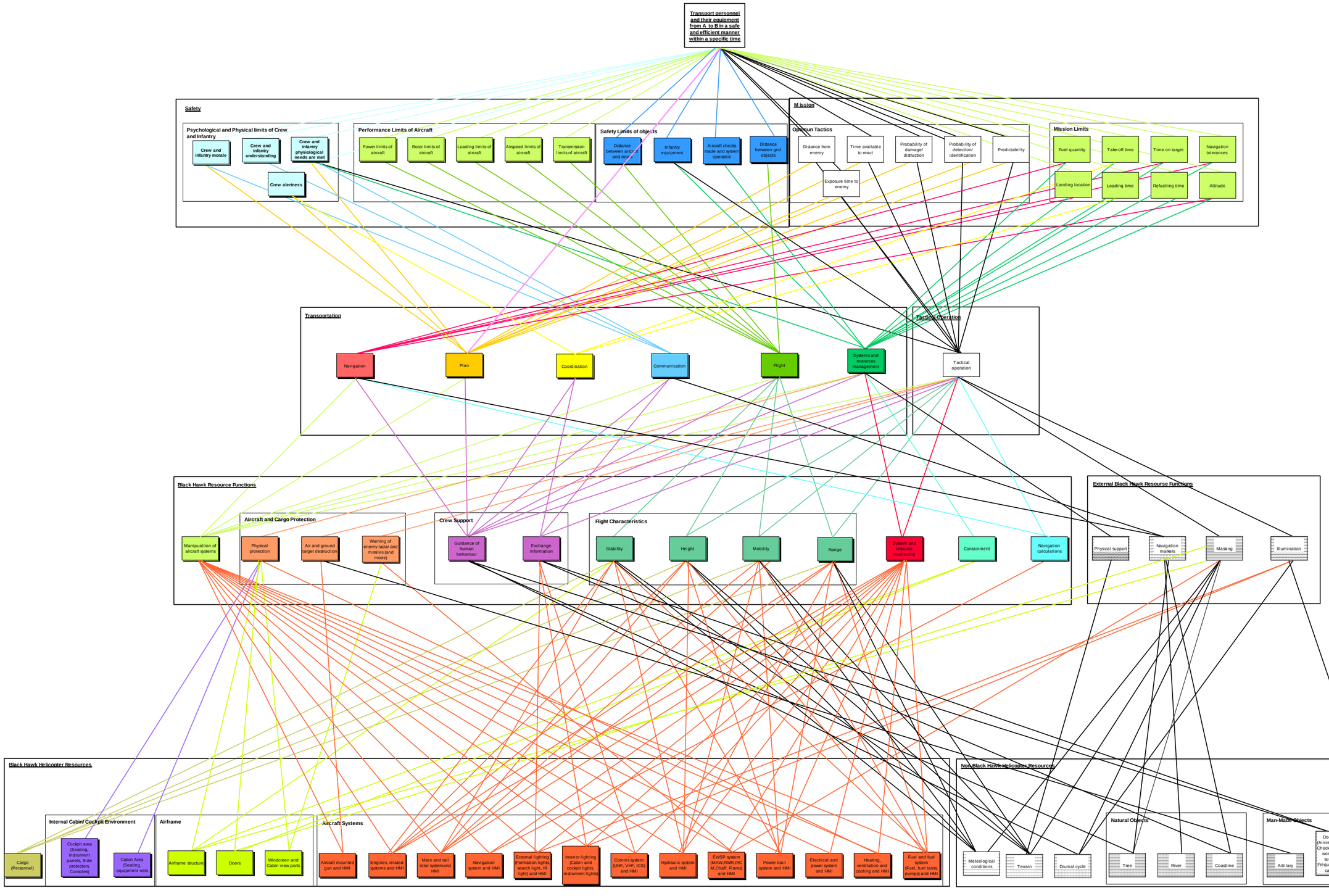


Figure 5-2 AH of the Black Hawk Airmobile (Patrol Insertion).

This page is intentionally blank.

Table 5-2 Abstraction-Decomposition Space (ADS). Each row of the table represents a different abstraction level. The columns represent different decomposition levels.

	System	Subsystem	Unit	Component
Domain Purpose	Transport personnel and their equipment from one point to another in a safe and efficient manner within a specified time			
Domain Values and Priorities		Safety Values Mission Values	Psychological and Physical limits of Crew and Infantry Operational Limits of Aircraft Safety Limits of Objects Optimum Tactics Mission Limits	Crew and infantry morale. Crew and infantry understanding Crew and infantry physiological needs Crew alertness Power limits of aircraft Rotor limits of aircraft Loading limits of aircraft Airspeed limits of aircraft Operation within transmission limits of aircraft Distance between aircraft and rotors Infantry equipment Aircraft checks made and systems operated Distance between ground objects Distance from enemy Exposure time to enemy Time available to react Damage/ destruction Detection/ identification Predictability Fuel quantity Landing location Take-off time Loading time for equipment and infantry Time on target Refuelling time Navigation tolerances ,Altitude.
Domain Functions		Transportation Tactical Operation	Navigation Plan Coordination Communication Flight Systems and Resources Management	

System	Subsystem	Unit	Component
		Tactical Operation	
Physical Functions	Black Hawk Resource Functions	Manipulation of Aircraft Systems	Physical protection
		System and Resource Monitoring	Air and ground target destruction
	External Black hawk Resource Functions	Containment	Warning of enemy radar and missiles
		Navigation Calculations	Guidance of human behaviour
		Aircraft and Cargo Protection	Exchange information
		Crew Support	Stability
		Flight characteristics	Height
		Physical Support	Mobility
		Navigation Markers	Range
		Masking	
		Illumination	
Physical Objects	Black Hawk Resources	Cargo	Cockpit area: Seats
	Non- Black Hawk Resources	Internal Cabin/ Cockpit Environment	Cabin area: Seats
			Airframe structure, Doors
		Airframe	Windscreen and cabin view ports
		Aircraft Systems	Aircraft mounted gun and HMI
		Metrological Conditions	Engine, related systems and HMI
		Terrain	Main and tail rotor system and HMI
		Diurnal Cycle	Navigation system and HMI
		Natural Objects	External lighting and HMI
		Man-made Objects	Interior lighting and HMI
			Communication system and HMI
			Hydraulic system and HMI
			EWSP system and HMI
			Power train system and HMI
			Electrical and power systems and HMI
			Heating, ventilation system and HMI
			Fuel and fuel system and HMI
			Tree, River
			Coastline
			Artillery, Documents, Night Vision Goggles

5.2.1.2 Black Hawk Airmobile (Patrol Insertion) RWR WDA – The Specific model

Following the production and validation of the Black Hawk Airmobile (Patrol Insertion) AH, the part of the AH that specifically represented the Electronic Warfare domain (and the RWR system more specifically) were presented separately. Figure 5-3 is the portion of the general AH that is primarily associated with RWR systems (an RWR system is a subsystem of the larger Electronic Warfare domain). Table 5-3 presents the same information as AH shown in Figure 5-3.

The parts of the Black Hawk Airmobile (Patrol Insertion) AH that were specifically selected as being related to the RWR system was determined by identifying the Electronic Warfare Self Protection System (the RWR forms part of that system) at the Physical Objects level of the AH and following the means-ends links to functions at the Physical Function level. The means-ends links were then followed from these functions to the Domain Functions, and so on until the Domain Purpose was reached.

Given that this analytic product is a subset of the general model it is reasonable to assume that it is valid.

This page is intentionally blank

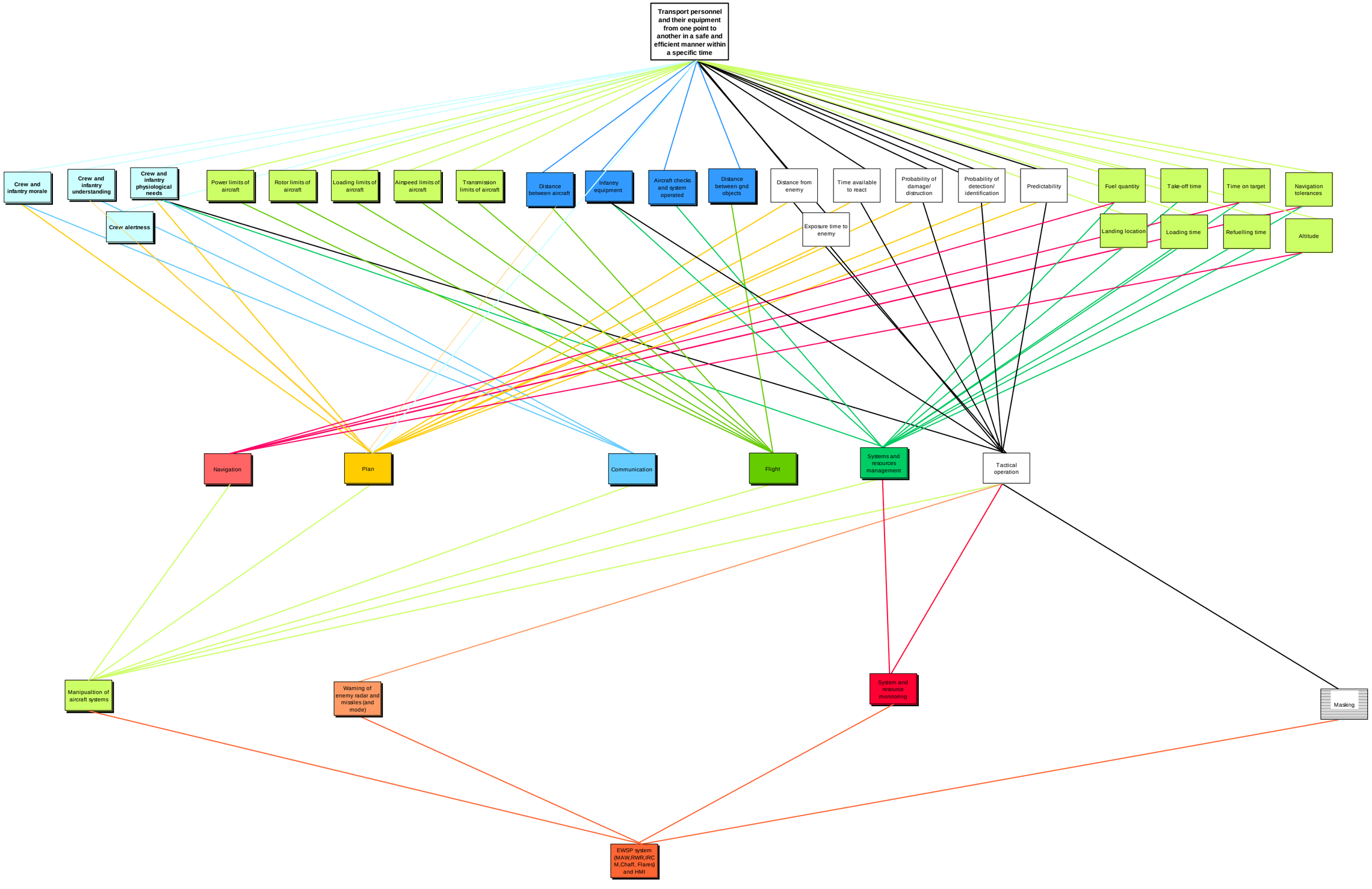


Figure 5-3 AH of the Black Hawk Airmobile (Patrol Insertion) RWR.

This page is intentionally blank

Table 5-3 ADS of the Black Hawk Airmobile (Patrol Insertion) RWR. The highlighting represents the level of granularity shown in the AH

	System	Subsystem	Unit	Component
Domain Purpose	**Transport personnel and their equipment from one point to another in a safe and efficient manner within a specified time**			
Domain Values and Priorities		Safety Values Mission Values	Psychological and Physical limits of Crew and Infantry Performance Limits of Aircraft Safety Limits of Objects Optimum Tactics Mission Limits	**Crew and infantry morale.** **Crew and infantry understanding** **Crew and infantry physiological needs** **Crew alertness** **Power limits of aircraft** **Rotor limits of aircraft** **Loading limits of aircraft** **Airspeed limits of aircraft** **Operation within transmission limits of aircraft** **Distance between aircraft and rotors** **Infantry equipment** **Aircraft checks made and systems operated** **Distance between ground objects** **Distance from enemy** **Exposure time to enemy** **Time available to react** **Damage/ destruction** **Detection/ identification** **Predictability** **Fuel quantity** **Landing location** **Take-off time** **Loading time for equipment and infantry** **Time on target** **Refuelling time** **Navigation tolerances** **Altitude.**

	System	Subsystem	Unit	Component
Domain Functions		Transportation	**Navigation**	
		Tactical Operation	**Plan**	
			Communication	
			Flight	
			Systems and Resources Management	
			Tactical Operation	
Physical Functions		Black Hawk Resource Functions	**Manipulation of Aircraft Systems**	Warning of enemy radar and missiles
		External Black hawk Resource Functions	**System and Resource Monitoring**	
			Aircraft and Cargo Protection	
			Masking	
Physical Objects		Black Hawk Resources	Aircraft Systems	**EWSP system and HMI**

5.2.1.3 Black Hawk Airmobile (Patrol Insertion) Control Task Analysis

Figure 5-4 presents the completed temporal coordination CTA. The temporal coordination CTA (TC-CTA) represents control tasks and activities that have to be performed but does not indicate the specific time when a task or activity is to be performed, i.e. it does not reflect a specific mission scenario. Hence, a double-headed horizontal arrow shows the limits of the possible start and end times for activities.

In the figure the control tasks and activities are labelled along the left vertical axis together with the crew allocated to that particular activity. Along the horizontal axis the mission phases (e.g. Loading Phase and Air Movement Phase) and significant mission events (e.g. arriving at the Initial Point, IP or Air Control Point, ACP [12]) are indicated. The right column indicates the equipment associated with activity. The figure also shows a possible sequence of aircrew activities associated with the "Detect threat" event (which is relevant to RWR operations). In the figure the aircrew and activity sequence is presented as a series of letters and numbers attached to vertical arrows that link the activities together.

For example, the figure shows that the control task "System Management" and the activity "Manage EW system" can occur during any of the three periods between Take-off and Landing. The figure also shows that a response to a threat event (shown as a dashed horizontal red line and labelled as "Detect threat") can occur only after the helicopter has departed from the pick-up (PZ) point. Once a threat is detected the Loadmaster (L), the Non-flying pilot (NFP, also known as the Aircraft Captain (AC)) and Pilot (FP) are all involved in activities associated with defeating the threat. These activities are performed in sequence. For example, the first activity that the Loadmaster and Aircraft Captain perform is "Surveillance of the airspace" (L-1, NFP-1). Once they complete this, the first activity that the Pilot performs is "Control aircraft in airborne flight" (FP-1).

[12] The Initial Point and Air Control Points are navigation points. The former is the point prior to the target; the latter are general navigation points.

A Microsoft Excel spreadsheet was used to record the main aspects of the temporal coordination CTA. The spreadsheet and the transcribed ARMY SMEs interviews that were used to construct the figure provide an auditable path that meets the requirement for reliability as discussed in Chapter 3. The TC-CTA was assessed by CWA SME and found to meet the construct, internal and external validity requirements.

This page is intentionally blank

Figure 5-4 Black Hawk Airmobile (Patrol Insertion) Control Task Analysis

This page is intentionally blank

5.2.2 Task-based analytic products

The task-based analytic products selected for this research program were chosen on the basis of an assessment of an RWR system given in Beevis (1999). That assessment (and subsequent confirmation by independent assessors) indicated that the task-based analytic products that best support an analysis of a medium complexity system at the Preliminary Definition Stage of the System Life Cycle were the following:

- Mission Narrative (MN)
- Function Flow Diagrams (FFD)
- Ad Hoc Function Allocation (FA),
- Time Line Analysis (TL) of the critical task.

The mission narrative descriptions were used for the mission analysis. The mission analysis provided information about high level human-system functions and mission constraints and variables. Function flow diagrams were for the function analysis. Function Flow diagrams indicate the sequence of the functions that must occur (and so must be evaluated). Ad Hoc Function Allocation was used to allocate functions. Ad hoc function allocation identifies which crew and system were responsible for any particular function. Time Line Analysis was used for the Task Analysis process. The time line analysis of the critical tasks was used to provide a baseline set of timing for the tasks.

5.2.2.1 Narrative mission description

Three products were produced to meet the requirements for a mission narrative:

- Table 5-4 shows the main phases associated with Patrol Insertion Airmobile missions in general.
- Table 5-5 shows details associated with the Loading and the Ground Tactical Phases.
- Figure 5-5 provides a schematic of the main events that may occur during a typical mission.

As can be seen from Table 5-4 the mission starts with a Military Appreciation Process (MAP). The MAP is the Planning Phase of the mission. During the process supporting plans are developed for the Ground Tactical Phase, Landing Phase, Air Movement Phase, Loading Phase and Staging Phase. As part of developing the plans, Intelligence Preparation of the Battlespace (IPB) is conducted. This entails conducting a detailed analysis of the enemy. One the MAP has been completed the Warning Order is issued to command elements of the forces concerned with the mission. The purpose of the warning order is to inform the force elements that they should be making themselves ready to conduct a mission. Reconnaissance of the troop Pick-up Zone (PZ), Landing Zone (LZ) will occur. During the Staging Phase troops move to their respective staging areas. Detailed planning is then conduced by each of the force elements for their specific mission during the Planning Phase. The order to proceed with the mission is then given. Each force element rehearses their specific mission in detail. During the Loading Phase all troops move to their PZ and are loaded onto the Black Hawk helicopters. During the Air

Movement Phase the troops are transported to their destination by the helicopters. At the destination (the landing zone, LZ) the helicopters land and the troops disembark and secure the LZ. This is the Landing Phase. The helicopters take off as soon as the troops have disembarked. While the helicopters are in the air the troops conduct their mission. This is the Ground Tactical Phase. Once the troops have secured their objective they are picked up and returned to base.

Table 5-4 Overview of the phases of the Airmobile (Patrol Insertion) mission. The main phases are shown in bold. Some phases may run concurrently

Mission Analysis – Patrol insertion
Military Appreciation Process (MAP)
Issue Warning Orders
Reconnaissance of possible PZs, LZs, routes and the area of the objective
Ground and aviation elements move to the staging areas **(STAGING)**
Detailed planning and coordination conducted **(PLANNING)**
Orders issued
Rehearse AMO
Move ground and aviation elements to the PZ and load **(LOADING)**
Air movement of troops. Enemy in the area of the LZ, objective and flight routes neutralises by air strikes, AFS, artillery or other ground elements **(AIR MOVEMENT)**
Force lands and secures LZ **(LANDING)**
Ground elements move to secure the objective **(GROUND TACTICAL PHASE)**
Extraction

Table 5-5 provides the details of the Loading and Ground Tactical phases that are most important for this program of research. The table provides the mission aim, details of the relevant phases, aircraft and system information, environmental conditions and, finally, events that may occur during the Air Movement Phase.

Figure 5-5 shows the main events that occur during the mission. Each of the mission phases is labelled. The Aircraft Control Points (ACPs) are navigation waypoints. The figure indicates the helicopter takes off and lands at the Forward Operating Base (FOB) during a typical mission. During the mission the aircraft transits through the Pickup Zone (PZ), ACPs, Dummy landing points (DLZ), Initial Point (IP) and lands at the Target (TGT). Dummy landing points (DLZ) are used to confuse the enemy. The action for these includes performing a landing, but not disembarking the troops, then taking off and resuming the mission. The aim of this tactic is to confuse the enemy on the real intention of the mission.

HE SME1 and HE SME2 found the narrative mission description to be valid (the SMEs full report is available from the author on request). HE SME 2 did note that the products provided the information that would typically be seen in a narrative mission description and "was complete". However, HFE SME2 specifically noted that although the information elements were included, they were presented as "…a mission definition than a narrative…". However, it was later acknowledged by the HE SME2 that the transcription of the interviews performed with aircrew provided the "narrative" (HFE SME2, personal communication). In addition, the mission narrative produced met the general output requirement for mission narratives in general. The requirement is, as Beevis (1999) notes; "The outputs of the techniques should be sufficiently detailed to identify the upper-level functions performed by the system" (p.40). Given that high level functions were identified

as a result of the outputs produced and the fact that HFE SME1 agreed the narrative mission description "was complete", the narrative mission description, as produced by the author, was deemed to be valid.

Table 5-5 Details of the Loading and Ground Tactical Phases of the Airmobile (Patrol Insertion) mission

Mission Characteristics	Details
Mission aim	The aim of this mission is to deliver troops and their equipment to a pre-briefed position, at a specific time, to conduct their specific tasking.
Mission details (Phases: Loading to Ground Tactical)	During the Loading Phase the aircraft and crew leave the forward operating base (FOB) and transit to the pick-up zone (PZ). At the PZ troops and their equipment are loaded onto the aircraft and the aircraft departs. During the Air Movement Phase the aircraft may encounter a pop-up threat and will take evasive action. After surviving the threat the aircraft navigates via two air control points (ACPs) to a dummy landing zone (DLZ), where the aircraft lands and takes-off. The aircraft navigates to the initial point (IP). On the way to the IP another pop-up threat may be encountered. During the Landing Phase the aircraft leaves the IP and lands at the target area (TGT). Once at the TGT troops and their equipment are unloaded (Ground Tactical Phase) and the aircraft departs. During the mission standard operating procedures (SOPs) are adhered to.
Aircraft configuration and mission specific equipment	The Black Hawk is crewed by its normal complement of four. It is configured for troop transport. Apart from its normal systems the aircraft is fitted with the Gemini Electronic Warfare system. The crew are using Night Vision Goggles (NVGs).
Environmental conditions	The mission is conducted at night. Level of precipitation and cloud cover may vary. The whole mission is conducted over land.
Events	Pop-up (ambush) threats (Infra-red, Radio Frequency surface to air missiles, gun systems) may occur during the Air Movement Phase. Aircraft systems may fail.

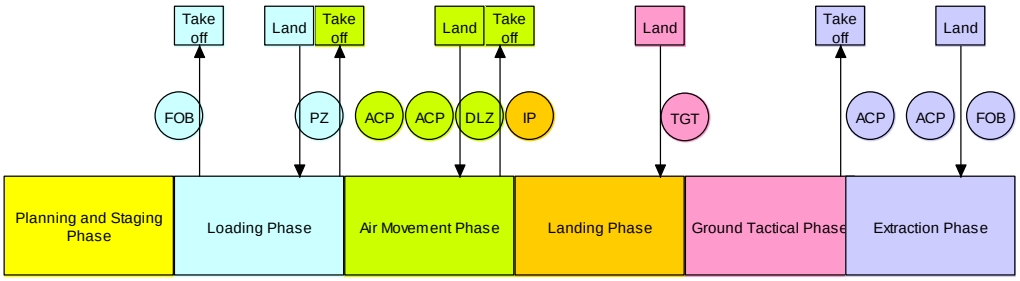


Figure 5-5 Significant points during the Airmobile Operation (Patrol Insertion) mission

5.2.2.2 Function Flow Diagrams and Ad Hoc Function Allocation

The actual function flow diagrams and ad hoc function allocation that were developed are shown in Appendix C. The figures presented below show the Phases, sub-phases, functions, activity and tasks associated with a part of the analysis conducted for the Airmobile Operation.

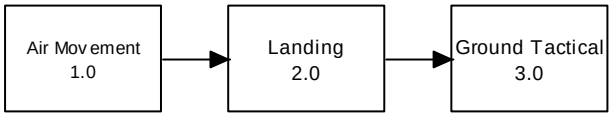


Figure 5-6 Main Airmobile Operations mission phases

Figure 5-6 shows the main phases that constitute an Airmobile Operation. These are Air Movement, Landing and Ground Tactical. The figure shows that the Air Movement Phase precedes the Landing Phase, which in turn precedes the Ground Tactical Phase.

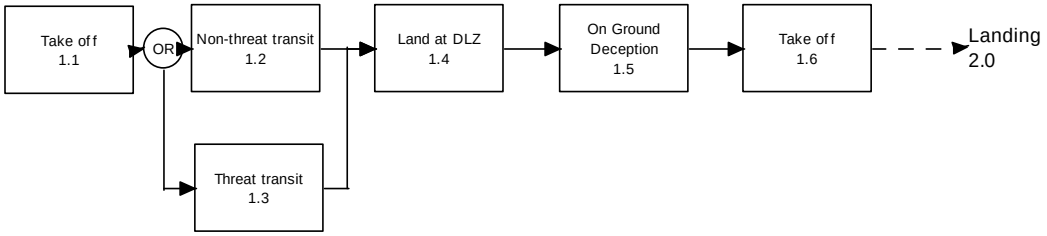


Figure 5-7 Sub-phases of Air Movement (1.0)

Figure 5-7 shows the sequence of sub-phases that make up the Air Movement phase. After take off the aircraft can either encounter or not encounter a threat. After landing at the dummy landing zone, and after the on ground deception has been completed, the aircraft takes off. The Landing Phase is then performed.

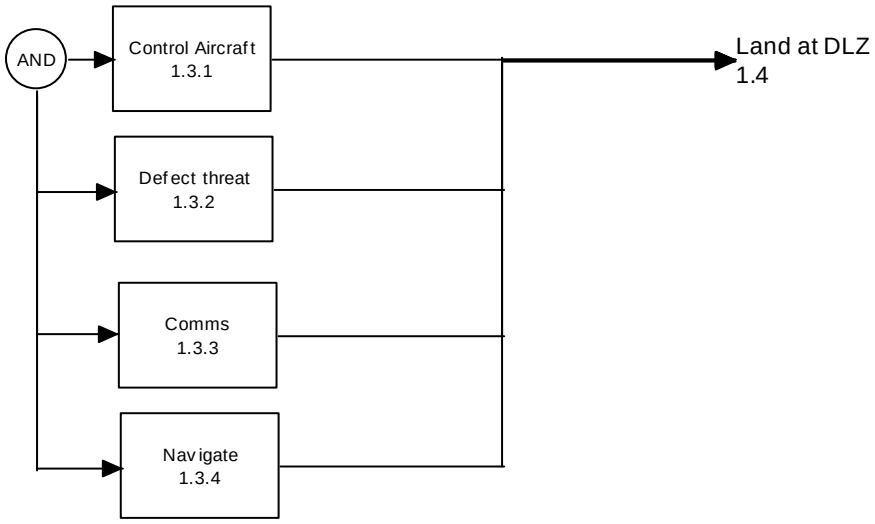


Figure 5-8 Functions of Transit with Threat (1.3)

Figure 5-8 shows that four functions are performed simultaneously during the Transit with Threat sub-phase. Once these are completed the Land at the Dummy Landing Zone sub-phase is then performed.

Once all the functions were identified the activities associated with the Defeat Threat function were identified and allocated to the aircrew and aircraft systems.

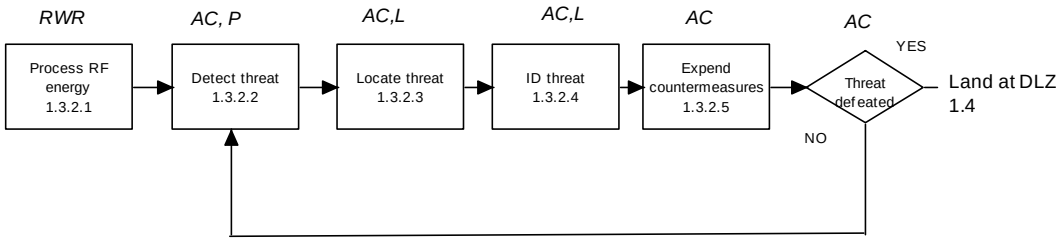


Figure 5-9 Activities for Defeat threat (1.3.2.).The allocation of the activity to the RWR, Aircraft Captain (AC), Pilot (P) and Loadmaster (L) is also shown.

Figure 5-9 presents the activities associated with the Defeat Threat function and the aircrew and system allocation against those activities. The figure indicates that it may take several attempts to defeat a threat and it is the Non-flying pilot (the Aircraft Captain) that makes the decision on whether the threat has actually been defeated.

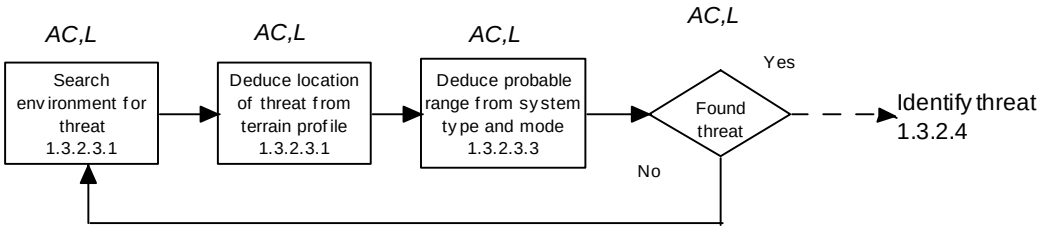


Figure 5-10 Tasks for Locate threat (1.3.2.3). The allocation of the task to the RWR, Aircraft Captain (AC), Pilot (P) and Loadmaster (L) is also shown.

Figure 5-10 shows the tasks associated with Locate Threat activity. The figure clearly indicates that the Non-flying pilot (the Aircraft Captain) and the Loadmaster are both involved and that both may or may not find the threat.

HE SME1 found that the function flow diagrams and function allocation were representative of the ones that Human Factors practitioners would produce. HFE SME2 commented on the presentation of the information rather than the information that the diagrams contained. It was concluded, therefore, that the function flow diagrams and function allocation were representative of the products from Human Factors practitioners. (The SMEs' full report is available from the author on request.)

5.2.2.3 Timeline Analysis of Critical Tasks

The timeline analysis of the critical tasks provides sequencing and timing information for the tasks that have been identified as being critical for workload or safety. The tasks that were identified as meeting these criteria are given in the first column of Table 5-6.

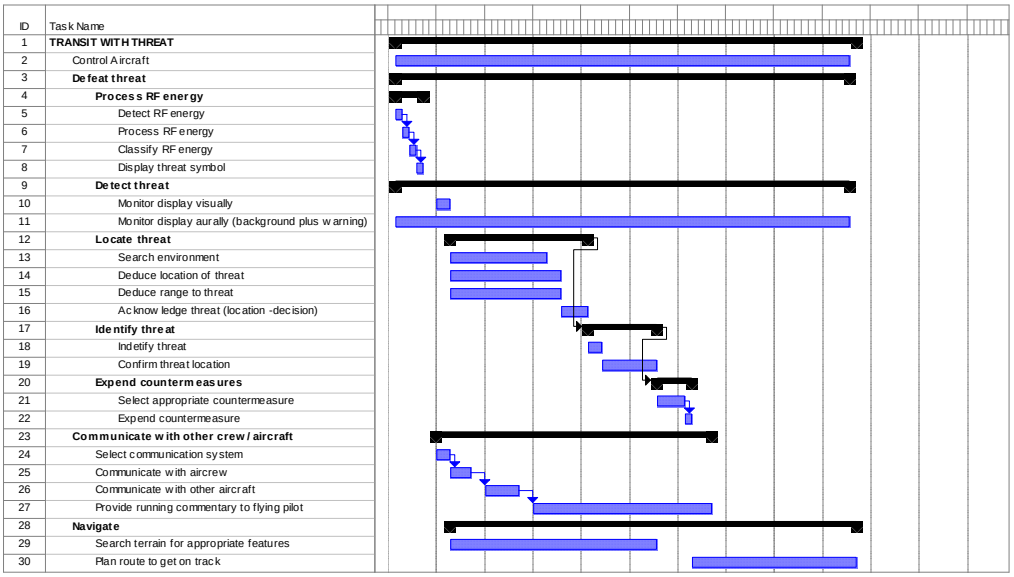
Table 5-6 Task timings for the Defeat threat function. The tasks are shown in relation to the parent activity.

Task and Activity	Time (sec)
Process RF energy	
Detect RF energy	0.25
Process RF energy	0.25
Classify RF energy	0.25
Display threat symbol	0.25
Detect threat	
Monitor display visually	0.5
Monitor display aurally (background and warning	24
Locate threat	
Search environment	2.5
Deduce location of threat	3
Deduce range to threat	3
Acknowledge threat (location- decision)	0.5
Identify threat	
Identify threat	0.5
Confirm threat location	1.5
Expend countermeasures	
Select appropriate countermeasure	0.5
Expend countermeasure	0.25
Communicate with other crew/ aircraft	
Select communication system	0.5
Communicate with aircrew	0.75
Communicate with other aircraft	1.25
Provide running commentary to flying pilot	5
Navigate	
Search terrain for appropriate features	5.5
Plan route to get on track	4.5

Figure 5-11 shows a screen shot of the Microsoft Project GANT chart that was used to present the information. The figure shows the hierarchical nature of the functions, activities and tasks and also shows the temporal relationships between tasks. Table 5-6 shows the duration (in seconds) of each of the tasks. The task durations shown are representative of the real times[13].

HFE SME1 and HFE SME2 found the timeline analysis to be representative of what Human Factors practitioners would produce. (The SMEs' full report is available from the author on request.).

[13] Given the security classification of this thesis it was not possible to use real task durations.



ID	Task Name
1	**TRANSIT WITH THREAT**
2	Control Aircraft
3	**Defeat threat**
4	**Process RF energy**
5	Detect RF energy
6	Process RF energy
7	Classify RF energy
8	Display threat symbol
9	**Detect threat**
10	Monitor display visually
11	Monitor display aurally (background plus warning)
12	**Locate threat**
13	Search environment
14	Deduce location of threat
15	Deduce range to threat
16	Acknowledge threat (location -decision)
17	**Identify threat**
18	Indetify threat
19	Confirm threat location
20	**Expend countermeasures**
21	Select appropriate countermeasure
22	Expend countermeasure
23	**Communicate with other crew / aircraft**
24	Select communication system
25	Communicate with aircrew
26	Communicate with other aircraft
27	Provide running commentary to flying pilot
28	**Navigate**
29	Search terrain for appropriate features
30	Plan route to get on track

Figure 5-11 Screen shot of the Microsoft Project GANT Chart for the Transit with Threat sub-phase.

5.3 Conclusion

The development of the task-based and constraint-based analytic products represents an important second step of the research program. This chapter described the analytic products that were produced and discussed the assessment of them by SMEs. The types of analytic products used for this research were selected on the basis of previous work. Each of the analytic products was constructed using established practice.

Chapter 4 discussed several ways to evaluate the analytic products for reliability and validity. For reliability it was concluded that, given the complexity of the domain in this program of work, the best approach would be to ensure that the process taken to construct the products was auditable and that the data used to construct the products was made available for inspection. For validity, it was concluded that independent SMEs could be used to assess the various types of validity.

In this chapter reliability assurance was achieved of the analytic products by ensuring that the process taken to construct the products was auditable and that the data used to construct the products is available for inspection. This chapter showed that the analytic products were found to have construct validity and internal validity by independent SMEs and the author. The chapter also showed that the constraint-based analytic products were found to have external validity by the author using a data set that had not been used to construct the products. Finally, the chapter showed that the task-based products were not formally assessed for external validity. On the basis of the assessments made by the independent SMEs it is concluded that the analytic products are valid and can be used in the measure-selection methods. Those methods are described next.

6. Selecting Measures for Evaluating the Test Case System

Chapter 5 described the task- and constraint-based analytic products that were developed, which is Stage 2 in evaluating methods for selecting measures for system evaluation. Figure 6-1 shows, again, the four stages of the research program. Stage 3, the subject of this chapter, is to develop and produce reliable and valid methods that will use the analytic products produced earlier to select measures.

As with the previous chapter, the concepts of reliability and validity are central. However, here I am concerned with reliability and validity of methods rather than analytic products. Once I have produced reliable and valid methods then I can be confident that the measures selected are representative of the ones that are seen in HEP and CWA.

In Chapter 4 the reliability of the measure-selection methods was discussed in terms of assurance rather than formal testing. In this chapter I will show that the reliability of the methods has not been tested formally but a strategy of reliability assurance has been put in place. I will also show that the validity of the methods have been tested using independent SMEs. Additionally, I will show that the measures produced by the task-based method were assessed for validity by independent SMEs.

In the following section (Section 6.2) the method that was used to develop the measure-selection methods is described. The measure-selection methods are then presented and discussed in Section 6.3. In Section 6.4 the measures that were derived from following the methods are presented. Sets of measures for the current RWR and the future RWR will be defined. The chapter will then conclude.

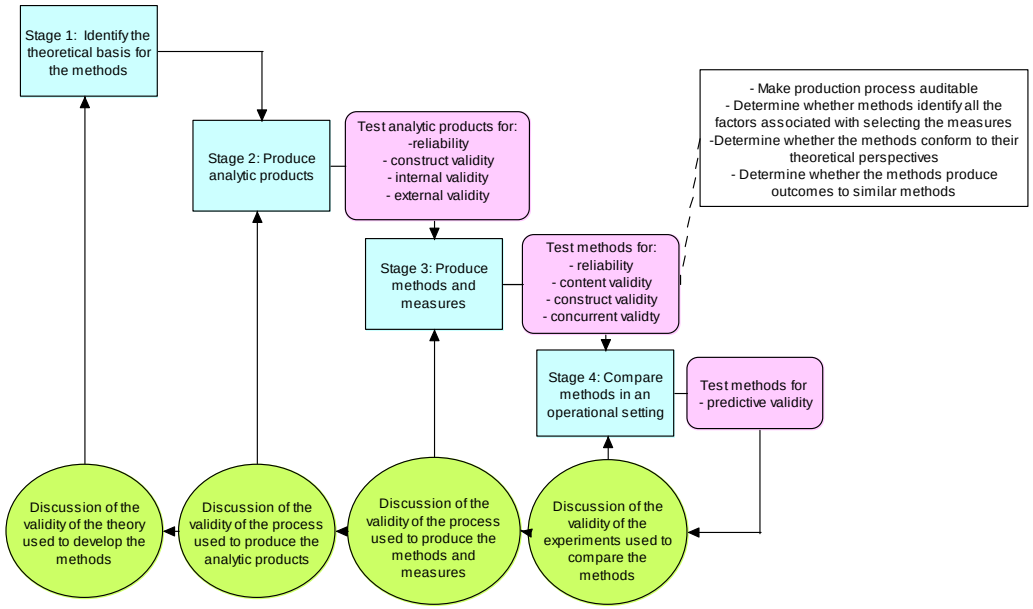


Figure 6-1 Four stages of the research program. In Stage 3, methods are developed for extracting measures (measure-selection methods) and the measures are produced. Chapter 6 focuses on the production of the measures.

6.1 Process used to develop measure-selection methods

The process taken to produce the methods involved reliability assurance and validity testing. The method of reliability assurance adopted by the author included the following activities. First, following the methods through and documenting all issues found. Second, specifying the methods in sufficient detail so that another reasonably knowledge person could probably follow it and produce similar results. Third, recording all the measures that were suggested as being sensitive to the system modification. The validity of the methods was assessed by independent SMEs.

