Australian Government
Department of Defence
Defence Science and
Technology Organisation

Usability Evaluation of Air Warfare Assessment & Review Toolset in Exercise Black Skies 2012

Julian Vince and Jessica Parker

Aerospace Division
Defence Science and Technology Organisation

DSTO-TR-2923

ABSTRACT

A usability evaluation of a suite of after action review tools was undertaken as part of a synthetic collective training research exercise (Exercise Black Skies 2012). The suite of tools represent a test-bed for enquiry into what qualities and features are useful in after action review tools used in the collective synthetic training context. This report provides a description of the usability design issues observed in the use of the tool suite and proposes opportunities for future development. The study makes use of two complementary approaches to usability: User Testing and the Systemic-Structural Theory of Activity.

RELEASE LIMITATION

Approved for public release

Published by

Aerospace Division
DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend, Victoria 3207 Australia

Telephone: 1300 333 362
Fax: (03) 9626 7999

AR-015-817
August 2013

APPROVED FOR PUBLIC RELEASE

Usability Evaluation of Air Warfare Assessment & Review Toolset in Exercise Black Skies 2012

Executive Summary

Exercise Black Skies 2012 (EBS12) was the latest in a series of simulation exercises facilitated by the Aerospace Division (AD) of the Defence Science and Technology Organisation (DSTO), aimed at developing a deeper appreciation of the benefits of synthetic team and collective training. Exercise Black Skies 2012 was held as an Exercise Pitch Black preparation opportunity for a Royal Australian Air Force (RAAF) 41 Wing Air Defence Ground Environment (ADGE) Air Battle Management (ABM) team and a 42 Wing Wedgetail mission crew.

The broad research aim of EBS12 was to develop and evaluate a simulated Pitch Black training environment along with supporting training technologies, to examine the benefits of virtual preparation for a live training exercise. Among the training technologies being evaluated was the Air Warfare Assessment and Review (AWAR) toolset. The AWAR toolset was designed to aid ABM and Wedgetail assessors in assessing the performance of their respective teams during missions, and providing feedback during After-Action Review (AAR).

This report describes the evaluation of the usability of the assessment and review functions of the AWAR tool. The impact of the use of the AWAR tool on the AAR learning process is reported separately. User Testing and Activity Theory approaches were applied in parallel to test the usability of AWAR and its effectiveness as an assessment and review system, and to inform recommendations for future design modifications to the AWAR toolset. Thirteen recommendations for improvements to the toolset were identified:

1. Revise hierarchy structure
2. Synchronise timing & employ timeline depiction of events
3. Add flexibility in time-stamping
4. Address Smartboard functionality problems
5. Allow comments without ratings and allow user to add comment before rating
6. Maximise the practical integration of AWAR assess function
7. Give the user the option of scoring with finer granularity
8. Improve recoverability by providing more exits
9. Reposition the 'Add New' button closer to the hierarchy
10. Provide structured training before an exercise begins
11. Adopt conformity with convention where possible
12. Hide less-utilised features in menus
13. Explore design features for reducing typing burden.

Authors

Julian Vince
Aerospace Division

Julian Vince is a psychologist and human factors researcher in the Crew Environments & Training branch of Aerospace Division. Julian's research interests centre on the application of constructivism and Activity Theory to issues of work and learning. He has a Master of Industrial/Organisational Psychology degree from Deakin university.

Jessica Parker
Aerospace Division

Jessica Parker is a training simulation researcher in the Air Operations Simulation Centre in the Crew Environments and Training branch of Aerospace Division. Jessica has completed a Bachelors Degree in Computer Science and a Bachelors Degree in Cognitive Science at La Trobe University, Melbourne (2009). She joined DSTO in 2010.

Contents

Glossary

AAR	After-Action Review
ABM	Air Battle Manager
AT	Activity Theory
ADGE	Air Defence Ground Environment
AOD	Air Operations Division
AWAR	Air Warfare Assessment & Review
CSUQ	Computer System Usability Questionnaire
DSTO	Defence Science and Technology Organisation
EBS	Exercise Black Skies
EBS08	Exercise Black Skies 2008
EBS10	Exercise Black Skies 2010
EBS12	Exercise Black Skies 2012
HCI	Human Computer Interaction
MECs	Mission Essential Competencies
QUIS	Questionnaire for User Interface Satisfaction
RAAF	Royal Australian Air Force
SSTA	Systemic-Structural Theory of Activity
SUS	System Usability Scale
TAARDIS	Tactical After-Action Review of Distributed Interactive Simulation
USAF	United States Air Force

This page is intentionally blank

1. Introduction

1.1 Background

The use of synthetic training environments for training collectives of teams may hold potential for overcoming some of the shortcomings of large and expensive live training exercises. However, the implementation of simulation-based training environments potentially invokes some complexities unique to these environments. In order to address the potential training risks of large and complex simulation training events, these events and associated toolsets must be trialled and evaluated.