To assure reliability the author developed and documented, using flow charts, two methods that could be used by an analyst with the help of the analytic products to select measures to use during system evaluation. In the case of the constraint-based method, the flow charts presented here incorporate the AH and TC- CTA analytic products that were described in the previous chapter. The information sources identified previously, and given in Table 5-1, as well as the Microsoft Access™ database that was identified earlier, were all used in the execution of the process. In the case of the task-based approach, the method was derived from consulting a number of HE design standards and other source material (see Table 6-2).

Once the task-based method had been developed it was assessed for validity by SMEs. HE SME1 and HE SME2 were consulted to see whether the method for selecting measures was truly representative of world best practice. In particular they were asked to assess whether the method was one that HE practitioners would develop and use and whether the measures that were produced from it were the measures that should be produced. The experts' conclusion was that it was. (The SMEs' full report is available from the author on request).

Once I was satisfied that the methods met the requirements for validity, they were used to produce the candidate measures for the evaluation and the measures recorded. In the next section each of the flowcharts will be presented and issues regarding them will be discussed. A brief outline of the process that the SMEs took to assess the methods is also given.

6.2 Description of measure-selection methods

In this section the constraint-based measure-selection method is described followed by the task-based measure-selection method. The process that the SMEs followed to assess the methods is then briefly discussed.

6.2.1 Constraint-based measure-selection method flowcharts

Figure 6-2 shows the method that was developed to select potentially sensitive measures using the AH. Figure 6-3 shows the method that was developed to select potentially sensitive measures using the temporal coordination CTA.

The measure-selection method flow charts are presented so that the steps that make up the method as a whole flow from top to bottom. Inputs are shown on the left of a step, whereas outputs are shown on the right. Outputs from earlier steps may become inputs to later steps. The steps are labelled sequentially with either the WDA or CTA prefix (for example, WDA1, WDA2…WDAn). Inputs and outputs labelled numerically (1…n).

Appendix D provides details about all the individual inputs and outputs of the method. The following paragraphs will describe how the method was used to select candidate measures for testing. The corresponding steps of the methods are also given in parenthesis.

For the purposes of this program of research the author was interested in the impact of a system modification associated with the RWR subsystem. The RWR subsystem is part of the "EWSP system and HMI" object. Hence, the RWR was the object selected from the AH (Figure 5-2). A modification to the RWR was possible because the evaluation was performed in the simulator—in reality, the simulator would be emulating an existing or specified RWR system.

The RWR system was modified to provide a greater degree of radar sensitivity (See Section 3.2.1 for details). One of the properties of the EWSP system and HMI object that is affected by the modification is range to target (see step WDA2 in Figure 6-2). Using that property all the objects, functions, values and priorities and purpose that are logically related and shown in the WDA are selected and recorded. Figure 6-4 shows only the objects, functions, values and priorities and purpose associated with the RWR system modification (see step WDA3). For each of the objects, functions, values and priorities and purpose properties an assessment is made of how the system modification would affect that property, and then recorded (see steps WDA4, WDA5, WDA6, WDA7, WDA8 and WDA9). Issues considered included whether the modification would change the property in some way and, if so, how. The assessment is made on the basis of SME advice and a detailed understanding of the airmobile domain.

Once all the properties that could be affected by the modification are identified and the effect of the modification noted (WDA10, WDA11), the properties are operationalised in the context of the domain, if necessary. (Note that some levels of the AH are operationalisations of the properties, whereas other levels are not). Following this the properties are parameterised so that that a clear definition of the data that should be collected is produced (this is vital for testing the measures) (WDA12). Table 6-1 shows the measures associated with the RWR sensitivity modification and Section 6.4.1 provides definitions for the measures. Once this part of the process is completed the second part of the process using the TC-CTA is initiated.

With the TC-CTA, once again the property of interest is range to target. Given that there different levels in the abstraction hierarchy it is important to consider where to "enter" the CTA method (CTA1). Range to target is from a WDA object on the Physical Object level of the AH. Hence, Figure 6-3 is "entered" at the point associated at CTA6. The significant event associated with range to target is Detect Threat (CTA11). The section of the TC-CTA is noted (CTA12). The activity that is directly related to this event is Manage EW systems (CTA13). The indirect activities, for the Aircraft Captain, are Surveillance of Airspace and

Navigate/ route find (CTA14). For each of the activities identified measures are selected based on existing guidelines (CTA15). Depending on where in the System Life Cycle the evaluation activity occurs previous studies may further inform the choice of measures. At this point, the constraint-based theory and the constraint-based method have constrained exactly what property needs to be measured. Once the measures are selected an assessment of how that measure will be affected is made and a list of activities (with associated measures) that are, and are not, predicted to change, results (CTA16).

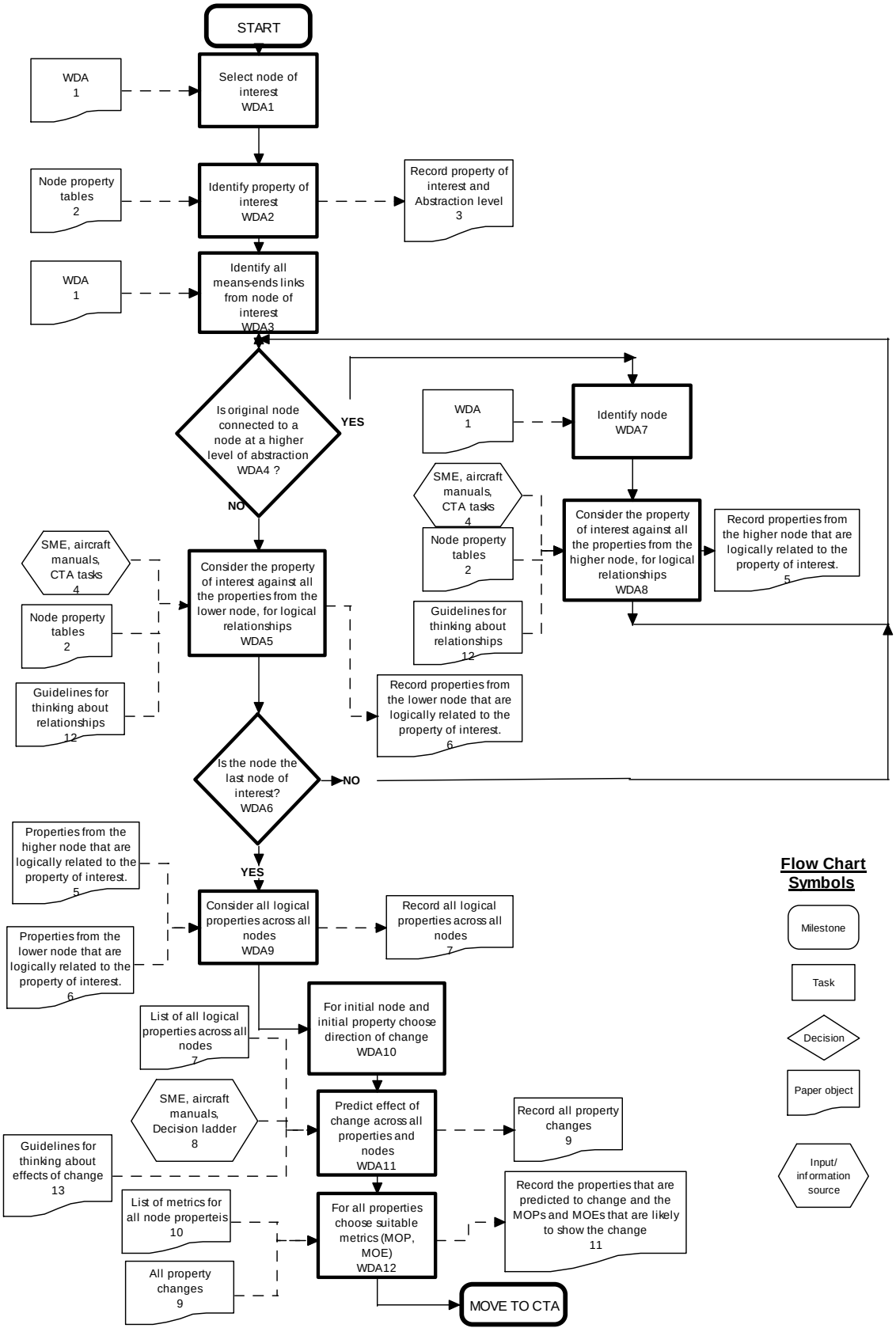


Figure 6-2 Method for selecting sensitive measures using the abstraction hierarchy

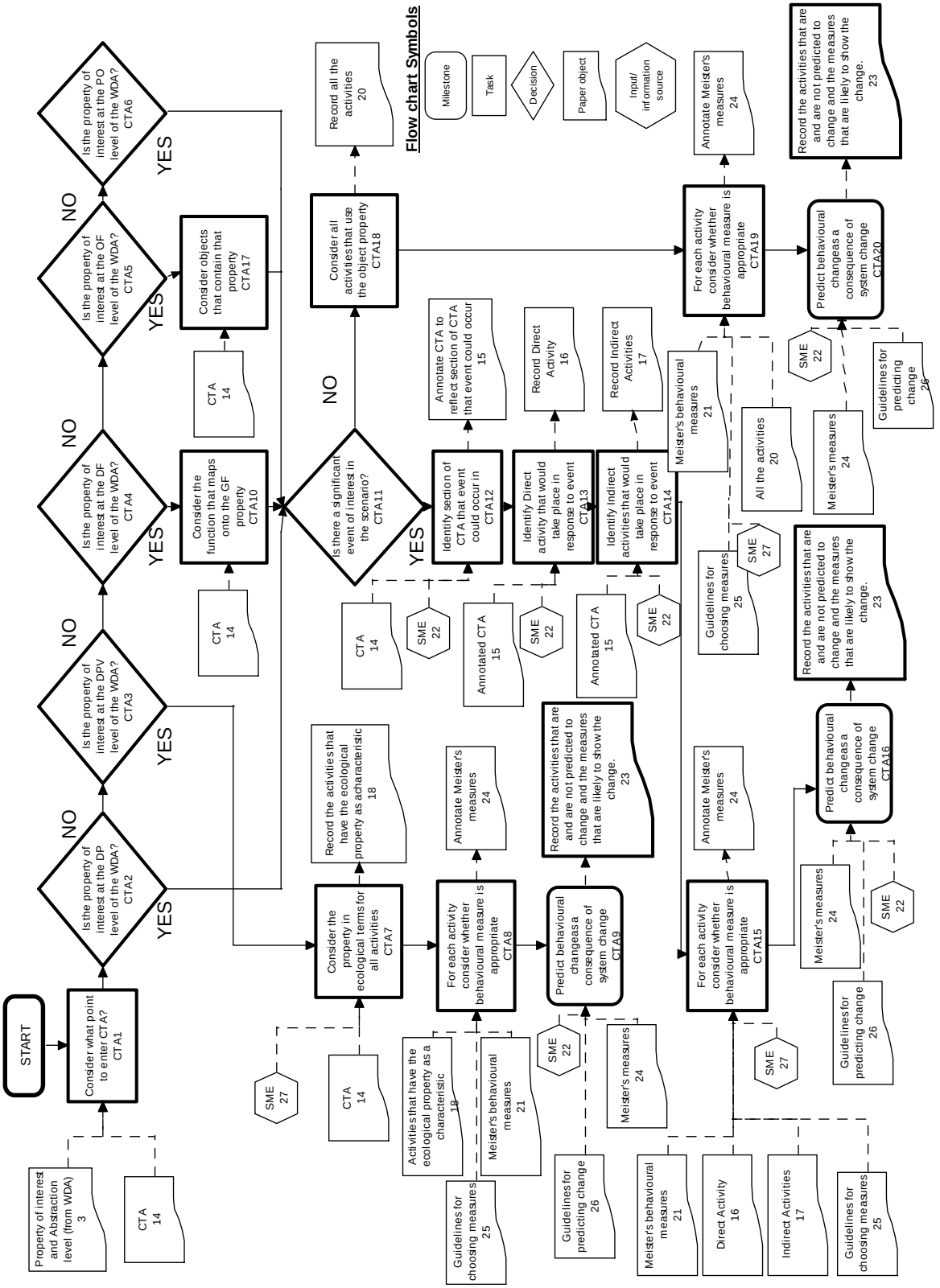


Figure 6-3 Method for selecting sensitive measures using the temporal Control Task Analysis

This page is intentionally blank

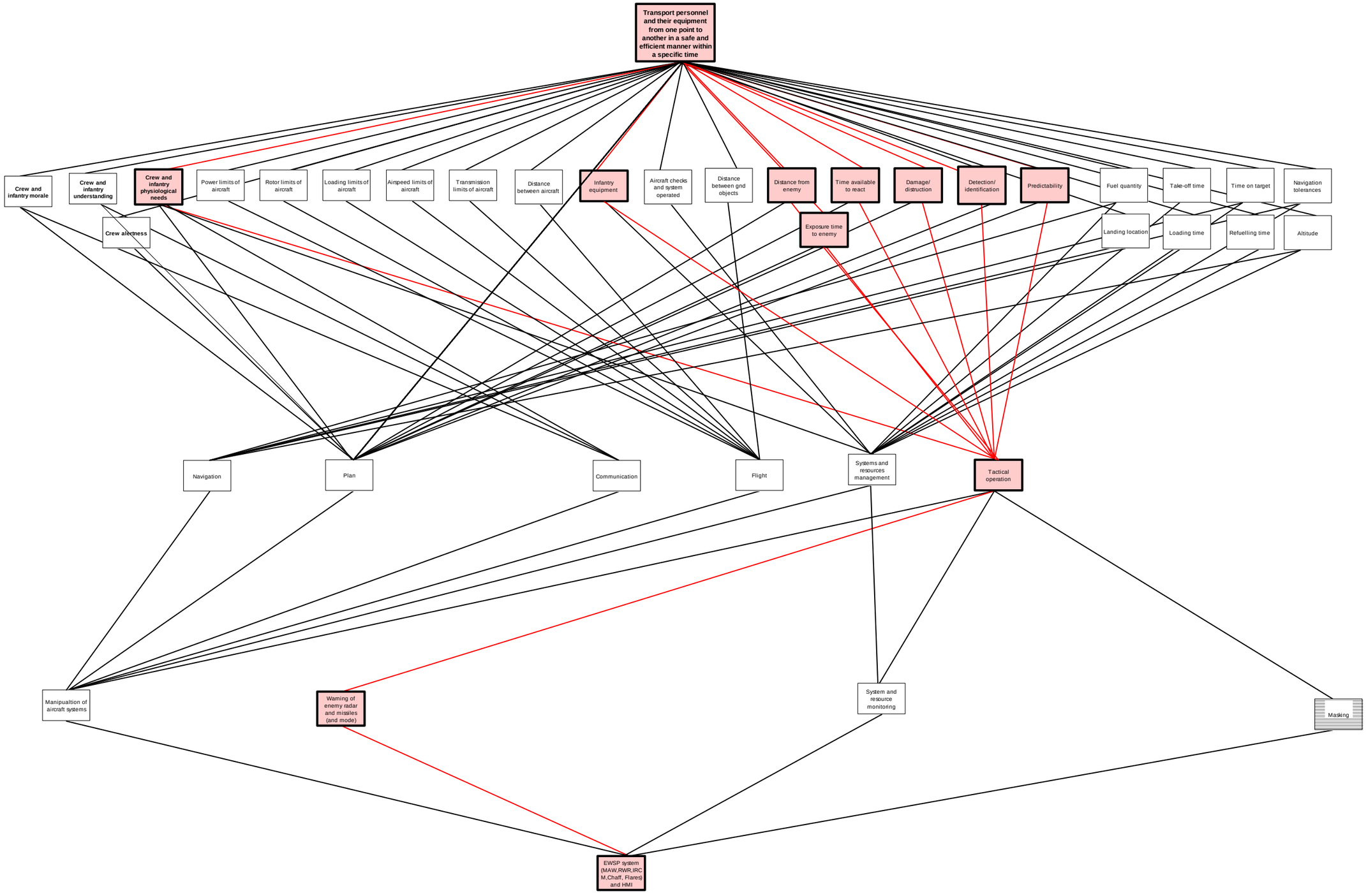


Figure 6-4 Black Hawk Airmobile (Patrol Insertion) RWR. The highlighted parts are related to RWR sensitivity

This page is intentionally blank

Table 6-1 AH objects, functions, values and priorities and purpose and associated measures related to RWR sensitivity

AH Objects, Functions, Values and Priorities, and Purpose	**Measure name** (see Section 6.4.1 for definitions)
Domain Purpose: Transport personnel and their equipment from one point to another in a safe and efficient manner within a specific time.	Mission Achieved
Domain Priority and Value: Crew and infantry physiological needs	Crew physiology
Domain Priority and Value: Infantry equipment	Infantry Equipment Undamaged
Domain Priority and Value: Distance from enemy	Distance to enemy
Domain Priority and Value: Detection/ identification	Probability of detection
Domain Priority and Value: Damage/ destruction	Probability of damage/ destruction
Domain Priority and Value: Exposure time to enemy	Exposure time to enemy
Domain Priority and Value: Time available to react	Maximise time available to react
Domain Priority and Value: Predictability	Minimise predictability
Domain Function: Tactical Operation – Surprise	Surprise
Domain Function: Tactical Operation – Timeliness	Timeliness
Domain Function: Tactical Operation – Tempo	Tempo
Domain Function: Tactical Operation – Understanding	Understanding
Domain Function: Tactical Operation - Decision making	Decision making
Physical Function: Warning of enemy radar and missiles (and mode) - Number of threats	Number of threats

6.2.2 Task-based measure-selection method flowchart

The task-based flow chart uses the task-based analytic products that were described in the previous chapter. The information sources identified previously, and given in Table 6-2, as well as the Microsoft Access™ database that was identified earlier, were used to execute the method.

Figure 6-5 shows the task-based method that was developed for evaluating a socio-technical system. One part of that method specifically relates to selecting sensitive measures. As with the previous flowcharts, the steps flow from top to bottom. Inputs are shown on the left of a step, whereas outputs are shown on the right. Outputs from earlier steps may become inputs to later steps. The steps are labelled sequentially with the HE prefix (for example, HE1, HE2…HEn). Inputs and outputs are labelled numerically (1…n). The following paragraphs will describe how the method was used to select measures that were potentially sensitive to the RWR modification.

The method begins with the appropriate task-based analytic product type being selected for use in the evaluation (step HE1). This is based on the SLC phase and type of system being evaluated (as described in Chapter 2). Each of the analytic products are then produced (steps HE2 to HE8 inclusive). A list of appropriate behavioural measures is then produced (step HE9) as a result of considering each measure against selection criteria (see Chapter 2 for typical lists of criteria). The measures can then be used as dependent variables in any empirical evaluation (steps HE10, HE11).

Table 6-2 HE material consulted in the development of the task-based method

HE guide or standard	Notes
Electronic Industries Alliance (EIA). (2002). Engineering bulletin. Human engineering principles and practices (HEB1)	Outline of Human Factors Engineering best practice based on 46855A
Beevis, D. (Ed.). (1999). Analysis Techniques for Human-Machine System Design: A report produced under the auspices of NATO Defense Research Group Panel 8 (CSERIAC-SOAR-99-01). Wright Patterson Air Force Base, OH: Crew System Ergonomics Information Analysis Center.	Survey of Human Factors Engineering techniques by system engineering phase
Kirwan, B. and Ainsworth, L.K. (Eds.) (1992). A Guide to Task Analysis. London: Taylor and Francis.	Guide to normative (task analytic) techniques.
DEF STAN 00-25(Part 12) / Issue 1. (1989). Human Factors for Designers of Equipment Part 12 – Systems. United Kingdom Ministry of Defence.	A guide for the integration of Human Factors Engineering practices into system design and evaluation.
Smode, A.F., Gruber, A. and Ely, J.H. (1962).The Measurement of Advanced Flight Vehicle Crew Proficiency in Synthetic Ground Environments. Behavioral Sciences Laboratory, Wright-	Guidelines for selecting performance measures aviation based tasks.

HE guide or standard	Notes
Patterson Air Force Base, Report Number MRL-TDR-62-2.	
MIL-HDBK-46855A. (1999) DOD Human Engineering Program Process and Procedures.	Human Factors Engineering practices and procedures to follow when designing and testing a system.
ASCC 61/116/12 (1996) Crew Performance Measurement.	A guide on how to measure human performance in aircraft design, test and evaluation. This document includes performance measures and criteria for selecting them.
RTO-TR-021 AC/323(HFM-018) TP/19. (2001). NATO Guidelines in Human Engineering Testing and Evaluation. North Atlantic Treaty Organization.	Recommended guidelines for accompanying Human Factors Engineering test and evaluation.
AIR-STD-61/116/13 (1996). The Application of Human Engineering to Advanced Aircrew Systems. United States Department of Defense.	A guide on how to measure human performance in aircraft design, test and evaluation. This document includes performance measures and criteria for selecting them. This is the same as ASCC 61/116/12 (1996) Crew Performance Measurement.
ISO 13407: Human-Centred Design Processes for Interactive Systems	This document provides guidance on human-centred design activities for computer-based interactive systems relating to all stages in the system life cycle. Although aimed at computer-based systems the guidance is seen to be applicable to non-computer based systems.
ISO 15288: Life Cycle Management - System Life Cycle Processes	The standard provides a common framework for describing the life cycle of systems. It also identifies a set of well-defined processes to facilitate project definition, control and improvement.
ISO 15504: Software Process Assessment	This document provides a management capability scale for assessing how well software engineering processes in a model are being performed. This scale can equally be applied to systems engineering and human factors process models.
Human Factors in Defence (SMI Conference Proceedings, June 2005)	This represents the state of the art in Human Factors Integration best practice with contributions from leading researchers in the US, Canada, UK and Europe.
Human Factors Integration (DTC Conference Proceedings, April 2004)	This is a series of case studies taken from the UK defence force (Air Force, Army Navy) and industry.

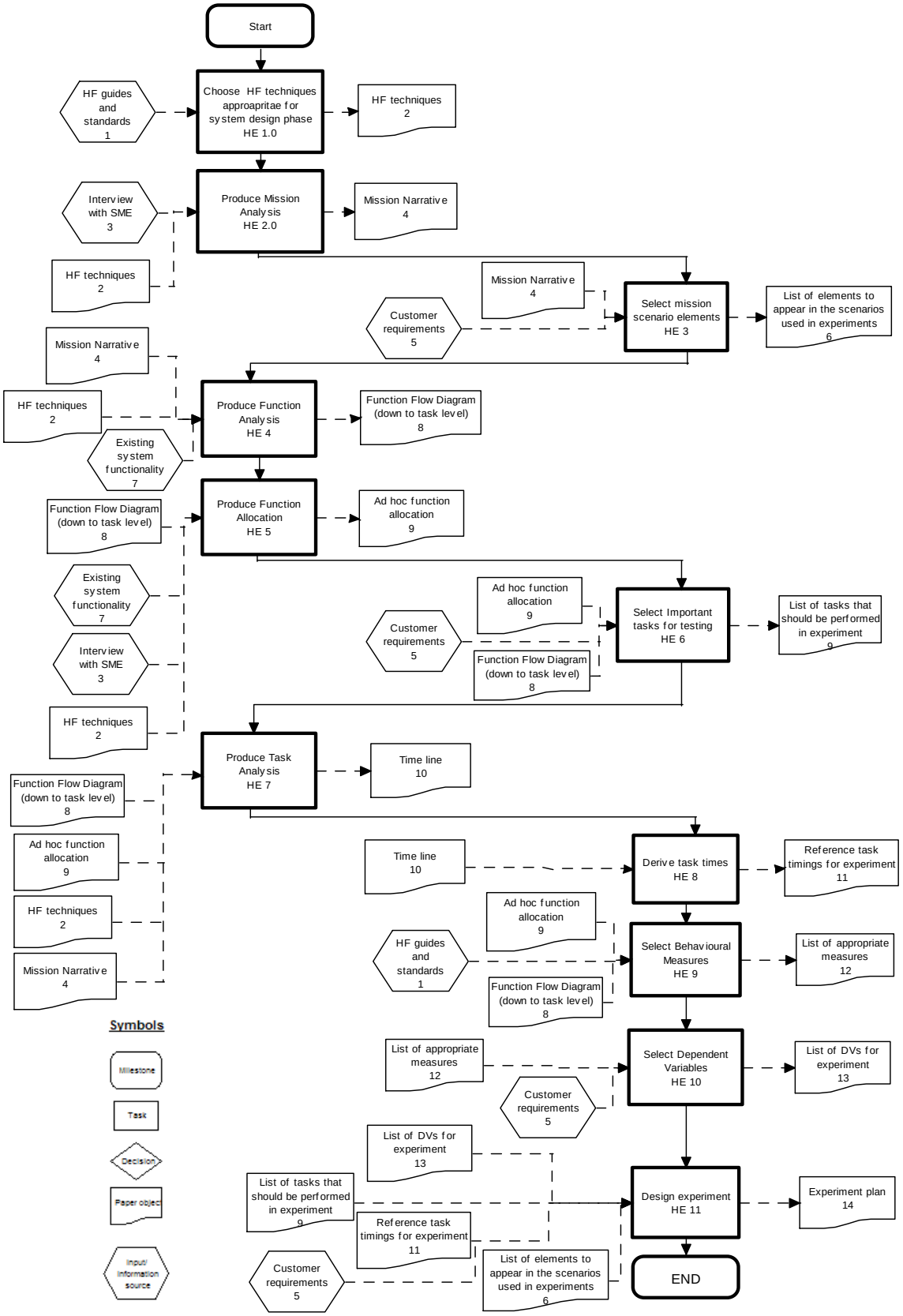


Figure 6-5 Task-based method for evaluating complex socio-technical systems

6.2.3 Subject Matter Expert process for assessing the Task-based method

In Chapter 4 it was argued that it is important to ensure that the task-based method used is valid and is representative of world best practice. The assessment by HE SME 1 and HE SME2 was that the method was valid. A summary of the approach taken by the SMEs follows.

1. I made available to HE SME1 and HE SME2 the measure-selection method that was developed, and a list of references that were consulted during the production of the method. I then asked them to judge whether the method was representative of what Human Factors practitioners would do.
2. HE SME1 and HE SME2 assessed whether the list of references were complete, i.e. they wanted to establish whether the references used were the ones that Human Factors practitioners would use to guide them in conducting a system evaluation. They considered each item and assessed whether the list as a whole represented a "good" cross-section of the material available.
3. HE SME1 and HE SME2 assessed whether the definition of the RWR as a medium complexity system was correct by using the classification in Beevis (1999) as a guide.
4. HE SME1 and HE SME2 assessed whether the inputs and outputs of each stage of the method were the ones that were most appropriate for a Preliminary Definition Phase of the System Life cycle again by using the classification in Beevis (1999) as a guide.
5. HE SME1 and HE SME2 assessed whether the flow chart accurately reflected the stages of a Human Engineering evaluation of a system by comparing the method to a wide range of civilian and military standards.

6.3 Selecting and refining sets of measures for testing

To produce the measures of performance, the author followed each measure-selection method described above, and recorded the measures of performance produced from each. Then it was necessary to validate and finalise the sets to be used in the two experiments.

For the task-based measures, national and international experts (HE SME1, HE SME2) were asked to comment on whether the performance measures were truly representative of the measures that any HE practitioner could produce. Their analysis revealed that the measures were representative in principle, but that more measures could be included. The additional measures that the experts identified had been excluded because they were seen to be representative of the measures from other CWA phases and given that this program of research is interested in two of the five CWA phases (WDA and CTA) it was not appropriate to include them (see Section 2.5.3 and also Chapter 9).

6.3.1 Measures for the RWR system

As a result of following each method, the initial sets of measures for the RWR system are:

- CWA-all measures = {CWA-based, all possible measures for the RWR}
- HE-all measures = {HE-based, all possible measures for the RWR}

Following the experts' assessments, sets of measures were produced for both the constraint-based and task-based measure-selection methods. These sets are:

- Constraint-testable measures = {Constraint-based empirically testable measures for the RWR}
- Task-testable measures = {Task-based, empirically testable measures for the RWR}.

Hence,

- Constraint-testable measures $\subset$ CWA-all measures
- Task-testable measures $\subset$ HE-all measures

Definitions for the Task-testable (task-based) and Constraint-testable (WDA-based and CTA-based) measures are given in Table 6-3, Table 6-4 and Table 6-5. The data collection rules for each of the measures are given in Sections 7.3.3.1 and 7.3.3.2 of the following chapter.

Table 6-3 Definitions of the WDA-based measures

WDA–based measures	Definition of the measures.
Mission Achieved	A mission is achieved when the aircraft lands at the planned location with the correct amount of fuel and at the correct time. This is a categorical measure. The mission is either successful or it is not. This measure is from the DP level of the WDA and reflects the importance to the crew of successfully achieving the mission. The question that this measure addresses is: Does the modified system provide information that increases the chance that the mission will be successfully achieved?
Infantry Equipment Undamaged	This is the level of damage sustained by the aircraft caused by a threat system (gun or missile). This is a categorical measure. The aircraft was either damaged or it was not. This measure is from the DVP level of the WDA and reflects the importance to the crew of not damaging the aircraft's cargo. The question that this measure addresses is: Does the modified system reduce the likelihood that the aircraft will be damaged by the threats?
Distance to enemy	This is the smallest distance between the aircraft and all of the priority threats that are detected. This measure is presented as a ratio of the distance to a detected threat over the distance to the closest threat (whether that threat was detected or not). That is, the raw distances are normalised for mission scenario. If the closest threat is detected then the score will be 1. If the detected threat is not the closest threat the score will tend to zero. This is a continuous measure. This measure is from the DVP level of the WDA and relates to the tactics that the crew use to select a route that is optimised to maximise the distance from all possible threat locations (detected and not detected). Hence, the data that are analysed will be from all threats (those that are displayed to

WDA–based measures	Definition of the measures.
	the crew and those that are hidden from the crew? The question that this measure addresses is: Does the modified system detect high priority threats further out than the unmodified system?
Probability of detection	This is the ratio of threats that detected the aircraft over the number of threats that could detect the aircraft at the TZP. This is a continuous measure. A value of 1 means that all threat that could detect the aircraft did detect the aircraft. A value of .75 indicates that 3 treats detected the aircraft when 4 could have detected the aircraft. A value of 0 indicates that no threat detected you when some should have. This measure is from the DVP level of the WDA and relates to the tactics that the crew use to select a route that is optimised to reduce the likelihood that the threat detects the aircraft. Hence, the data that are analysed will be from all threats (those that are displayed to the crew and those that are hidden from the crew – because the RWR may detect a threat before a threat detect the aircraft). NB: this looks at highest priority threats. The question that this measure addresses is: Does the modified system reduce the amount of threats that detect the aircraft?
Probability of damage/ destruction	This is the mode of the highest priority threat. This measure is from the DVP level of the WDA and relates to the tactics that the crew use to select a route that is optimised to increase the chance that when a threat system detects the aircraft the threat system will be in a benign mode. NB: this looks at highest priority threats. The question that this measure addresses is: When the aircraft is detected what mode is the threat in?
Exposure time to enemy	This is the time that the aircraft is being detected contiguously by any priority threats. This is a continuous measure. This measure is from the DVP level of the WDA and relates to the tactics that the crew use to select a route that is optimised to minimise the length of time that any threat detects the aircraft. The question that this measure addresses is: Does the use of the modified system result in reduced exposure to threats?
Surprise	Percentage of threats displayed truthfully. This is a continuous measure. This measure is from the DF level of the WDA and relates to the ability of the system to provide accurate (truthful) information about high priority threats. The question that this measure addresses is: Does the modified system resolve threats that are located together?
Timeliness	Minimum time to a displayed priority threat. This is a continuous measure. This measure is from the DF level of the WDA. The question that this measure addresses is: Does the modified system provide information about a threat earlier than an unmodified system?
Number of threats	The percentage of threats displayed on the RWR. This is a continuous measure. This measure is from the PF level of the WDA. The question that this measure addresses is: Does the modified system detect and display all the threats to the aircrew?
Tempo	This refers to a specific real world RWR property and may be loosely defined as the rate at which actions take place with respect to time (how long do actions take) and space (distance covered). This rate may be relative to other actions or absolute. This measure is from the DF level of the WDA. The question that this measure addresses is: does the modified system provide information about a threat at an increased rate relative to the unmodified system?
Understanding	This refers to a specific real world RWR property and may be loosely defined as the process of collecting information from a variety of sources and making sense of this information in tactical environment. This measure is from the DF level of the WDA.

WDA–based measures	Definition of the measures.
	The question that this measure addresses is: does the modified system provide information about a threat using appropriate sources of information?
Decision making	This refers to a specific real world RWR property and may be loosely defined as the process of choosing one alternative course of action from many courses of action in a tactical environment. This measure is from the DF level of the WDA. The question that this measure addresses is: does the modified system select an appropriate response form the set of all responses?
Maximise time available to react	This refers to a specific real world RWR property and may be loosely defined as the time taken to react to a threat. This measure is from the DV&P level of the WDA. The question that this measure addresses is: does the modified system react to a threat faster than the unmodified system?
Minimise predictability	This refers to a property of the threat and may be loosely defined as the degree to which an enemy can anticipate the outcome of actions. This measure is from the DV&P level of the WDA. The question that this measure addresses is: does the modified system result in a reduced level of anticipation by the threats of the aircraft behaviour?
Crew physiology	This refers to a property of the threat and may be loosely defined as the degree to which crew experience physiological changes (blood chemistry changes). This measure is from the DV&P level of the WDA. The question that this measure addresses is: does the modified system result in changed blood chemistry of the aircrew?

Table 6-4 Definitions of the unique CTA-based measures

Unique CTA-based measures for the Control Task "Manage EW System"	Definition of the measures
Control task priority	A control task is a higher priority than another control task if it is not interrupted by the latter. The question that this measure addresses is: Does the Aircraft Captain interrupt the control task when using the modified system?
Control task frequency	This measure records the number of times that the control task occurred. The question that this measure addresses is: Does the Aircraft Captain respond to threat events less with the modified system?

Table 6-5 Definitions of the task-based and non-unique CTA-based measures

Task–based and non-unique CTA-based measures	Definition of the measure
Choice of procedure -overall behaviour rating	The choice of the tactic that the Aircraft Captain (AC) adopted in response to a threat was assessed against standard operating procedures (SOPs) and mission debrief information. The measure is scored 1 if the tactic is consistent with SOPs or a successful defence of the tactic by the aircrew in the mission debrief occurs, and is scored 0 if not. The question that this measure addresses is: Does the use of the modified system affect whether the correct tactic is selected?

Task–based and non-unique CTA-based measures	Definition of the measure
Analysis of crew behaviour relevance -overall behaviour rating	The behaviour of the aircraft captain (e.g. locating present position) was assessed against SOPs and mission debrief for the particular tactic observed against a threat. The measure is scored 1 if the behaviour is consistent with SOPs or a successful defence by the aircrew in the mission debrief occurs, and is scored 0 if not. The question that this measure addresses is: Does the use of the modified system affect whether the correct procedure for the tactics is performed?
Observation of system state - Overall Communication rating	This measure recorded whether the Aircraft Captain reported the threat to pilot. The measure is scored 1 if the Aircraft Captain communicates to the pilot that a threat is being displayed, and is scored 0 if not. The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain reports a threat to the Pilot?
Accuracy in identifying stimuli - Display related	This measure recorded whether the Aircraft Captain looked at the RWR when a threat appeared. The measure is scored 1 if the Aircraft Captain correctly looked at the RWR, and is scored 0 if not. The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain notices a threat on the display?
Accuracy of detection of stimulus change over time - Display related	This measure recorded whether the Aircraft Captain identified (by reporting to the pilot) that a particular threat had changed (mode change of same threat) and is scored 1 or 0. The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain detects a change of a threat property using the display?
Accuracy in estimating parameters of threat - Display related	This measure recorded whether the Aircraft Captain correctly identified (by reporting to the pilot) the type, and clock code, e.g., SA8, 4 o'clock, of the threat. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly articulates the threat properties to the Pilot?
Accuracy in estimating position of threat - Display related	This measure recorded whether the Aircraft Captain correctly identified (by reporting to the pilot) the position (terrain feature) of the threat. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly estimates the position of a threat?
Accuracy in estimating distance of threat - Display related	This measure recorded whether the Aircraft Captain correctly located the distance of the threat (by reporting to the pilot) within a 5km error margin. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly estimates the distance of the threat from the aircraft?
Accuracy in identifying stimuli - Not display related	This measure recorded whether the Aircraft Captain correctly identified (by reporting to the pilot) the position of the threat given to him by the pilot. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly identifies a threat?
Accuracy in estimating parameters of threat - Not display related	This measure recorded whether the Aircraft Captain correctly identified (by reporting to the pilot) the type (SA8, etc) of the threat given to him by the pilot. This is a categorical measure (1, 0). The question that this measure

Task–based and non-unique CTA-based measures	Definition of the measure
	addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly reports the properties of a threat that is observed in the world to the Pilot?
Accuracy in response selection - Not display related	This measure recorded whether the Aircraft Captain correctly respond to the threat (by instructing the Pilot to manoeuvre the aircraft) based the mode/ type of the threat given to him by the pilot. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly instructs the Pilot to manoeuvre the aircraft in response to a threat that has been observed in the world?
Accuracy in confirmation of threat - Flying pilot related	This measure recorded whether the Aircraft Captain correctly confirmed threat locality as identified by the pilot. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly confirm the location of a threat reported by the Pilot?
Analysis of crew behaviour relevance - Flying pilot related	This measure recorded whether the Aircraft Captain correctly acknowledged (by reporting to the pilot) that a threat had been identified by the pilot. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly acknowledges the Pilot's report of a threat?
Choice of procedure - Flying pilot related	This measure recorded whether the acknowledgement procedure that the Aircraft Captain adopted was correct for the situation. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain uses the correct response to acknowledge the Pilot's report of a threat?
Analysis of crew behaviour relevance - Flying pilot related	This measure recorded whether the Aircraft Captain behaviour was relevant for the situation. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain's behaviour is appropriate in response to the Pilot's report of a threat?
Rating of performance adequacy - Aircraft captain self assessment	This measure recorded the rating that Aircraft Captain gave his own performance (by verbal report to the pilot during the flight phase, or to the experimenter during the debrief phase). This was a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain is likely to assess his own behaviour?
Accuracy in response select - Action	This measure recorded whether the Aircraft Captain correctly instructed the pilot to expend the correct countermeasure. This is a categorical measure (1, 0). The question that this measure addresses is: Does the use of the modified system affect whether the Aircraft Captain correctly instructed the pilot to expend the correct countermeasure?
Time taken to defeat threat	This measure recorded the time taken for a threat to be defeated.
Time taken detect threat (perceive threat icon)	This measure recorded the time taken for the Aircraft Captain to detect a threat icon on the RWR.