Exercise Black Skies 12 (EBS12) was the latest in a series of simulated collective training exercises conducted by the Defence Science and Technology Organisation under DSTO task AIR07/327 (Support to Air Force Training Capability and Projects), to address client requirements set forth by Air Force Headquarters. These requirements relate to the development of tools and processes for conducting collective synthetic mission rehearsal. The objectives of EBS12 built upon the outcomes of previous exercises such as the Pacific Link exercises (e.g., Best, Hasenbosch, Skinner, Crane, Burchat, Finch, Gehr, Kam, Shanahan & Zamba, 2007), Exercise Black Skies 08 (e.g., Shanahan, Best, Finch, Stott, Tracey, & Hasenboch, 2009), and Exercise Black Skies 10 (Stevens, Crone, Temby, Best & Simpkin, 2011; Stott, Best, & Shanahan, 2011).

Exercise Black Skies 10 (EBS10) attempted to extend the application of synthetic collective training tools and processes to new air warfare contexts. EBS10 comprised two separate human-in-the-loop simulation activities, run in the Air Operations Simulation Centre, DSTO Melbourne. The first week was aimed at the evaluation of simulation-based mission preparation for Close Air Support (CAS) teams, comprising of F-18 pilots and Joint Terminal Attack Controllers (JTACs) (Stevens, Crone, Temby, Best & Simpkin, 2011). The second week was aimed at the evaluation of simulation-based mission preparation for an Air Battle Management (ABM) team (Stott, Best, & Shanahan, 2011).

EBS12 attempted to further expand the bounds of Exercise Black Skies simulated collective training events in terms of scale. The expanded scale of the collective simulation training exercise was achieved through the joint exposure of a Royal Australian Air Force (RAAF) 41 Wing Air Defence Ground Environment (ADGE) ABM team and a RAAF 42 Wing Wedgetail mission crew to a simulated Exercise Pitch Black training environment.

The broad research aim of EBS12 was to develop and evaluate a simulated Pitch Black training environment along with supporting training technologies, in order to examine the benefits for a collective of teams resulting from virtual preparation for live training exercises. Among the training technologies being evaluated was the Air Warfare Assessment and Review (AWAR) toolset. The AWAR toolset is a test-bench for evaluating technical and procedural aids in the provision of assessment of training audience performance and subsequent feedback through an After-Action Review (AAR) following the training mission. AWAR was first introduced for Exercise Black Skies 08 as a spreadsheet-based performance assessment and review tool (Tracey, Hasenbosch, Vince,

Pope, Stott, Best, Shanahan & Finch; 2009). The ability to review video and radio communications replay was added for EBS10 and EBS12.

AARs are important because they give the training audience an opportunity to frame their subjective understanding of mission performance with reference to a more objective 'ground truth' understanding of the events. An AAR ground truth that contains more objective information than the participants had available to them during the execution of the mission is important because perceptions and memories of events can be distorted (Goldberg & Meliza, 1993). AWAR was in part developed to aid in representing an objective ground truth to a training audience.

There were two perspectives taken with regard to the evaluation of the AWAR tool. These were: (1) an evaluation of the usability of the assessment and review functions of the AWAR tool, and (2) the impact of the AWAR tool on the AAR learning process. This report deals solely with the evaluation of AWAR tool usability. The evaluation of AWAR tool impact on AAR learning processes is reported separately.

1.2 Air Warfare Assessment and Review (AWAR) tool

1.2.1 AWAR tool description

AWAR was developed as a test-bed to help understand what kind of features and processes in a toolset will support assessors in making evaluations of team performance and imparting feedback - as well as what kinds of features detract from these efforts.

AWAR was designed to support assessment and review in three ways: (1) provide a platform for improved validity and reliability in assessment of team performance; (2) provide a mechanism to assist in the automatic recording of ratings, and prevent problems associated with double handling when transferring pen-and-paper rating into electronic format for storage; and (3) provide a platform on which AARs could use objective memory aids in the form of audio-visual recordings to help provide feedback to the training audience (Tracey, et. al., 2009).

RAAF ABM and Wedgetail assessors made use of AWAR by observing trainees in a simulated or live exercise (such as Exercise Black Skies (EBS) or Pitch Black), and recorded ratings and comments for the team they assess. Ratings are given according to a set of pre-defined criteria called a 'hierarchy', which is divided into high-level and low-level goals. AWAR time-stamps each rating and comment. These time-stamps act as bookmarks to help assessors reference events for audio-visual replay and discussion during the AAR.

It is important to note that the AWAR tool itself and the assessment hierarchy are separate parts of the tool. The AWAR is a vessel into which any assessment criteria may be inserted. The assessment hierarchy is a test-bed in itself, which enables the exploration of the development of criteria for the purposes of assessment in collective training (Hasenbosch & Best, 2007). There are three assumptions underlying the hierarchy's design: (1) as each assessor is made to use the same criteria to rate teams, the system promotes

greater consistency between instructors; (2) the hierarchy compels instructors to consider a broader range of evaluation criteria than a more ad-hoc approach (Tracey, et. al., 2009); and (3) a standard AWAR hierarchy also allows instructors to make fair comparisons between different teams being rated against the same hierarchy, or the same team being rated on different missions.