Task–based and non-unique CTA-based measures	Definition of the measure
Time taken to locate threat in the environment	This measure recorded the time taken for the Aircraft Captain to locate (visually) a threat in the environment.
Time taken to initiate a movement to search for a threat in the environment	This measure recorded the time taken for the Aircraft Captain to initiate a movement to search for a threat in the environment.
Time taken to deduce location of threat	This measure recorded the time taken for the Aircraft Captain to deduce the location of a threat.
Time taken to deduce range to threat	This measure recorded the time taken for the Aircraft Captain to deduce the range of a threat.
Time taken to select appropriate countermeasure	This measure recorded the time taken for the Aircraft Captain to select an appropriate countermeasure.
Time taken to expend countermeasure	This measure recorded the time taken for the Aircraft Captain to expend a countermeasure.
Time taken to select communication system	This measure recorded the time taken for the Aircraft Captain to select a communication system.

6.4 Conclusion

This chapter has presented the two methods for selecting measures: a constraint-based measure-selection method and a task-based measure-selection method. Independent SMEs judged the task-based method as representative of the state of the art. The constraint-based method was developed by the author. Using these methods, sets of measures were produced. Independent SMEs judged that the task-based measures were representative of the measures that HE practitioners would produce but that additional measures could be included. The additional measures were not included because they were not equivalent to measures from WDA and CTA. The exclusion of these measures from testing is recognised as a general limitation of this program of research and is discussed further in Chapter 9.

In the following chapter the measures derived from the task-based and constraint-based methods are tested for sensitivity using a current RWR system and the methods assessed for suitability for use in operational settings.

7. Experiment 1: Comparing Methods using a Current System

The previous chapter has described the task-based and constraint-based methods developed to select measures for system evaluation. These methods were used to obtain a set of task-based measures and a set of constraint-based measures for testing.

This chapter covers part of the material at Stage 4 of Figure 7-1. The chapter describes an experiment that compares the methods in two ways. First, the experiment compares the number of measures that the methods suggest as being sensitive to a modification in a current system that is used in the field (from now on the measures will be termed "low-level dependent variables"). The modification in question is a change in the range over which a threat can be detected by the RWR. To be considered sensitive, a low-level dependent variable should show a statistical difference between the two system conditions (unmodified and modified). Second, the experiment compares the methods on their suitability for use in operational settings. Comparing the methods in terms of measure sensitivity and suitability provides an indication of the predictive validity of the methods.

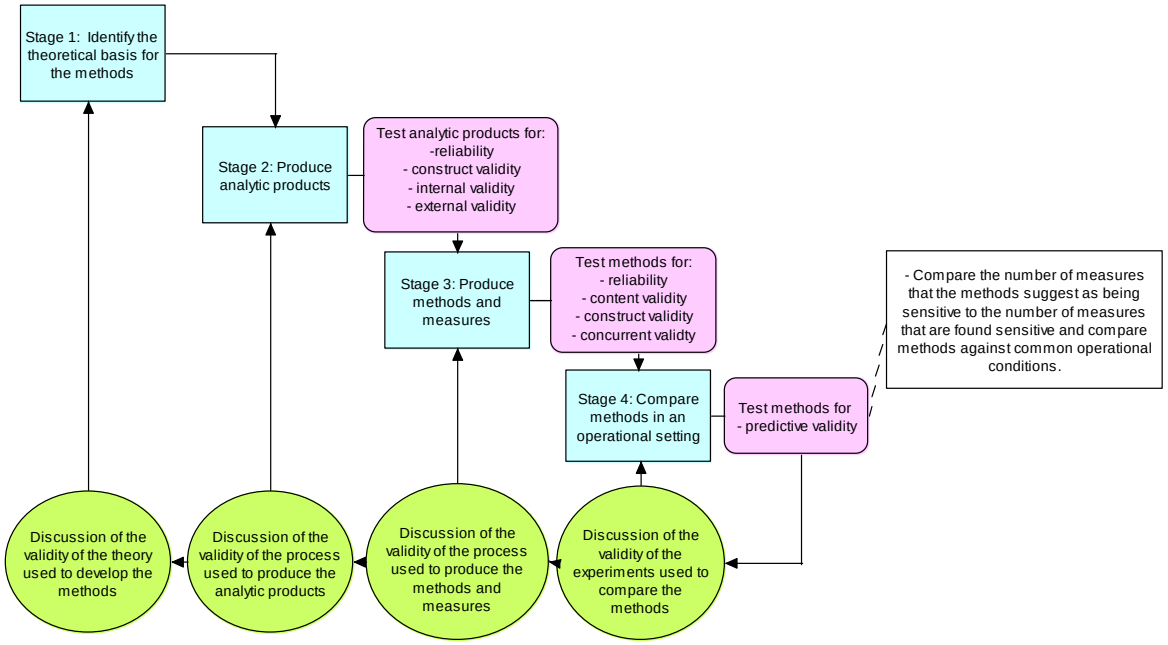


Figure 7-1 Four stages of the research program. The present chapter focuses on Stage 4, in which the measure-selection methods are compared for their relative sensitivity to a change in a Current technical system (RWR)

In the next section the background and aims for this experiment are outlined. The experiment is then reported and implications for this program of research stated. Finally, the chapter draws conclusions.

7.1 Background, aims and hypotheses

In Chapter 2 it was reported that the task-based method of selecting measures using guidelines had been widely used for evaluating systems. It was also shown that some authors (e.g. Kantowitz, 1992) expressed concern that using guidelines to select measures was not appropriate for operational (field) evaluations. Other authors (e.g. Muckler and Severn, 1992) argued that theory should guide measure selection. CWA embodies a constraint-based approach to system evaluation. CWA has a strong theoretical basis and

has been used to select measures in laboratory-based experiments. However, CWA has not been used to select measures for operational evaluations.

Chapter 2 also raised the issue that although both CWA and task-based methods should be able to predict sensitive measures for system evaluation for a current system, an empirical test was needed. The chapter also indicated that the methods should also be compared on how suitable they are for selecting measures for operational evaluations.

In Chapter 4 the importance of assessing the predictive validity of the methods was raised. It was concluded that the only method for assessing predictive validity was to conduct an experiment that compared predictions made about the sensitivity of the measures to the results gained in the experiment and compare the methods on how suitable they are for use in field settings. Chapter 6 described the task-based and constraint-based measure-selection methods for use in the empirical tests.

The aim of Experiment 1 is to test the relative predictive validity of the two measure-selection methods. The predictive validity of each measure-selection method will be tested by evaluating whether the low-level dependent variables that the measure-selection methods suggest are sensitive to the system modification and whether the measure-selection methods lead to low-level dependent variables that are suitable for use in operational settings.

As was discussed in Chapter 4 there are five hypotheses that are related to the issues of measure sensitivity and five hypotheses that are related to the issues of method suitability. These hypotheses are tested in Experiment 1 and are given again below.

Measure sensitivity:

H1 None of the task-based and none of the constraint-based low-level dependent variables are sensitive to the system modification.

H2 All the task-based and all the constraint-based low-level dependent variables are sensitive to the system modification.

H3 Significantly more of the constraint-based low-level dependent variables than the task-based low-level dependent variables will be sensitive to the system modification.

H4 Significantly more of the task-based low-level dependent variables than the constraint-based low-level dependent variables will be sensitive to the system modification.

H5 Some of the task-based and constraint-based low-level dependent variables will be sensitive to the system modification.

Method suitability:

H6 The task-based and constraint-based low-level dependent variables will be affected by all of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H7 The task-based and constraint-based low-level dependent variables will not be affected by all of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H8 The constraint-based low-level dependent variables will not be affected by some of the following as the task-based low-level dependent variables will be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H9 The task-based low-level dependent variables will not be affected by some of the following as the constraint-based low-level dependent variables will be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H10 The task-based low-level dependent variables and the constraint-based low-level dependent variables will not be affected by some of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

The experimental method that I used to test the hypotheses is presented in the following section.

7.2 Method

In this section the important biographical data of the aircrew that participated in the experiments are given. The apparatus and the materials used in the experiments are also described.

7.2.1 Participants

Two serving members of the Australian Army took part in the experiment. The two participants formed one helicopter crew that consisted of a Flying Pilot (FP) and an Aircraft Captain (AC). Both participants held the rank of Captain. The participants were volunteers and were different from those used in the exploratory experiment (Appendix A). Both crewmen are considered experienced RWR operators and have flown together on training exercises. However, neither crewman had experience of the Black Hawk helicopter simulator used in the experiment.

The Aircraft Captain had 1500 flight hours flying Black Hawk. He had completed the Army Electronic Warfare (EW) course, had experience with the Joint Electronic Warfare Operational Support Unit and had developed threat counteractions for EW systems. He had relevant operational experience.

The Flying Pilot had 2500 flight hours flying Black Hawk. He had completed the Army EW course, and had experience with the Joint Electronic Warfare Operational Support Unit. He had relevant operational experience.

During the experiment the aircrew used their own flight gear, including helmets, gloves and flight clothes.

7.2.2 Apparatus and materials

In this section the RWR system and Aircrew Mission Debriefing Tool used in this experiment is described. The other apparatus and materials are described in Chapter 4 (Research Design). The RWR display used in Experiment 1 represents a plan view of the aircraft and threat radar emitters (see Figure 7-2). The display is centred on the aircraft and shows the relative bearing, type and mode of the surrounding threat radar emitters. Emitter modes are presented as unique icons and displayed relative to the centre of the display. For example, Figure 7-2 shows that there are three SA-8 emitters operating. The emitter at 7:30 is in surveillance mode, the emitter at 9:30 is in tracking mode and the emitter at 3:30 has lock-on on the aircraft and may fire a weapon at any stage.

When a new threat radar emitter is detected a "New guy" auditory tone sounds and the icon is underlined for two seconds to alert the aircrew that the display has new information on it. When an emitter is lost (perhaps because line of sight no longer exists) the symbol is displayed for four seconds. After four seconds the symbol appears as stippled for a further six seconds. If the emitter is still not detected then it will disappear from the display. If the emitter is reacquired the two second tone will sound and the icon will be underlined. A maximum of eight emitters are displayed at any one time on the RWR display.

It is important to note that the distance between the emitter icon and the centre of the display is not related to the actual distance of the emitter to the aircraft but is related to the mode of the emitter – emitters appearing closer to the centre of the display are a greater threat than those appearing further away. Bearing to the emitter may be directly inferred from the display. Emitter locations are displayed relative to the nose of the aircraft, with the nose being the 12 o'clock position on the display. Hence, a threat icon appearing in the three o'clock position is on a bearing of 090 ("zero nine zero") and a threat emitter appearing at the nine o'clock is on a bearing of 270 ("two seven zero") from the aircraft.

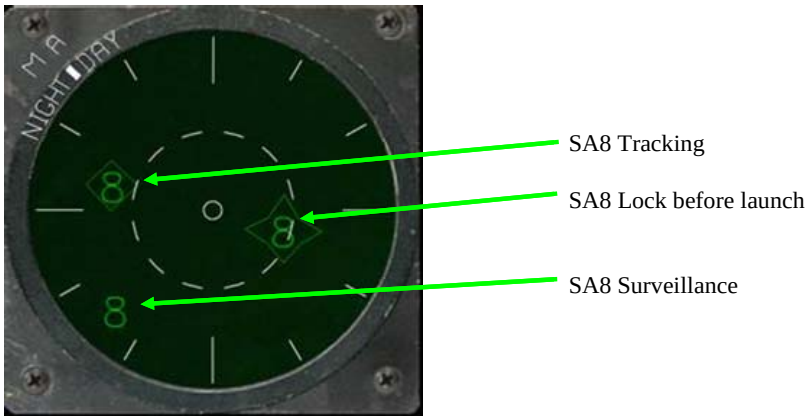


Figure 7-2 Current RWR display

One important aspect of the threat-aircraft interaction is that it takes time for the radar wave to travel from the threat to the aircraft and then back to the threat. It is possible therefore for the aircraft to detect the radar before the threat detects the aircraft. In fact, it is this characteristic that is exploited by many aircrews in operational situations.

Chapter 4 has described the modified Critical Decision Method (CDM) and Direct Observation data-collection methods used in the experiments. The Aircrew Mission Debriefing Tool was developed to facilitate the use of those methods. The tool consists of multiplexing software and hardware that is used to record the following data sources collected during the mission flight phase:

- Video output from faceLAB™ eye tracking software (video channel)
- Video and auditory output from a camera place in the cockpit used to record the Aircraft Captain's behaviour
- Video repeat of the flight instruments
- Video output of a video camera placed behind the cockpit
- Video and auditory repeat of the aircraft RWR display
- Video and auditory repeat of the threat picture
- Video repeat of a "chase plane" view
- Video repeat of the Scenario Toolkit And Generation Environment (STAGE) software.

These data were then synchronised and recorded to video tape. Four data sources were multiplexed to each of two video recorders (see Figure 7-3). The video recorders and data projector could then replay the mission flight phases and flight phase events of a mission to the aircrew.

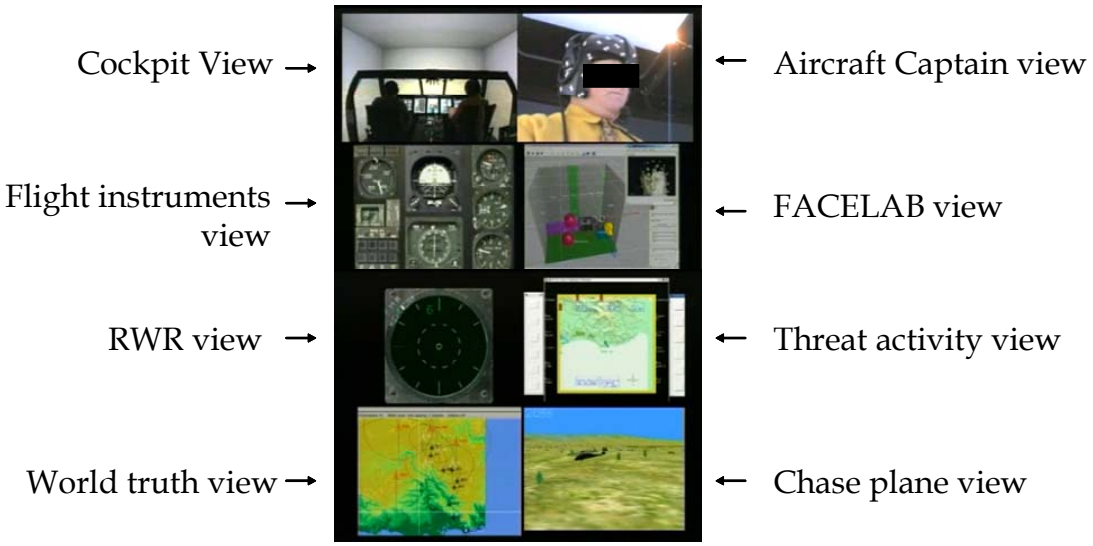


Figure 7-3 Aircrew Mission Debriefing Tool (AMDT). Eight screens are used to display the Black Hawk aircraft as it flew though the visual scene.

7.2.3 Design

The experiment was conducted as an n=1 (one aircrew consisting of an Aircraft Captain and Flying Pilot) within-subject design, where the aircrew completed 10 separate mission scenarios.
Table 7-1 shows the experimental design. The table shows that the within-subject independent variables are measure-selection method (constraint-based and task-based) and system modification (unmodified and modified). The low-level dependent variables (the measures selected from the measure-selection methods) are shown as, for example, "Constraint-based low-level dependent variable 1" and "Task-based low-level dependent variable 1" and are detailed in the following sections.

The high-level dependent variables that will be used for hypothesis testing are:

- *Number of low-level dependent variables that are statistically significant.* This is defined as the number of low-level dependent variables that are statistically sensitive to the system modification.
- *Percentage of low-level dependent variables for which the collection of data is limited by simulation resources.* This is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because system models (for example, a property of the RWR system, a property of the world, and a property of the helicopter and mission) cannot be developed, divided by the total number of measures.
- *Percentage of low-level dependent variables for which the collection of data is limited by the data collection method used.* This is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because the data collection method used to collect the data for the low-level dependent variables was not adequate (because of problems associated with gathering data points manually or automatically), divided by the total number of low-level dependent variables.
- *Percentage of low-level dependent variables for which data is limited by the number of data gathering opportunities.* This is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because the number of data collection opportunities was not sufficient to meet statistical protocols (for example, if the number of data points was less than 10 or if data were not present in all categories - if categorical data were collected), divided by the total number of low-level dependent variables.
- *Percentage of measures for which data could not be collected because of theory.* This is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because theory restricted the collection of data, divided by the total number of low-level dependent variables. In other words, if it can be shown that data could not be collected because the underlying theoretical perspective excluded that data, then this high-level dependent variable is appropriate.

Table 7-1 Experimental design for Experiment 1

	Measure-selection method			
	Constraint-based		**Task-based**	
	System modification		System modification	
	Unmodified	**Modified**	**Unmodified**	**Modified**
Current RWR (Expt 1)	Constraint-based low-level dependent variable 1 ... Constraint-based low-level dependent variable 42		Task-based low-level dependent variable 1 ... Task-based low-level dependent variable 25	

A total of 12 mission scenarios were used for the study. Two were used for training and the remaining 10 were used for data collection. Five missions were run under each system modification condition (modified, unmodified). A number of mission variables were balanced across the mission scenarios in order to make the scenarios equivalent. The mission variables balanced were: threat density, threat position, pop-up (ambush) distance and number of threats in a scenario.

The choice of what mission variables to include and balance in the mission scenarios were determined by interviewing SMEs and adhering to the requirement to have a simulation environment that did not limit the behaviour of the aircrew and RWR system and the requirement for statistical testing (see Crone et al, 2007; and also Appendix E). The process used to identify what mission variables to include is now briefly described:

- The SME was interviewed using a semi-structured interview technique.
- The SME was asked to describe what Air Mobile Insertion missions are and then asked to describe in detail, using real-life examples, operational tactical procedures. During the interview the researcher used a series of prompts that were aimed at eliciting information on the important characteristics for the scenarios.
- All interview material was recorded and later transcribed.
- Analysis of the interview data revealed that the main characteristics that the scenarios should have were, a mixture of radar-based threat types, a mixture of threats whose locations are known by the aircrew and threats whose locations are not known about (ambush threats), a mixture of threats that the RWR can identify (SA-8, SA-6, etc) and threats that the RWR cannot identify (unknowns).
- The interview data also revealed that scenarios should occur over terrain with a wide range of topographical features because the performance of the RWR would change as a consequence and threats modes should be accurately replicated.
- Using the framework shown in Crone et al (2007) the mission variables were combined to stimulate all the "survivability rings" (categories of human-system behaviour). For example, having a threat close to the aircraft would be a stimulus for the human-system behaviour associated with the ring "if seen don't be engaged". The mission variables were balanced so that all the "rings" would be stimulated.

7.2.3.1 Constraint-based low-level dependent variables

Descriptions of the constraint-based low-level dependent variables are given in Chapter 6. The rules used for collecting data depend on whether the low-level dependent variables require categorical or continuous data. For the categorical low-level dependent variables of "Mission Achieved" and "Infantry Equipment Undamaged" the number of data points was limited to one per mission, or in other words, five per system modification condition (i.e. five data points were collected in the system modified condition and five data points collected in the system unmodified condition). The rule was to include all the data in the analysis.

For the continuous WDA low-level dependent variables of "Distance from enemy", "Probability of detection/ identification", "Probability of damage/ destruction", "Surprise", and "Timeliness" the rule adopted was intended to reflect the idea that the RWR is designed to detect a "change" in the environment. Therefore it is important to capture the data associated at the moment that a change in the environment occurred, which is what was done. Any further change may be as a consequence of aircrew behaviour and was therefore excluded from analysis. This rule also ensured that the data points collected met the statistical analysis requirement for data independence. From a practical point of view this means that a data point would only be used for a low-level dependent variable if:

- The source threat was encountered (i.e. displayed on the RWR) in the current frame (that is, one 60 Hz, or 0.016 second portion of the data run) but was not encountered in the previous frame.
- The source threat had been seen in the previous frame but had a different threat mode in the current frame and was still the highest priority.

For the remaining continuous WDA low-level dependent variables of "Exposure time to enemy (Continuous exposure time, Total exposure time)" and "Display of threat information (Type, Location, Lethality, Number)", the data were collected continuously, frame-by frame, across the entire mission.

In addition, because more than one threat could be displayed on the RWR at any one time, it was decided to use the data associated with the highest priority threat. The priority level of the threats is available from the author on request and was based on recommendations from SMEs.

The data collection rule adopted for CTA low-level dependent variables reflected the theoretical underpinning that control tasks represent what needs to be achieved in the work domain. Hence, the author attempted to collect all quantitative data and qualitative data when a threat event occurred. Qualitative data included transcriptions taken from the mission briefing, mission flight and mission debriefing and video from the flight phase.

7.2.3.2 Task-based low-level dependent variables

The task-based low-level dependent variables that were tested are defined in Chapter 6. All the low-level dependent variables that were tested relate to aircraft captain behaviour. Hence, video data of the aircraft captain during the mission flight phase was used. Interview transcriptions from the mission briefing and mission debriefing phases were

also used where appropriate – generally, in assessing whether a specific sequence of behaviour (for example, a tactic) was correct.

7.2.3.3 Experimental Task

The aircrew were required to complete an Air Mobile Insertion mission. After initial training, the mission included a briefing phase, a flight phase and debriefing phase.

Mission briefing phase. During the mission briefing phase the aircrew were given an operational briefing (including the reason for the mission, approximate threat locations, and other standard mission briefing elements) and a limited opportunity to modify the mission plan. The crew were instructed to complete the mission, i.e. to land the aircraft at the predetermined landing site.

Flight phase. During the mission flight phase, the aircrew were given a set amount of fuel and were asked to land with 20% fuel above the absolute minimum level. They were required to follow the mission plan, perform aircraft system checks as per standard operating practices (SOPs) and monitor the RWR system and counter any threats (including operating the chaff dispense system[14]) as required. They were also instructed to mark the location of any threats that they encountered on the map provided. The aim of each flight phase was to land at the predefined landing zone within the time and fuel constraints. The flight phase was deemed a failure if they failed to land with fuel greater than 20% above the minimum initial fuel level, if they crashed the aircraft, or if they exceeded the time limit.

Mission debriefing phase. After the mission flight phase, the crew undertook a lengthy mission debrief. The mission debrief was used to elicit information from the aircrew on the conduct of the mission flight phase. The average time for one mission was about 1.5 hours (20 minutes for the mission briefing, 40 minutes for the flight phase and 30 minutes for the mission debriefing). This means that during the experiment (including training) the aircrew spent approximately a total of four hours engaged in Mission Briefing Phase, eight hours engaged in the Flight Phase and six hours engaged in the Mission Debriefing Phase.

7.3 Procedure

The experiment was performed over three days. On the first day of the experiment the aircrew were welcomed, given an Occupational Health and Safety briefing, and then briefed on the contents of the briefing pack (available from the author on request). Questions from the aircrew were answered on an ad hoc basis. During this time the faceLAB™ eye tracking tool was calibrated to the Aircraft Captain's features. Once this was completed, two participant-paced training sessions were conducted (one with the RWR modification and one without the RWR modification). The training sessions were conducted in the same manner as the experimental sessions and included a limited mission briefing phase, mission flight phase and mission debrief phase. The experimental sessions were conducted on the second and third days.

[14] Chaff is used to confuse an enemy radar system and is dispensed by the crew using a button on the collective.

7.3.1 Training phase

During the training flight phase (which lasted approximately three hours), the aircrew were briefed on the differences between the simulator systems and operational Black Hawk systems (including the chaff and flare dispensing system, navigation systems, aircraft warning systems, and communication systems). The aircrew were shown that some of the navigation system display modes were not available and if selected the display screen would remain blank. They were instructed to fly the aircraft at approximately 50ft, using standard operating procedures and were tasked to take off, practice threat avoidance techniques and land the helicopter. Only when the aircrew had experienced all threats, flight conditions, consequences of their actions (for example, damaging the aircraft during a "hard" landing) and expressed satisfaction with their flight performance, were they allowed to progress to the experimental sessions. During the practise sessions the participants were able to request information about the simulation and play "what-if" games. It should be noted that the flight model used in the simulation was very representative of the Black Hawk aircraft. The aircrew were made aware of any departure from normal flight by the information on the flight instruments and by "cracking" appearing over the instruments (indicating that the aircraft had crashed). For example, if the aircrew exceeded engine torque limits the torque indicator moved into the red zone and if the aircraft landed with too high a rate of descent all the flight instruments appeared cracked.

An important part of the training flight phase was to ensure that the aircrew was aware of, and adapted to, the visual differences between the field environment and the simulated world. The visual differences occur because the simulated world is projected onto the screens from a point behind them. In the case of the front screen, this point is approximately three metres in front of, and to the centre of, the aircrew. This is at odds with the field environment where the visual scene is focused at infinity, in front of each of the aircrew. The visual difference results in the aircrew "feeling" that they are flying the simulator with yaw (side slip) even when they are actually flying in a balanced (no yaw) state. Aircrew were briefed on the discrepancy and then allowed to fly the aircraft until they had adapted to the difference, i.e. until they flew the simulator straight and level with no yaw.

7.3.2 Experimental sessions

Once the training session was completed the experimental sessions commenced. These included a mission briefing phase, a flight phase and a mission debrief.

7.3.2.1 Mission briefing phase

During the mission briefing phase the aircrew were presented with 1:100000 scale maps that had a route marked on it. They were then briefed on the specifics of their mission including: threat locations (an area was indicated where threats were likely), and the mission objective and constraints (i.e. fuel, time and flight profile). The aircrew were not allowed to modify the route but could annotate the map with timing, fuel, navigation, threat and other operational information as per normal practice. The crew were then briefed on the technical properties of the modified and unmodified RWR.

During the briefing phase the aircrew were asked describe "out loud" what they were thinking. Various questions were asked by the author to elicit information about route planning decisions, characteristics of the environment that the aircrew considered important, and threat avoidance techniques. All audio and video of the discussions were recorded.

7.3.2.2 Flight phase

Once the briefing phase was completed the aircrew proceeded to the mission flight phase. The aircrew were seated in the cockpit and instructed to indicate to the control room when they were ready to begin the flight phase. Once they were ready the simulator was "released" and they could take-off. Data recording was then started. The flight phase was terminated when the aircrew landed the aircraft safely and indicated that the mission was over or when the aircrew terminated the mission because of damage to the aircraft sustained as a consequence of a threat engagement or when the aircraft crashed.

During the flight phase the author used a mission flight phase template (available from the author on request) to record any mission events of interest. Such events included threat engagements, and decisions points made by the Aircraft Captain to depart from the pre-planned route or landing zone and decisions to terminate the mission.

7.3.2.3 Mission debriefing phase

Once the flight phase was terminated the aircrew proceeded to the mission debriefing phase. Video and audio recording equipment was used to record all subsequent discussions. For each event of interest, the aircrew was asked to recount the main aspects of the event. They were then shown a replay of the event using the Aircrew Mission Debriefing Tool and a modified Critical Decision Methodology semi-structured interview technique was used to elicit information from the aircrew. Aircrew were then given the opportunity to ask questions on any aspect of the flight phase.

Once the debriefing session was completed the crew were given the opportunity to have a break (approximately 20 minutes) before the next mission commenced.

7.3.3 Final wrap-up

At the end of the experiment participants were debriefed on the objectives of the experiment and thanked for their participation. Letters expressing the positive contribution that they made to the research program were then sent to their Commanding Officer.

7.4 Results and discussion

The aim of Experiment 1 is to test the comparative predictive validity of the task-based and constraint-based methods for producing low-level dependent variables. If the methods produce sensitive measures and the measures are suitable for operational settings then they may be judged as having predictive validity.

In the following sections the findings from the experiments are presented under headings representing these two aims. The results show that the constraint-based method has a greater degree of predictive validity in terms of the sensitivity of the measures produced than the task-based method, but that there is no statistical difference between the two measure-selection methods. The results also show that the constraint-based method and the task-based method do not have predictive validity in terms of suitability because they suggest low-level variables that are affected by simulation resource limitations, data collection method limitations, opportunities for data gathering and theory.

7.4.1 Assessing sensitivity of measures

On the basis of the results there is no statistical evidence of a difference between the methods in terms of the numbers of sensitive low-level dependent variables that they suggest. In terms of the hypotheses tested none can be accepted. However, given that there is some evidence to suggest that the WDA-method has a higher degree of predictive validity than the CTA-and task-based methods the following one is hinted at.

H5 Some of the task-based and constraint-based low-level dependent variables will be sensitive to the system modification.

The results from Experiment 1 are shown in three tables. Table 7-2 summarises the results for the WDA-based variables. Table 7-3 summarises the results for the unique CTA-based variables and Table 7-4 summarises the results of the remaining CTA-based and task-based variables. Finally Table 7-5 provides a summary of the variables that are used to compare the methods. (Note: observational data, the interview transcription, and the statistical tests conducted on the low-level dependent variables are available from the author on request.)

Table 7-2 reveals that of the 15 WDA–based low-level dependent variables, five were sensitive to the difference between the modified and modified RWR, and 10 were not sensitive. The table also indicates that interviews, observational data and interview transcription information provided evidence to indicate that some of the low-level dependent variables (seven of 15) had apparent validity (in this thesis Apparent validity is shown when a measurable property (of the domain) appears in data (from the actor) via discussion and/ or observation and/or inference from the transcription data.

Table 7-2 Summary of WDA results comparing performance with and without the RWR system modification (Experiment 1).

Low-level dependent variables suggested by WDA method	Summary of statistical test (result, test used, statistic, summary)	Assessment of apparent validity
Probability of detection	Significant result Mann-Whitney U, U = 2541, N_1 = 81, N_2 = 121, p = 0.000 The probability of being detected by the threats was higher when the helicopter was fitted with the unmodified RWR compared to when it was fitted with the modified RWR.	Yes

Low-level dependent variables suggested by WDA method	Summary of statistical test (result, test used, statistic, summary)	Assessment of apparent validity
Probability of damage/ destruction	Significant result Pearson X^2 (3, N = 202) = 10.45, p = 0.02 The probability of the helicopter being damaged (or destroyed) was higher when the helicopter was fitted with the unmodified RWR compared to when it was fitted with the modified RWR.	Yes
Surprise	Significant result Mann-Whitney U, U = 2454.5, N_1 = 60, N_2 = 101, p = 0.04 The unmodified RWR provided a greater degree of accuracy about the location of coincident threats than the modified RWR.	Yes
Timeliness	Significant result ANOVA F(1,159) = 9.63, p = 0.002 The unmodified RWR detected threats later (resulting in less time before the threat was encountered) than the modified RWR.	No
Number of threats	Significant result Kolmogorov-Smirnov (p<0.001) The unmodified RWR detected fewer threats than the modified RWR.	No
Mission achieved	Not significant 4/5 missions were achieved with the modified RWR compared to 3/5 with the unmodified RWR.	Yes
Infantry equipment undamaged	Not significant Damage was seen in 1/5 missions with the modified RWR compared to 0/5 with the unmodified RWR.	Yes
Distance to enemy	Not significant Mann-Whitney U There was no difference in the distance from the aircraft when a threat was detected between conditions.	Yes
Exposure time to enemy	Not significant Contiguous time exposed - not enough data in each category for Chi-squared test. Observations - there is a greater frequency of shorter periods of time during which the aircraft was exposed to the threats when it was fitted with a modified RWR. Additionally, when the aircraft was fitted with a modified RWR there is a greater frequency of longer periods of time spent exposed to the threats.	Yes
Tempo	No data collected – RWR property associated with the measure could not be modelled in the simulation environment.	No
Understanding	No data collected – RWR property associated with the measure could not be modelled in the simulation environment.	No

Low-level dependent variables suggested by WDA method	Summary of statistical test (result, test used, statistic, summary)	Assessment of apparent validity
Decision making	No data collected – RWR property associated with the measure could not be modelled in the simulation environment.	No
Maximise time available to react	No data collected – RWR property associated with the measure could not be modelled in the simulation environment	No
Minimise predictability	No data collected – Threat property associated with the measure could not be modelled in the simulation environment.	No
Crew physiology	No data collected – Data could not be collected within the constraints on the program of research.	No

Table 7-3 shows that of the two unique[15] low-level dependent variables for the control task "Operating RWR system in response to threat", neither showed a significant statistical difference. The table also indicates that only one of the low-level dependent variables had apparent validity.

Table 7-4 shows that there was no significant effect of the RWR system modification as measured by the low-level dependent variables that are common to both the task-based and CTA-based methods. The table also shows that nine of 25 had apparent validity.

Table 7-3 Summary of unique CTA results comparing performance with and without the RWR system modification (Experiment 1).

Low-level dependent variables suggested by the CTA method for the control task activity "Manage EW System"	Summary of statistical test (result, test used, statistic, summary)	Assessment of apparent validity
Control task activity priority	Not significant Chi-squared The control task was interrupted 73% of all opportunities (n=67) when the helicopter was fitted with the modified RWR and 79% when the helicopter was fitted with the unmodified RWR (n=49).	Yes
Control task activity frequency	Not significant The control task was performed in 23% of all opportunities (n=318) during the modified condition and 15% (n=413) during the unmodified condition.	No

[15] The other CTA-based variables are reported in the common task-based and CTA results section.

Table 7-4 Summary of common task-based and TC-CTA results comparing performance with and without the RWR system modification (Experiment 1).

Low-level dependent variables suggested by the task-based and CTA method	Summary of statistical test (result, test used, statistic, summary)	Assessment of apparent validity
Observation of system state - Overall Communication rating	Not significant Chi-squared Modified – Aircraft Captain (AC) correctly communicated threat to pilot 73% of all instances (n=118); Unmodified - AC correctly communicated threat to pilot 66% of all instances (n=115)	Yes
Accuracy in identifying stimuli - Display related	Not significant Chi-squared Modified - AC detected threat 92% of all instances (n=118); Unmodified - AC detected threat 93% of all instances (n=115)	Yes
Accuracy in estimating parameters of threat - Display related	Not significant Chi-squared Modified - AC correctly identified threat properties 67% of all instances, n=118; Unmodified - AC detected threat properties 60% of all instances, n=115	Yes
Choice of procedure - overall behaviour rating	Not significant Modified - 100% of all procedures were that the AC performed were correct; Unmodified - 100% of all procedures were correct. (Modified n=11, Unmodified n=30)	Yes
Accuracy of detection of stimulus change over time - Display related	Not significant Modified - AC detected a change in the threat 66% of all instances; Unmodified – No data Note: Modified n=3, Unmodified n=0	Yes
Accuracy in estimating position of threat - Display related	Not significant Modified – AC correctly estimated the position of the threat 100% of all opportunities; Unmodified – AC correctly estimated the position of the threat 66% Note: Modified n=6, Unmodified n=5	Yes
Accuracy in estimating distance of threat - Display related	Not significant Modified – AC correctly estimated the distance of the threat 50% of all opportunities; Unmodified – AC correctly estimated the distance of the threat 100% Note: Modified n=3, Unmodified n=3	Yes

Low-level dependent variables suggested by the task-based and CTA method	Summary of statistical test (result, test used, statistic, summary)	Assessment of apparent validity
Rating of performance adequacy - Aircraft captain self assessment	Not significant Modified – No data; Unmodified - AC rated own performance 100% adequate Note: Modified n=0, Unmodified n=2	Yes
Accuracy in response select - Action	Not significant Modified – AC correctly expended countermeasure 100% of all opportunities; Unmodified- AC correctly expended countermeasure 100% Note: Modified n=1, Unmodified n=2	Yes
Analysis of crew behaviour relevance - overall behaviour rating	No comparison possible – No data	No
Accuracy in identifying stimuli - Not display related	No comparison possible – No data	No
Accuracy in estimating parameters of threat - Not display related	No comparison possible – No data	No
Accuracy in response selection - Not display related	No comparison possible – No data	No
Accuracy in confirmation of threat - Flying pilot related	No comparison possible – No data	No
Choice of procedure - Flying pilot related	No comparison possible – No data	No
Analysis of crew behaviour relevance - Flying pilot related	No comparison possible – No data	No
Time taken to defeat threat	No comparison possible – No data	No
Time taken detect threat (perceive threat icon)	No comparison possible – No data	No

Low-level dependent variables suggested by the task-based and CTA method	Summary of statistical test (result, test used, statistic, summary)	Assessment of apparent validity
Time taken to locate threat in the environment	No comparison possible – No data	No
Time taken to initiate a movement to search for a threat in the environment	No comparison possible – No data	No
Time taken to deduce location of threat	No comparison possible – No data	No
Time taken to deduce range to threat	No comparison possible – No data	No
Time taken to select appropriate countermeasure	No comparison possible – No data	No
Time taken to expend countermeasure	No comparison possible – No data	No
Time taken to select communication system	No comparison possible – No data	No

One aim of this program of research is to compare task-based and constraint-based measure-selection methods on whether each can provide low-level dependent variables that are sensitive (show a statistical significant difference) to the system modification. A reasonable way to compare the two methods is on the percentage of low-level dependent variables that prove to be sensitive. Such a statistic indicates the relative predictive validity of each method. Table 7-5 shows the percentage of low-level dependent variables that were found to be statistically sensitive to the system modification.

Table 7-5 Summary of the results for Experiment 1 that report on the statistical significance of the dependent variables.

Method type	% of low-level dependent variables that are statistically sensitive
Constraint	12% (5/42)
WDA	33% (5/15)
CTA	0% (0/27)
Task	0% (0/25)

The results of this comparison reveal that 12% (5/42) of the constraint-based low-level dependent variables were statistically sensitive to the system modification and that none (0/25) of the task-based low-level dependent variables were statistically sensitive to the system modification for a current system. An exploratory chi-squared test was conducted on the number of low-level dependent variables that were found to be statistically significant. The analysis revealed that there was no significant difference between the measure-selection methods, $X2 (1, N = 67) = 3.22$, $p = 0.07$.

From the results it seems that although more constraint-based low-level dependent variables were sensitive to the modification than the task-based low-level dependent variables the result was not significant, i.e. there is no difference between the measure-selection methods.

Table 7-5 shows that the number of measures that were sensitive to the system modification is low.

Table 7.2, Table 7-3 Table 7-4 and show that some low-level dependent variables were sensitive (statistically significant) but did not have apparent validity whereas others were not sensitive but did have apparent validity. Apparent validity is shown when a measurable property (of the domain) appears in data (from the actor) via discussions and/or observation and/or inference from the interview transcription data.