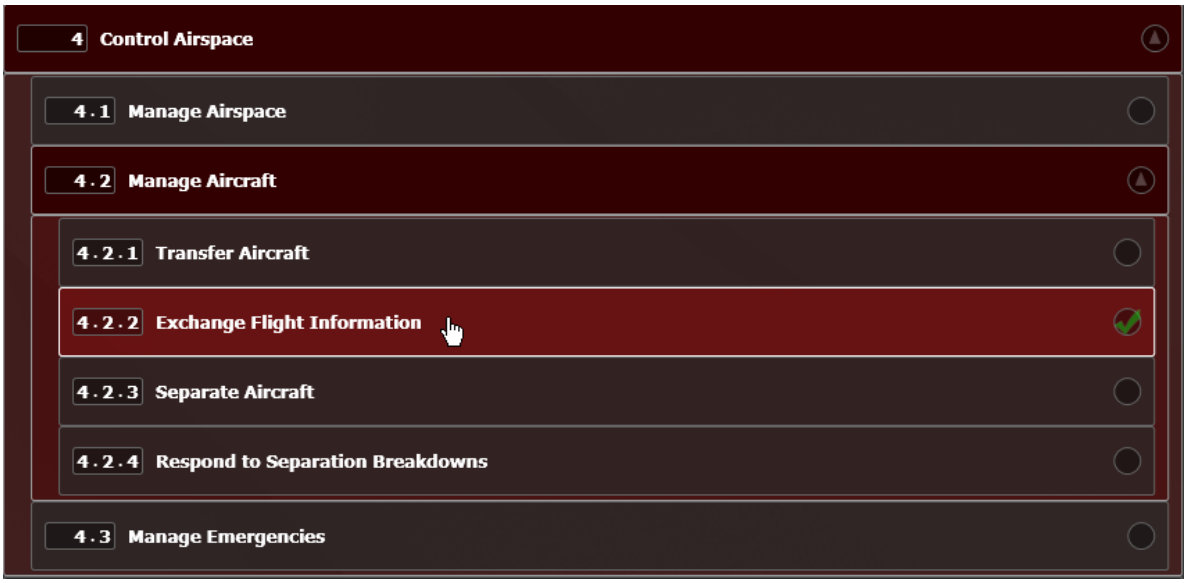

Figure 1. Selection of AWAR assessment hierarchy criteria from EBS12.

A sample hierarchy is shown in Figure 1. The instructors are permitted to assess either at the high level (e.g. "Control Airspace"), or at the lower level for more detailed assessment (e.g. "Exchange Flight Information").

DSTO developed the AWAR tool initially as an electronic spreadsheet and more latterly as purpose-built software to improve the assessment and review process of teams during training exercises. In essence, the EBS12 software version of AWAR was designed to develop a database of video timestamps and associated ratings and assessor comments. During the AAR phase, the AWAR database of timestamps are used as a controller for locating, playing, stopping and pausing video recorded via the Tactical After-Action Review of Distributed Interactive Simulation (TAARDIS) video-logging tool. The two software components were not developed to automatically function together, and required a knowledgeable person to set-up the two tools so that they would work together.

Figure 2. Review and playback-control interface for AWAR review mode.

During the EBS12 AARs, AWAR generated a review and playback-control interface that the assessor could use to drive the audio-visual replay and encourage discussion (Figure 2). AWAR summarised the ratings given during the assessment phase according to the preference of the assessor (giving preference to high or low-level ratings). To focus on a particular comment or rating, the assessor replayed the exercise from a bookmarked point, time-stamped during the assessment of the mission when the rating and comment was made. Bookmark selection and replay was done via the review and playback interface, which aims to reduce the requirement for assistance from a person managing the AWAR desktop computer to manage selections and screen views.

1.3 Evaluation of AWAR

The approach to the general evaluation of the pre-EBS12 versions of AWAR was based on feedback from the assessors that used it, as well as the collective training research team. With the development of the software-based tool for EBS12, it was decided that it was time to apply more formal methods to the evaluation of the AWAR tool usability. Discussion between the authors of this report revealed a difference in preferred approaches to usability assessment, with one preferring a User Testing approach, and the other interested in usability assessment from an Activity Theory (AT) perspective. As the two approaches were seen as complementary but with different emphases, it was decided to conduct the AWAR usability analysis by applying the two orientations separately and comparing the outcomes of each where their results differed. The Activity Theory-based approach has been argued to augment the relative technical emphasis toward on-screen elements of the User Testing approach; by also accounting for the experience, motives and

goals the user brings to the use of a tool (Vrazalic, 2003). An outline of the User Testing and AT approaches is provided.

1.3.1 User Testing Approach

There are a number of attributes that any quality software should possess. The first and most obvious attribute of good software is its functionality. A piece of software has good functionality when it conforms closely to the specifications detailed by the client or end user, and more generally, when it achieves the functions it was built for – that is, it allows the user to do what they want to do with it (Pressman, 2005). This concept is sometimes called fitness for purpose (Nielsen, 1993). Other characteristics of good software defined by Pressman (2005) are: reliability – the proportion of time the software is available for its intended use, as opposed to down time or recovery time; efficiency – how many of the computer's resources must be used to perform a certain function; portability – how well the software can be used on different platforms and how easily it may be transported between them; maintainability – how easily errors in the software can be isolated and corrected; and usability – the ease of use and learnability of the software. It is this last characteristic, usability, upon which the User Testing analysis of this report is focused.