Two questions need to be asked to establish the reasons why the numbers of low-level variables that were sensitive were low and also to establish why not all sensitive variables had apparent validity. These questions are: (1) whether the experiments were designed correctly, and (2) whether the low-level dependent variables tested were correctly operationalised.

7.4.1.1 Were experiments designed correctly?

To determine whether the experiments were designed correctly it is important to consider whether there was sufficient data to make valid conclusions. A good way to assess this is to determine the number of low-level dependent variables that had enough data points for a valid statistical test to be performed. If the number of data points in any condition is less than 10 or if data were not present in all categories (if categorical data were collected) the variable is classed as not having sufficient data. To determine the relative difference between the measure-selection methods, a useful statistic is the percentage of low-level dependent variables with sufficient data for statistical testing.

Table 7-6 shows the results in terms of the number of low-level dependent variables that did and did not have enough data for statistical testing. The first column shows the percentage of low-level dependent variables with sufficient data for statistical testing. This statistic reflects variables that had the required number of data points to meet the requirements for statistical testing. The next three columns show low-level dependent variables with insufficient data for statistical testing—that is, variables that did not have the required number of data points to meet the requirements for statistical testing. Percentage of low-level dependent variables with insufficient data is further broken down into percentage of low-level dependent variables with some data (at least one data point is

required but the total number of data points is less than required for statistical testing) and percentage of low-level dependent variables with no data.

Table 7-6 Summary of the results for Experiment 1 showing the percentage of low-level dependent variables with and without sufficient data for statistical analysis. The information for the low-level dependent variables with insufficient data is also shown as the percentage of low-level dependent variables with some data and no data.

Method type	% of low-level dependent variables with sufficient data for statistical testing	Low-level dependent variables with insufficient data for statistical testing		
		% of low-level dependent variables with insufficient data	% of low-level dependent variables with some data but still insufficient	% of low-level dependent variables with no data
Constraint	29% (12/42)	71% (30/42)	27% (8/30)	73% (22/30)
WDA	40% (6/15)	60% (9/15)	33% (3/9)	66% (6/9)
CTA	22% (6/27)	78% (21/27)	24% (5/21)	76% (16/21)
Task	16% (4/25)	84% (21/25)	24% (5/21)	76% (16/21)

Table 7-6 shows that 29% (12/42) of the constraint-based low-level dependent variables and 16% (4/25) of the task-based low-level dependent variables had sufficient data for statistical testing. As a corollary, 71% (30/42) of the constraint-based low-level dependent variables and 84% (21/25) of the task-based low-level dependent variables did not have sufficient data for statistical testing (see the second column of Table 7-6). If an exploratory chi-squared test is conducted on the number of low-level dependent variables with sufficient data for statistical testing the result indicates that there no difference between the measure-selection methods, $X2\ (1, N = 67) = 1.36\ p = 0.24$.

Given these results it is reasonable to ask whether if more data points could have been collected for the low-level dependent variables, would a better indication of whether the method predicted the sensitivity of the measures correctly have resulted. The answer to this question is now discussed in terms of the inherent properties of the low-level dependent variables. Specifically, I examine the tendency for the some of the constraint-based low-level dependent variables (WDA-based) to measure the ecological properties of the world and the CTA and task-based low-level dependent variables to measure behavioural properties trigged by events in the world. In the following paragraphs three groups of variables will be described: low-level dependent variables with sufficient data for statistical testing, low-level dependent variables with some data and low-level dependent variables with no data. Within each of these groups WDA, CTA and the task-based low-level dependent variables are discussed.

The properties of the low-level dependent variables are important for the amount of data that can be collected. If the results of the constraint-based low-level dependent variables that had sufficient data for statistical testing are considered, it can be seen that 40% (6/15) of the WDA low-level dependent variables and 22% (6/27) of the CTA low-level dependent variables had sufficient data for analysis. The six WDA low-level dependent variables that have sufficient data for statistical analysis are:

- Probability of detection,
- Timeliness,

- Distance to enemy,
- Probability of damage/ destruction,
- Number of threats and
- Surprise.

Table 7-3 and Table 7-4 show that the six CTA-based low-level dependent variables are:

- Control task priority,
- Control task frequency,
- Choice of procedure -overall behaviour rating,
- Observation of system state - Overall Communication rating,
- Accuracy in identifying stimuli - Display related and
- Accuracy in estimating parameters of threat - Display related.

Chapter 2 showed that the WDA low-level dependent variables are associated with ecological properties of the work domain. The CTA low-level dependent variables represent measures associated with the behavioural inputs and outputs of the tasks ("black boxes"). The results from this experiment suggest that there may be a clear distinction between low-level dependent variables that measure properties of the work domain and low-level dependent variables that measure properties of behavioural tasks. To the extent that the present data are representative, it seems that data can be collected for proportionally more of the low-level dependent variables that described the ecological properties of the work domain than for the low-level dependent variables that reflect the behavioural tasks.

Table 7-6 shows that 16% (4/25) of the task-based low-level dependent variables had sufficient data to analyse statistically. These low-level dependent variables are:

- Choice of procedure -overall behaviour rating,
- Observation of system state - Overall Communication rating,
- Accuracy in identifying stimuli - Display related and
- Accuracy in estimating parameters of threat - Display related.

Closer examination of the results reveals that data collection for these low-level dependent variables was triggered by events (threat changes) outside the control of the Aircraft Captain. In other words, when the RWR displayed a threat the Aircraft Captain was required to respond to it (because standard operating procedures require a response) and behavioural data were collected.

Turning to the data from the low-level dependent variables that had some data (that is low-level dependent variables that had a least one data point but not enough data for statistical testing, see third column of Table 7-6) it can be seen that three WDA-based low-level dependent variables had limited data collection opportunities. These variables are:

- Mission achieved,
- Infantry equipment undamaged and
- Exposure time to enemy.

Two variables Mission achieved and Infantry equipment undamaged had limited data because the number of missions that were run (5 per system modification condition) limited the number of data points that could be collected. For Exposure time to enemy opportunities to collect data for all the exposure time categories was not available. On the surface these results seem to indicate an issue with experimental design. However, as I will show in Section 7.4.2 the issue is not with experimental design, but the suitability of these low-level dependent variables for operational system evaluation.

Table 7-4 shows that there are five low-level dependent variables common to CTA and the task-based method that do have some data, but not enough for statistical testing. These low-level dependent variables are:

- Accuracy of detection of stimulus change over time - Display related,
- Accuracy in estimating position of threat - Display related,
- Accuracy in estimating distance of threat - Display related,
- Rating of performance adequacy - Aircraft captain self assessment and
- Accuracy in response select – Action.

All these low-level dependent variables had external (world) stimuli to which the Aircraft Captain was required to respond. The fact that some data for five of the variables representing the Aircraft Captain's behaviour could not be collected was not due to the design of the simulation environment, but rather to the influence of events (conditions) in the environment. The events (e.g. a threat being observed directly in the world) and the behaviour associated with the events had been observed in the exploratory experiment. Conditions for the events to take place were included in the requirements for this experiment but were not observed in the actual experiment.

Looking at the final column of Table 7-6, it can be seen that six WDA low-level dependent variables and 16 low-level dependent variables common to CTA and the task-based method had no data. The six WDA low-level dependent variables are:

- Tempo,
- Crew physiology,
- Understanding,
- Decision making,
- Maximise time available to react and
- Minimise predictability.

For five of the WDA low-level dependent variables (Tempo, Understanding, Decision making, Maximise time available to react and Minimise predictability) data could not be collected because of the limitations imposed by the simulation environment and this is discussed in Section 7.4.2. Data for the remaining low-level dependent variable (Aircrew physiology) could not be collected because of experiment resource limitations which are discussed in Section 7.4.2.

There are 16 low-level dependent variables common to CTA and task-based methods that had no data. Of these, seven were related to an event in the world that did not occur. The seven low-level dependent variables are:

- Analysis of crew behaviour relevance -overall behaviour rating,
- Accuracy in identifying stimuli - Not display related,
- Accuracy in estimating parameters of threat - Not display related,
- Accuracy in response selection - Not display related,
- Accuracy in confirmation of threat - Flying pilot related,
- Choice of procedure - Flying pilot related and
- Analysis of crew behaviour relevance - Flying pilot related.

Analysis of the results from the experiment revealed that data could not be collected for these low-level dependent variables because behaviour by the Aircraft Captain that would have been captured by these low-level dependent variables was not observed. The failure to observe the behaviour was not a function of the data collection methods – the behaviour was not observed because it was not present in the experiment even though it was present in the exploratory experiment. The reasons for this are discussed in Section 7.5.2.4.

For the remaining nine task-based low-level dependent variables the lack of data may be best explained as an issue associated with the data collection methods which is discussed in Section 7.4.2. The nine low-level dependent variables are all related to measuring time:

- Time taken to defeat threat,
- Time taken detect threat (perceive threat icon),
- Time taken to locate threat in the environment,
- Time taken to initiate a movement to search for a threat in the environment,
- Time taken to deduce location of threat,
- Time taken to deduce range to threat,
- Time taken to select appropriate countermeasure,
- Time taken to expend countermeasure and
- Time taken to select communication system.

From an analysis of the results it seems that the lack of data for some of the task-based and CTA low-level dependent variables was not because of poor experimental design but because of the relationship between such low-level dependent variables and events in the environment. For some of the WDA low-level dependent variables the lack of data may be attributed to limitations of the resources available for modelling the system and the type of data collection method used. These limitations are discussed in Section 7.4.2.

7.4.1.2 Were low-level dependent variables operationalised correctly?

The percentage of low-level dependent variables with apparent validity is a good measure of the number of low-level dependent variables that were operationalised correctly. In the following paragraphs two groups of variables will be described; low-level dependent variables with apparent validity and variables without apparent validity. Within each of these groups of variables, the WDA, CTA and task-based low-level dependent variables are discussed in that order.

Table 7-7 Summary of the results for Experiment 1 that report on the apparent validity of the variables

Method type	% of low-level dependent variables with apparent validity	% of low-level dependent variables with no apparent validity
Constraint	40% (17/42)	60% (25/42)
WDA	46% (7/15)	53% (8/15)
CTA	37% (10/27)	63% (17/27)
Task	36% (9/25)	64% (16/25)

In this experiment a low-level dependent variables is considered to be operationalised correctly if apparent validity can be shown.

Given the above criteria, Table 7-7 shows that 40% (17/42) of the constraint-based low-level dependent variables had apparent validity compared to 36% (9/25) of the task-based low-level dependent variables, and were therefore operationalised correctly. Again this suggests that 60% (25/42) of the constraint-based low-level dependent variables and 64% (16/25) of the task-based low-level dependent variables were not operationalised correctly.

From the constraint-based method results, shown in Table 7-7, 46% (7/15) of the WDA low-level dependent variables had apparent validity. The WDA low-level dependent variables that had apparent validity are:

- Probability of detection,
- Probability of damage/destruction,
- Surprise,
- Distance to enemy,
- Mission achieved,
- Infantry equipment undamaged, and
- Exposure time to enemy.

These low-level dependent variables correctly reflected the ecological properties of the work domain.

The unique CTA low-level dependent variable Control task priority had apparent validity, as did the following low-level dependent variables that are common to the CTA and task-based method:

- Choice of procedure -overall behaviour rating,
- Observation of system state - Overall Communication rating,
- Accuracy in identifying stimuli - Display related,
- Accuracy in estimating parameters of threat - Display related,
- Accuracy of detection of stimulus change over time - Display related,
- Accuracy in estimating position of threat - Display related,
- Accuracy in estimating distance of threat - Display related,
- Rating of performance adequacy - Aircraft captain self assessment and
- Accuracy in response selects – Action.

In the case of the common CTA and task-based low-level dependent variables there were events in the environment that required the Aircraft Captain to respond to.

Looking now at the low-level dependent variables that did not have apparent validity it can be seen that there are 46% (7/15) WDA low-level dependent variables that do not have apparent validity. Six of these low-level dependent variables (Tempo, Understanding, Decision making, Maximise time available to react, Minimise predictability and Crew physiology) were excluded from the experiment because of system modelling and resources constraints and will be discussed in Section 7.4.2. Two of the low-level dependent variables (Timeliness and Number of threats) were incorrectly operationalised and will be briefly discussed now.

The WDA low-level dependent variable Timeliness is interpreted in an operational context, as "the ability of the system to provide information about a threat in a period of time when the aircrew can make best use of that information". The ecological property of the RWR that was determined to form the basis of the low-level dependent variable Timeliness was the speed at which the RWR detected and displayed threats. Timeliness was then defined in terms of the "minimum time to a displayed priority threat". Timeliness was therefore operationalised in terms of the speed at which the RWR detected and displayed threats. The results indicated that the average speed for the modified system was 0.177sec and the average speed for the unmodified system was 0.099sec, a difference of 0.078sec. From a statistical point of view the results were significant, but from the crew's point of view the difference was too small to be noticeable, and had no effect on how they reacted to threats. In other words it is likely that the processing speed of the RWR was the wrong ecological property to operationalise.

The low-level dependent variable Number of threats is defined as "the percentage of threats displayed on the RWR". Once again this low-level dependent variable showed a statistically significant difference between the modified and unmodified system. However, it was found that although it was important for the crew to gather information on the number of threats during a mission they crews did not require (or use) comparative information between RWR system types.

The result that some WDA low-level dependent variables were incorrectly operationalised has implications for the WDA-based measure selection method and specifically points to a requirement to involve SMEs when the ecological properties of the system are being determined. Without involvement of SMEs there may be a risk that low-level dependent variables are erroneously used during system evaluation activities.

The unique CTA low-level dependent variable Control task frequency did not have apparent validity. At no time did the aircrew communicate that the amount of times that they encountered a threat was related to the RWR system modification or was an important factor to them in the conduct of their mission. One again this points to the use of SMEs to determine the important ecological property that should form the basis of the low-level dependent variable.

There are 16 low-level dependent variables common to CTA and task-based methods that had no data. Of these seven were related to an event in the world that did not occur. The seven low-level dependent variables are:

- Analysis of crew behaviour relevance -overall behaviour rating,
- Accuracy in identifying stimuli - Not display related,
- Accuracy in estimating parameters of threat - Not display related,
- Accuracy in response selection - Not display related,
- Accuracy in confirmation of threat - Flying pilot related,
- Choice of procedure - Flying pilot related and
- Analysis of crew behaviour relevance - Flying pilot related.

The relationship between events occurring and some CTA and all task-based low-level dependent variables was discussed in the previous section and also in Section 7.5.2. For the remaining nine task-based low-level dependent variables the lack of data may be best explained as an issue associated with the data collection methods and this is discussed in Section 7.4.2. The nine low-level dependent variables are all related to measuring time and are:

- Time taken to defeat threat,
- Time taken detect threat (perceive threat icon),
- Time taken to locate threat in the environment,
- Time taken to initiate a movement to search for a threat in the environment,
- Time taken to deduce location of threat,

- Time taken to deduce range to threat,
- Time taken to select appropriate countermeasure,
- Time taken to expend countermeasure,
- Time taken to select communication system.

In general given that the analytic products were initially validated (see Chapter 5) and the simulation environment was designed and to emulate the ecological properties of the world (where possible) the aircrew seem more likely to talk, or provide apparent validity, for the WDA-based low-level dependent variables. The fact that some of the CTA and task-based low-level dependent variables had apparent validity is attributed to the occurrence of events that were necessary to stimulate that low-level dependent variable. For the low-level dependent variables that had no apparent validity there was no event to stimulate the behaviour.

In summary, it seems that incorrect operationalisation of the task-based low-level dependent variables did not account for the lack of predictive validity of the task-based method. The low-level dependent variables suggested by task-based method were simply not observable. This represents a limitation of task-based method: the method has poor predictive validity because it incorrectly suggested low-level dependent variables that should be sensitive to the system modification. On the other hand it seems that incorrect operationalisation of the constraint-based variables did account for the reduced of predictive validity of the constraint-based method.
From the preceding results and discussion it seems that although there is no statistical difference between the two measure-selection methods it seems that the constraint-based measure-selection method may show slightly more predictive validity (as measured by the number of sensitive variables) than the task-based method. This difference between the two methods does not occur because the experiments were designed poorly (in terms of data points available and operationalisation of the variables). The lack of predictive validity of the task-based method occurs because the variables that were suggested by the method require events to stimulate them; when those events are not present the low-level dependent variables are of no use. It is suggested that the predictive validity of the WDA-based measure-selection method is improved by involving SME in the process of operationalisation of the low-level dependent variables.

7.4.2 Assessing suitability of methods

Table 7-8 presents the results from the experiment in terms that describe whether the methods are suitable for use in operational settings. From the results it can be seen that the task-based low-level dependent variables are not affected by simulation resource limitations. The table also shows that the constraint-based low-level dependent variables are affected by all the high-level dependent variables.

On the basis of this result the following hypothesis comparing the task-based and constraint-based measure-selection methods is accepted:

H9 The task-based low-level dependent variables will not be affected by some of the following as the constraint-based low-level dependent variables will

be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

The results of this experiment will now be discussed in the context of each of the high-level dependent variables. In the following paragraphs each high-level dependent variable will be defined and then discussed in the context of the constraint-based (WDA, CTA) method and then the task-based method. But first two exploratory Chi-squared tests are conducted. The first is used to ascertain if there is a statistical difference between the measure-selection methods on the number of low-level variables that could not be tested (see Table 7-8). The results revealed that there was no difference between the measure-selection methods, $X2$ (1, N = 67) = 1.36, p = 0.24. The second is used to ascertain if there is a statistical difference between the measure-selection methods on the number of low-level variables were both sensitive and suitable. The results of this second test revealed that there was no difference between the measure-selection methods ($X2$ (1, N = 16) = 1.23, p = 0.27). Table 7-9 shows the summary data for this test.

Table 7-8 Summary of the results for Experiment 1 that report on method suitability

Method type	% of low-level dependent variables that could not be tested	Method suitability high-level dependent variables			
		% of low-level dependent variables for which data could not be collected because of simulation resources limitations	% of low-level dependent variables for which data could not be collected because of data collection method limitations	% of low-level dependent variables for which data not be collected because of the number of data gathering opportunities.	% of low-level dependent variables for which data not be collected because of theoretical limitations
Constraint	71% (30/42)	14% (6/42)	21% (9/42)	19% (8/42)	17% (7/42)
WDA	60% (9/15)	40% (6/15)	0% (0/15)	20% (3/15)	0% (0/15)
CTA	78%(21/27)	0% (0/27)	33% (9/27)	19% (5/27)	26% (7/27)
Task	84% (21/25)	0% (0/25)	36% (9/25)	20% (5/25)	28% (7/25)

Table 7-9 Number of low-level dependent variables that are sensitive and suitable

Method type	Number of low-level variables that were both sensitive and suitable	Number of low-level variables that were suitable and not sensitive
Constraint	3/12	9/12
WDA	3/6	3/6
CTA	0/6	6/6
Task	0/4	4/4

7.4.2.1 Are low-level dependent variables affected by simulation resources?

Percentage of low-level dependent variables for which the collection of data is limited by simulation resources is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because system models (for example, a property of

the RWR system, a property of the world, and a property of the helicopter and mission) cannot be developed, divided by the total number of low-level dependent variables. Table 7-8 shows that the WDA method suggests six low-level dependent variables that could not be tested because of limitations associated with resources needed to model the properties of the RWR and environment. These low-level dependent variables are:

- Tempo,
- Understanding,
- Decision making,
- Maximise time available to react,
- Minimise predictability and
- Crew physiology.

The four low-level dependent variables, Decision making, Tempo, Understanding and Maximise time available to react, represent important internal processes of the RWR and could not be modelled with the resources available. Minimise predictability represents a property of threat systems and again could not be modelled with the resources available. Crew physiology data could not be collected because of resource limitations. There were no CTA and task-based low-level dependent that could not be modelled.

The results from this analysis are important because they illustrate the specific problem of evaluating complex systems in simulation environments. Ultimately evaluation is limited by the resources available. Given that it was only the parts of the work domain specifically related to RWR and threat that could not be modelled; it can be concluded that this is a limitation of the WDA method.

7.4.2.2 Are low-level dependent variables affected by the data collection method used?

Percentage of low-level dependent variables for which the collection of data is limited by the data collection method used is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because the data collection method used to collect the data for the low-level dependent variable was not adequate (because of problems associated with gathering data points manually or automatically), divided by the total number of low-level dependent variables.

The WDA-based method did not suggest any low-level dependent variables that were adversely affected by the type of data collection method used. The lack of data for nine low-level dependent variables common to the CTA and the task-based method is directly attributed to the data collection method used. These low-level dependent variables are:

- Time taken to defeat threat,
- Time taken detect threat (perceive threat icon),
- Time taken to locate threat in the environment,
- Time taken to initiate a movement to search for a threat in the environment,
- Time taken to deduce location of threat,

- Time taken to deduce range to threat,
- Time taken to select appropriate countermeasure,
- Time taken to expend countermeasure and
- Time taken to select communication system.

All these low-level dependent variables are associated with measuring time periods. During the experiment it was impossible to identify accurately the start and end time of the cognitive and physical processes simply through observation. Given that the experiment emulated the evaluation of a complex system in field conditions no method is suitable. It is exceptionally difficult to measure cognitive processes if one is not using a standard experimental paradigm in a laboratory setting. The same point is made by Hennessy (1990) who argued that it is mistaken to try to collect (and analyse) some types of quantitative behavioural data using an experimental paradigm in an operational (simulator) setting and by Vicente (1999) who argues that using task-based data collection methods is problematic because of the propensity to "miss" behaviours in the field. Therefore, it seems that some of the CTA and task-based behavioural measures suggested by the methods are simply not suited for evaluating operational systems. The implication is that any method that identifies potentially sensitive low-level dependent variables must provide some guidance on the efficacy of those low-level dependent variables. Simply providing a list of low-level dependent variables is not adequate. Without such guidance analysts could still be relying on personal experience to select low-level dependent variables; analysts could be selecting measures (low-level dependent variables) "solely on the basis of the expertise and experience of the individual tester" (Charlton and O'Brien, 1996).

7.4.2.3 Are low-level dependent variables affected by the number of data gathering opportunities?

Percentage of low-level dependent variables for which data is limited by the number of data gathering opportunities is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because the number of data collection opportunities was not sufficient to meet statistical protocols (for example, if the number of data points was less than 10 or if data were not present in all categories - if categorical data were collected), divided by the total number of low-level dependent variables.

There are three WDA-based low-level dependent variables that could not be tested because the number of data points was limited by the number of data gathering opportunities in the experiment. These low-level dependent variables are:

- Mission achieved,
- Infantry equipment undamaged, and
- Exposure time to the enemy.

For two low-level dependent variables Mission achieved and Infantry equipment undamaged the data collected was limited by the number of mission runs (one data point per mission). Both low-level dependent variables are concerned with aircraft survivability

– given that aircrew are trained not to get shot at or shot down it could be expected that these variables would not be sensitive to any system modification even if the trials were repeated many times. What is more, the definitions for these variables point to them being typical of measures of effectiveness (see Glossary). It seems that the WDA-based method incorrectly suggested measures of effectiveness (MOE) as being sensitive to the system modification. It also seems that the method did not account for the low number of mission runs. This suggests that the method should be modified to take into account the relationship between the properties of field evaluations (for example, low numbers of trials) and low-level dependent variable properties (whether the low-level dependent variable is typical of a MOE or measure of performance). In that way only low-level dependent variables sensitive to the system modification would be identified.

For the low-level dependent variable Exposure time to enemy it seems that the aircrew adopted range behaviours (tactics) that were dependent on the system modification and tactical situation. It is possible that a broader distribution of data, in a wider range of tactical situations, would result given more opportunities to collect data. It seems that the WDA-method incorrectly identified low-level dependent variables that would be sensitive. However, the experiment would have to be repeated to determine this conclusively.

There are five low-level dependent variables common to CTA and the task-based method for which the number of data point collected was limited by the number of data gathering opportunities. These low-level dependent variables are:

- Accuracy of detection of stimulus change over time - Display related,
- Accuracy in estimating position of threat - Display related,
- Accuracy in estimating distance of threat - Display related,
- Rating of performance adequacy - Aircraft captain self assessment and
- Accuracy in response select – Action.

For all of these low-level dependent variables there was some data but not enough for data for analysis. From an analysis of the data it seems that the CTA and task-based method incorrectly suggested these low-level dependent variables would be sensitive to the system modification. However, the experiment would have to be repeated to determine this conclusively.

7.4.2.4 Are low-level dependent variables affected by theory?

Percentage of low-level dependent variables for which data could not be collected because of theory is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because theory restricted the collection of data, divided by the total number of low-level dependent variables.

There were no WDA-based low-level dependent variables for which the underlying constraint- based perspective limited the collection of data. However, data for seven low-level dependent variables common to CTA and the task-based method were affected by underlying theory. These low-level dependent variables are:

- Analysis of crew behaviour relevance -overall behaviour rating,
- Accuracy in identifying stimuli - Not display related,
- Accuracy in estimating parameters of threat - Not display related,
- Accuracy in response selection - Not display related,
- Accuracy in confirmation of threat - Flying pilot related,
- Choice of procedure - Flying pilot related and
- Analysis of crew behaviour relevance - Flying pilot related.

These low-level dependent variables are limited by "theory" – data can only be collected if an event that stimulates that low-level dependent variable is present in the mission. In the missions observed in the exploratory experiment there were instances where events were observed that could provide data for these low-level dependent variables, but the same event was not observed in the main experiment. For example, in the case of the low-level dependent variable "accuracy in estimating parameters of the threat – not display related" the event related to this, "being able to see a threat out of the window", was never experienced by the Aircraft Captain even though physical representations of the threats were included in the scenarios and located in positions that could have been over flown by the Black Hawk during the experiment. The reliance on using normative behaviour to guide system evaluation is a limitation.

7.5 Are analytic products valid?

Chapter 4 indicated that the results of the experiments could provide information to comment on the validity of the analytical products. Experiment 1 provided the opportunity to test the external validity of the analytical products that were used in the measure-selection methods. The external validity of the analytic product is established if it can be shown that the product represents the domain that it was developed from. In our case the analytic products were developed initially from interviews from SMEs and the results from Experiment 1 provide another data set to assess whether the products represent the ecological properties and tasks that are important to the Aircraft Captain. A good way to do this is by comparing the numbers of low-level dependent variables that the products suggest with the ones that the Aircraft Captain indicates are important. Given that the previous section has indicated that some of the variables could not be tested, because no data could be collected, it is important to only consider the low-level dependent variables for which some data could be collected. Hence, the external validity of the products can be calculated as the numbers of low-level dependent variables that have apparent validity, over the sum of the number of low-level dependent variables with some data plus the number of low-level dependent variables that were "incorrectly" suggested. For a low-level dependent variable to be "incorrectly" suggested it must be shown that the low-level dependent variable clearly did not represent an ecological property or task in the domain. To make comparisons between analytic products easier external validity may be expressed as a percentage. Hence, if the value tends to zero the products have no validity –ecological properties and tasks were incorrectly identified as

being important to the Aircraft Captain. If the value is 100% then the products are valid - the products have correctly identified ecological properties and tasks (embodied as variables) that are important to the Aircraft Captain.

Table 7-10 shows the external validity scores and indicates that products have a range of validity scores. In the case of the constraint-based products a combined score of 85% is shown. The table also shows that the modified AH has a score of 78% and the TC-CTA has a score of 91%. The modified AH score reflects the fact that the ecological properties of two low-level dependent variables (Timeliness and Number of threats) were incorrectly selected by the author. In the case of the TC-CTA product the ecological property of one low-level dependent variable (Control task frequency) was incorrectly selected by the author as important to the Aircraft Captain. The table shows that all the tasks (100%) that were observed in the experiment were correctly identified using the task-based products. Chi squared test reveals that there is no significant difference in the external validity of the analytic products ($X2$ (1, N = 30) =1.67, p = 0.20).

It is concluded that in general the analytical products that the methods used were valid, but that some improvements can be made through better selection of the ecological properties of the constraint-based variables.

Table 7-10 External validity of the analytic products. The number of low-level dependent variables with apparent validity over the sum of the number of correct and incorrect low-level dependent variables expressed as a percentage for each analytical product.

Analytic products	External validity score
Constraint-based	85% (17/20)
AH	78% (7/9)
TC-CTA	91% (10/11)
Task -based	100% (10/10)

7.6 Conclusion

This chapter has presented Experiment 1, which was designed to compare the predictive validity of the constraint-based and task-based measure-selection method for a current system using two tests: (1) the sensitivity of the low-level dependent variables selected under each method to system modification, and (2) the suitability of the methods for use in operational evaluations.

The first test of predictive validity was achieved by evaluating whether the low-level dependent variables that the methods suggest are sensitive to the system modification, i.e. whether the low-level dependent variables showed a statistically significant difference between the two conditions (modified RWR and unmodified RWR). The results indicate that although there was no statistical difference between the measure-selection methods, some of the constraint-based low-level dependent variables are sensitive to the RWR modification. In particular, some of the WDA-based low-level dependent variables but none of the CTA-based low-level dependent variables were statistically significant. Finally, none of the task-based low-level dependent variables were sensitive to the RWR system modification. In summary, although there is no conclusive proof that there is a difference

between the two methods the constraint-based WDA method shows some degree of predictive validity on this test, but the task-based method does not.

The second test of predictive validity was achieved by assessing whether the low-level dependent variables, suggested by the methods, were affected by four common pragmatic conditions that are typical of operational settings. The assessment found that the task-based low-level dependent variables were not affected by resource limitations, but were affected by data collection methods, data gathering opportunities and theory. The constraint-based low-level dependent variables, as a whole, were affected by resource limitations, data collection methods, data gathering opportunities and theory. However, the low-level dependent variables that the WDA method suggested were not affected by data collection methods and theory and the CTA low-level dependent variables were not affected by resource limitations, but were affected by data collection methods, data gathering opportunities and theory. The results also revealed that there was no difference between the methods in terms of the number of low-level dependent variables that were both sensitive and suitable. In summary, no method showed predictive validity because they were all affected by some common operational conditions (resource limitations, data collection method limitations, low number of data gathering opportunities or theory).

These results from this experiment are important because they indicate that for a current complex socio-technical system:

- The task-based method does not produce sensitive low-level dependent variables for operational system evaluation.
- The constraint-based method produces some sensitive low-level dependent variables (the WDA-based method produces some low-level dependent variables, the CTA-based method does not).
- The task-based method and the CTA-method were found to be more suitable (in terms of the resources needed to use the methods) than the WDA-based method for evaluating complex socio-technical systems in the operational setting tested. However, the low-level dependent variables produced by the WDA-based method are not affected by data collection and theoretical concerns.

To improve the constraint-based method for use during system evaluation it is suggested that:

- The constraint-based method, specifically the WDA method should be modified to better indicate how the properties of the operational context should be operationalised.
- The WDA method should be modified to identify important properties of low-level dependent variables and in particular identify whether the low-level dependent variables is a measure of effectiveness or a measure of performance.
- The results of the experiment have indicated that the WDA method relies on system models that reflect the ecology of the work domain so that low-level dependent variables are identified that are sensitive to system modifications. This means that the cost of resources (financial and human resources) needed to use the WDA method in operational settings should be carefully considered.

- The CTA-based method should be modified to ensure that operationalisation of the properties of the control tasks in the simulator is performed correctly.

The next chapter will examine whether the pattern of results seen here is echoed for a future RWR system.

8. Experiment 2: Comparing methods with a Future System

Experiment 1 revealed that although there was no statistical difference between the measure-selection methods on the number of sensitive low-level dependent variables that they suggested, some WDA-based low-level dependent variables were sensitive to a modification for a system with which aircrew were familiar, but none of the task and CTA low-level dependent variables tested were sensitive. However, Experiment 1 also revealed that the task-based low-level dependent variables and the CTA-based low-level dependent variables were affected by the conditions commonly experienced in operational settings (e.g. resource limitations, data collection method limitations, low number of data gathering opportunities and limitations of theory).

Experiment 2 is an investigation of whether constraint-based and task-based dependent variables are sensitive to a modification of an RWR system of which the aircrew had no experience – a future system. Experiment 2 will also investigate whether the methods are suitable for evaluation of future systems in operational settings, i.e. whether the low-level dependent variables that they suggest are affected by the conditions commonly experienced in operational settings. Figure 8-1 shows that, as for Experiment 1, Experiment 2 lies at Stage 4 of the research program.

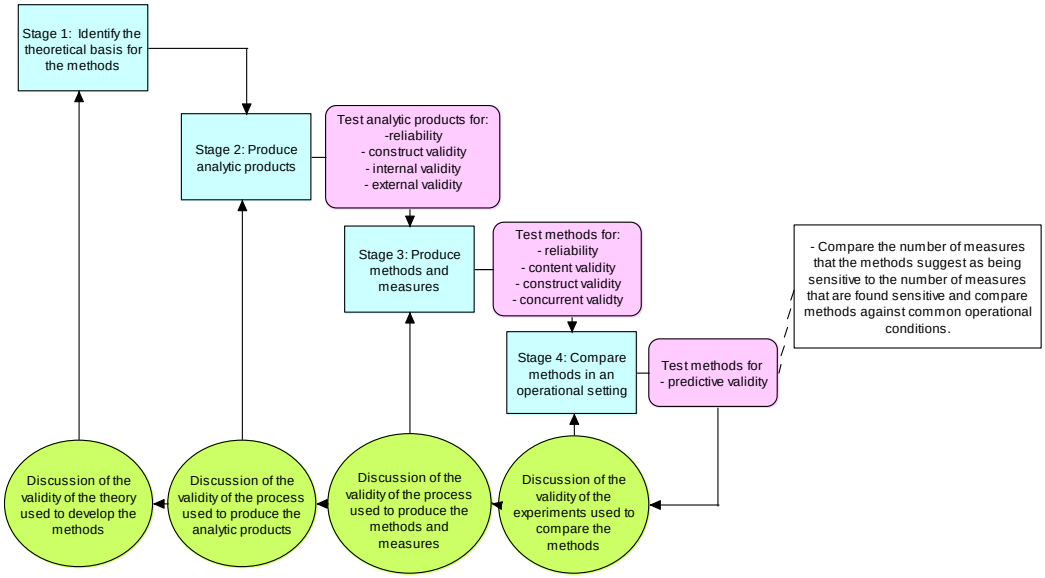


Figure 8-1 Four stages of the research program. The present chapter focuses on Stage 4, in which the measure-selection methods are compared for their relative sensitivity to a change in a future technical system (RWR).

The factor that distinguishes the future RWR system used in Experiment 2 from the RWR modification used in Experiment 1 is the presence of information about the range (distance) of the threat to the aircraft on the display, as well as information about the threat's position relative to the aircraft and terrain. The principal modification is the change in the range over which a threat can be detected by the RWR.

To be considered sensitive, a low-level dependent variable should show a statistical difference between the two system conditions (unmodified and modified). To be suitable a method must suggest low-level dependent variables that are not affected by resource limitations, data collection method limitations, low number of data gathering opportunities and limitations of theory. Comparing the methods in terms of measure sensitivity and suitability provides an indication of the predictive validity of the methods.

In the next section the background and aims for this experiment are outlined. The experiment is then reported and implications for this program of research stated. Finally, the chapter draws conclusions.

8.1 Background, aims and hypotheses

The aim of Experiment 2 is to test the relative predictive validity of the two measure-selection methods. The predictive validity of each measure-selection method will be tested by evaluating whether the low-level dependent variables that each measure-selection method suggests are sensitive to the system modification and whether each measure-selection method is suitable for use in operational settings.

For Experiment 2, as for Experiment 1, there are five hypotheses that are related to the issues of measure sensitivity and five hypotheses that are related to the issues of method suitability.

Measure sensitivity:

H1 None of the task-based and none of the constraint-based low-level dependent variables are sensitive to the system modification.

H2 All the task-based and all the constraint-based low-level dependent variables are sensitive to the system modification.

H3 Significantly more of the constraint-based low-level dependent variables than the task-based low-level dependent variables will be sensitive to the system modification.

H4 Significantly more of the task-based low-level dependent variables than the constraint-based low-level dependent variables will be sensitive to the system modification.

H5 Some of the task-based and constraint-based low-level dependent variables will be sensitive to the system modification.

Method suitability:

H6 The task-based and constraint-based low-level dependent variables will be affected by all of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H7 The task-based and constraint-based low-level dependent variables will not be affected by all of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H8 The constraint-based low-level dependent variables will not be affected by some of the following as the task-based low-level dependent variables will be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H9 The task-based low-level dependent variables will not be affected by some of the following as the constraint-based low-level dependent variables will be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

H10 The task-based low-level dependent variables and the constraint-based low-level dependent variables will not be affected by some of the following: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

The experimental method that was used to test the hypotheses is presented in the following section.

8.2 Method

In the following sections the main biographical details of the participants, the apparatus and materials used during the experiment, and the experimental design will be described.

8.2.1 Participants

Two serving members of the Australian Army took part in the experiment. The two participants formed a helicopter crew that consisted of a Flying Pilot (FP) and an Aircraft Captain (AC). Both participants held the rank of Captain. The participants were volunteers and were different from those used in the exploratory experiment and Experiment 1. The Aircraft Captain was considered an experienced RWR operator [16].

The Aircraft Captain had 150 flight hours flying Black Hawk. He had significant experience in several other helicopter types: Merlin, 400 flight hours; Lynx, 300 flight hours; Wessex, 2400 flight hours; and Gazelle 150 flight hours. He had significant Electronic Warfare (EW) experience including being an EW instructor and tactics

[16]The term "experienced" is used in a relative rather than absolute sense. The Aircraft Captain was an experienced RWR operator because unlike other Australian Army Aircraft Captains he had EW operational experience and was an EW instructor.

instructor. He had relevant operational experience. The Flying Pilot had 1000 Black Hawk flight hours. During the experiment the aircrew used their own flight gear including helmets, gloves and flight clothes.

8.2.2 Apparatus and materials

The apparatus and materials used in this experiment are the same as those used in the first experiment. Only the RWR was different.

Figure 8-2 shows the RWR display that was used in this experiment. The main features of this display are:

- The actual position of the threat is shown relative to terrain.
- Once detected a threat is not lost. The icon will remain visible throughout the duration of the mission. However, if the threat is at a distance greater than that covered by the display, the icon will have a broken outline.
- The threats are shown as red icons. They are labelled and their radar detection ranges and weapon engagement zones are shown as red rings.
- Dark blue areas indicate that the aircraft is currently in line of sight of a threat.
- The terrain features are shown.
- The aircraft is shown in blue in the middle of the display.
- There is no error associated with the display of the threat's position (unlike the Current RWR system).