Usability is an extremely important characteristic of a software system. A system's usability will play a vital role in the success of a system, both in terms of its uptake into the environment for which it was designed, and the degree to which its intended users enjoy and continue to use it. Ensuring a high degree of usability in a system will encourage users to continue using the system to its greatest effect. A system that is intuitive, allows users to recover easily from errors, conforms to other software norms, and is aesthetically pleasing will always be preferred over a system that is awkward, clunky and difficult for new users to adopt due to defiance of conventions. Because of the importance of usability, thorough usability testing is a vital step in the development of any software system (Nielsen, 1993).

User testing generates a list of system features that are satisfactory and unsatisfactory (Nielsen, 1993). This list should be dynamic and should inform development. For this reason, usability testing should be an iterative process that is conducted throughout the development cycle of the software. Significant changes to the software's interface or functionality should always be tested for their usability qualities. This will help to ensure that significant changes to the interface or functionality will be avoided as the software approaches completion (Nielsen, 1993).

Very few systems will have novice or expert users exclusively. Therefore, usability testing should be conducted with a variety of users, ranging from novice to expert users, to ensure a broad cross section of skill levels is accounted for in the system design. This will ensure that the test group is a representative sample of the user population (Nielsen, 1993). Every user starts using the system as a novice – if the system is difficult to learn to use, the likelihood that the novice user will continue to use it will be reduced. Usability testing should also be conducted under a variety of conditions, but should focus on the context in which the system will most often be used – that is, testing should not only be conducted with a contrived set of tasks, but with the additional pressures of normal use (Nielsen, 1993). This may include elements such as time pressure or large data sets.

Usability testing can be performed subjectively and objectively. An ideal analysis of usability would include both subjective and objective measures. By analysing both subjective and objective metrics, the tester can ascertain both the participants' attitudes to a piece of software, as well as how successfully they actually performed tasks using the software, including areas of particular difficulty. It is also possible to see how many errors were made, if and how the user recovered from those errors, and whether the user reacted positively or negatively towards the system while using it.

Nielsen (1993) describes the various types of objective features of software use that may be employed in usability testing. As mentioned above, testing should be performed at various stages of the software development cycle and hence it may be established, from the performance of the user, whether the current iteration is a statistically significant improvement on the previous one. Some measures listed by Nielsen (1993) that may inform the objective usability of a software system include: the number of errors made; the number of commands or features that are not employed by the user; or the time taken to complete various tasks.

Though many different scales have been developed for testing usability subjectively, testers may be tempted to develop their own questionnaire for usability. This can be dangerous for a variety of reasons. A scale developed for one-time use is at risk of being less valid and less reliable than others. It is advisable therefore to use a scale that, while not developed for the specific system being tested, is developed to assess usability in a general sense. Usually, these scales will have been used a number of times, for different types of software, and will have had statistical testing performed on them to show that they have acceptable levels of validity and reliability. Some well-validated examples of usability scales include the Questionnaire for User Interface Satisfaction (QUIS) (Chin, Diehl & Norman, 1987), the Computer System Usability Questionnaire (CSUQ) (Lewis, 1995) and the System Usability Scale (SUS) (Brooke, 1996).

The SUS is a scale for testing usability that was designed to be simple and quick to administer. It returns a final score out of 100 to allow for easy comparison between users and test sequences. Bangor, Kortum and Miller (2008) evaluated the SUS over 206 studies and found that it was indeed a valid (single significant factor for all ten statements), and reliable (Cronbach alpha for internal consistency = 0.91) tool for measuring usability, and that it was useful for a wide variety of applications. The SUS was chosen over other tools due to the combination of its robustness as a questionnaire as well as its brevity and ease of use. The SUS is attached in Appendix A.

1.3.2 Activity Theory Approach

AT may be argued to employ a view of usability that begins from a broader socio-technical perspective than the User Testing approach. This broader socio-technical perspective provides scope for consideration of: (1) the nature of the needs and goals the subject is attempting to meet; (2) the sociocultural and technical context in which the subject is trying to meet the goals of their activity; (3) the way in which physical and mental processes are supported and influenced by the available tools; and (4) how historical social and technical factors influence the developmental transformations of activities over time

(Kaptelinin & Nardi, 2006). Nielsen's (1993) concepts of fitness for purpose, robustness, performance, interoperability, maintainability and usability are also areas of concern within an Activity Theoretical approach to software evaluation.

Software tools are treated by AT in much the same way as any other tool. Tools are important in AT because they are seen to mediate how people are able to make some change to an object in attempting to attain a goal. A software tool may be evaluated as possessing high usability if it allows a user to achieve their desired goal without having to substantially reformulate their approach to the goal.

AT-based evaluation tools and approaches such as the "Activity Checklist" (Kaptelinin, Nardi & Macaulay, 1999) have been proposed for human-computer interaction (HCI) design and evaluation contexts. Evaluation tools such as checklists and questionnaires arguably suffer from too great a degree of abstractedness when the aim is to analyse the situated cognitive and behavioural processes a user exhibits in employing novel software and hardware tools. Video analysis has been adopted as a means for developing detailed and situated analyses of activity in real world and laboratory settings.