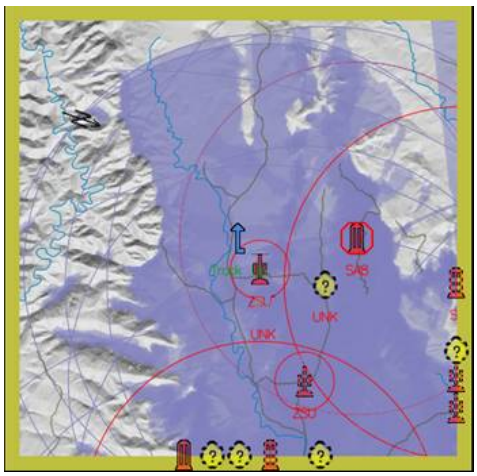


Figure 8-2 Future radar warning receiver (RWR) system

8.2.3 Design

The design of Experiment 2 (see Table 8-1) was the same as Experiment 1.

Table 8-1 Experimental design for experiment 2

	Measure-selection method			
	Constraint-based		**Task-based**	
	System modification		System modification	
	Unmodified	**Modified**	**Unmodified**	**Modified**
Future RWR (Expt 2)	Constraint-based low-level dependent variable 1 ... Constraint-based low-level dependent variable 42		Task-based low-level dependent variable 1 ... Task-based low-level dependent variable 25	

8.2.3.1 Constraint-based low-level dependent variables

The constraint-based low-level dependent variables tested were the same as in Experiment 1.

8.2.3.2 Task-based low-level dependent variables

The task-based low-level dependent variables tested were the same as in Experiment 1.

8.3 Procedure

Experiment 2 was conducted over a period of three days. The procedure was the same as in Experiment 1 except that the aircrew were informed that the RWR provided perfect threat location information.

8.4 Results and discussion

The aim of Experiment 2 is to test the comparative predictive validity of the measure-selection methods. If the methods produce sensitive low-level dependent variables and are suitable for operational settings then they may be judged as having predictive validity.

Findings are presented under headings representing these two aims. The results show that the constraint-based method and the task-based methods have a similar but low degree of predictive validity in terms of the sensitivity of the low-level dependent variables produced but that there is no statistical difference between the two measure-selection methods. The results also show that the constraint-based method and the task-based method do not have predictive validity in terms of suitability because they suggest low-level variables that are affected by simulation resource limitations, data collection method limitations, opportunities for data gathering and theory.

8.4.1 Assessing sensitivity of variables that methods suggested

On the basis of the results there is some evidence to support the following hypothesis, but the evidence is minimal:

H5 Some of the task-based and constraint-based variables are sensitive to the system modification.

The results from Experiment 2 are shown in three tables. Table 8-2 summarises the results for the WDA-based low-level dependent variables. Table 8-3 summarises the results for the unique CTA-based low-level dependent variables and Table 8-4 summarises the results of the remaining CTA-based and task-based low-level dependent variables. Finally Table 8-5 provides a summary of the low-level dependent variables that are used to compare the methods. (Note: observational data, the interview transcription, and the statistical tests conducted on the low-level dependent variables are available from the author on request.)

Table 8-2 reveals that of the 15 WDA–based low-level dependent variables, one showed a statistically significant result and 14 did not. The table also indicates that interviews and observational data provided evidence to indicate that some of the dependent low-level dependent variables (seven of 15) had apparent validity.

Table 8-2 Summary of the results for the WDA-based low-level dependent variables (Experiment 2)

Low-level dependent variables suggested by WDA	Summary of statistical test (result, test used, statistic, summary)	Apparent validity
Probability of damage/ destruction	Significant result X^2 (3, N = 171) = 10.62, p= 0.01 Modified system detected less instances of higher threat modes	Yes
Mission Achieved	Not a significant result 5/5 missions were achieved with the modified system compared to 4/5 with the unmodified system	Yes
Infantry Equipment Undamaged	Not a significant result Damage was seen in 0/5 missions with the modified system compared to 0/5 with the unmodified system	Yes
Distance to enemy	No statistical test – Property of display prevents data being collected	Yes
Probability of detection	No statistical test – Property of display prevents data being collected	Yes
Exposure time to enemy	No statistical test – Property of display prevents data being collected	Yes
Number of threats	No statistical test – Property of display prevents data being collected	Yes
Timeliness	No statistical test – Property of display prevents data being collected	No
Surprise	No data collected – Not an RWR property.	No
Tempo	No data collected – RWR property associated with the measure could not be modelled in the simulation environment	No
Understanding	No data collected – RWR property associated with the measure could not be modelled in the simulation environment	No

Low-level dependent variables suggested by WDA	Summary of statistical test (result, test used, statistic, summary)	Apparent validity
Decision making	No data collected – RWR property associated with the measure could not be modelled in the simulation environment	No
Maximise time available to react	No data collected – RWR property associated with the measure could not be modelled in the simulation environment	No
Minimise predictability	No data collected – Threat property associated with the measure could not be modelled in the simulation environment	No
Crew physiology	No data collected – Data could not be collected within the constraints on the program of research.	No

Table 8-3 shows that of the two unique dependent low-level dependent variables for the control task "Operating RWR system in response to threat", neither showed a significant statistical difference between the modified and unmodified system. The table also indicates that one of the low-level dependent variables had apparent validity.

Table 8-3 Summary of the unique CTA results (Experiment 2)

Low-level dependent variables suggested by CTA for the Control task "Manage EW system"	Summary of statistical test (result, test used, statistic, summary)	Apparent validity
Control task priority	Not significant The Control task "manage EW systems" was interrupted in 49% (54/111) instances of threat engagements in the modified system condition compared to 58% (73/126) in the unmodified system.	Yes
Control task frequency	Not significant The Control task "manage EW systems" occurred in 40% (177/442) instances of threat engagements in the modified system condition compared to 38% (199/521) in the unmodified system.	No

Table 8-4 indicates that of the 27 low-level dependent variables that were common to both the task-based and CTA method, one showed a statistical significant result and had apparent validity. In total nine of the 27 had apparent validity.

Table 8-4 Summary of the common task-based and CTA low-level dependent variables results (Experiment 2)

Low-level dependent variables suggested by HE and CTA	Summary of statistical test (result, test used, statistic, summary)	Apparent validity
Accuracy in estimating parameters of threat - Display related	Significant X^2 (1, N = 353) = 4.26, p= 0.04 (Modified - AC correctly identified threat properties	Yes

Low-level dependent variables suggested by HE and CTA	Summary of statistical test (result, test used, statistic, summary)	Apparent validity
	83% of all instances; Unmodified - AC detected threat properties 74% of all instances)	
Observation of system state - Overall Communication rating	Not significant X^2 (1, N = 353) = 11.68, p = 0.73 (Modified - AC correctly communicated threat to pilot 88% of all instances; Unmodified - AC correctly communicated threat to pilot 87% of all instances)	Yes
Accuracy in identifying stimuli - Display related	Not significant X^2 (1, N = 353) = 0.33, p =0.56 (Modified - AC detected threat 95% of all instances; Unmodified - AC detected threat 94% of all instances)	Yes
Accuracy in estimating position of threat - Display related	Not significant (Modified – AC correctly estimated the position of the threat 100%; Unmodified – AC correctly estimated the position of the threat 50%)	Yes
Accuracy in response select - Action	Not significant (Modified n=10, AC correctly expended countermeasure 100%; Unmodified n=36, AC correctly expended countermeasure 100%)	Yes
Choice of procedure -overall behaviour rating	Not significant No data – display property, coding reliability	Yes
Accuracy of detection of stimulus change over time - Display related	Not significant No data – display property, coding reliability	Yes
Accuracy in estimating distance of threat - Display related	Not significant No data – display property, coding reliability	Yes
Analysis of crew behaviour relevance -overall behaviour rating	Not significant No data – display property, coding reliability	No
Accuracy in identifying stimuli - Not display related	No comparison possible – No data	No
Accuracy in estimating parameters of threat - Not display related	No comparison possible – No data	No
Accuracy in response selection - Not display related	No comparison possible – No data	No
Accuracy in confirmation of threat - Flying pilot related	No comparison possible – No data	No
Analysis of crew behaviour	No comparison possible- No data	No

Low-level dependent variables suggested by HE and CTA	Summary of statistical test (result, test used, statistic, summary)	Apparent validity
relevance - Flying pilot related		
Choice of procedure - Flying pilot related	No comparison possible – No data	No
Rating of performance adequacy - Aircraft captain self assessment	No comparison possible – No data	No
Time taken to defeat threat	No comparison possible – No data	No
Time taken detect threat (perceive threat icon)	No comparison possible – No data	No
Time taken to locate threat in the environment	No comparison possible – No data	No
Time taken to initiate a movement to search for a threat in the environment	No comparison possible – No data	No
Time taken to deduce location of threat	No comparison possible – No data	No
Time taken to deduce range to threat	No comparison possible – No data	No
Time taken to select appropriate countermeasure	No comparison possible – No data	No
Time taken to expend countermeasure	No comparison possible – No data	No
Time taken to select communication system	No comparison possible – No data	No

One aim of this program of research is to compare task-based and constraint-based measure-selection methods in terms of the extent to which each can provide low-level dependent variables that are sensitive to the system modification. As with Experiment 1 a reasonable way to compare the two methods is on the percentage of low-level dependent variables that prove to be sensitive. Such a statistic indicates the relative predictive validity of each method. Table 8-5 shows the percentage of low-level dependent variables that were found to be statistically sensitive to the system modification.

The results of this comparison reveal that 5% (2/42) of the constraint-based low-level dependent variables were statistically sensitive to the system modification and that 4% (1/25) of the task-based low-level dependent variables were statistically sensitive to the system modification for a future system. An exploratory chi-squared test was conducted on the number of low-level dependent variables that were found to be statistically

significant. The analysis revealed that there was no significant difference between the measure-selection methods, X2 (1, N = 67) = 0.02, p=0.88.

Table 8-5 Summary of the results for Experiment 2 that report on the statistical significance of the low-level dependent variables

Method type	% of low-level dependent variables that are statistically sensitive
Constraint	5% (2/42)
WDA	7% (1/15)
CTA	4% (1/27)
Task	4% (1/25)

Table 8-2, Table 8-3 and Table 8-4 show that some low-level dependent variables were sensitive (statistically significant) but did not have apparent validity, whereas others were not sensitive but did have apparent validity.
As with Experiment 1 two questions need to be asked to establish why the numbers of low-level dependent variables that were sensitive were low and also why not all sensitive low-level dependent variables had apparent validity: (1) were the experiments designed correctly? And (2) were the low-level dependent variables tested correctly operationalised?

8.4.1.1 Were experiments designed correctly?

To determine whether the experiments were designed correctly it is important to consider whether there was sufficient data to make valid conclusions. A good way to assess this is to determine the number of variables that had enough data points for a valid statistical test to be performed. If the number of data points is less than 10 (in any condition) or if data were not present in all categories (if categorical data were collected) the variable is classed as not having sufficient data.

To determine the relative difference between the measure-selection methods a useful statistic is the percentage of low-level dependent variables with sufficient data for statistical testing.

Table 8-6 Summary of the results for Experiment 2 showing the percentage of low-level dependent variables with and without sufficient data for statistical analysis.

Method type	% of low-level dependent variables with sufficient data for statistical testing	Low-level dependent variables with insufficient data for statistical testing		
		% of low-level dependent variables with insufficient data	% of low-level dependent variables with some data but still insufficient	% of low-level dependent variables with no data
Constraint	19% (8/42)	81% (34/42)	44% (15/34)	56% (19/34)
WDA	7% (1/15)	93% (14/15)	43% (6/14)	57% (8/14)
CTA	26% (7/27)	78% (21/27)	20% (4/20)	80% (17/20)
Task	16% (4/25)	84% (21/25)	43% (9/21)	57% (12/21)

Table 8-6 shows the results in terms of the number of low-level dependent variables that did and did not have enough data for statistical testing. The first column is the percentage of low-level dependent variables with sufficient data for statistical testing. This statistic reflects low-level dependent variables that had the required number of data points to meet the requirements for statistical testing. The remaining columns refer to low-level dependent variables with insufficient data for statistical testing. The second column is the percentage of variables with insufficient data. This statistic reflects low-level dependent variables that did not have the required number of data points to meet the requirements for statistical testing. The third column is percentage of variables with some data but still insufficient (at least one data point is required but the total number of data points is less than required for statistical testing) and a fourth column is percentage of variables with no data.

Table 8-6 shows that 19% (8/42) of the constraint-based low-level dependent variables and 16% (4/25) of the task-based low-level dependent variables had sufficient data for statistical testing, with the corollary that 81% (34/42) of the constraint-based low-level dependent variables and 84% (21/25) of the task-based variables did not have sufficient data for statistical testing (see the second column of Table 8-6). If an exploratory chi-squared test is conducted on the number of low-level dependent variables with sufficient data for statistical testing it can be seen that there is no difference between the measure-selection methods, $X2(1, N=67)=0.01$, $p=0.75$.

Given these results it is reasonable to ask whether, if more data points had been collected for the low-level dependent variables, would a better indication of whether the method could predict the sensitivity of the measures correctly have resulted. The answer probably lies in the inherent properties of the low-level dependent variables. Those properties are the tendency for some of the constraint-based low-level dependent variables (WDA-based) to measure the ecological properties of the world and for the TC-CTA and task-based low-level dependent variables to measure behavioural properties trigged by events in the world. In the following paragraphs three groups of low-level dependent variables will be described; low-level dependent variables with sufficient data for statistical testing, low-level dependent variables with some data and low-level dependent variables with no data. Within each of these groups WDA, CTA and the task-based low-level dependent variables are discussed.

Table 8-6 shows that both the constraint-based low-level dependent variables and the task-based low-level dependent variables have similar proportions of low-level dependent variables that have sufficient data for statistical testing: 19% (8/42) for the constraint-based low-level dependent variables and 16% (4/25) for the task-based low-level dependent variables. There is one WDA-based low-level dependent variable that had sufficient data and this was Probability of damage/ destruction. Both of the unique CTA-based low-level dependent variables have sufficient data for statistical testing. These are Control task priority and Control task frequency.

There were four low-level dependent variables common to the CTA and task-based method that have sufficient data for statistical testing. These low-level dependent variables are:

- Accuracy in estimating parameters of threat- display related,
- Observation of system state - Overall Communication rating,
- Accuracy in identifying stimuli - Display related,
- Accuracy in response selection – Action.

In general, these low-level dependent variables are the same ones as seen in Experiment 1.

Unlike the results to Experiment 1 there does not seem to be a relationship between what the low-level dependent variables measure (ecological aspects or event-based behavioural aspects) and whether the low-level dependent variables have sufficient data for statistical testing. However, as will be discussed in Section 8.4.2 the reason why only one WDA low-level dependent variables had sufficient data for analysis was because of limitations associated with system modelling and resources.

The results shown in Table 8-6 reveal that there are six WDA- based low-level dependent variables that have some data. For two of these low-level dependent variables, Mission Achieved and Infantry equipment undamaged, the number of data points collected was limited by the number of data gathering opportunities, and the reasons for this is discussed in Section 8.4.2.
For the remaining four low-level dependent variables (Probability of detection, Number of threats, Distance to enemy, Exposure time to enemy) there was some data, but not enough for statistical testing because of the limitations associated with the display properties and the data collection methods used. Section 8.4.2 discusses this in detail.

There are 3 low-level dependent variables common to the CTA and task-based method that have some data. These low-level dependent variables are:

- Choice of procedure -overall behaviour rating,
- Accuracy of detection of stimulus change over time - Display related, and
- Accuracy in estimating distance of threat - Display related.

The reason for the lack of data is attributed to the interaction between the display properties and the data collection methods used. Section 8.4.2 discusses this in detail.

There are eight WDA-based low-level dependent variables for which no data could be collected:

- Tempo,
- Crew physiology,
- Understanding,
- Decision making,

- Max. time available to react,
- Minimise predictability,
- Surprise and
- Timeliness.

The first six have no data because of resource limitations (see Section 8.4.2). Surprise was incorrectly suggested by the WDA-method as potentially being sensitive. This may mean that the analytic product that the WDA-method used, specifically the AH, was not valid. Timeliness had no data due to limitations associated with the data collection methods and this is discussed in Section 8.4.2.

There are 17 low-level dependent variables common to the CTA and task-based methods that have no data. Seven of these have no data because of the theoretical perspective that they embody. These low-level dependent variables are:

- Accuracy in identifying stimuli - Not display related,
- Accuracy in estimating parameters of threat - Not display related,
- Accuracy in response selection - Not display related,
- Accuracy in confirmation of threat - Flying pilot related,
- Choice of procedure - Flying pilot related,
- Analysis of crew behaviour relevance - Flying pilot related and
- Rating of performance adequacy - Aircraft captain self assessment.

Section 8.4.2 discusses how theory, in particular the emphasis on representing normative behaviour, limits the data than can be collected. Data for the remaining ten low-level dependent variables is limited by the data collection methods used. The low-level dependent variables affected are:

- Analysis of crew behaviour relevance -overall behaviour rating,
- Time taken to defeat threat,
- Time taken detect threat (perceive threat icon),
- Time taken to locate threat in the environment,
- Time taken to initiate a movement to search for a threat in the environment,
- Time taken to deduce location of threat,
- Time taken to deduce range to threat,
- Time taken to select appropriate countermeasure,
- Time taken to expend countermeasure and
- Time taken to select communication system.

Section 8.4.2 discusses the impact that data collection methods have on the collection of data for these low-level dependent variables.

From an analysis of the results it seems that the lack of data for some of the low-level dependent variables was not because of poor experiment design but because of the relationship between task-based and CTA low-level dependent variables and events in the environment, i.e. there is a limitation on the use of these low-level dependent variables for system evaluation that is attributable to theory. This is consistent with the results from Experiment 1. For some of the WDA constraint-based low-level dependent variables the lack of data may be attributed to limitations imposed by the simulation (operational) environment, resource limitations and the type of data collection method used. This is again consistent with the results from Experiment 1.

8.4.1.2 Were the low-level dependent variables operationalised correctly?

The percentage of low-level dependent variables with apparent validity is a good measure of the number of low-level dependent variables that were operationalised correctly. In the following paragraphs two groups of low-level dependent variable sets will be described; low-level dependent variables with apparent validity and low-level dependent variables without apparent validity. Within each of these groups of low-level dependent variables the WDA, CTA and task-based low-level dependent variables are discussed in that order.

Table 8-7 Summary of the results for Experiment 2 that report on the apparent validity of the low-level dependent variables

Method type	% of low-level dependent variables with apparent validity	% of low-level dependent variables no apparent validity
Constraint	38% (16/42)	62% (26/42)
WDA	46% (7/15)	53% (8/15)
CTA	33% (9/27)	67% (18/27)
Task	32% (8/25)	68% (17/25)

In this experiment a low-level dependent variable is considered to be operationalised correctly if apparent validity can be shown, either by the aircrew stating it as a goal or by the author observing instances of the behaviour that the variable represents on at least one occasion.

Given the above criteria, Table 8-7 shows that 32% (8/25) of the task-based low-level dependent variables had apparent validity, and were therefore operationalised correctly, compared to 38% (16/42) of the constraint-based low-level dependent variables. Again this suggests 68% (17/25) of the task-based low-level dependent variables and 62% (26/42) of the constraint-based low-level dependent variables were not operationalised correctly. This is consistent with the results of the first experiment.

From the constraint-based method results shown in Table 8-7 46% (7/15) of the WDA low-level dependent variables had apparent validity. The WDA low-level dependent variables that had apparent validity were:

- Probability of detection,
- Probability of damage/ destruction,

- Distance to enemy,
- Mission achieved,
- Infantry equipment undamaged,
- Exposure time to enemy and
- Number of threats.

These low-level dependent variables correctly reflected the ecological properties of the work domain.

One unique CTA low-level dependent variable Control task priority had apparent validity, as did the following low-level dependent variables that are common to the CTA and task-based method:

- Choice of procedure -overall behaviour rating,
- Observation of system state - Overall Communication rating,
- Accuracy in identifying stimuli - Display related,
- Accuracy in estimating parameters of threat - Display related,
- Accuracy of detection of stimulus change over time - Display related,
- Accuracy in estimating position of threat - Display related,
- Accuracy in estimating distance of threat - Display related,
- Accuracy in response selects – Action.

In the case of the common CTA and task-based low-level dependent variables there were events in the environment to which the Aircraft Captain was required to respond to.

Looking now at the low-level dependent variables that did not have apparent validity it can be seen that there are 53% (8/15) of WDA low-level dependent variables that do not have apparent validity. Seven of these low-level dependent variables (Tempo, Understanding, Decision making, Maximise time available to react, Minimise predictability, Crew physiology and Surprise) were excluded from the experiment because of system modelling and resources constraints and will be discussed in Section 8.4.2. Data for Timelines could not be collected because of the interaction between the display properties of the future RWR and data collection methods (see Section 8.4.2).

The unique CTA variable Control task frequency did not have apparent validity. At no time did the aircrew communicate that the amount of times that they encountered a threat way was related to the RWR system modification or was an important factor to them in the conduct of their mission. This result is consistent with Experiment 1 and points to poor operationalisation of the low-level dependent variable and the use of SMEs when the low-level dependent variables are being operationalised.

There are 17 low-level dependent variables common to CTA and task-based methods that had no data. Of these seven were related to an event in the world that did not occur. The seven low-level dependent variables were:

- Accuracy in identifying stimuli - Not display related,
- Accuracy in estimating parameters of threat - Not display related,
- Accuracy in response selection - Not display related,
- Accuracy in confirmation of threat - Flying pilot related,
- Choice of procedure - Flying pilot related,
- Analysis of crew behaviour relevance - Flying pilot related and
- Rating of performance adequacy - Aircraft captain self assessment.

The relationship between events occurring and data for some CTA and all task-based low-level dependent variables is discussed in Section 8.4.2.

For the remaining 10 task-based low-level dependent variables, the lack of data may be best explained as an issue associated with the data collection methods and this is discussed in Section 8.4.2. Nine low-level dependent variables are all related to measuring time and are:

- Time taken to defeat threat,
- Time taken detect threat (perceive threat icon),
- Time taken to locate threat in the environment,
- Time taken to initiate a movement to search for a threat in the environment,
- Time taken to deduce location of threat,
- Time taken to deduce range to threat,
- Time taken to select appropriate countermeasure,
- Time taken to expend countermeasure,
- Time taken to select communication system.

The one remaining low-level dependent variable, Analysis of crew behaviour relevance - overall behaviour rating, is related to crew behaviour.

From the analysis it seems that, in general, poor operationalisation of the low-level dependent variables was not a factor that limited the amount of data collected. It seems for the WDA low-level dependent variables data were limited because of system modelling and resource limitations. In the case of the common CTA and task-based low-level dependent variables the main factors that limited the collection of data were the data collection methods that were used and the theoretical perspective that the variables embodied.

From the preceding results and discussion it seems that the limited degree of predictive validity of the constraint-based and task-based method is not because the experiments were designed poorly (in terms of data points available and operationalisation of the variables). The lack of predictive validity of the task-based and CTA-based methods is because the low-level dependent variables that were suggested by the methods require events to stimulate them; when those events are not present the variables are of no use. The task-based and CTA-based low-level dependent variables were also found to be susceptible to issues associated with the interaction between the display properties and data collection methods. In the case of the WDA-based method it seems that the limited predictive validity is because of system modelling and resource issues and the interaction between the display properties and data collection methods.

8.4.2 Assessing suitability of methods

On the basis of the results the following hypothesis appears to be supported:

H9 The task-based low-level dependent variables will not be as affected by some of the following as the constraint-based method will be: simulation resources, data collection methods used, the number of data gathering opportunities and limitations from theory.

Table 8-8 presents the results from the experiment in terms of the high-level dependent variables that describe whether the methods are suitable for use in operational settings. From the results it can be seen that the low-level dependent variables from the task-based method are affected by the data collection methods and theory. The table also shows that the constraint-based low-level dependent variables are affected by resource limitations, data collection methods, data gathering opportunities and theory.
The results of this experiment will now be discussed in the context of each of the high-level dependent variables. In the following paragraphs each high-level dependent variable will be defined and then discussed in the context of the constraint-based (WDA, CTA) method and then the task-based method. But first two exploratory Chi-squared tests are conducted. The first is used to ascertain if there is a statistical difference between the measure-selection methods on the number of low-level variables that could not be tested (see Table 8-8). The results revealed that there was no difference between the measure-selection methods, $X2\ (1, N = 67) = 0.01\ p = 0.75$. The second is used to ascertain if there is a statistical difference between the measure-selection methods on the number of low-level variables that were both sensitive and suitable. The results of this second test revealed that there was no difference between the measure-selection methods ($X2\ (1, N = 6) = 0.38, p = 0.54$). Table 8-9 shows the summary data for this test.

Table 8-8 Summary of the results for Experiment 2 that report on method suitability

Method type	% of low-level dependent variables that could not be tested	Method suitability high-level dependent variables			
		% of low-level dependent variables for which data could not be collected because of simulation resources limitations	% of low-level dependent variables for which data could not be collected because of data collection method limitations	% of low-level dependent variables for which data not be collected because of the number of data gathering opportunities.	% of low-level dependent variables for which data not be collected because of theoretical limitations
Constraint	81% (34/42)	14% (6/42)	43% (18/42)	5% (2/42)	17% (7/42)
WDA[17]	93%(14/15)	40% (6/15)	33% (5/15)	13% (2/15)	0% (0/15)
CTA	78%(21/27)	0% (0/27)	52% (14/27)	0% (0/27)	26% (7/27)
Task	84% (21/25)	0% (0/25)	56% (14/25)	0% (0/25)	28% (7/25)

Table 8-9 Number of low-level dependent variables that are sensitive and suitable

Method type	Number of low-level variables that were both sensitive and suitable	Number of low-level variables that were suitable and not sensitive
Constraint	1/2	1/2
WDA	1/1	0/1
CTA	0/1	1/1
Task	1/4	3/4

8.4.2.1 Are low-level dependent variables affected by simulation resources?

Percentage of low-level dependent variables for which the collection of data is limited by simulation resources is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because system models (for example, a property of the RWR system, a property of the world, and a property of the helicopter and mission) cannot be developed, divided by the total number of low-level dependent variables. Table 8-8 shows that the WDA method suggests six low-level dependent variables that could not be tested because of limitations associated with resources and with modelling the properties of the RWR and environment. These low-level dependent variables were:

- Tempo,
- Understanding,
- Decision making,
- Maximise time available to react,
- Minimise predictability and
- Crew physiology.

[17] It should be noted that the number of WDA variables that could not be tested is given as 14 but that the total number of the WDA variables (across all criteria) is 13. The variable, *Surprise*, has not been included in any of the criteria. *Surprise* is discussed in relation to the validity of the AH in Section 8.6.

The four low-level dependent variables, Decision making, Tempo, Understanding and Maximise time available to react, represent important internal processes of the RWR and could not be modelled with the resources available. Minimise predictability represents a property of threat systems and again could not be modelled with the resources available. Crew physiology data could not be collected because of resource limitations. There were no CTA and task-based measures that could not be modelled.

As with Experiment 1 the results illustrate the specific problem of evaluating complex systems in simulation environments. Ultimately evaluation is limited by what can be modelled and the resources available. Given that it was only the parts of the work domain specifically related to RWR and threat that could not be modelled, it can be concluded that this is a limitation of the WDA method.

8.4.2.2 Are low-level dependent variables affected by the data collection method used?

Percentage of low-level dependent variables for which the collection of data is limited by the data collection method used is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because the data collection method used to collect the data for the low-level dependent variable was not adequate (because of problems associated with gathering data points manually or automatically), divided by the total number of low-level dependent variables.

There were five WDA-based low-level dependent variables for which data could not be collected (Table 8-8). These were:

- Distance to enemy,
- Probability of detection,
- Number of threats,
- Exposure time to enemy and
- Timeliness.

Data could not be collected because of the interaction of the RWR display properties and data sampling rules. The display property means that once a threat is detected it is always shown on the display, i.e. it does not disappear from the display. The data sampling rules (described in Chapter 3) are designed to meet statistical requirements for data independence. Data independence was achieved in the experiments, and data collected, if the following two rules were met. First, a data point would only be collected if the source threat was encountered (i.e. displayed on the RWR) in the current frame (that is, one 60 Hz or 0.016 second portion of the data run) but was not encountered in the previous frame. Second, a data point would only be collected if the source threat had been seen in the previous frame but had a different threat mode in the current frame and was still the highest priority. Given that a threat was always displayed on the RWR (after the first time it was encountered) there was only ever one instance where the rules were met and a data point collected. In all subsequent frames the rules were broken and no data could be validly collected. There is no easy technical solution to this problem.

The lack of data for 13 low-level dependent variables common to the CTA and the task-based method is directly attributed to the data collection method used (Table 8-8). Nine of these low-level dependent variables are associated with measuring periods of time. These were:

- Time taken to defeat threat,
- Time taken detect threat (perceive threat icon),
- Time taken to locate threat in the environment,
- Time taken to initiate a movement to search for a threat in the environment,
- Time taken to deduce location of threat,
- Time taken to deduce range to threat,
- Time taken to select appropriate countermeasure,
- Time taken to expend countermeasure and
- Time taken to select communication system.

As with Experiment 1 it was impossible to identify accurately the start and end of cognitive and physical processes simply through observation. Given that the experiment emulated the functioning of a complex system in field conditions no method is suitable. It is exceptionally difficult to measure cognitive processes if one is not using a standard experimental paradigm in a laboratory setting. As was discussed in Experiment 1 the same point is made by Hennessy (1990) who argued that it is mistaken to try to collect (and analyse) some types of quantitative behavioural data using an experimental paradigm in an operational (simulator) setting.

Analysis of the remaining five low-level dependent variables revealed that the coding could not be performed reliably. These are:

- Choice of procedure -overall behaviour rating,
- Accuracy of detection of stimulus change over time - Display related,
- Accuracy in estimating distance of threat - Display related
- Accuracy in estimating position of threat – Display related, and
- Analysis of crew behaviour relevance -overall behaviour rating

The reason for this was because the RWR display displayed the threats continuously once they were initially detected. This meant that it was impossible for the author to judge whether the Aircraft Captain's behaviour was targeted at a specific threat (since it was likely that there was more than one threat displayed at any time).

On the surface this represents a limitation of the experiment: a limitation of the observation and data collection method. However, in general, determining what threat is being attending to using observation techniques alone is extremely difficult in simulation

experiments18. Currently, there is no method to "tag" a threat. Methods that could be used (for example, instructing the Aircraft Captain to identify what threat he is attending to) run the risk of compromising the ecological validity of simulation experiments and confounding the results for other variables.

8.4.2.3 Are low-level dependent variables affected by the number of data gathering opportunities?

Percentage of low-level dependent variables for which data is limited by the number of data gathering opportunities is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because the number of data collection opportunities was not sufficient to meet statistical protocols (for example, if the number of data points was less than 10 or if data were not present in all categories, if categorical data were collected), divided by the total number of low-level dependent variables. There are two WDA-based low-level dependent variables that could not be tested because the number of data points was limited by the number of data gathering opportunities in the experiment (Table 8-8). These were:

- Mission achieved, and
- Infantry equipment undamaged.

The reasons given in Experiment 1 for why the data were limited are relevant to Experiment 2 as well. They include the failure of the WDA method to account for the low number of trials typically seen in operational evaluation and the failure of the WDA method to account for characteristics of the variables, i.e. whether the variables are typical of a measure of performance or measure of effectiveness.

The CTA or task-based methods did not suggest any variables for which data were limited by data gathering opportunities.

8.4.2.4 Are low-level dependent variables affected by theory?

Percentage of low-level dependent variables for which data could not be collected because of theory is defined as the number of low-level dependent variables that could not be tested for statistical sensitivity because of theoretical constraints, divided by the total number of low-level dependent variables.

There were no WDA-based low-level dependent variables for which the underlying constraint- based perspective limited the collection of data. However, data for seven low-level dependent variables common to CTA and the task-based method were affected by underlying theory (Table 8-8). These are:

- Analysis of crew behaviour relevance -overall behaviour rating,
- Accuracy in identifying stimuli - Not display related,
- Accuracy in estimating parameters of threat - Not display related,

[18] Note, analysis of the FACELAB (eye tracking) data revealed that the FACELAB equipment could not provide the required level of resolution for the author to discern what the Aircraft Captain was looking at.

- Accuracy in response selection - Not display related,
- Accuracy in confirmation of threat - Flying pilot related,
- Choice of procedure - Flying pilot related, and
- Analysis of crew behaviour relevance - Flying pilot related.

These low-level dependent variables are limited by theory because data could not be collected for the behaviour that these variables represent. Data could not be collected because the behaviour was not observed in the experiment even though it was observed in the exploratory one and the conditions for the behaviour were present. This observation seems to imply that the reliance on using normative behaviour to guide the selection of dependent variables is a limitation because the behaviour is not guaranteed to occur.

8.5 Are the analytic products valid?

As with Experiment 1, the data from Experiment 2 can be used to establish the external validity of the analytic products. The external validity of the products can be calculated as the numbers of low-level dependent variables that have apparent validity, over the sum of the number of low-level dependent variables with some data plus the number of low-level dependent variables that were "incorrectly" suggested. For a low-level dependent variables to be "incorrectly" suggested it must be shown that the low-level dependent variables clearly did not represent an ecological property or task in the domain. To make comparisons between analytic products easier external validity may be expressed as a percentage. Hence, if the value tends to zero the products have no validity –ecological properties and tasks were incorrectly identified as being important to the Aircraft Captain. If the value is 100% then the products are valid - the products has correctly identified ecological properties and tasks (embodied as low-level dependent variables) that are important to the Aircraft Captain.

Table 8-10 shows the external validity scores and indicates that the constraint-based products have a combined score of 89%. The figure shows that the modified AH has a score of 88% and the TC-CTA has a score of 90%. The modified AH score reflects that fact that the variable Surprise was incorrectly suggested as being important to the Aircraft Captain. It was found that Surprise was incorrectly suggested because the future RWR always displayed the position of the threats accurately (unlike the current RWR that occasionally displayed the bearing of the threats erroneously, see Section 3.2). In the case of the TC-CTA analytical product the score of 90% reflects that one variable Control task frequency was incorrectly suggested as being important to the Aircraft Captain by the author. The table shows the task-based products correctly identified all the tasks that were observed in the experiment. Chi squared test reveals that there is no significant difference in the external validity of the analytic products ($X2$ (1, N = 26) = 0.96, p = 0.33).

It is concluded that in general the analytical products that the methods used were valid, but that improvements can be made through better selection of the ecological properties of the constraint-based variables.

Table 8-10 The external validity of the process and analytic products. The number of low-level dependent variables with apparent validity over the sum of the number of correct and incorrect low-level dependent variables expressed as a percentage for each analytical product.

Analytic products	External validity score
Constraint	89% (16/18)
AH	88% (7/8)
TC-CTA	90% (9/10)
Task (all)	100% (8/8)

8.6 Conclusion for Experiment 2

Experiment 2 was designed to test the predictive validity of the two methods (constraint-based and task-based) using two tests: the sensitivity of the variables to system modification and the suitability of the methods for use in operational evaluations for a future system.

The first test of predictive validity was achieved by evaluating whether the variables that the methods suggest are sensitive to the system modification, i.e. whether the measures showed a statistically significant difference between the two conditions (modified RWR and unmodified RWR). The results indicate that there was no significant difference between the two measure-selection methods; however, some of the constraint-based and some of the task-based low-level dependent variables are sensitive to the system modification. It is concluded that both the constraint-based and task-based methods show some degree predictive validity on this test.

The second test of predictive validity was achieved by assessing the methods against four high-level dependent variables. The assessment found that the task-based and the CTA-based method suggested low-level dependent variables that were not affected by resource limitations and data gathering opportunities but were affected by data collection methods and limitations from theory. It was also found that the WDA method suggested low-level dependent variables that were not affected by limitations from theory but the low-level dependent variables were affected by resource limitations, data collection methods and data gathering opportunities. It is concluded that the task-based method and the constraint-based method do not have predictive validity because they are both affected in some way by conditions commonly seen in operational settings.

These results from this experiment are important because they indicate that for a future complex socio-technical system:

- The task-based and constraint-based methods produce sensitive low-level dependent variables for operational system evaluation.
- The task-based and the CTA-based method for selecting low-level dependent variables are more suited than the WDA-based method to evaluate future complex socio-technical systems.

- To improve the constraint-based method for use during system evaluation it is suggested that:
- The constraint-based method, specifically the WDA method should be modified to better indicate how the properties of the operational context should be operationalised.
- The WDA method should be modified to identify important properties of low-level dependent variables and in particular identify whether the low-level dependent variables are measures of effectiveness or a measures of performance.
- The experiment results have indicated that the WDA method relies on system models that reflect the ecology of the work domain so that low-level dependent variables are identified that are sensitive to system modifications. This means that the cost of resources (financial and human resources) needed to use the WDA method in operational settings should be carefully considered.
- The CTA-based method should be modified to ensure that operationalisation of the properties of the control tasks in the simulator is performed correctly.

8.7 Comparing results of Experiment 1 and Experiment 2

In this section the results from Experiment 1 and Experiment 2 are compared. In the first part the methods are compared with respect to the sensitivity of the variables that they suggested. In the second part the methods are compared with respect to how suitable they are for system evaluation. The aim of this section is to identify any differences between the methods.

8.7.1 Comparing measure-selection methods on sensitivity

To be sensitive a variable must show a statistically significant difference between the unmodified and modified system. Table 8-11 shows the results from Experiment 1 and Experiment 2 in terms of the percentage of low-level dependent variables that were sensitive for each of the methods. In the paragraphs that follow the results concerning the task-based method are discussed first, then the results concerning the constraint-based method.

The results show that the task-based method does not produce sensitive low-level dependent variables for evaluating a current system. The results also show the task-based method does produce sensitive low-level dependent variables for evaluating a future system.

Table 8-11 The sensitivity of the variables that the methods suggest. Results from Experiments 1 and 2.

Method type	Experiment 1 (Current system): % of low-level dependent variables that are statistically sensitive	Experiment 2 (Future system): % of low-level dependent variables that are statistically sensitive
Constraint	12% (5/42)	5% (2/42)
WDA	33% (5/15)	7% (1/15)
CTA	0% (0/27)	4% (1/27)
Task	0% (0/25)	4% (1/25)

The results for Experiment 1 showed that the task-based method incorrectly suggested low-level dependent variables that should be sensitive to the system evaluation. The method suggested low-level dependent variables for which data could not be collected for either of two reasons: (1) the event necessary to trigger the behaviour that the variable was measuring was not observed even though the conditions for the event to occur were present and (2) the data collection methods were not suited for use in field evaluations.

The results for Experiment 2 indicated that out of all the variables that the task-based method suggested should be sensitive to the system evaluation, only one variable was sensitive. The results also indicated that, as with Experiment 1, the reason why variables that were not sensitive was because the events required to trigger the behaviours were not observed. Unlike Experiment 1 some variables could not be tested because of the interaction between the display properties of the system and the data sampling rules.

Turning now to the constraint-based method, the data show that the WDA-based method produces some sensitive low-level dependent variables for evaluating a current system but the CTA-based method does not (see Table 8-9). The WDA-based and the CTA-based methods each produced one sensitive low-level dependent variable for evaluating a future system.

In Experiment 1 the WDA-based variables correctly suggested five low-level dependent variables that that were found to be sensitive. The WDA-method incorrectly suggested low-level dependent variables that would be sensitive. By "incorrectly suggested" it is meant that (1) the method suggested some low-level dependent variables that are measures of effectiveness, rather than only suggesting variables that are measures of performance, and (2) some low-level dependent variables were suggested when the conditions used in that specific operational setting meant that they could never be sensitive, i.e. the resources were not available to model some of the RWR properties in sufficient detail for data to be collected yet the method did not take into account the impact that limited resources would have. In Experiment 1 no CTA-based low-level dependent variables were found to be sensitive, either because events that triggered them were not observed, or the data collection methods were not suited for use in the evaluations.

The results from Experiment 2 indicated that one WDA-based and one CTA-based low-level dependent variable were sensitive. Many low-level dependent variables were found to be not sensitive because of any of the following factors: the low number of data

gathering opportunities, the interaction between the display properties and the data sampling rules, simulation modelling and resource limitations, and limitations associated with the data collection methods used.

Comparing the results from the two experiments, it seems that the task-based method is not good for suggesting low-level dependent variables that are sensitive to a system modification for a current system because the variables suggested depend on known (predicted) system behaviour occurring; when that predicted behaviour does not occur, no data can be collected for the variable. This is less true for the WDA low-level dependent variables because the data collected is not related to instances of a specific type of behaviour. The task-based method and the low-level dependent variables it suggests rely on data collection methods that were originally designed for laboratory based system evaluation but are not suitable for field evaluations. The task-based method is also not good for suggesting low-level dependent variables for evaluating future systems, again because the variables that are suggested are dependent on known (predicted) system behaviour occurring. When that predicted behaviour does not occur no data can be collected for the variable.

From the results it seems that the constraint-based method is better than the task-based method for selecting sensitive low-level dependent variables for evaluating a current system. However, closer inspection reveals that that the WDA-based method incorrectly suggested some low-level dependent variables as being sensitive. They were incorrectly suggested because they had properties that identified them as measures of effectiveness rather then measures of performance (measures of effectiveness are generally insensitive to sub-system modifications, like the RWR modification) and therefore it is extremely unlikely that they would be sensitive.

It was also found that the WDA low-level dependent variables are more susceptible to simulation resource limitations than the task-based low-level dependent variables. The results suggest that the CTA-based method failed to produce sensitive low-level dependent variables because of the reliance on known system behaviour.

It seems that the constraint-based method, specially the WDA-based method correctly suggests more sensitive variables than the task-based method for evaluating current systems. It also seems that there is no difference in the sensitivity of the measures that the two methods suggest when evaluating future systems.

Looking at the results of the Experiment 1 and 2 as a whole, it is clear that the number of statistically sensitive low-level dependent variables is very low. Two possible reasons are given. The first reason may be that the comparison between the measure-selection methods was limited to a constraint-based method, which represented two of the five CWA phases, and which included some overlap with some task-based low-level dependent variables. It is possible that sensitive low-level dependent variables "exist" in the excluded CWA phases and in the remaining task-based low-level dependent variables. This point if further discussed in Chapter 9.

The second reason may be that the two methods are actually very poor at predicting sensitive low-level dependent variables. However, this can only be explored by modifying

the methods on the basis of the results of the experiment and retesting them. Again this point is further discussed in Chapter 9.

8.7.2 Comparing measure-selection methods on suitability

Table 8-12 shows the results from Experiment 1 and Experiment 2 in terms of the percentage of low-level dependent variables for which data could not be collected for each of the high-level dependent variables. In the paragraphs that follow, the results concerning the task-based method are discussed first, then the results concerning the constraint-based method.

The task-based method suggests low-level dependent variables that are not affected by resource limitations when used to evaluate a current system and suggests low-level dependent variables that are not affected by resource limitations and data collection methods when used to evaluate a future system.

Analysis of the results for Experiment 1 (Current system) reveals that the task–based method suggests low-level dependent variables that are affected by data collection methods, the number of data gathering opportunities and limitations of theory. In contrast to Experiment 1, the results for Experiment 2 (evaluation of a future system) show that the task-based method is not affected by resources available for modelling the system and the number of data gathering opportunities, but is affected by the data collection methods used and theoretical limitations.

Analysis of the results for the constraint-based measure-selection method reveals that for both current and future systems, the constraint-based method suggests low-level dependent variables that are affected by resource limitations, data collection methods, data gathering opportunities and theory. However, when the WDA method is used to suggest low-level dependent variables for the evaluating a current system, the low-level dependent variables are not affected by data collection methods and theoretical limitations but are affected by resources available for modelling the system and the number of data gathering opportunities. The results also reveal that the CTA-based method suggests low-level dependent variables that are not affected by simulation resources, and when it is used to evaluate a future system it suggests low-level dependent variables that are not affected by data collection methods and resource limitations.

From the results there seems to be a difference between the two methods and this depends on whether the methods are being used to evaluate a current or future system. The results show that the WDA-based method is better suited (less affected by the operational conditions) than the task-based and CTA-based methods when used for evaluating current systems. However, it also seems that the task-based and CTA-based methods are more suited than the WDA-based method when evaluating future complex systems.

Table 8-12 The suitability of the methods. Results from Experiment 1 and 2.

Experiment	Methods	Method suitability high-level dependent variables			
		% of low-level dependent variables for which data could not be collected because of simulation resource limitations	% of low-level dependent variables for which data could not be collected because of data collection method limitations	% of low-level dependent variables for which data could not be collected because of the number of data gathering opportunities.	% of low-level dependent variables for which data could not be collected because of theoretical limitations
Expt 1	Constraint	14% (6/42)	21% (9/42)	19% (8/42)	17% (7/42)
Current	WDA	40% (6/15)	0% (0/15)	20% (3/15)	0% (0/15)
RWR	CTA	0% (0/27)	33% (9/27)	19% (5/27)	26% (7/27)
	Task	0% (0/25)	36% (9/25)	20% (5/25)	28% (7/25)
	Constraint	14% (6/42)	43% (18/42)	5% (2/42)	17% (7/42)
Expt 2	WDA[19]	40% (6/15)	33% (5/15)	13% (2/15)	0% (0/15)
Future	CTA	0% (0/27)	52% (14/27)	0% (0/27)	26% (7/27)
RWR	Task	0% (0/25)	56% (14/25)	0% (0/25)	28% (7/25)

The fact that there seems to be a difference between the two methods and this depends on whether the methods are being used to evaluate a current or future system may be an indication that an uncontrolled variable is affecting the methods differently. If the data is considered it is clear that the number of WDA low-level dependent variables that are affected by data collection methods increase from Experiment 1 to Experiment 2 (0 to 5) and the task low-level dependent variables also increase (9 to 14). This change is also accompanied by a change in the number of task-based low-level dependent variables that are affected by data gathering opportunities (5 to 0). Closer inspection of all these low-level variables reveals that the changes between Experiment 1 and Experiment 2 are because data could not be collected in Experiment 2 because of the effect of the future system's display on data gathering rules. For example, in the case of the WDA low-level dependent variables, automated data (e.g. distance to threat) could not be collected because the display always showed the position of the threat, which meant that the statistical requirement for data to be independent was broken (see Section 7.2.3). For the task-based low-level dependent variables observation data (e.g. identifying where the Aircraft Captain was looking) could not be collected. As was discussed in Section 8.5.2.2, the impact of the future system's display on data collection is not an easy issue to resolve but one that should be addressed in the future before it can be concluded that "system display" is an uncontrolled variable.

[19] It should be noted that the number of WDA low-level dependent variables that could not be tested is given as 14 but that the total number of the WDA low-level dependent variables (across all high- level dependent variables) is 13. The low-level variable, *Surprise*, has not been included in any of the high- level variables. *Surprise* is discussed in relation to the validity of the AH in Section 8.6.

8.8 Conclusion

The small difference between the sensitivity of constraint-based and task-based measures seen in Experiment 1 is absent in Experiment 2. This chapter has also presented a comparison between the results from Experiment 1 and Experiment 2.

The results of that comparison revealed the following:

- For a current system the WDA-based method is seen to have more relative predictive validity than the task- and CTA- methods in terms of the number of sensitive low-level variables it predicts. It was also seen that the WDA-method seems to be more suitable than the CTA-method and task-method for use in operational settings but because all the methods suggest low-level variables that are affected by common operational conditions it is concluded that none have predictive validity.
- For a future system there is little difference in the number of sensitive low-level dependent variables that the methods suggest. There does seem to be a difference between the methods in terms of their suitability for use in operational settings. It seems that the task-based and CTA-based methods are more suitable than the WDA- based method for evaluating future complex systems. However, because all the methods suggest low-level variables that are affected by common operational conditions it is concluded that none have predictive validity.

The implications of these findings are discussed in the next chapter.

9. General Discussion and Conclusion

The aim of this thesis was to compare the predictive validity of two methods for selecting measures for use in the evaluation of current and future systems: the task-based method commonly seen in Human Engineering and the constraint-based method developed from Cognitive Work Analysis. The two methods were compared on two aspects of predictive validity. The first aspect was whether the measures suggested by each method were sensitive to a system modification. The second was whether the methods were suitable for use in operational settings.

The results suggested that for a modification to a current system the WDA-based method has slightly more relative predictive validity than the task- and CTA- methods in terms of the number of sensitive measures it predicts, but not significantly more. It was also seen that the WDA-method was more suitable than the CTA-method and task-method for use in operational settings because it was least affected by the conditions. However, this was not significant. It can be concluded that because both methods suggest measures that are affected by common operational conditions none have predictive validity.

For a future system there is little difference in the number of measures that the methods suggest. In terms of the suitability of the methods for use in operational conditions the

results hint that the task-based and CTA-based methods are more suitable than the WDA-based method for evaluating future complex systems. However, because all the methods suggest measures that are affected by common operational conditions it is concluded that none have predictive validity.

Closer analysis of the results from both experiments revealed that the measures sensitive to the system modification were the ones that operationalised the ecological properties of the domain or that reflected behaviours that the aircrew were required to perform. The results also revealed that measures that were not sensitive were linked to behaviours that were not observed. Although this seems obvious, it was less likely to observe behaviours for the measures suggested by the task-based and CTA-based method than for the measures suggested by the WDA-based method. Given that there was no fully-articulated method for deriving measures from Cognitive Work Analysis prior to this study, an outcome of this research has been to produce a method that has been initially tested for reliability and validity.

This chapter covers the findings of the program of research. In the following section a summary of the research is given and the original outcomes discussed. The theoretical implications are then presented. Finally, the limitations and areas for future research are discussed.

9.1 Summary of research

The conceptual framework that was followed in this thesis is shown in Figure 9-1. Each of the stages shown will be briefly reviewed and important points highlighted. Some of those points will then be discussed in more detail in the sections that follow this one.
Stage 1 (see Chapter 2) of the process was to review of the literature with the aim of identifying the theoretical foundations and methods that have been used by the Human Engineering (HE) and Cognitive Work Analysis (CWA) communities to evaluate complex socio-technical systems in operational settings. It was found that the Human Engineering method was based on the analysis of tasks and the application of guidelines to select measures that could be used for system evaluation and was widely used. It was also found that methods that the CWA community used were based on the idea that analysing the constraints acting on the system, rather than tasks, would result in better system evaluation outcomes, especially for future systems. However, there was no method specified that could be used in operational settings. The outcome of Stage 1 was to identify the following.

- A comparison of the HE method and a CWA method was needed and this comparison would be particularly beneficial to analysts involved in complex system procurement activities.
- A CWA method was needed for the comparison as a HE one existed.
- A complex socio-technical system should be evaluated.
- The methods compared should represent activities typical of analysts involved in complex system procurement activities.

- The methods should be compared on two important aspects, whether they produce measures that are sensitive to system modifications and whether the methods were actually suitable for use in operational settings. These aspects were seen to be indicators of the methods' predictive validity.

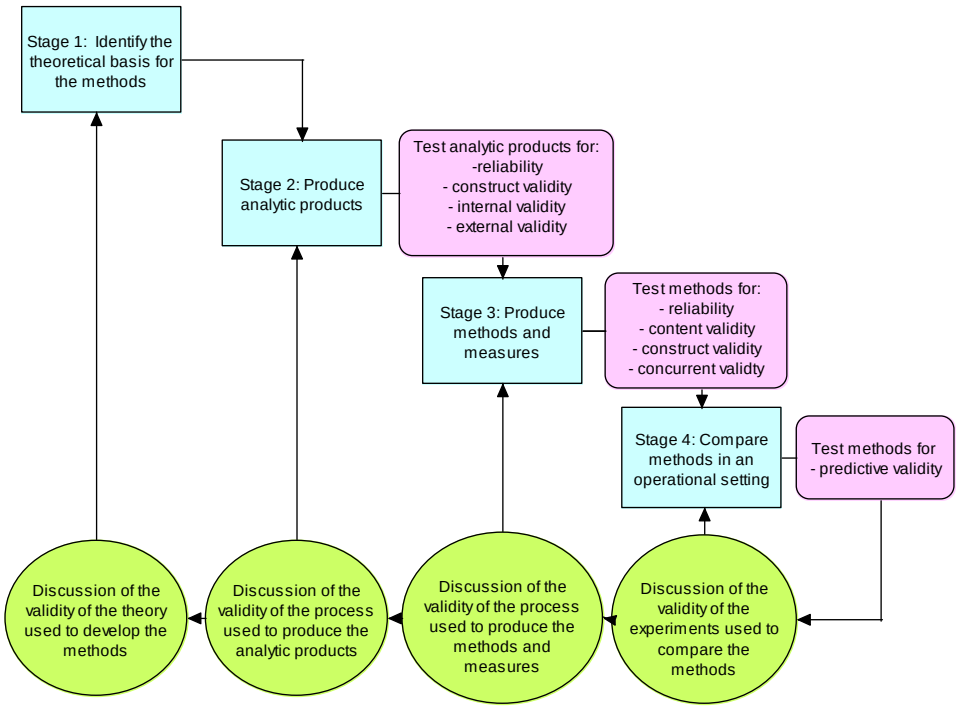


Figure 9-1 The four-stage process undertaken in this thesis to compare the HE (task-based) and CWA (constraint-based) methods for selecting measures to evaluate the impact of system changes

Given the outcomes of Stage 1 an integrated research program was designed (shown in Figure 9-1 and described in Chapter 4) to compare the two methods. The research design that was developed included the following.

- The use of a single system test case (a radar warning receiver; RWR) (see Chapter 3).
- The development of HE (task-based) and CWA (constraint-based) analytic products (Stage 2 – see Chapter 5).
- The development of the task- and constraint-based methods for extracting measures (Stage 3 – see Chapter 6).
- Testing the measures selected via empirical experiment (Stage 4 – see Chapter 7 and 8).

In Stage 2 (Chapter 5) the task-based and constraint-based analytic products were identified and developed. Producing reliable and valid analytic products for the two approaches was a critical step in the comparison because the analytic products would be used in the methods that follow. Important points from this stage include the following.

- The products were selected on the basis of previous research.
- They were constructed using established practice.
- They were tested for internal validity by the author and independent SMEs from the CWA and HE communities (both national and international).
- They were tested for construct validity by the author and a CWA expert.
- They were tested for external validity by the CWA expert.
- A process of assurance was used to meet some of the requirements for reliability.

The outcome of assessing the products by the experts was that they were found to be valid.

In Stage 3 (Chapter 5) the methods were produced, assessed for reliability and validity and the measures selected. The measures were also assessed for validity. Important points from this stage include the following.

- An original constraint-based method was produced that represented the Work Domain Analysis and Control Task Analysis phases of Cognitive Work Analysis. The Strategies Analysis, Social Organisation Analysis and Worker Competencies Analysis were not represented.
- A task-based method was produced that was based on HE standards and NATO committee recommendations and represented world's best practice.
- The task-based method was tested for content validity by the author and independent SMEs from the HE community (both national and international) - the method was found to be valid on this dimension.
- The task-based method was tested for construct validity by the author. The method was not found to be valid because no firm theoretical basis for it was found. The constraint-based method was also tested and found to be valid because it had a strong theoretical foundation that was recognised by the CWA community.
- The task-based method was tested for concurrent validity by independent SMEs from the HE community (both national and international). The method was found to be valid on this dimension. The constraint-based method was not found to be valid by the author because it was original; there are no similar CWA methods to compare it to.
- It was decided that predictive validity of the methods could be achieved by experiment.
- A process of assurance was used to meet some of the requirements for reliability.
- The task-based measures were assessed for validity (representativeness) by the independent SMEs from the HE community (both national and international) and found to be valid.

In Stage 4 (chapters 7 and 8) the methods were tested. In Chapter 7, (Experiment 1) the sensitivity of the task-based set of measures to a modification in a current RWR was compared with the sensitivity of the constraint-based set of measures to the same change.

It was also important to see whether the methods were suitable for use in an advanced simulation setting for a system that was familiar to the aircrew- a current system. The results to Experiment 1 revealed the following.

- There is no statistical difference between the task-based and constraint-based measure-selection methods in terms of the sensitivity of the measures that they suggest (X2 (1, N = 67) = 3.22, p = 0.07) although there is a trend in that direction.
- Some measures produced by Work Domain Analysis, were sensitive to the modification of the current RWR system, i.e. the WDA method showed some predictive validity.
- No measures produced by the task-based and Control Task Analysis were sensitive, i.e. the methods did not show predictive validity.
- There was no statistical difference between the two measure-selection methods in terms of their suitability for use in the operational setting used in the experiment (X2 (1, N = 67) = 1.36, p = 0.24).
- The task-based method suggested measures that were not affected by simulation resource limitations and the WDA-based method suggested measures that were not affected by the data collection methods used and theory.
- WDA, CTA and task-based method suggested measures that were not affected by the number of data gathering opportunities.
- In terms of the predictive validity of the measure-selection methods no method showed predictive validity because they were all affected by some of the common operational conditions.

The results of Experiment 2 suggest the following.

- There is no statistical difference between the task-based and constraint-based measure-selection methods in terms of the sensitivity of the measures that they suggest (X2 (1, N = 67) = 0.02, p = 0.88).
- The constraint-based method and the task-based methods have a similar but low degree of predictive validity.
- There was no statistical difference between the two measure-selection methods in terms of the suitability for use in the operational setting used in the experiment (X2 (1, N=67) = 0.01, p =0.75.).
- The WDA method suggested measures that were affected by simulation resource limitations, data collection method limitations and the number of data gathering opportunities.
- The WDA method suggested measures that were not affected by limitations of theory.
- That the CTA and task method suggested measures that were affected by data collection methods and theory but not affected by simulation resource limitations and the number of data gathering opportunities.

- In terms of the predictive validity of the measure-selection methods no method showed predictive validity.

9.2 Significant and original outcomes of research

There were several original outcomes of this research which are given in the sections that follow.

9.2.1 Comparison of methods for system evaluation

Although some research has compared task-based and constraint-based methods in general (for example, Hoffman and Militello, 2009; Hajdukiewicz and Vicente, 2004; Miller and Vicente, 1999, 1998), no research has specifically compared them for their ability to produce sensitive measures of performance and effectiveness and for their suitability for use in operational settings. This program of research is the first time that measure-selection methods have been empirically evaluated for both a current and a future system.

This evaluation is important for a number of reasons. First, the evaluation provides a clear alternative to the task-based method for selecting measures. Second, the evaluation extends knowledge about the similarities and differences between task-based and constraint-based methods. For example, many of the measures suggested by CTA are the same as those suggested by the task-based method whereas the WDA measures are unique. Third, the research has identified factors that affect the suitability of the task-based and constraint-based methods, such as the resources that are available to model the test system, the data collection methods used, the number of data gathering opportunities and the theory that the methods are based on.

9.2.2 Development of WDA-CTA measures framework

The work reported in this thesis has also provided an original framework to describe the relationship between various measures derived from Work Domain Analysis and Control Task Analysis. The framework is particularly suited to operational evaluation environments. Although Vicente (1999) presented a theoretical approach for identifying classes of variables from CWA analytic products (e.g. a WDA should identify system state variables) the work in this thesis provides a method of selecting measures within those classes. The framework does this by mapping WDA and CTA properties onto specific measures that are useful for system evaluators.

The framework is important for a number of reasons, even though the present thesis does not yet offer strong empirical evidence for its effectiveness. First, it provides a means to represent or model the implications of a change within the work domain (as represented by WDA) or activity domain (as represented by CTA) that may result from a system change. For example, as I have shown in this thesis, measures from the work domain and activity domain are sensitive to a system modification. Second, the framework can be used to represent or model the implications of a change between the work domain and activity domain. Third, the framework provides a structure on which the measures from the remaining CWA phases (SOA, SA and WCA) could be "attached" to form an integrated

network of measures that cover all the system from all points of view. Fourth, every relationship between measures in the framework is testable making the validity of the framework itself testable. Fifth, and related to the last point, by using the framework together with the AH and TC-CTA, analysts can identify and select measures that are important for system evaluation. In other words the framework provides a theoretical grounding (construct validity) for the selection of measures a priori, and stands in contrast to guideline-based approaches. Sixth, the framework is generalisable, in principle, to any complex operational socio-technical system in actual operation. Given these advantages, further development and empirical testing of the framework seems warranted.

9.2.3 Development of constraint-based method

The constraint-based method presented in this thesis for selecting measures derived from CWA is unique. An initial method was presented and tested in Experiment 1 and Experiment 2. In this section I will briefly argue that the method should be developed further and then list a number of recommendations.

The development of the method is important for three main reasons. First, it shows that ecological properties of the world can be measured in operational human-machine (intentional) systems. This is an extension of the work of Vicente (1999), Xinyao, Lau, Vicente and Carter (2002), who predominantly assessed laboratory-based microworld (causal) systems.

Second, the method focuses attention on the problem of evaluation. As was discussed in Chapter 2 most researchers in the field have not focused on evaluation, but instead on applications of CWA to system design in general.

Third, and despite the limitations outlined in another section, a constraint-based method has been presented that can be used during system procurement. The results of the experiments have indicated that novel measures for system evaluation may be obtained. It is notable that these measures were not identified using the task-based method.

On the basis of the results from the experiments, four recommendations were made for improvements to the constraint-based method. Those recommendations have been incorporated and shown in Figure 9-2 and Figure 9-3. The figures show the WDA based method and the CTA based method that together form the constraint-based method. The areas where the modifications are implemented are highlighted. The four main recommendations are now discussed.

The first recommendation (R1) is that the WDA method should be modified to identify the importance of operationalising the properties of the simulation environment. In the experiments it was shown that some of the WDA measures were not operationalised in terms that reflected the ecological properties of the work domain and therefore were found to be not sensitive to the system modification. The use of SMEs to identify the ecological properties of the work domain and then suggest ways to operationalise those properties in terms that can be measured should mitigate this problem.

The second recommendation (R2) is that the WDA method should be modified to identify important properties of measures and, in particular, to identify whether the measure is a measure of effectiveness (MOE) or a measure of performance (MOP). As has been noted, a measure of performance is a measure that reflects the result of a test and is related to hardware, software and human characteristics such as probability of detection, false alarm rates and human reaction times. A measure of effectiveness is a measure that is used to reflect how well a human or system meets a criterion. Although the WDA-CTA measures framework clearly identifies the differences between MOEs and MOPs this distinction has not been integrated into the method. By providing guidelines for the identification of MOPs and MOEs measures should be selected that are sensitive. For example, the measure Mission achieved was suggested as being sensitive by the WDA method but found not to be sensitive by experiment. If the guideline "only choose measures that are operationalisations of the physical properties of the subsystem" was included in the method, Mission achieved (clearly not operationalisation of a physical property of the subsystem) would not have been selected.

The third recommendation (R3) is that the method should be modified to take into account the importance of estimating the financial and human resources needed to model complex systems. The results of the experiment have indicated that the WDA method relies on system models that reflect the ecology of the work domain to identify variables that are sensitive to system modifications. This means that the cost of resources (financial and human resources) necessary for use of the WDA method in operational settings should be carefully considered.

The fourth recommendation (R4) is that the CTA-based method should be modified to ensure that operationalisation of the properties of the control tasks is performed correctly. As with the WDA-based method it is recommended that SMEs to be used to identify the ecological properties of the activity domain and then suggest ways to operationalise those properties in terms that can be measured.

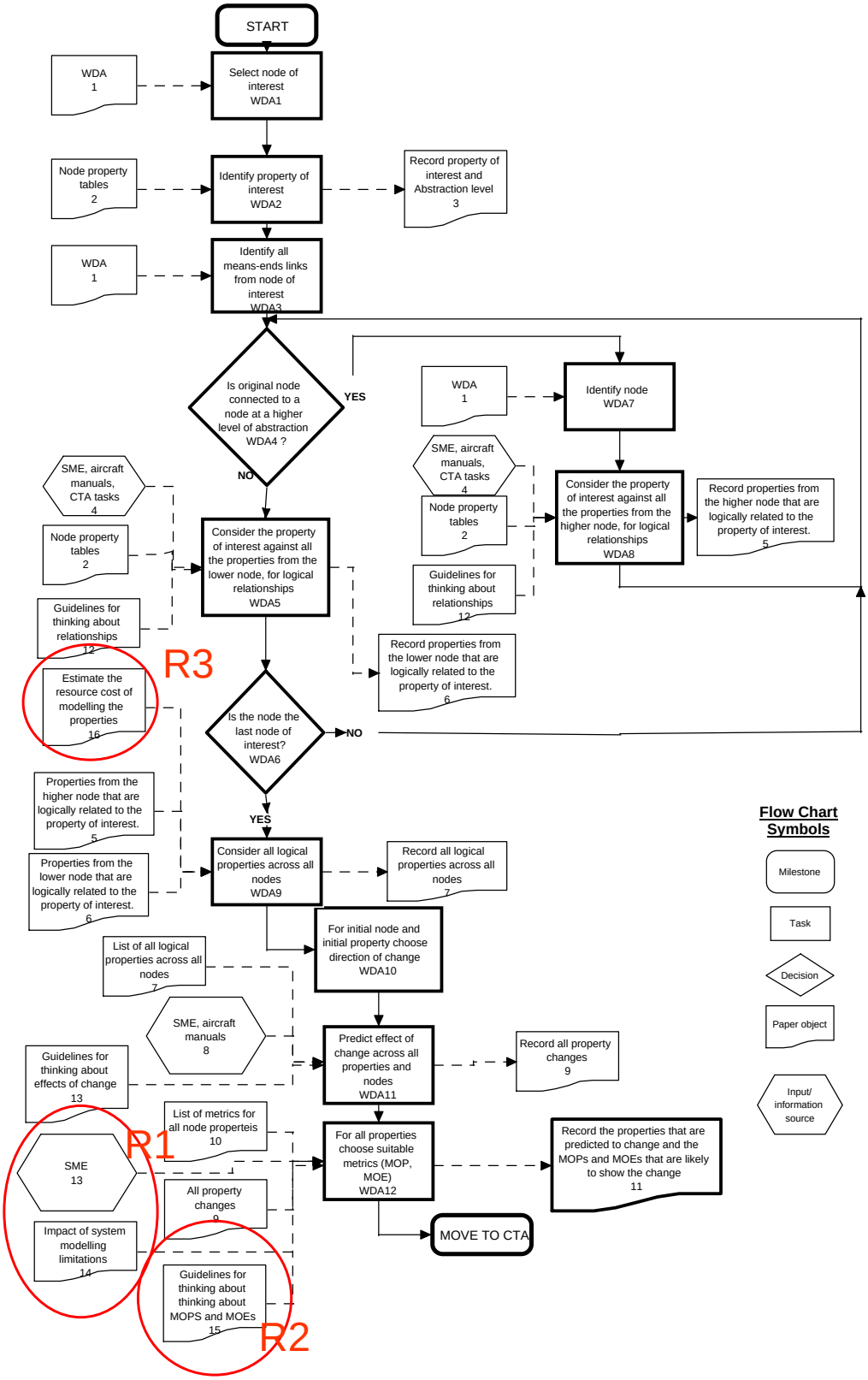


Figure 9-2 WDA- based method

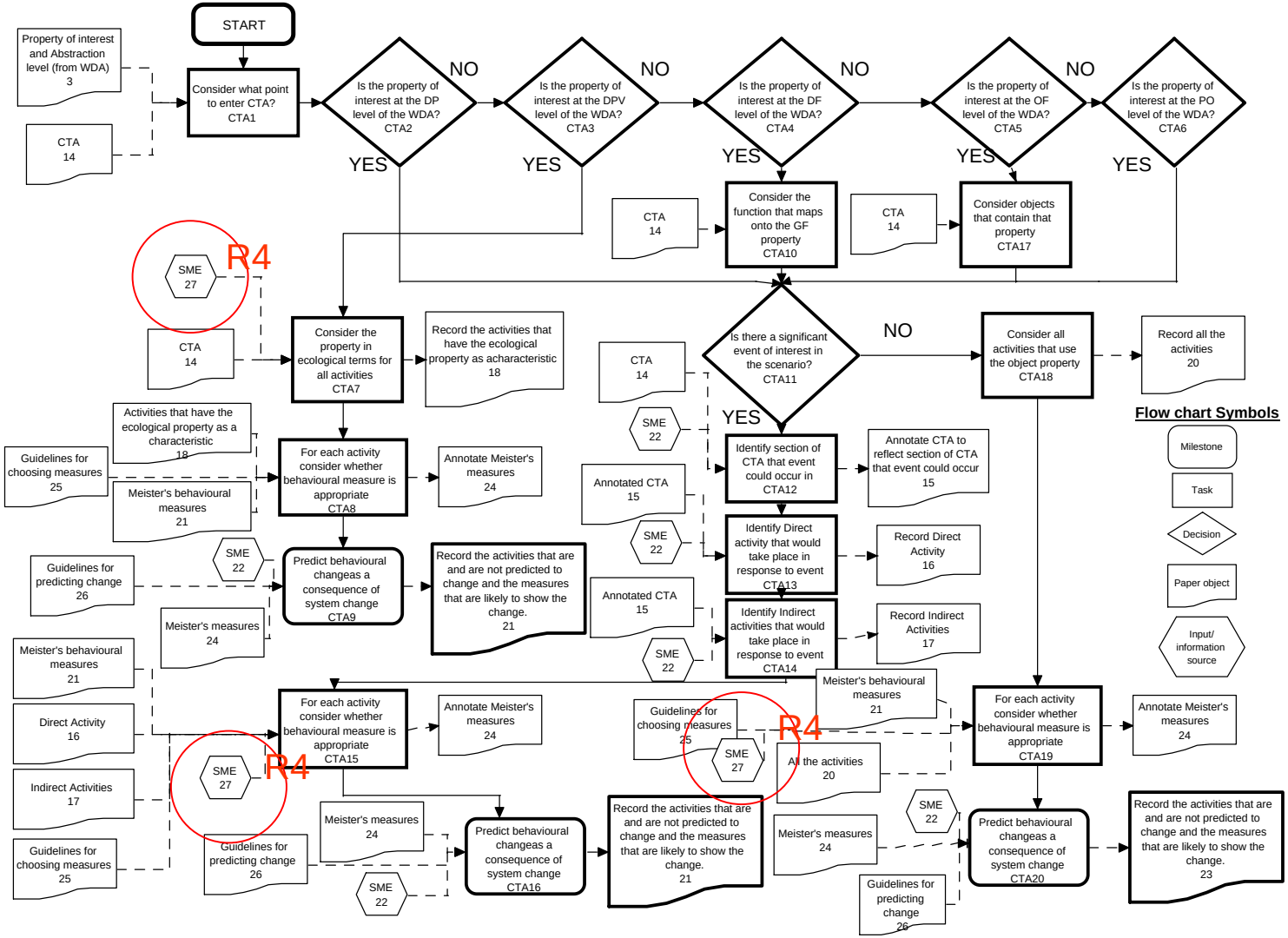


Figure 9-3 CTA based method

9.2.4 Contributions to system evaluation

In this section I will briefly outline the contributions that this thesis makes to system evaluation in general.

9.2.4.1 Properties of candidate measures are important

When planning to evaluate a system it is important to consider how different measures will work in the evaluation context. The results of the experiments indicate that some measures suggested by the methods were not suited for evaluating complex socio-technical systems in complex operational environments. Rather than being a question of whether the measure was suggested by the task-based or constraint- based method, the main issue was whether the measure had properties that made it valid in the operational test environment. Hence, some of these measures appear to have inherent properties that preclude them from simulation-based studies, preventing them from helping to deliver procurement advice on the basis of simulator studies. For example, the "time" variables required the author to accurately identify the start and finish of a cognitive process, which was impossible in the simulator but which may be possible under more tightly constrained laboratory conditions.

9.2.4.2 Future system is not necessarily more complex than current

The results from the two experiments provided some initial results to suggest that thinking about systems in terms of "current" and "future" may not be useful. The terms were used here to indicate whether the functions provided by a system are familiar to the aircrew (a current system) or unfamiliar (a future system). In the experiments, however, the future system had a display that provided a view of the world that was simpler than the current system. It seemed that the current system's display was more complex than the future system display. It was more complex because it provided a greater degree of what Vicente (1999) terms "mediated interaction". This means that users of the current system had to bring to bear significant cognitive resources to make sense of what was being viewed – more resources than the users of the future system had to use. The implications of this is that a system should not be defined in terms of just the familiarity of the operators with the system functions, but also how those functions are presented to the operators. From the point of view of evaluating a system the choice of which method should be used to select sensitive measures may be dependent on more than the functions the system possesses.

9.2.4.3 Categories of measures

An important contribution to system evaluation, that supported previous work, was that the WDA developed for the constraint-based method identified different types of measures from different levels of abstraction. This means that an analyst evaluating a system at the object level of abstraction, for example, would be guided to the types of measures at that are relevant to that level, rather than perhaps choosing a measure that has no relevance.

9.3 Theoretical implication: Let theory guide measure selection

This thesis has compared two methods for selecting measures for evaluating a system: the first a constraint-based method that is theoretically grounded and the second a task-based method that is based on the application of guidelines. This provides us with a "case study" that helps us judge whether using theory is suitable for selecting measures.

Kantowitz (1992) provides us with a useful starting point for this discussion. In his paper he identifies several benefits of using theory. Each of these is given below in italics, and is then considered.

Theory allows for interpolation of results where data cannot be collected, or where limited data points can be gathered. By having theory, predictions about how a variable will affect results can be made. The development of the WDA-CTA framework presented earlier in this thesis was a necessary step for the production of the constraint-based methods. Without the framework the methods could not be developed. As was seen, the methods provided a mechanism for the analyst to consider the possible effects of the system modification on the measures. With the task-based method, where a theory was not presented, measures were selected by independent practitioners and confirmed as being suitable for use that eventually could not be used. Hence, using guidelines offered no reliable indication of where data should be collected. Additionally, it was seen that WDA-CTA framework offered a means to play "what if" games concerning the areas of probable impact of system modification.

Predictions based on theory can be used in the design of a system, before the system is built. Clearly a real system was not built. However, the program of research emulated the system design process in the sense that a system was proposed (the unmodified RWR), a modification proposed (increase RWR sensitivity), the effect of the modification predicted (variables were identified that should be sensitive to the modification) and the modification tested (changes in the variables were statistically tested).

In the case of the evaluation of the current system, albeit not the future system, the constraint-based method not only identified measures that could and sometimes did show the effect of the system modification but the measures that were identified were predicted before the system was "built". In other words, the method predicted some specific measures that would be affected by the system modification before the system was tested. The task-based method on the other hand, with its dependence on guidelines and normative system behaviour, did not predict what variables would be affected.

Theory can be used to aid measurement and system design. The research confirmed that theory can be used to aid measurement and system design. It was shown that the WDA-CTA framework and constraint-based method provided a clear process linking WDA objects, functions, values and priorities and purpose to specific properties and then to specific measures; and linking CTA activities to task-specific or task-generic measures. In this way it was possible to operationalise the constraint-based properties as measures. In the case of the task-based method there is no direct link between behaviour and the specific measures. The use of theory provided a clear aid to measurement.

Theory can be used as a means for representing normative human behaviour or system performance. This benefit raises several issues. First, I would argue that there is no valid theory for normative behaviour – the measures were chosen on the basis of guidelines and this benefit is not applicable. Second, the real benefit in using theory is for representing formative behaviour. The constraint-based framework and method are independent of events that specify normative behaviour. The predictive power of the constraint-based framework comes from its ability to handle "non-normal" behaviour. For example, in the experiments the constraint-based method suggested the WDA measure exposure time to the enemy as being sensitive to the system's modification. Although the measure was not found to be statistically sensitive (because of the low numbers of data points that could be used in the statistical test), it was found to reflect the tactics that the aircrew were employing to threats. The tactics involved "dashing" (flying fast in the helicopter) across open land in view of a threat. The aircrew were able to do this safely because the threat was too far away to get a "shot off' within the time it took them to cover the open ground. They did not perform this behaviour when the threat was close to them. Purposely exposing the helicopter to a threat is non-normal behaviour. The results of the analysis of the data for exposure time to the enemy suggested that the aircrew performed this behaviour more often with the modified RWR (it could detect the threat far away) than with the unmodified RWR (it could only detect threat close by) and this was confirmed through interview with the aircrew. Hence, exposure time to the enemy was "sensitive" to the non-normal behaviour "dashing". In the case of the task-based measures none were statistically sensitive to the RWR modification and none suggested any relationship between "dashing" and whether the RWR was modified or not. For example, a typical measure, broadly equivalent to the WDA measure in terms of representing tactics, is Choice of procedure – overall behaviour rating. If this measure is considered it seems that it was not sensitive to the tactic because the aircrew always responded that their action was correct irrespective of the RWR condition. The measure cannot distinguish between "dashing" with a modified RWR and not "dashing" with an unmodified RWR.

9.4 Limitations

There are four areas that limit the results from this program of work. The first limitation is that although the reliability of the constraint-based analytic products and methods presented is "assured" by the presence of a well-defined and replicable process, a more powerful assessment would be to test reliability formally. Second, the strategy used for testing the predictive validity of the methods needs modification. Third, the number of participants used in the experiments was limited. Fourth, only two out of the five CWA phases were incorporated into the constraint-based method. Each of these will now be discussed.