Bødker's (1996) research in AT-based video analysis of HCI activity developed the concept of breakdowns and focus-shifts as indicators of tool-use events potentially entailing usability problems. A breakdown is where work activity is interrupted by unexpected or inappropriate tool behaviour. A focus-shift is regarded as a change in the nature of the operations undertaken to work towards a goal. Focus-shifts occur when a person needs to change their mental focus from a relatively automatized and unconscious approach to working on an object, and needs to work at a more conscious and deliberate level. For example, focus-shifts occur when a software user needs to achieve a work-around, because the original approach does not seem to be satisfactorily meeting the goal of the activity.

The concept of studying breakdowns and focus-shifts in video analysis of HCI activity has been further extended from Cultural-Historical AT, as employed by Bødker, to one based in the Systemic-Structural Theory of Activity (SSTA) (Harris, 2004). SSTA draws more directly on Russian applied ergonomic research, and places a greater emphasis on the psychological aspects of the subject undertaking activity. This shift in emphasis, while not disregarding the utility of an understanding of the cultural-historical context of tool mediated activity, arguably has advantages in developing tool design to support the user.

SSTA proposes that as human activity unfolds, the subject of that activity adjusts their strategies and goals in response to the comparison of the desired goal and the emergent result. Human activity is argued to be an integrated and logically ordered system of motor and cognitive actions directed by mechanisms of self-regulation (Bedny, Seglin & Meister, 2000). The approach to the analysis of human activity reflects the integrated and logically ordered nature of activity (Bedny & Karwowski, 2007; Bedny & Meister, 1997).

SSTA-based analysis of video data generally relies on repeated observation and the development of integrated analyses of the recorded events (Harris, 2004). SSTA makes use of three interrelated methods: (1) the parametrical, which focuses on the parameters of the activity through techniques such as breakdown and focus-shift analyses; (2) the

morphological, which relies on the description of structure of activity in terms of discrete actions and operations; and (3) the functional, which focuses on describing the nature of self-regulation processes.

In this study, these three approaches were applied in the first two stages of SSTA analysis: (1) qualitiative description; (2) algorithmic analysis. These stages involve setting out a qualitative description of the events under analysis, and then constructing a step-by-step analysis of discrete motor and cognitive actions and operations. These analyses are described in Section 2.2.5.2.

2. Method

2.1 Participants

The participants for this usability study were two experienced RAAF ABM instructors (41 Wing) and one experienced Wedgetail instructor (42 Wing). All participants had extensive debriefing experience in providing AARs in their respective Wings. One ABM assessor and one Wedgetail assessor had no experience using AWAR for assessment or providing AARs. A DSTO-based ABM assessor had extensive experience using of AWAR and had a development role in the creation of the AWAR hierarchy for EBS12.

2.2 Apparatus and Procedure

2.2.1 AWAR hardware

The AWAR toolset has two modes: (1) an "Assess" mode to aid the assessment of the training audience; and (2) the "Review" mode, to aid the subsequent AAR. The assessors made use of the AWAR assessment toolset via standard desktop computers, including a mouse and keyboard and two monitors, as per Figure 3. The workstation was placed in the vicinity of the training audience being assessed. AWAR was used on one of the two workstation screens, with the other screen used for mission related displays.

The assessors made use of the AWAR review toolset via two Smartboard PG-363 touchscreens. As can be seen in Figure 4, the screen on the right of the AAR room controlled the AWAR review software. The assessor was able to touch the screen in order to select AWAR review content and functionality. The screen on the left of the room presented video recordings of training audience 'scope' content. The scope content was a recording of each ABM or Wedgetail tactical display. The assessor was able to select and expand the recorded training audience scope video in order to present feedback.

2.2.2 Participant familiarisation with AWAR

No formal training in AWAR was conducted prior to EBS12 for the two assessors who were not involved in the development of AWAR. Instead, the assessors new to AWAR were given an informal general orientation to the software prior to their first use. The assessors who were inexperienced with AWAR were also supplied with the AWAR criteria hierarchy some weeks prior to the beginning of EBS12 for familiarisation. Both these assessors stated that they had not had time to adequately engage with the hierarchy content before arrival at EBS12.

2.2.3 Data Collection Overview

SUS questionnaire data were collected for the assess function of AWAR. Video data was captured for both the assessment and review functions. Open-ended feedback sheets, known as the "Top 3, Bottom 3" collected the three best and worst general features of the simulation experience of the day, and provided the assessors with the opportunity to provide general feedback on the AWAR as they saw fit.

2.2.3.1 Administration of the System Usability Scale (SUS) and Feedback Sheets

The SUS questionnaire was included in the EBS12 Questionnaire Booklet and completed as part of the broader EBS12 data gathering session after a mission. Further feedback on the AWAR tools could also be volunteered by the assessors in the daily "Top 3, Bottom 3" feedback comments sheet included with the data collection booklets. A non-interventional approach was taken to usability testing. Participants were to interact with the software under normal simulation exercise pressures. The SUS was administered as part of the wider program of twice-daily measurement sessions. The SUS was administered during the second and fourth days' afternoon assessment sessions to gather subjective usability perspectives. This was done to assess usability after only a short period of usage and once again after the users had slightly more experience.

2.2.3.2 Collection of Video Data

Assessor participants were filmed using Canon XA10 professional video cameras. Video data was recorded on 16-Gb Secure Digital memory cards (SD cards).