First, Chapter 4 describes the concepts of reliability and validity as they apply to the analytic products. The chapter makes it clear that in this program of research I sought to assess the reliability of the analytic products by using material from multiple sources and ensured that the steps taken in the production of the analytic products were auditable and replicable. Although this might be an acceptable approach, a better approach would have been to use independent assessors to produce the products and then calculate inter-rater reliability. Chapter 4 also describes the concepts of reliability and validity as they apply to

the measure-selection methods. The chapter shows that the reliability of the methods was assured rather than formally tested and that the validity was assessed on a number of dimensions by the author and by independent national and international SMEs. The measure-selection methods were then followed by the author to select the measures. The task-based measures were then assessed for validity by the national and international SMEs. Again, although this is an acceptable approach a better approach would have been to have independent assessors (who are blinded to the purpose of the study) use the methods to produce the measures. The measures produced by the independent assessors then could have been compared to the measures that the author produced. The use of independent assessors in this way has the added benefit of removing potential bias arising from the author's previous experience with the task-based and constraint-based perspectives to system evaluation.

Second, the strategy that was described in Chapter 4 for testing the predictive validity of the measures could be improved. In this program of research, the measures that were tested were the ones that the methods suggested would be sensitive to the system modification. Although this is an acceptable approach, a more powerful approach would be also to test measures that were suggested by the method as not being sensitive. If the measures suggested as sensitive were found to be sensitive and the converse, then I would be in a very strong position to claim that the method was valid. However, in order to perform such a test I would have to identify all possible measures for all objects, functions, priorities and values, and purposes. Given the size of the domain of interest this is an impossible task. An alternative approach would be to identify a group of measures that should not be sensitive and test them for sensitivity. If the measures were found to be not sensitive, but were demonstrably sensitive in other contexts, I would be in a better position to claim the products and methods as valid. Related to this is the idea that the transcription data could help us identify measures that have not been suggested by the methods. The existing research method identified the importance of using the transcription material to test the variables for apparent validity, but content analysis could help us identify further measures. If additional measures were found I could say that the method did not capture all the measures and therefore needed refining.

Third, another limitation was the use of single aircrew. Chapter 4 has made it clear that I was limited to conducting the experiments with single aircrew because I needed to examine experienced aircrews that had experience of RWRs, and I was restricted by the resources available for the program of research. Although the strategy adopted focussed on collecting a high number of data points (where applicable) and so had good internal validity, it is noted that the experiments are limited in terms of external validity. The use of a small number of aircrew may have consequences in terms of the number of measures found sensitive. It may be that by testing the methods with more aircrew the number of constraint-based measures found to be sensitive would not change. This is because a change in the number of aircrew would not in itself affect the ecology of the domain. However, it may also be the case that if the methods were tested with more aircrew the number of task-based measures that would be found sensitive could change. This is because the range of aircrew behaviours observed may change. Of course this is only one proposition, further testing would be needed.

The final limitation is the fact that the constraint-based method only included two of the five CWA phases and this limited the task-based measures that could be compared. The choice of which of the five phases should be used was based on previous research. That research showed that WDA and CTA were most suitable for system evaluation activities and were the ones that had been used most frequently. On the basis of that research the WDA-CTA framework and the constraint-based measure-selection method were developed. Now that the constraint-based method has been initially tested it is important to develop a method that includes all the CWA phases so that all the task-based measures can be compared. Without that comparison it is a possibility that those excluded measures are sensitive to the system modification.

9.5 Further research

Five main areas of future research are identified. The first focuses on generalising the results to other systems. The second focuses on extending the CWA method to include the other CWA phases so that all the HE and CWA measures can be captured and compared. The third area focuses on extending the CWA framework to take into account possible time and state changes in the domain of interest. The fourth area focuses on other system evaluation perspectives. Finally, the fifth area focuses on practical ways of reducing the time needed to select the measures. In the following paragraphs each of the areas for future research will be discussed.

First, pathways to generalisation should be explored. In this thesis the measure-selection methods have been used to select measures for a single test case. The test case was a modification to a RWR (changing the range over which the RWR could detect a threat) that was fitted to a helicopter operating an Airmobile (Patrol Insertion) mission. It is possible that the range detection property was not sufficiently strong or appropriate to draw out the differences between the methods. However, it could be claimed that this is exactly the sort of system modification that is made and so the results are valid. Until the methods are tested using another system, that proposition cannot be verified.

In a more general sense, from a theoretical perspective the constraint-based measure-selection method should apply to other complex intentional domains. The method described in the thesis was designed to be independent of any particular domain. I used the test system to provide an initial opportunity for testing the method. However, until research using other domains is conducted the idea that the method is applicable to other domains remains untested.

Second, further aspects of the CWA framework should be brought to bear on the problem. Vicente (1999) noted that WDA could be used to identify system state variables, CTA could be used to identify variables that identify what people do, SOA could be used to identify variables that can be used to study the coordination across multiple actors, and WCA could be used to identify variables that can be used to study worker competencies in specific conditions. Clearly, the constraint-based method presented in this thesis considered only WDA and CTA. By focussing on these phases a large group of HE (task-based) measures were excluded from the comparison. For a complete comparison between HE and CWA methods, all the CWA phases should be considered.

Third, in this use of the constraint-based method and the subsequent analysis of the experimental data, it is assumed that the measures identified are independent of time or the state of the work domain. Further research should be directed at investigating whether the sensitivity of particular measures are related to time or system state, or both.

Fourth, Chapter 2 presented the case that the two dominant approaches to evaluating complex socio-technical systems are the task-based perspective and the constraint-based perspective. Although this is true, there are other perspectives that are used in the evaluation of systems such as Naturalistic Decision Making, Cognitive System Engineering and Ethnography. In general, although all of these perspectives use the same data collection methods as the task-based and constraint-based methods, the lens through which the data are viewed may reveal unique measures that are important for complex system evaluation. For example, data collected during a system evaluation exercise may be viewed from a task-based lens and point to system performance issues that are attributable to a specific task. For example, the task was not performed correctly by an operator. The same data may be viewed from a constraint-based perspective and reveal an issue attributable to constraints acting on the system. For example, the same operator may be reacting to a unique (previously unidentified) state change of a property in the domain. The same data may be viewed from an ethnographic perspective and reveal an issue to do with organisational pressures placed on the system. For example, the same operator may have been under organisational pressures to attend to other tasks. This example illustrates that comparisons of these and other perspectives is needed to fully understand the subtleties associated with evaluating complex systems and should be considered in future research.

Finally, practical ways of reducing the time it takes to produce the constraint-based measures should be investigated so that they can be used during a system evaluation activity. Whilst it is speculated that the constraint-based measures can be reused for other systems and that these systems should be evaluated in the same domain as the original system anecdotal evidence, gained during the research program, suggests that most of the time needed to use the constraint-based method was taken in producing the analytic products on which the method acted rather than in deriving the measures. From a practical perspective, tools and techniques should be developed to reduce the time taken to produce the analytic products.

9.6 Future vision for community of practice of using a constraint-based perspective

The problem of how to test complex socio-technical systems in operational settings will not be quickly resolved. This thesis provides a small insight into just how the complex that problem is. In the future, socio-technical systems will become more complex; they will be evaluated in operational settings, and analysts will want evaluation to be no more complicated than it is now. In the next few paragraphs I describe some observations I have made throughout this research program that provide a glimpse of a possible future. I group my observations under a number of statements.

In the future it will be just as easy to conduct a constraint-based analysis of a system as it is to conduct a task-based analysis. One of my first observations in this program of research was that it was easier to describe the work domain in task-based terms than constraint-based terms. First, one has an everyday "understanding" of what a task is. Second, previous researchers had detailed exactly what should information should be used about tasks, how to acquire that information and then how to use that information. For the constraint-based description of the work domain it was less clear what a "constraint" was and there were subtle differences in the types of constraints identified by CWA. There was also less explicit guidance on what information to gather, how to gather it, and how to use it. I hope that difficulty will change as research programs such as the one described in this thesis make the term "constraint" easier to understand by describing it the context of evaluating an operational system with which practitioners are familiar. It certainly became easier to identify constraints over the course of this research.

In the future we will view work domains with different "lenses". Currently, researchers involved in system evaluation tend to view work domains through one lens, the task-based lens, but there is another lens, the constraint-based lens. The research described in this thesis has described some problems and benefits with both perspectives. Given that there are some shortcomings of the task-based perspective, we should look for ways of mitigating those shortcomings. In the future we will be armed with knowledge that will help us view work domains from both viewpoints.

In future the constraint-based perspective will inform the system evaluation activity. The constraint-based perspective will inform system evaluation in three areas: (1) the design and specification of the simulation, (2) the achievement of scenario equivalence, and (3) data handling.

First, the constraint-based perspective may change how simulators are used. The work domain analysis (in the form of the AH and ADS) that was produced helped to specify the objects and their functions to the system engineers who configured the simulator. By using the AH and ADS as a definition of the requirements for the simulator, the author and the engineers discussed the specifications for the simulator configuration and capabilities. During this process, we used the actual AH and ADS analytic products to describe what the impact would be of having or not having an object in the simulated world (both in terms of the resources, cost and human resources, and impact on other parts of the simulation). By using the CTA it was possible to describe to the engineers what activities would be affected by including or removing an object in the simulator. For example, having trees of a specified height and type in the simulator was necessary because it could be shown that the type of tree indicated to the aircrew an approximate height from the ground, which was necessary for low-level flight.

Second, scenarios need to have some level of equivalence if contrasts are to be performed. The understanding that the work domain could be represented in ecological terms was important for designing mission scenarios that were equivalent at some level. While investigating what an "ecological" view of the work domain meant and how that could be represented, I discovered that the aircrew viewed their work domain (the RWR Black Hawk Airmobile operations) as a series of "survivability rings" (Crone et al, 2007), that they "moved between" during a mission. For example, they would spend much of their

time in the "don't be seen ring" and other time in the "don't be hit" ring. I was able to design scenarios that "forced" aircrew into various rings, so that I could count the instances when they were in a ring. This made it possible to judge whether a scenario was equivalent not solely by the number of threats present (although this was balanced) but also by the number of times during a mission the aircrew "visited" each ring.
Finally, the constraint-based approach affected data handling. Adopting a constraint-based perspective to system evaluation necessitated much larger data storage and more powerful data manipulation protocols that had been previously been used in task-based evaluations. For example, during a typical mission flight in excess of 200 variables were collected at a rate of 60Hz for a period of 40 minutes. The large amount of data required had an impact on storage and also on manipulation to be used by desk-top statistical packages.

9.7 Conclusions

Analysts involved in system evaluation activities are faced with the problem of how to select measures that should be used to evaluate a system. The research reported in this thesis has compared two methods: a Human Engineering task-based method was compared to a novel Cognitive Work Analysis constraint-based method. The comparison was made on the basis of whether the methods showed predictive validity, i.e. whether the measures that were suggested by the two methods were statistically sensitive to a modification of a current and future system and whether they were suitable for use in operational settings.

The results from a series of experiments revealed that although a statistical significant difference was not found, the constraint-based method provided more measures that would be sensitive to a current system modification than the task-based method. The results also showed that there was no difference between the methods when used to select measures to evaluate a future system. Larger and more sensitive experiments are needed for a definitive answer about the benefits of the constraint-based method. Future developments of the constraint-based method were proposed that included testing the methods using different systems and incorporating all the CWA phases into an analytical framework.

Although the results did not find a statistically significant difference between the two methods the research has provided initial support to the idea that a constraint-based perspective should be considered as a viable alternative to the task-based perspective for evaluating complex socio-technical systems.

References

Annett, J., Duncan, K., Stammers, R. & Grey, M. (1971). *Task analysis.* London: HMSO.

Annett, J. (2002). A note on the validity and reliability of ergonomics methods. *Theoretical Issues in Ergonomics Science, 3*, 228-232.

AIR-STD-61/116/13 (1996). *The Application of Human Engineering to Advanced Aircrew Systems.* United States Department of Defense.

ASCC (1996). *Advisory Publication 61/116/12 Crew Performance Measurement*. Air Standardization Coordinating Committee.

Australian Army (2001). *Australian Army Land Warfare Doctrine,* Part Three, Volume 3, Pamphlet No 3, Airmobile Operations.

Baber, C. & Stanton, N. A. (1994). Task analysis for error identification: a methodology for designing error tolerant consumer products, *Ergonomics, 41*, 1607-1617.

Beevis D. (1999). (Ed.) *Analysis techniques for human-machine system design*. Crew Systems Ergonomics/ Human Systems Technology Information Analysis Centre. Wright-Patterson Air Force Base. Ohio.

Benda, P. J. & Sanderson, P. M. (1999). New technology and work practice: Modelling Change with cognitive work analysis. *Proceedings of the 7th IFIP TC13 conference on Human-Computer Interaction (INTERACT99)*, Edinburgh, Scotland.

Benda, P. J. & Sanderson, P. M. (1998). Towards a dynamic model of adaptation to technological change. *Proceedings of OzCHI98,* Adelaide, Australia.

Bisantz, A. M., & Burns, C. M. (2009). (Eds.) *Applications of Cognitive Work Analysis.* CRC Press. Taylor and Francis Group. Boca Raton, Fl.

Bisantz, A. M., Roth, E., Brickman, B., Gosbee, L., Hettinger, L. & McKinney, J. (2003). Integrating cognitive analyses in large-scale system design process. *International Journal Human-Computer Studies, 58*, 177-206.

Burns, C. M., Enomoto, Y. & Montahan, K. (2009). A cognitive work analysis of cardiac care nurses performing teletriage. In Bisantz, A. M., & Burns, C. M. (2009) (Eds.) *Applications of Cognitive Work Analysis.* CRC Press. Taylor and Francis Group. Boca Raton, Fl.

Burns, C. M., Bryant, D. J., & Chalmers, B. A. (2001). Scenario mapping with work domain analysis. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*. Minneapolis.

Campbell, E., & Herdmen, C. M. (2003). Development and evaluation of prototyped new and advanced head-down displays for the CF188 Fighter: Part 1. *Proceedings of the 12th International Symposium of Aviation Psychology.* Dayton OH.

Chapanis, A. (1959). *Research techniques in human engineering*. Baltimore: John Hopkins Press.

Charlton, S. G. & O'Brien, T. G. (1996). The role of Human Factors testing and evaluation in systems development. In O'Brien, T. G., Charlton, S. G. (Eds) *Handbook of Human Factors testing and Evaluation*. Lawrence Earlbaum, New Jersey.

Crone, D., Sanderson, P. & Naikar, N. (2007). Studying complex human-system behaviour: Human-in-the-loop simulation requirements. *Proceedings of the 51st Annual Meeting of the Human Factors and Ergonomics Society.* 1-5 October. Baltimore, MA.

Crone, D., Sanderson, P., Naikar, N. & Parker, S. (2007). Selecting Sensitive Measures of Performance in Complex Multivaraible Environments. *Proceedings of SimTECT 2007.* Brisbane.

Crone, D., Sanderson, P. & Naikar, N. (2003). Using Cognitive Work Analysis to Develop a Capability for the Evaluation of Future Systems. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society.* Denver, CO.

Crystal, A. & Ellington, B. (2004). Task analysis and human-computer interaction: approaches, techniques, and upper levels of analysis. P*roceedings of the tenth Americas Conference on Information System.* New York.

Diaper, D., & Stanton N. (2004a). (Eds.) *Task analysis for human-computer interaction.* Lawrence Erlbaum Associates, London.

Diaper, D., & Stanton N. (2004b). Wishing on a sTar: The future of task analysis. In Diaper, D., & Stanton N. (2004). (Eds.) *Task analysis for human-computer interaction*. Lawrence Erlbaum Associates, London.

DEF STAN (1989). Defence Standard 00-25 (Part 12)/Issue 1. *Human Factors for Designers of Equipment - Systems*. United Kingdom Ministry of Defence.

DEF STAN (2008). Defence Standard 00-250: 2008, *Human Factors of Designers of Systems.* United Kingdom Ministry of Defence.

MIL-HDBK-46855A (1999). *Human Engineering Program Process and Procedures.* Department of Defense.

EIA (2002). Electronic Industries Association, EIA/ HEB1, *Human Engineering – Principles and Practices,* Engineering Department, Arlington, VA.

Gawron, V., J. (2000). *Human Performance Measures Handbook*. Lawrence Erlbaum Associates. Mahwah, New Jersey.

Hajdukiewwicz, J. R. & Vicente, K. J. (2004). A theoretical note on the relationship between work domain analysis and task analysis. *Theoretical Issues in Ergonomics Science, 5, 6*, 527-538.

Hertzberg, F. (1966). *Work and the nature of man*. Cleveland: World Publishing.

Hori. S., Vicente, K. J., Shimizu, Y., & Takami, I. (2001). Putting Cognitive Work Analysis to Work in Industry Practise: Integration with IOS 13407 on Human- Centred Design. *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society.* Minneapolis.

Hoffman, R. R. & Militello (2009). *Perspectives on cognitive task analysis. Taylor and Francis Group.* New York, NY.

Hoffman, R. R., Crandall, B., & Shadbolt, N. (1998). Use of the Critical Decision Method to elicit expert knowledge: A case study in the methodology of cognitive task analysis. *Human Factors, 40*, 254-276.

Hennessy, R. T. (1990). Practical Human Performance Testing and Evaluation. In Booher, H. A. (Ed.). *Manprint: An Approach to Systems Integration*. Van Nostrand Reinhold, New York.

ISO (1999). *Human-Centred Design Processes for Interactive System (ISO13407: 1999)*. Geneva: International Organisation for Standardization.

ISO (2008). *Life Cycle Management - System Life Cycle Processes (ISO 15288: 2008)*. Geneva: International Organisation for Standardization.

Jenkins, D. P., Stanton, N. A., Salmon, P. M. & Walker, G.H. (2009a). *Cognitive Work Analysis: Coping with Complexity.* Ashgate Publishing Limited. Surrey, England

Jenkins, D. P., Stanton, N. A., Salmon, P. M. & Walker, G.H. (2009b). Using work domain analysis to evaluate the impact of technological change on the performance of complex socio-technical systems. *Theoretical issues in Ergonomics Science*, 1-14.

Kantowitz, B. H. (1992). Selecting Measures for Human Factors Research. *Human Factors, 34 (4)*, 387-398.

Klein, G. A., Calderwood, R. & MacGregor, D. (1989). Critical Decision Method for Eliciting Knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, Vol 19,* 3, May/ June.

Kilgore, R. M., St-Cyr, O. & Jamieson, G. A (2009). From work domain to worker competencies: A five-phase CWA. In Bisantz, A. M., & Burns, C. M. (2009) (Eds.) *Applications of Cognitive Work Analysis*. CRC Press. Taylor and Francis Group. Boca Raton, Fl.

Kirwan, B. & Ainsworth, L. K. (Eds) (1992). *A guide to task analysis.* Taylor and Francis Ltd. London.

Meister, D. (1999a). *The history of human factors and ergonomics.* Lawrence Erlbaum, London.

Meister, D. (1999b). Measurement in Aviation Systems. In D. Garland, J. Wine and V. Hopkin (eds). Handbook of Human Factors. Lawrence Erlbaum.

Miller, G.A., Galanter, E. & Pribram, K.H. (1970) *Plans and the structure of behaviour.* Holt, Reinhart and Winston, London.

Miller, A. M. & Vicente, K. J. (1999). Task Versus Work Domain Analysis Techniques: A Comparative Analysis. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting.*

Miller, A. M. & Vicente, K. J. (1998). Toward an integration of task- and work domain analysis techniques fro human-computer interface design. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting.*

MIL-HDBK-46855A (1999). *DOD Human Engineering Program Process and Procedures.* Department of Defense.

Muckler, F. A., & Seven, S. A. (1992). Selecting Performance Measures: "Objective" versus "Subjective" Measurement. *Human Factors, 34 (4*), 441-455.

Naikar, N., Moylan, A., & Pearce B. (2006). Analysing activity in complex systems with cognitive work analysis: Concepts, guidelines, and case study for control task analysis. *Theoretical Issues in Ergonomics Science, 7, 4*, 371-394.

Naikar, N., Hopcroft, R. & Moylan, A (2005). *Work domain analysis; Theoretical concepts and methodology*. DSTO Technical Report, DSTO-TR-1665, Defence Science and Technology Organisation, Melbourne, Australia

Naikar, N., & Sanderson, P. (2001). Evaluating system design proposals with work domain analysis. *Human Factors, 43(4)*.

Naikar, N., & Sanderson, P. (2000a). Evaluating design proposals with work domain analysis. *Proceedings of the 44rd Annual Meeting of the Human Factors and Ergonomics Society*. San Diego, CA.

Naikar, N., & Sanderson, P. (2000b). Temporal Coordination control task analysis for analysing human-system integration. *Proceedings of the 44rd Annual Meeting of the Human Factors and Ergonomics Society*. San Diego, CA.

Naikar, N. & Sanderson, P. M. (1999). Work domain analysis for training-system definition and acquisition. *The International Journal of Aviation Psychology 9(3)*, 271-290

NATO (2001). *NATO Guidelines in Human Engineering Testing and Evaluation.* RTO-TR-021 AC/323(HFM-018) TP/19. North Atlantic Treaty Organization.

Pearce, F. (1990). *Manned systems engineering in US NAVSEA*. Presentation to the 7th meeting of NATO AC/243 Panel-8/RSG.14. Washington, DC: US Sea Systems Command.

Pfautz, J. D. & Pfautz, S. L. (2009). Methods for the analysis of social and organisational aspects of work domain. In Bisantz, A. M., & Burns, C. M. (2009) (Eds.) *Applications of Cognitive Work Analysis*. CRC Press. Taylor and Francis Group. Boca Raton, Fl.

Rasmussen, J. (1983). Skills, rules and knowledge; signals, signs, and signals, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*, 257-266.

Rasmussen, J. (1981).Models of mental strategies in process plant diagnosis. In J. Rasmussen and W. B. Rouse (Eds.), *Human detection and diagnosis of system failures* (pp. 241-258). New York: Plenum.

Rasmussen, J. (1980). The human as a system component. In Smith, H. T. & Green, T. R. G. (Eds.), *Human Interaction with Computers.* London: Academic.

Rasmussen, J. (1974). *The human data processor as a system component: Bits and pieces of a model.* (Riso-M-1722). Roskilde, Denmark: Danish Atomic Energy Commission.

Rasmussen, J., Pejtersen, A. M. & Goodstein, L. P. (1994). *Cognitive Systems Engineering.* John Wiley and Sons, Inc. New York

Sanderson, P. M., & Naikar, N. (2000). Temporal Coordination Control Task Analysis for analysing Human System Integration. *Proceedings of the Joint Meeting of the Human Factors and Ergonomics Society and the International Ergonomics Association (IEA2000/ HFES2000)*. Santa Monica, CA: HFES. Vol 1, pp206-209.

Sanderson, P. M., Naikar, N., Lintern, G. & Goss, S. (1999). Use of Cognitive Work Analysis across the System Life Cycle: Requirements to Decommissioning. *In the Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society.* Houston, Tx.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive Systems Engineering.* New York: Wiley.

Smode, A.F., Gruber, A. & Ely, J.H. (1962).*The Measurement of Advanced Flight Vehicle Crew Proficiency in Synthetic Ground Environments. Behavioral Sciences Laboratory,* Wright-Patterson Air Force Base, Report Number MRL-TDR-62-2.

Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C. & Jenkins, D. P. (2005). *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate. Aldershot, UK.

Stanton, N. A. & Young, M. S. (1999) *A Guide to Methodology in Ergonomics: Designing for Human Use*. Taylor and Francis, London.

STANAG 3994. *Application of HE to Advanced Aircraft Systems*. NATO.

Swain, A. & Weston, L. M. (1988). An Approach to the diagnosis and misdiagnosis of abnormal conditions in the post-accident sequences in complex man-machine systems, in L. P. Goodstein, H. B. Anderson and S. E. Olsen (eds), *Tasks, Errors and Mental Models* (London: Taylor and Francis), 209-229.

Taylor, F. (1911). *Scientific Management.* New York: Harper and Row.

Vicente, K. (1999). C*ognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-Based Work*. Lawrence Erlbaum Associates. Mahweh NJ.

Wilson, J. R. & Corlett, E., N. (Ed) (1995). *Evaluation of human work.* Taylor & Francis Ltd. London.

Xiao, T., Sanderson, P.M., Mooij, M. & Fothergill, S. (2008). Work domain analysis for assessing simulated worlds for ATC studies. *Proceedings of the 52nd Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors & Ergonomics Society.

Xinyao, Y., Lau, E., Vicente, K. & Carter, M. (2002). Toward theory-driven, quantitative performance measurement in ergonomics science: the abstraction hierarchy as a framework for data analysis. *Theoretical Issues in Ergonomics Science, 3 (2)*, 124-142.

Yin, R. K. (2009). *Case Study Research.* (4th ed.) Newbury Park, CA: Sage Publications.

Appendix A: Exploratory Experiment

Introduction

This experiment was designed to test the infrastructure used in Experiment 1 and Experiment 2 so that both the task-based and constraint-based methods for system evaluation can be compared. The comparison can only be achieved if both known and novel aircrew and system behaviour can be supported in the experiments.

Several high-level requirements were identified. These requirements were designed to ensure that all aspects of the experiments (the experimental protocol, simulation hardware and software, scenario design and implementation, data collection and data storage devices) were fit for purpose.

The high-level requirements that were implemented and tested in this experiment are:

- Requirement 1: The simulation environment used for the experiments should have high ecological validity and should not artificially constrain behaviour. In other words when aircrew operate (or fly) the Black Hawk helicopter simulator it should behave as a real Black Hawk and the visual database must provide the level of detail required for tactical flying.
- Requirement 2: The experimental task should also have high ecological validity. In other words, the task that the aircrew is asked to perform (Airmobile Insertion Operation) should be familiar to them and the behaviour of aircraft and threat systems should be doctrinally accurate. The experimental task should be designed to allow known behaviour (as described in standard operational procedures) and also novel behaviour to be implemented. This is particularly important in the design of the mission scenarios.
- Requirement 3: The experiment protocol should not artificially limit the observed system and human behaviour. In other words, data should be collected from all mission phases (Briefing, Flight and De-brief).
- Requirement 4: The experiment design, data collection and analysis should be as wide ranging as possible so that a particular work analysis viewpoint or stakeholder is not favoured. The experiments and data collected must be designed to accommodate both known and unknown system and human behaviour. Quantitative and qualitative data collected should be objective. The exploratory experiment should provide an initial assessment of the low-level dependent variables.

The exploratory study offered an opportunity to identify potential problems and rectify them before Experiment 1 and Experiment 2 were conducted, rather than provide an outcome to a hypothesis per se. The description of the exploratory experiment follows.

Method

In the following sections the important biographical details of the aircrew are given. The apparatus and materials are also described.

Participants

Two serving members of the Australian Army took part in the exploratory experiment. The two participants formed one helicopter crew that consisted of a Flying Pilot (FP) and an Aircraft Captain (AC). Both participants held the rank of Captain. The Aircraft Captain had 1100 flight hours, 800 of which were flying Black Hawk, had been a Troop Commander of a Black Hawk squadron, was qualified as a 'C' Category Pilot and NVG Captain. He had overseas operational experience using the RWR equipment used in the simulation. The Flying Pilot had 200 flight hours. All of which were in the Squirrel training helicopter.

Apparatus and materials

The apparatus and materials used are the same as those described in Section 7.3.2.

Design

This experiment was conducted as a within-subjects design with n=1. The design of this experiment reflected the requirement to fully test all aspects of the experiment rather than test an experimental hypothesis per se. Hence, this section departs from traditional experiment design sections in that low-level dependent variables are not discussed in detail. However, all other aspects of the experiment design will be discussed including the data that were used to derive the low-level dependent variables.

A total of ten missions were run with two being used for training. The remaining eight missions were used for data collection. The data categories that were collected are given in below. The independent variable was RWR sensitivity (high, low). Five missions were run under each RWR sensitivity condition. Seven missions were flow with one crewman acting as Aircraft Captain while the other acted as the Flying Pilot. These roles were reversed for the remaining one mission.

Data collected
The data collected can be broken up into two main categories, data relating to participant behaviour and system data. Both categories of data were synchronised and collected every 16 milliseconds.

The participant data included a number of variables from faceLAB™ (the frequency that the Aircraft Captain looked at the helicopter instruments, navigation map (positioned on his knee), and the "outside" world) and all video, audio and cockpit button presses. Button presses were recorded for the chaff and flare dispensing system, navigation systems, aircraft warning systems, and communication systems.

The system data that was collected included the following:

- threat and aircraft interaction information displayed to the crew (the time when a threat signal is displayed to the crew via the RWR system (both visually and aurally); the threat's behaviour (mode) on a continuous basis – time and position stamped; whether the aircraft had been damaged by a weapon at what level of damage; and the chaff usage (how much and when)).
- aircraft data (6 degrees of freedom of the helicopter (x, y, z, heading, pitch, roll); position and velocity; latitude and longitude of helicopter; airspeed – true and indicated; pedal, cyclic, collective position; altitude – radar and barometric; fuel levels and fuel flow; chaff and radio button states and cockpit warning information.
- threat data (6 degrees of freedom of lead aircraft (x, y, z, heading, pitch, roll); latitude and longitude of threats; threat mode information – freq, scan info, tracking; Missile information and whether line of sight to the helicopter existed.
- Navigation data (buttons pressed, current waypoint information – heading to waypoint; waypoint number; distance to waypoint and position of waypoints).

Mission Scenarios
Ten mission scenarios were used during this study and were designed to have high ecological validity.

Information used to produce the constraint-based and task–based analytical products and additional information from a SME was used to produce the scenarios. The SME was interviewed using a semi-structured interview technique. The SME was asked to describe what Air Mobile Insertion missions are and then asked to describe in detail, using real-life examples, operational tactical procedures. During the interview the researcher used a series of prompts that were aimed at eliciting information on the important characteristics for the scenarios. All interview material was recorded and later transcribed. Analysis of the interview data revealed that the main characteristics that the scenarios should have were; a mixture of radar-based threat types, a mixture threats whose locations are known by the aircrew and threats whose locations are not known about (ambush threats), a mixture of threats that the RWR can identify (SA-8, SA-6, etc) and threats that the RWR cannot identify (unknowns). The interview data also revealed that scenarios should occur over terrain with a wide range of topographical features because the performance of the RWR would change as a consequence and threats modes should be accurately replicated.

In addition to having high ecological validity it was also important that the scenarios were designed to meet good experimental design requirements and to reduce possible confounding effects (for example, participant learning) so that the data could be analysed statistically. Hence, the following conditions were balanced across missions:

- **Threat density**. This refers to the number of threats located in the RWR resolution error region (see section…). High threat density refers to the case where there is more than one threat of the same type located together. Low threat density refers to the case where there is only one threat in a certain location.

- **Threat distance.** This refers to the distance from the pre-planned route that the threat is located and is related to whether the aircraft is within weapon engagement range (near) or whether the aircraft is outside the weapon engagement range (far).
- **Pop-up activation distance**. This refers to the distance from pre-planned route that the pop-up threat or ambush is activated. If the pop-up is activated when the aircraft is within weapon range then the condition is "near". If the pop-up is activated outside weapon range then the condition is "far".
- **Pop-up activation mission stage.** This refers to the stage in the mission when the pop-up (ambush) is activated. An ambush could occur close to the mid-point of the mission or toward the end of a mission.
- **Number of threats in the missions and events.** Each mission had the same number of threats that the aircrew could detect. At least one weapon launch against the aircraft should occur during each mission.

Once the scenarios had been designed they were flown by the researcher to see whether the events (for example, an ambush) could still occur over a range of local alternative flight paths. For example, a scenario may have a route running to the right of a hill that had a threat concealed behind it. The event would be that the threat would engage the aircraft as it passed it. In this case various routes around the hill, to the left and across the top, would be flown to see whether the event could still occur. If the event did not occur the threat would be relocated until a position was found that supported the event. The mission scenarios were then allocated to the high or low sensitivity condition and used in the experimental task.

Experimental task
The aircrew were required to complete an Air Mobile Insertion mission. The mission included a briefing phase, a flight phase and debriefing phase.

During the mission briefing phase the aircrew were given an operational briefing (including the reason for the mission, approximate threat locations, and other standard mission briefing elements) and a limited opportunity to modify the mission plan. The crew were instructed to complete the mission, i.e. to land the aircraft at the predetermined landing site.

During the mission flight phase, the aircrew were given a set amount of fuel and were asked to land with 20% fuel above the absolute minimum level. They were required to follow the mission plan, perform aircraft system checks as per standard operating practices (SOPs) and monitor the RWR system and counter any threats (including operating the chaff dispense system20) as required. They were also instructed to mark the location of any threats that they encountered on the map provided. The aim of each flight phase was to land at the predefined landing zone within the time and fuel constraints. The flight phase was deemed a failure if they failed to land with fuel greater than 20% above

[20] Chaff is used to confuse an enemy radar system and is dispensed by the crew using a button on the collective.

the minimum initial fuel level, if they crashed the aircraft, or if they exceeded the time limit.

After the mission flight phase, the crew undertook a lengthy mission debrief. The mission debrief was used to elicit information from the aircrew on the conduct of the mission flight phase. The average time for one mission was about 1.5 hour (20 minutes for the mission briefing, 40 minutes for the flight phase and 30 minutes for the mission debriefing).

Procedure

The experiment was performed over three days. On the first day of the experiment the aircrew were welcomed, given an Occupational Health and Safety briefing, and then briefed on the contents of the briefing pack (available from the author on request). Questions from the aircrew were answered on an ad hoc basis. During this time the faceLAB™ eye tracking tool was calibrated to the Aircraft Captain's features. Once this was completed, two participant-paced training sessions were conducted (one with the RWR modification and one without the RWR modification). The training sessions were conducted in the same manner as the experimental sessions and included a limited mission briefing phase, mission flight phase and mission debrief phase. The experimental sessions were conducted on the second and third days.

Training phase
During the training flight phase, the aircrew were briefed on the differences between the simulator systems and operational Black Hawk systems. They were instructed to fly the aircraft using SOPs and were tasked to take off, practice threat avoidance techniques and land the helicopter. Only when the aircrew had experienced all threats, flight conditions, consequences of their actions (for example, damaging the aircraft during a "hard" landing) and expressed satisfaction with their flight performance, were they allowed to progress to the experimental sessions[21]. During the practise sessions the participants were able to request information about the simulation and play "what-if" games.

An important part of the training flight phase was to ensure that the aircrew was aware of, and adapted to, the visual differences between the field environment versus the simulated world. The visual differences occur because the simulated world is projected onto the screens from a point behind them. In the case of the front screen, this point is approximately three metres in front of, and to the centre of, the aircrew. This is at odds with the field environment where the visual scene is focused at infinity, in front of each of the aircrew. The visual difference results in the aircrew "feeling" that they are flying the simulator with yaw (side slip) even when they are actually flying in a balanced (no yaw) state. Aircrew were briefed on the discrepancy and then allowed to fly the aircraft until they had adapted to the difference, i.e. flew the simulator straight and level with no yaw[22].

[21] It is noted that a self assessment of performance on a task is generally not a reliable measure of competence. However, it was decided to accept the Aircraft Captain's recommendation given his qualifications and his previous role as a pilot examiner.
[22] It should be noted that the aircrew had no previous experience with the simulator and that simulation sickness was not reported by them.

Experimental sessions

Once the training session was completed the experimental sessions commenced. These included a mission briefing phase, a flight phase and a mission debrief.

- **Mission briefing phase.** During the mission briefing phase the aircrew were presented with 1:100000 scale maps that had a route marked on it. They were then briefed on the specifics of their mission including: threat locations (an area was indicated where threats were likely), and the mission objective and constraints (i.e. fuel, time and flight profile). The aircrew were not allowed to modify the route but could annotate the map with timing, fuel, navigation, threat and other operational information as per normal practice. The crew were then briefed on the technical properties of the modified and unmodified RWR. During the briefing phase the aircrew were asked describe "out loud" what they were thinking. Various questions were asked by the author to elicit information about route planning decisions, characteristics of the environment that the aircrew considered important, and threat avoidance techniques. All audio and video of the discussions were recorded.
- **Flight phase.** Once the briefing phase was completed the aircrew proceeded to the mission flight phase. The aircrew were seated in the cockpit and instructed to indicate to the control room when they were ready to begin the flight phase. Once they were ready the simulator was "released" and they could take-off. Data recording was then started. The flight phase was terminated when the aircrew landed the aircraft safely and indicated that the mission was over or when the aircrew terminated the mission because of damage to the aircraft sustained as a consequence of a threat engagement or when the aircraft crashed. During the flight phase the author used a mission flight phase template (available from the author on request) to record any mission events of interest. Such events included threat engagements, and decisions points made by the Aircraft Captain to depart from the pre-planned route or landing zone and decisions to terminate the mission.
- **Mission debriefing phase.** Once the flight phase was terminated the aircrew proceeded to the mission debriefing phase. Video and audio recording equipment was used to record all subsequent discussions. For each event of interest the aircrew was asked to recount the main aspects of it. They were then shown a replay of the event using the Aircrew Mission Debriefing Tool and a modified Critical Decision Methodology semi-structured interview technique was used to elicit information from the aircrew. Aircrew were then given the opportunity to ask questions on any aspect of the flight phase. Once the debriefing session was completed the crew were given the opportunity to have a break (approximately 20 minutes) before the next mission commenced.

Final wrap-up

At the end of the experiment participants were debriefed on the objectives of the experiment and thanked for their participation. Letters expressing the positive contribution that they made to the research program was then sent to their Commanding Officer.

Results and discussion

The introduction has made it clear that the aim of this experiment was to test the equipment, hardware and software performance, and the experimental protocol that would be applied to Experiment 1 and 2 and that in order to do that a number of requirements were developed. In this section each of the four high level design requirements will be presented and discussed in the context of the study. 24 recommendations for Experiment 1 and 2 are given with an indication whether they would be implemented. The recommendations were articulated using a framework described in Crone et al 2007.

Requirement 1: The simulation environment used for the experiments should have high ecological validity and should not artificially constrain behaviour.
During the experiment fuel quantity was used to limit the time taken to complete the flight phase and also and amount of distance that the aircraft could be flown. If the aircraft ran out of fuel it could crash. It was found, however, that at the end of some flight phases the fuel level was so low that it affected the handling characteristics of the aircraft to a degree that was not normal in a real aircraft. The recommendation for the future experiment was that a new mechanism be found to limit the time available to the aircrew to complete the mission. The solution found was to set a minimum fuel amount that the crew had to land with. If the minimum limit was exceeded the flight phase would be "failed". The use of an absolute minimum fuel level is standard practice. This recommendation will be implemented in the experiments.

The aircrew also commented on that the flight controls (cyclic, collective and pedals) did not "feel" like that real ones. The recommendation was that they would be adjusted to make them more realistic by increasing the friction on them. This recommendation will be implemented in the experiments.

The aircrew commented on the fact that some of the simulated world navigation terrain features were too similar to distinguish and should be made more distinguishable. In addition, they commented that they should be briefed on the visual differences of these features. The recommendation was that the aircrew would be provided with screen shots of these features and allowed top fully experience them during the training phase of the future experiments. This recommendation will be implemented in the experiments.

During three separate mission flight phases the simulation systems froze. This was deemed to be very disruptive to the aircrew because it "un-immersed" the participants from the simulation environment. Restarting the simulation during low-level flight, after a system freeze, was especially disruptive to the aircrew because of the increased likelihood that the aircraft would crash (the aircrew required some time to be re-immersed in the simulation). A number of system bugs were subsequently identified associated with the system freezes. The recommendation was that these bugs should be corrected so that they would not occur during the main experiments. This recommendation will be implemented in the experiments.

During training the participants experienced some difficulties compensating for the visual differences between the CUBE and the real-world. This difference resulted in the aircrew

flying the Black Hawk straight and level but with yaw and providing navigation commands that were "a clock code out". For example, it was observed that Aircraft Captain would indicate to the Pilot that the required valley was in the 11 O'clock, however, from the position of the Pilot the required valley was in the 12 O'clock. It was recommended that training should be included that highlighted the problems associated with providing navigation instructions. This recommendation will be implemented in the experiments.

During the mission debriefs the aircrew commented that the CUBE was a good simulation environment and they found themselves immersed in the simulated world very quickly. However, they commented that the visual scene was distorted at the joins of the CUBE's display surfaces and when they noticed this they became aware of the simulation. The solution for this was found to be relatively simple and the recommendation was that the Black Hawk cockpit be moved back from the front of the CUBE by a couple of centimetres. By doing this the frame around the windscreen of the cockpit masked the surface joins. This recommendation will be implemented in the experiments.

Missile smoke trails and gun muzzle flashes were seen to be an important simulation feature as they would be used by the aircrew to indicate that a weapon had been launched. The RWR does not give an indication of a weapon launch, only that a weapon could be launched. Hence, knowledge that a weapon has been fired is only available if the aircrew see a visual indication. The recommendation was that missile smoke trails be included in the simulation. This recommendation will be implemented in the experiments.

As with all simulations of the real world the number of individual objects that can be rendered is related to computing power. This means that for a given computer system a simulated world with a large number of objects (high degree of fidelity) results in the maximum display area that is less than a real world simulation with a low number of objects (low degree of fidelity). This relationship had an important impact on the design of the simulated world. During the study the number of objects was "set" so that they could be rendered at a maximum distance of five kilometres from the current helicopter position. By setting the distance to five kilometres a high degree of visual fidelity could be achieved. The aircrew acknowledged the limitation of the visual environment and commented that during low level flight, over undulating terrain the maximum distance of five kilometres was similar to their normal operational "area of interest". However, for flat areas of land, where the aircrew could potentially see for many kilometres a maximum distance of five kilometres was not sufficient. As a result of this it was recommended that the mission routes for the subsequent experiments should be designed to take place over mainly undulating terrain. This recommendation will be implemented in the experiments.

The visual database (the virtual world) that was used for this experiment was found to be too limited in terms of the physical size of the area that could be flown over. On a number of occasions the route that the aircrew took significantly departed from the route that was initially planned by the researcher this resulted in the limits of the visual database being reached, i.e. the aircraft flew off the world. Not surprisingly, this was very disruptive to the aircrew. It was recommended that the size of visual database be changed to accommodate route deviation of five to ten kilometres from the pre-planned route. This recommendation will be implemented in the experiments.

During the flight phase the aircrew expressed confusion in identifying some topographical features especially secondary roads and creeks. This was especially disruptive because aircrew use these features to navigate. It was recommended that these features were made more distinguishable for the other experiments. This recommendation will be implemented in the experiments.

Requirement 2: The experimental task should have high ecological validity.
Aircrew commented that the location of navigation waypoints supplied by the researcher were not appropriate. The standard aircrew practice is to position navigation waypoints away from built up areas and associate them with natural features such as river bends and other topographical features. The recommendation for the next experiment was to change the locations of the navigation waypoints. This recommendation will be implemented in the experiments.

The aircrew indicated that the maps that were used for mission planning and for navigation during the flight phase were not sufficiently detailed and were not readable under the red-filtered cockpit lighting. The recommendation was to increase the detail and readability of the maps. This recommendation will be implemented in the experiments.

The aircrew pointed out that in typical real life mission there is a degree of background communication chatter that is heard over the intercom. The aircrew indicated that the effect of the chatter was to raise workload because they were required to monitor it for their call sign and potentially important information. This background chatter was not present during the exploratory experiment and it detracted from the realism of the task. It was recommended that background communication chatter be included in the next experiment. The aircrew also recommended that mission specific radio calls (for example, when the aircraft crosses navigation) are made part of the experiment task. This recommendation will not be implemented in the experiments because of the difficulty in generating background chatter and specific radio calls.

Correctly simulating the type and behaviour of the threat systems was seen to be very important to the aircrew. During the experiment the behaviour of the threat systems was not the same as real-life and resulted in comments from the aircrew expressing dissatisfaction with the simulation. They also commented that the use of missile systems against helicopters was probably not realistic. It was recommended that a number of changes be made to make the threat systems more realistic. These included increasing the time taken to "reload" the missiles of the threat systems and changing the time (the amount of seconds) that the threat systems took to change mode and including radar guided anti-aircraft artillery (AAA) as a threat system. This recommendation will be implemented in the experiments.

The aircrew commented on the level of difficulty of the mission scenarios (in all of the scenarios the aircrew encountered multiple threats) and the fact that only one Black Hawk helicopter was taking part in the mission. The aircrew commented that the level of difficulty would be "very high to extreme" and the scenario probably would not be attempted in real-life unless "Ministerial endorsement" was given. It was recommended that the briefing be supplemented with information that indicated to the aircrew that

although the missions were atypical they had been "endorsed by the Minister". This recommendation will be implemented in the experiments.

Requirement 3: The experiment protocol should not artificially limit the observed system and human behaviour.
One aircrew (one Aircraft Captain and one Pilot) participated in this experiment. The discussions with the crew during the briefing and debriefing revealed that it was very important to get at least one crew member who was experienced with RWR systems and understood the nature of the various threat systems. This crew member should be the Aircraft Captain as he was responsible for navigation and tactical decision making. It was also revealed that as long as the Pilot had passed his basic training then he could be relied on to fly the Black Hawk simulator with the degree of proficiency required for the flight phase. This recommendation will be implemented in the experiments.

Given that one of the experiments' requirements was to have high ecological validity it was decided that the consequences of bad airmanship (possible crashing) would be allowed as this would deter the pilot from deliberately exceeding the handling limits of the aircraft or flying at the terrain height. However, given the limited number of mission runs it was recommended that threat weapon systems would not be allowed to catastrophically damage the helicopter. Instead, it was decided to limit the damage that the weapon could cause to the aircraft to a degree where the aircrew could operate the aircraft albeit in a limited form. By doing this useful data could still be collected even though "mission success" had been compromised. This recommendation will be implemented in the experiments.

The results of the exploratory experiment indicated that the aircrew required more time during the briefing session to become familiar with the proposed mission route and should be allowed to make changes to that route as appropriate based on their training and operational experience. It was recommended that more time be allocated to the briefing part of the mission and aircrew should be allowed to make route changes. This recommendation will be implemented in the experiments.

Requirement 4: The experiment design, data collection and analysis should be as wide ranging as possible so that a particular work analysis viewpoint or stakeholder is not favoured.
Difficulty was experienced in synchronising the two video recorders used for collecting the qualitative video data from the Aircrew mission debriefing tool (AMDT) during several debriefing phases. It was also found that the aircrew used a limited set of the data sources when responding to the researcher. The data sources the crew found were the most useful were the views from the RWR display, the view of the threat activity, the view of the Black Hawk and the view showing the truth of the world. The data sources least useful were the view from behind the cockpit, the view of the participant, the view of the flight instruments and the view of FACELAB. It was recommended that the debrief facility be improved so that the video playback was easier. This recommendation will be implemented in the experiments.

As part of the mission debriefing the aircrew were asked to complete SART questionnaire. However, the aircrew expressed difficulty in completing the questionnaire because they

found it difficult to rate all their mission experiences. It was recommended that the questionnaire was not used as the reliability of the data could not be relied upon. This recommendation will be implemented in the experiments.

It was also found that statistical packages, for example, Statistica, could not analyse the quantity of data in its current form without a significant degree of pre-processing (one mission flight phase would typically result in a data file that was several hundred megabits in size). It was recommended that a process be developed to convert the data into a form that desk-top statistical packages could analyse. This recommendation will be implemented in the experiments.

There were ten mission scenarios used, two for training and eight for data collection. Each of the missions was made up of a limited mission briefing, mission flight phase and detailed mission debriefing. Analysis of the data revealed that there should be more than eight mission scenarios used for data collection and aircrew should be allowed to have a more extensive mission briefing. It was recommended that ten missions be used for data collection and that aircrew should be afforded opportunity to significantly plan their route and change the pre-planned one (provided by the researcher) if it was not "realistic". This recommendation will be implemented in the experiments.

It was only after the mission was completed could an assessment be as to whether the variables had been successfully balanced across missions. This was because the aircrew did not keep to the pre-planned route. For example, a mission was designed to have a threat be detected "close" to the aircraft. However, given the nature of the route that the aircrew took the threat was actually detected "far" from the aircraft. The problem was therefore how to compare missions for equivalency. It was recommended that a mechanism be found to test the actual route of the aircrew to ensure that the conditions were still balanced. This was particularly important for the proposed statistical analysis of the quantitative data. This recommendation will be implemented in the experiments.

Conclusion

The Exploratory Experiment was designed to test the infrastructure for an empirical comparison of the constraint-based and task-based methods.

Four main high level design requirements were implemented in this experiment. Of the 24 recommendations suggested to improve the experiment only one could not be implemented because of resource (cost and time) limitations. The result of the implementation of the requirements will be to provide a test environment that accurately represents the important ecological properties of a Black Hawk Helicopter fitted with a RWR system conducting an Airmobile Operation.

Appendix B: Example of the Database

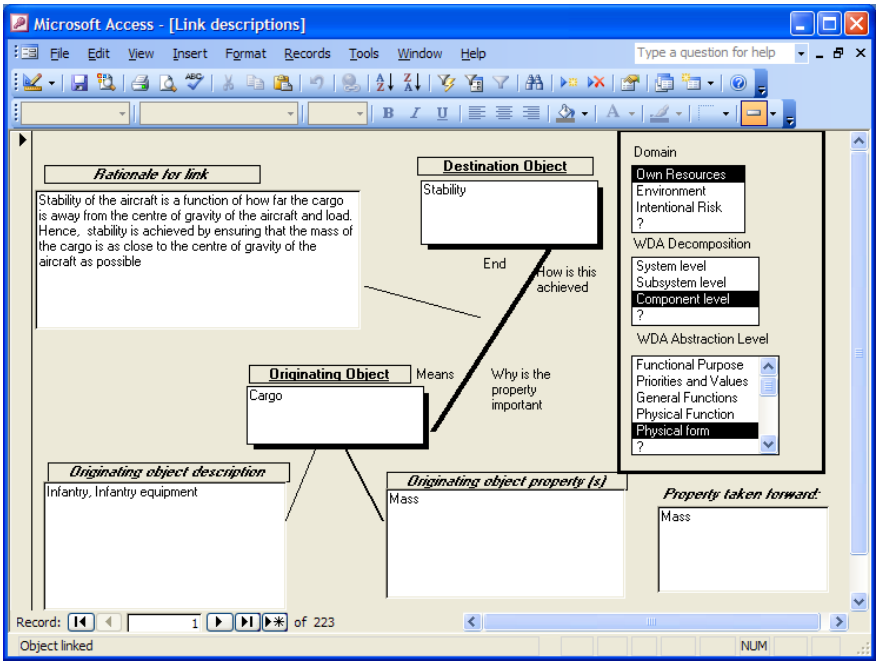


Figure B 1 Screen shot of AH database

Figure B 1 shows an example of the information used to construct the AH and ADS.

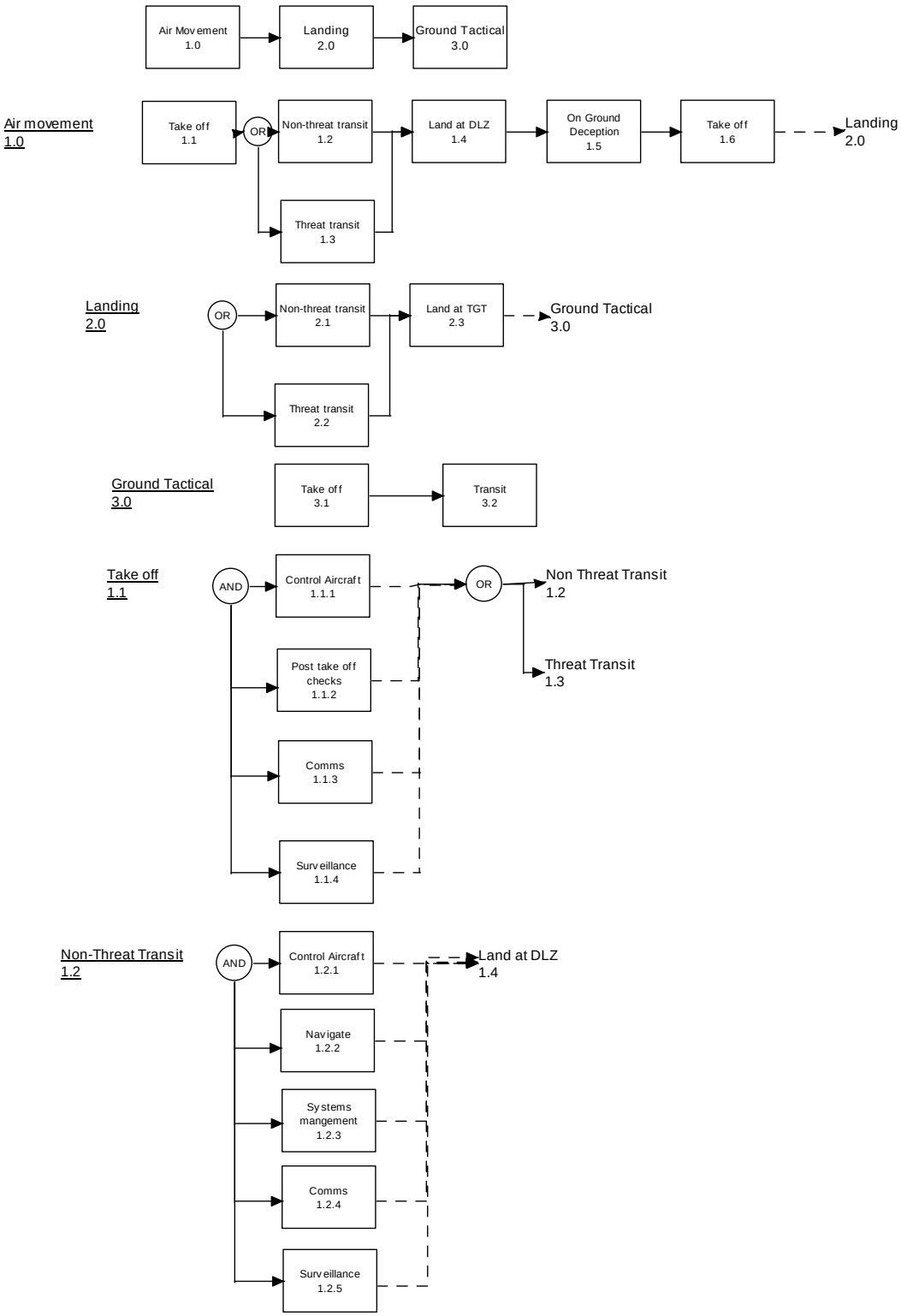
The form contains fields for:

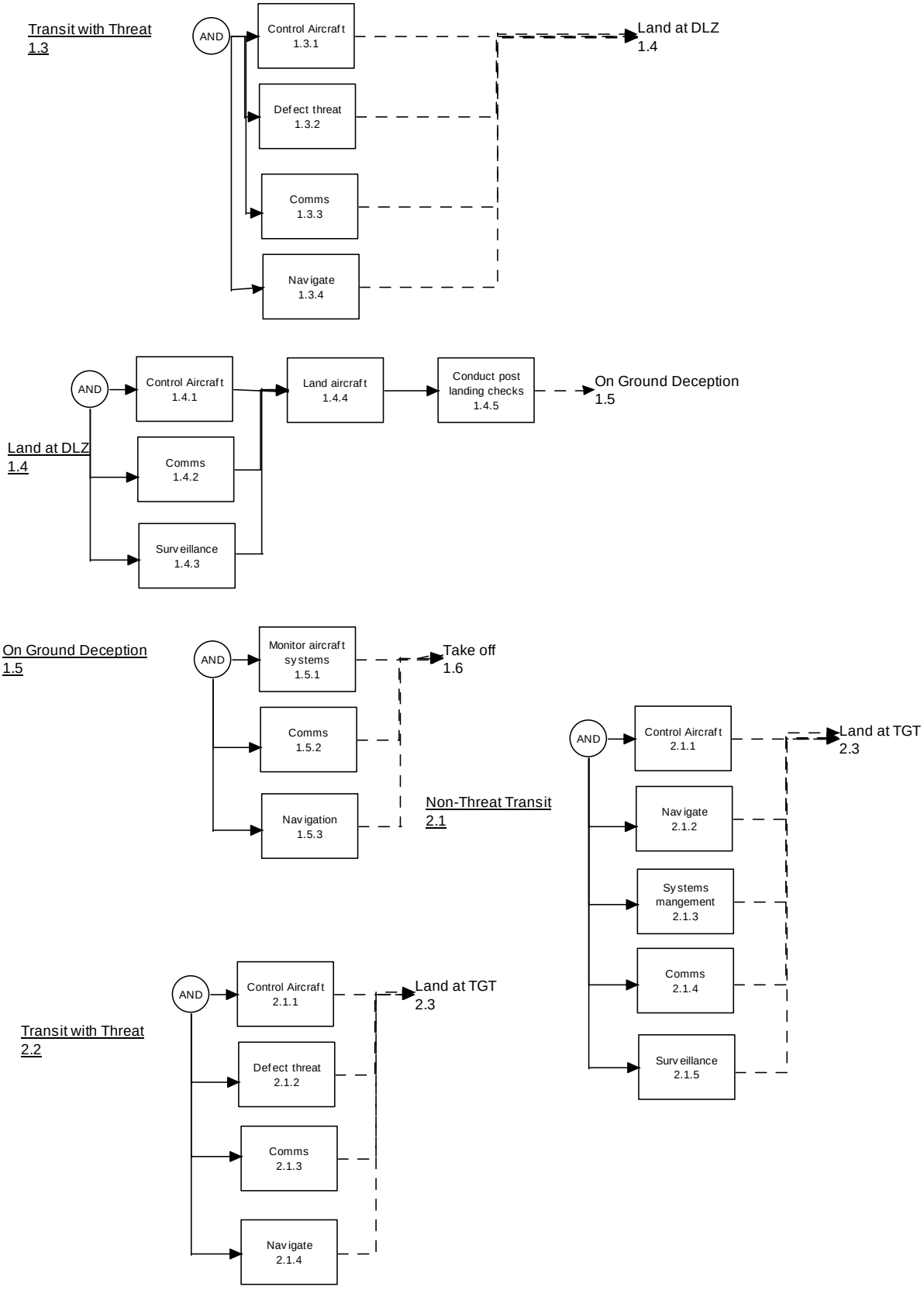
- The node of interest, "Originating Object";
- The description of the node of interest, "Originating object description";
- The properties of interest, "Originating object property (s)";
- The property that is the focus of the form, "Property taken forward";
- The domain that the node of interest is from, "Domain"
- The decomposition level that the node of interest is from, "WDA Decomposition";
- The abstraction level that the node of interest if from, "WDA Abstraction Level"
- The second node of interest, "Destination Object";
- The reason why first node is linked to the second using the property of interest, "Rationale for link"

The example shows that "Cargo" form a means-end relationship with "Stability" through the property "Mass". The reason why the two are related is given as:
"Stability of the aircraft is a function of how far the cargo is away from the centre of gravity of the aircraft and load. Hence, stability is achieved by ensuring that the mass of the cargo is as close to the centre of gravity of the aircraft as possible."

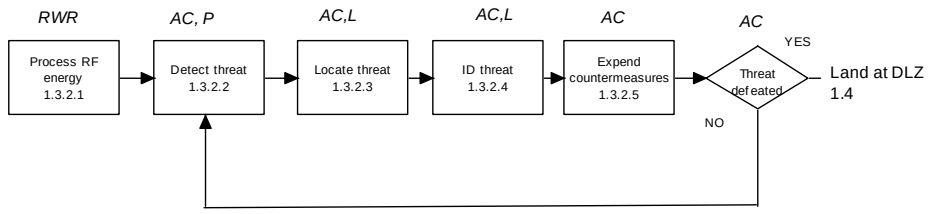
The example also indicates that "Cargo" is from the "Own resources" domain and is a "component" of the system and is derived from the "physical form" layer of the abstraction hierarchy.

Appendix C: Function Flow Diagrams

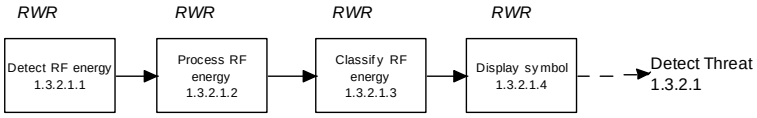



Transit with Threat
1.3
AND
Control Aircraft
1.3.1
Defect threat
1.3.2
Comms
1.3.3
Navigate
1.3.4
Land at DLZ
1.4
AND
Control Aircraft
1.4.1
Comms
1.4.2
Surveillance
1.4.3
Land aircraft
1.4.4
Conduct post
landing checks
1.4.5
On Ground Deception
1.5
Land at DLZ
1.4
On Ground Deception
1.5
AND
Monitor aircraft
systems
1.5.1
Comms
1.5.2
Navigation
1.5.3
Take off
1.6
Non-Threat Transit
2.1
AND
Control Aircraft
2.1.1
Navigate
2.1.2
Systems
mangement
2.1.3
Comms
2.1.4
Surveillance
2.1.5
Land at TGT
2.3
Transit with Threat
2.2
AND
Control Aircraft
2.1.1
Defect threat
2.1.2
Comms
2.1.3
Navigate
2.1.4
Land at TGT
2.3

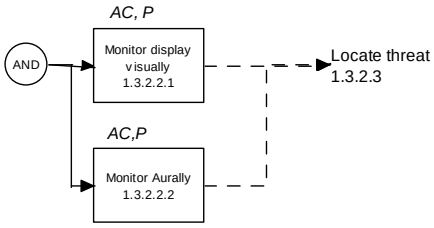
Defeat Threat
1.3.2



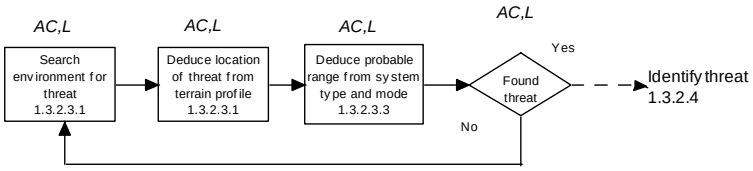
Process RF energy
1.3.2.1



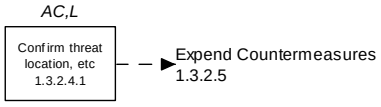
Detect Threat
1.3.2.2



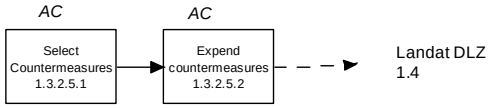
Locate threat
1.3.2.3



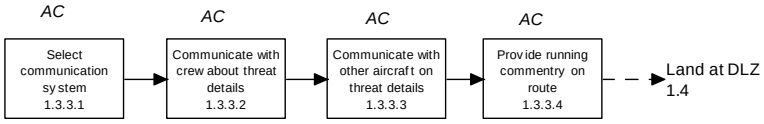
Identify threat
1.3.2.4



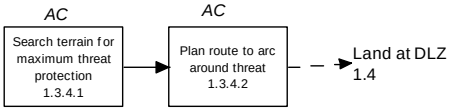
Expend Countermeasures
1.3.2.5



Communicate
1.3.3



Navigate
1.3.4



Appendix D: Input and Outputs

This appendix provides details on all the individual inputs and outputs of the constraint-based method. The list numbers are used in this section to identify the inputs, outputs and tasks and correspond to those used in the actual method.

Inputs and Outputs for the WDA method

1. The modified AH analytical product.
2. Object, function, values and priorities and purpose property tables. These tables are produced for each object, function, values and priorities and purpose in the AH. The tables list the WDA object, function, values and priorities and purpose and identify separate properties of the object, function, values and priorities and purpose. For example, Table E1 shows the four properties for the object Warning of enemy.

Table D1 Example properties table for the node Warning of enemy

Id	Properties for warning of enemy
1	Display threat information - type
2	Display threat information - location
3	Display threat information - lethality
4	Display threats (number of)

3. Record property of interest and abstraction level. The analyst records what property of the node that is of interest and the abstraction level that it is sourced from.
4. SME, aircraft manuals, CTA tasks. SMEs should be system operators who have had significant experience of the system of interest in operational conditions. The aircraft manuals should include; aircraft specification manuals, aircraft flight manuals, aircrew check lists; and other aircrew related products. The whole CTA is used.
5. Record properties from the higher node that are logically related to the property of interest. The analysts records the properties from the related nodes (on a higher level of abstraction) using the node property tables (See point 2) that are logically related to the property of interest.
6. Record properties from the lower node that are logically related to the property of interest. The analysts records the properties from the related nodes (on a lower level of abstraction) using the node property tables (see point 2) that are logically related to the property of interest.
7. List all logical properties across all nodes. The analyst produces a list of all properties from each of the nodes that are logically related. For example, Table E2 shows the properties for five of the WDA nodes. Each node is represented by a separate column; each logical relationship is represented by the rows.

Table D2 Example showing eight logical relationships between different objects, functions, values and priorities and purpose of the Black Hawk Air mobile mission work domain.

	AH levels				
Logical relationship number	**Physical Object - RWR**	**Physical function - warning of enemy**	**Domain function – tactical operation**	**Domain value or priority- mission values**	**Domain purpose- transport troops**
1	RWR sensitivity	Display threats (number of)	Surprise	Max distance from enemy	Transport troops and equipment safely
2	RWR sensitivity	Display threats (number of)	Surprise	Min exposure time to enemy	Transport troops and equipment safely
3	RWR sensitivity	Display threats (number of)	Surprise	Min probability of damage/ destruction	Transport troops and equipment safely
4	RWR sensitivity	Display threats (number of)	Surprise	Min probability of detection/ identification	Transport troops and equipment safely
5	RWR sensitivity	Display threats (number of)	Timeliness	Max distance from enemy	Transport troops and equipment safely
6	RWR sensitivity	Display threats (number of)	Timeliness	Min exposure time to enemy	Transport troops and equipment safely
7	RWR sensitivity	Display threats (number of)	Timeliness	Min probability of damage/ destruction	Transport troops and equipment safely
8	RWR sensitivity	Display threats (number of)	Timeliness	Min probability of detection/ identification	Transport troops and equipment safely

8. Same as 4.
9. Record all property changes. The analyst then records the predicted affect of the system modification for each of the properties. For example, the affect of modifying the RWR (increasing the RWR sensitivity) may result in: a greater number of threats being displayed; more surprise; an increase in the distance to the enemy; and more chance that the troops and equipment will be transported safely.
10. List the measures for all node properties. To generate the measures the analyst selects the ecological properties of objects, functions, priority and values and purposes of the WDA and tasks of the CTA that are important to the Aircraft Captain and second, identifies the measures to be associated with the ecological or task properties. The measures can be developed from analysing information from actors from the work domain (interview data and documents). For example, the Domain Function Tactical Operations has several ecological properties, one of which may be identified by Black Hawk aircrew as the time it takes for information to be transferred from its source to its destination. Based on the information from the Black Hawk crew "The time taken for data transfer from one location to another" may be considered as a valid measure of the Tactical Operations function.

11. Record the properties that are predicted to change and the MOPS and MOEs that are likely to show the change. The analyst records the measures/ variables that are likely to be sensitive to the system modification.
12. Guidelines for thinking about relationships. The guidelines for thinking about the relationships are:
    - Use the CTA. In particular identify the Direct and Indirect tasks associated with the event "Detect threat". The direct control task used will be the one that if of interest in the evaluation (in our case "Manage EW system").
    - Select the Physical Object property "RWR sensitivity", remembering that this reflects the change in range that the RWR can detect a threat.
    - For each property identified earlier consider how a change in detection range may affect the property in the context of the control task.
    - Indicate how the property may change using simple codes (e.g. better, worse; or an up arrow or down arrow; red or green).
13. Same as 12

Inputs and Outputs for the CTA method

14. The modified TC-CTA analytical product.
15. Annotate CTA. The MTC-CTA is annotated to indicate the section(s) that the event could take place in. In our case the event is "Detect threat".
16. Record Direct Activity. The Direct Activity is the activity that is primarily related to the event. In our case the Direct Activity is "Manage EW system and HMI".
17. Record Indirect Activity. Record all activities that have to be performed while the Primary Activity is being performed, i.e. identify all the activities that should be attended to. Ignore all the other activities.
18. Activities that have the ecological property as a characteristic. Record the activities that have the ecological property of interest. For example, an ecological property may be time. Hence, all activities that are characterised by time (for example, Operate Aircraft - changing the speed of the aircraft affects time taken to cover a distance) should be noted.
19. None
20. Record all the activities. Record all the activities that use the object (system).
21. Meister's behavioural measures. The list of behavioural measures identified by Meister (1995).
22. SMEs should be system operators who have had significant experience of the system of interest in operational conditions.
23. Record the activities that are and are not predicted to change and the measures that are likely to show a change.
24. Annotate Meister's measures. Record which of the measures are suitable for use in the evaluation.

25. Guidelines for choosing the measures. The guidelines are:
26. Choose measures that reflect what the aircrew do (Vicente, 1999),
27. Use the Direct and Indirect activities to consider the appropriateness of each of the measures.
28. Guidelines for predicting change. The guidelines are:
29. Assess each measure in the context of the activity and property of interest,
30. Indicate that a measure may be sensitive to the system change if a case can be demonstrated that shows that in a similar situation a similar system would behave in the way predicted.

Appendix E: Mission Scenarios for Experiment 1 and 2

The mission scenarios that were used during this study were designed to have high ecological validity. Information used to produce the WDA, CTA and Task Analysis, with additional information from a SME was used. The SME was interviewed using a semi-structured interview technique. The SME was asked to describe what Air Mobile Insertion missions are and then asked to describe in detail, using real-life examples, operational tactical procedures. During the interview the researcher used a series of prompts that were aimed at eliciting information on the important characteristics for the scenarios. All interview material was recorded and later transcribed. Analysis of the interview data revealed that the main characteristics that the scenarios should have were; a mixture of radar-based threat types, a mixture threats whose locations are known by the aircrew and threats whose locations are not known about (ambush threats), a mixture of threats that the RWR can identify (SA-8, SA-6, etc) and threats that the RWR cannot identify (unknowns). The interview data also revealed that scenarios should occur over terrain with a wide range of topographical features because the performance of the RWR would change as a consequence and threats modes should be accurately replicated.

In addition to having high ecological validity it was also important that the scenarios were designed to meet good experimental design requirements and to reduce possible confounding effects (for example, participant learning) so that the data could be analysed statistically. Hence, the following conditions were balanced across missions:

- Threat density. This refers to the number of threats located in the RWR resolution error region. High threat density refers to the case where there is more than one threat of the same type located together. Low threat density refers to the case where there is only one threat in a certain location.
- Threat distance. This refers to the distance from the pre-planned route that the threat is located and is related to whether the aircraft is within weapon engagement range (near) or whether the aircraft is outside the weapon engagement range (far).
- Pop-up activation distance. This refers to the distance from pre-planned route that the pop-up threat or ambush is activated. If the pop-up is activated when the aircraft is within weapon range then the condition is “near”. If the pop-up is activated outside weapon range then the condition is “far”.
- Pop-up activation mission stage. This refers to the stage in the mission when the pop-up (ambush) is activated. An ambush could occur close to the mid-point of the mission or toward the end of a mission.
- Number of threats in the missions and events. Each mission had the same number of threats that the aircrew could detect. At least one weapon launch against the aircraft should occur during each mission.

Once the scenarios had been designed they were flown in the simulator by the researcher to see whether the events (for example, an ambush) could still occur over a range of local alternative flight paths. For example, a scenario may have a route running to the right of a hill that had a threat concealed behind it. The event would be that the threat would engage

the aircraft as it passed it. In this case various routes around the hill, to the left and across the top, would be flown to see whether the event could still occur. If the event did not occur the threat would be relocated until a position was found that supported the event. The mission scenarios were then allocated to the high or low sensitivity condition and used in the experimental task.

Table E1 shows how the variables were balanced for each of the scenarios. In addition to these variables a similar number of threats and threat types were present in each of the scenarios. Analysis of the table shows that not all of the theoretical combinations were implemented. From a theoretical perspective the minimum number of mission scenarios needed for complete balancing was 48 (2xdensity conditions, 2xthreat position conditions, 2xpop-up threat position conditions, 3xpop-up activation distance conditions, 2xthreat type). However, many of these theoretical scenarios were deleted because they had no basis in the real-world, would not be accepted by the participants, and therefore would reduce the ecological validity of the missions. For example, a mission scenario that was the product of combining conditions in a logical manner would have a threat pop-up (ambush) position near to the aircraft and be activated at the start of the mission. However, this would be unacceptable to the aircrew as it would not reflect an operational situation given the type of threats used.

Another reason why not all possible scenarios were used was because of practical constraints. For example, the time available to complete the experiment could not exceed the crew availability. Hence, the mission scenarios that were finally used represented a compromise of a number of different factors including, ensuring the conditions were equivalent, ensuring the ecological validity of the missions and the practical (time) constraints imposed on the experiment.

One the scenarios were initially designed and tested a five step process was followed. The first step in the process was to use the framework described in Crone, Sanderson, Naikar, & Parker (2007) to modify the existing mission scenarios. The second step was to fly each scenario to see whether an "event" would still occur over a range of alternate (local) flight path. For example, a scenario may have a route running to the right of a hill that had a threat concealed behind it. The event would be that the threat would engage the aircraft as it passed it. In this case various routes around the hill, to the left and across the top, would be flown to see whether the event could still occur. If the event did not occur the threat would be relocated until a position was found that supported the event. The third step involved a "robot" flying each of the routes at an altitude of 50ft (the altitude specified by SMEs as typical of these missions) and the number of threat modes "experienced" by the robot analysed. The aim of this was to ensure that the route was not biased toward one of the sensitivity conditions. For example, if the frequency (or occurrence) of the threat mode was not the same in both high and low sensitivity conditions the mission scenario was deemed to be biased and would be redesigned. Threat mode was used because it provided a good indication of whether each of the engagement rings was supported.

The fourth step entailed allocating each mission scenario to the modified or not modified condition. The fifth step entailed a post experiment analysis of the routes that the aircrew had actually flow.

The fifth step was completed after the mission routes had been flown by the participants and after the experiment data were collected. This step was necessary because some aircrew deviated from the planned route and it was necessary to test whether the route that was taken had resulted in a significant change to the distance that a threat had been encountered. If the distance had changed then the unmodified RWR sensitivity condition may have had threats that were too far away to detect. This would have confounded the results. The variable that was used to test the scenarios was the "distance to the first encountered threat". This variable had two categories: Near and Far[23]. A chi-squared test was used to compare the missions that were to be used in the modified and unmodified system trials against the two conditions. The results are briefly discussed.

Table E1 Scenario definitions

Mission Scenario order	RWR condition	Threat Density	Threat position[24] (Distance of general threats to target)	Pop up position[24] (Distance from track when activated)	Number of unknowns
6 (Training)	Unmodified	High	Far	Near (Mid point)	Low
	Modified	High	Far	Near (Mid point)	High
1 (Training)					
2	Modified	High	Far	Near (End point)	High
8	Unmodified	Low	Near	Far (Mid point)	Low
3	Modified	Low	Near	Far (Mid point)	High
11	Unmodified	Low	Far	Far (End point)	Low
4	Modified	Low	Near	Far (End point)	High
5	Unmodified	High	Far	Near (End point)	Low
10	Modified	Low	Far	Far (End point)	High
7	Unmodified	Low	Near	Far (End point)	Low
9	Modified	High	Far	Near (Mid point)	High
12	Unmodified	Low	Far	Far (Mid point)	Low

Test results for experiment 1

The results indicated that there was no significant difference in the distance to the first encountered high priority threat for the two system conditions (X2 (1, N=23) = 0.43, p=0.5099). It was concluded that even after the deviation from the route by the aircrew the missions that were used for the modified system condition were still equivalent to the unmodified system ones.

Test results for experiment 2

The results indicated that there was no significant difference in the distance to the first encountered high priority threat for the two system conditions (X2 (1, N=22) = 0.35, p = 0.55). It was concluded that even after the deviation from the route by the aircrew the missions that were used for the modified system condition were still equivalent to the unmodified system ones.

[23] These categories reflect an interaction between the threat type and mode of operation. A threat may be physically near to the aircraft but may be scored "far" because it engaged the aircraft at its maximum distance. Whereas a second threat may be a large distance away from the aircraft and scored "near", because it engaged the aircraft at the minimum distance.
[24] Near = within missile range; far = inside tracking range

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Selecting Measures to Evaluate Complex Sociotechnical Systems: An Empirical Comparison of a Task-based and Constraint-based Method

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

David J. Crone

5. CORPORATE AUTHOR

DSTO Defence Science and Technology Organisation
Fairbairn Business Park Department of Defence
Canberra ACT 2600 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-RR-0395	AR-015-654	Research report	July 2013

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
	ERP 07/398	CJOAD	225	76

13. DSTO Publications Repository

http://dspace.dsto.defence.gov.au/dspace/

14. RELEASE AUTHORITY

Chief, Joint and Operations Analysis Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS

Systems analysis, task analysis, cognitive work analysis, performance measures, measures of effectiveness

19. ABSTRACT

Researchers need better measures for evaluating human performance in complex socio-technical systems. A constraint-based and a task-based method were compared for their effectiveness in identifying measures that would be sensitive to a system modification. Working in an advanced simulator, aircrews conducted tactical missions with or without a modification to a specific aircraft system. Across two experiments there was no significant difference between the methods in the sensitivity or suitability of the measures that they suggested. Nonetheless, observations made during the program of research suggested that the constraint-based method for identifying measures is a viable alternative to the task-based method.