Video of each of 9 assessment sessions was recorded. A single tripod-mounted camera was placed slightly behind and to one side of each assessor workstation. A representation of the camera placement is shown in Figure 3. This enabled the researchers to view which onscreen tools the assessors were attending to, as well as whether the assessors were using other tools such as notepads or discussion with other whiteforce participants in order to undertake their role. Audio data was recorded via the camera's internal microphone. Screen capture of the assessors' AWAR screens was also recorded using Techsmith 'Camtasia' (Version 8.0) screen capture software. This enabled the researchers to view each assessor's use of the AWAR tool. The screen capture and camera video and audio recordings were analysed to observe the use of the software in the exercise context.

Figure 3. Camera placement at assessor workstation.

Each of 4 AAR sessions using the AWAR review tools were recorded. A representation of the AAR room layout and equipment placement is shown in Figure 4. Three cameras were positioned around the AAR room. One camera was focussed from a gantry over the seated training audience on the AWAR review screen. This camera also recorded audio data from a directional microphone attached separately to the gantry to gather assessor speech from the AWAR review screen area. A second camera was placed at the back of the AAR room to take in the assessor, the AAR display screens, a whiteboard and a rear-view of the training audience. This second camera recorded audio data via a non-directional microphone hung over the training audience from the gantry. A third camera was placed at the front of the AAR room to one side in order to capture the responses of the training audience. This camera also recorded audio data from two directional microphones placed at each front corner of the AAR room to capture training audience verbal responses.

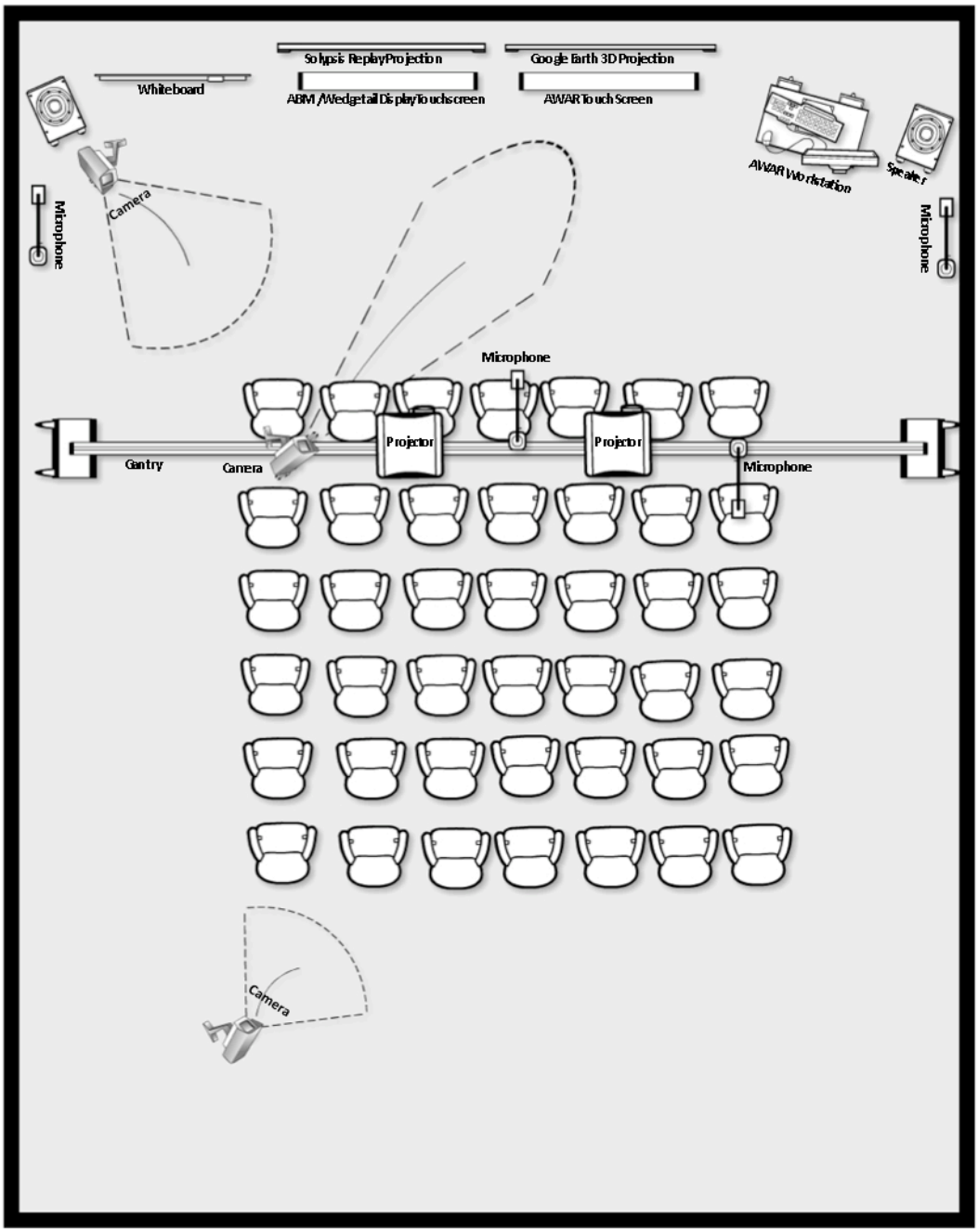

Figure 4. AAR room layout and equipment placement.

Video material was transferred to an Apple Mac Pro (OS 10.6.7) computer and edited using Final Cut Pro 7. Video data from the assessment sessions was edited to create a synchronised side-by-side "picture-in-picture" presentation of the assessor at their workstation next to the screen capture of their use of the AWAR assessment tools for each assessment session. Video data from the three cameras in the AAR sessions was edited to

create a synchronised "picture-in-picture" presentation of the three active video camera views.

2.2.4 Analysis of Video Data

2.2.4.1 User Testing Approach

Video analysis was conducted in three ways to identify when users made errors or had any difficulty with the AWAR software. Firstly, screen capture (including mouse movement and typing) was analysed to determine problems the users had during direct interaction with the software. Secondly, video of the user in situ at their workstation was analysed to identify problems users had when interacting with the entire system, including their dual monitors, keyboard, mouse, and any other tools they used during the exercise. Thirdly, conversational references to AWAR, and verbal expressions of satisfaction or frustration while using AWAR, were captured with audio data recorded during the exercise.

2.2.4.2 Systemic-Structural Theory of Activity Approach

The SSTA analysis of video data was undertaken in three stages: (1) qualitative description was used to develop the parametrical analysis of each AWAR assessment and AAR review session (description of events, particularly breakdowns and focus shifts); (2) a morphological description was developed through the algorithmic analysis of behavioural and cognitive actions (step-wise description of motor and cognitive actions); (3) a functional analysis description of the macro-level processes of self-regulation was developed by comparing the morphological description of actions and the parametrical description of breakdowns and focus-shifts. This functional analysis was aimed at identifying the nature of the evaluation of goal conditions that impact on the formation of task strategy.

In undertaking the parametrical method, the video of the assessment and review AWAR functions was initially viewed with the DSTO-based assessor in order to establish: (1) a narrative of the sequence of events; (2) an understanding of the procedures and jargon used by the assessors; (3) the identification of any less obvious breakdowns and focus-shifts that might slip the attention of a non-expert observing a specialised field of activity. A second viewing without the DSTO-based assessor was undertaken in order to refine these observations.

This parametrical analysis was developed with reference to what was known about the relevant assessor-role and assessment tool-use experience of each of the individual assessor participants. This analysis included the level of experience the assessor had in using tools like AWAR, what other approaches to assessment they had, as well as their general ABM/Wedgetail domain instructor background. This individual-psychological background of each participant was used as an aid in contextualising their assessment and AAR-related behaviour.

Transcriptions of the video were developed that supported the development of stepwise morphological-algorithmic descriptions of motor and cognitive actions (See Tables 4 & 5). These algorithmic descriptions of motor and cognitive actions were informed by the

breakdown and focus-shift observations drawn from the initial parametrical analysis. This allowed functional analysis descriptions of the self-regulatory processes to be developed. Functional analysis development entailed: (1) analyzing the relationship between the goals of each cognitive or motor action in relation to the broader assessment or review task-goal; (2) analyzing the utility of the input information arising from the AWAR toolset, particularly where a breakdown or focus-shift occurred; (3) identifying what issues are relevant to the evaluation of goal conditions and their impact on task strategy.

3. Results

The results of the usability evaluation of the AWAR software are divided into two sections. The first section pertains to the SUS questionnaire and the daily "Top 3/Bottom 3" feedback booklet comments. The second section pertains to observations made from the video recordings from both the User Testing and SSTA approaches on both the assessment and review AWAR functions.

3.1 User Testing: System Usability Scale Questionnaire & Top 3/Bottom 3 User Comments

3.1.1 System Usability Scale overall score

As this is the first time the SUS has been used to assess the usability of AWAR, the battery of SUS scores collected during EBS12 may not be compared with any previous version's scores, and hence must be analysed in isolation. The SUS creator, Brooke (1996), has not advised how to use SUS as a global rating scale in isolation of comparative scores, so it not feasible to assess the SUS scores properly without performing a comparative study on a later version.

This problem is solved in part by the advice of Bangor, Kortum and Miller (2008), who, on assessing the SUS scores of software used in a wide variety of contexts and applications, determined the approximate grading scale in Table 1.

Table 1. Software assessment based on SUS rating.

Score Range	Rating
90+	Superior products with high maturity
High 70's to High 80's	Better
Low-Mid 70's	Passable
Less than 70	Continued development and improvement of software essential; inferior at best.

As each score attained by AWAR – both for the Assess function and the Review function – is less than 70, we may conclude that this system is in need of further development. This

further development may take a number of different directions, as detailed in the AWAR design recommendations (Section 4.1).

3.1.2 System Usability Scale participant scores

The overall SUS score for the 'Assess' and 'Review' function for each of the two SUS scale administrations was calculated for each participant. The responses for the Assess function ranged from 55.00 to 65.00. The responses for the Review function ranged from 57.50 to 65.00. The results are shown in Figures 5 and 6.

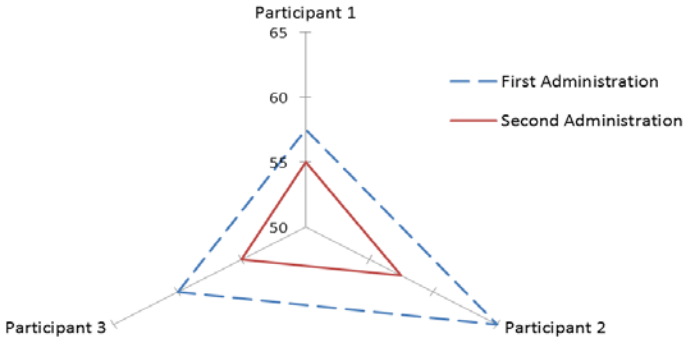

Figure 5. SUS scores for the AWAR 'Assess' for participants categorised by questionnaire administration.

As can be seen in Figure 5, each of the assessors' initial experience of the AWAR 'Assess' function degraded between the first and second SUS administration. This was surprising given that Participant 2 had been involved in the development of the AWAR.

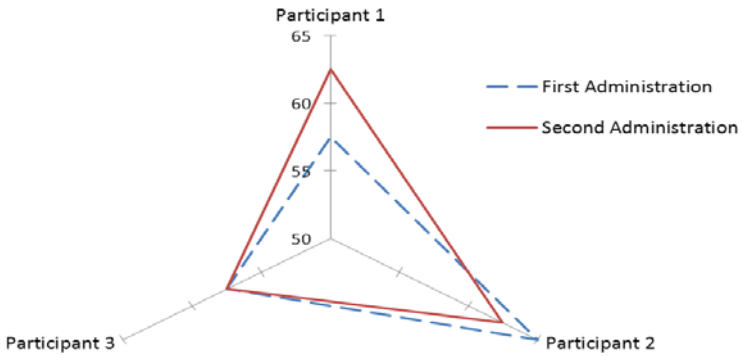

Figure 6. SUS scores for the AWAR 'Review' for participants categorised by questionnaire administration.

As can be seen in Figure 6, Participant 1's SUS score indicates an improved appreciation for the AWAR Review function, while the other participants' scores indicate that their appreciation diminished or remained stable.

3.1.2.1 Changes between first and second administrations in Assess item scores

Participant 1 felt on the early SUS administration that the Assess functionality of AWAR would not be learned quickly by most users, but late in the week they indicated that it would be quicker to learn than they previously thought.

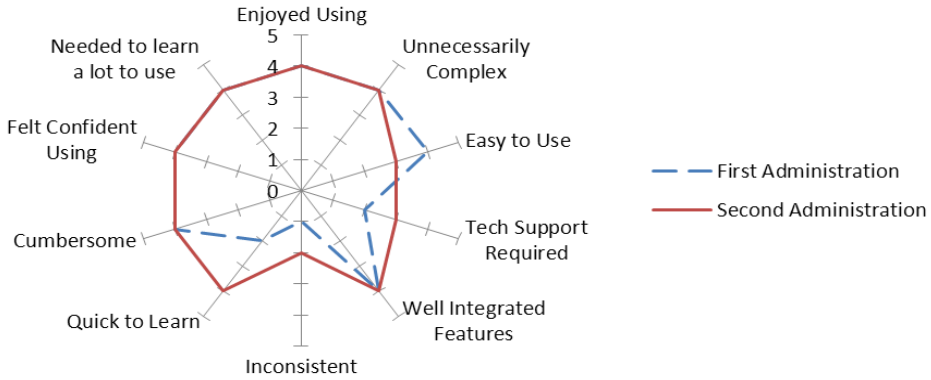

Figure 7. Change in SUS item scores for 'Assess' function between first and second administrations for Participant 1.

Participant 2 displayed the opposite change in opinion, in that they believed that the tool would be slower to learn than they originally thought. They also indicated on the early SUS administration that there was a considerable amount they needed to learn to use the software effectively, but after using it for a few days, their opinion changed and they felt they did not need to learn so much.

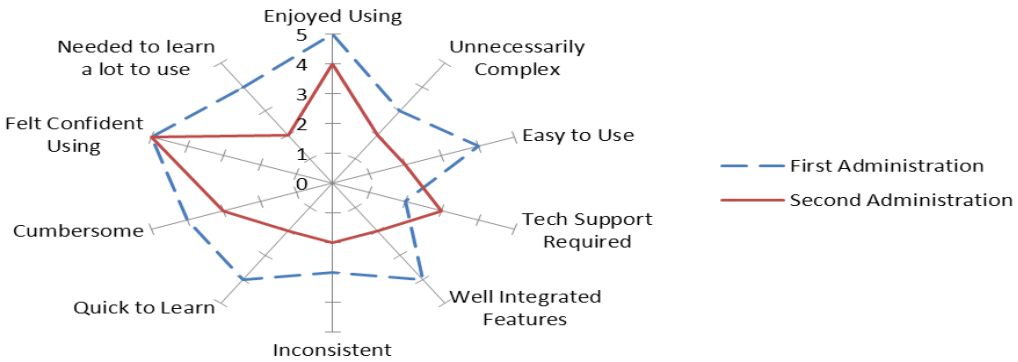

Figure 8. Change in SUS item scores for 'Assess' function between first and second administrations for Participant 2.

Participant 3 thought that the software was less complex at the late week SUS administration than after their initial exposure. However, their feeling of it being cumbersome to use also increased through the week.

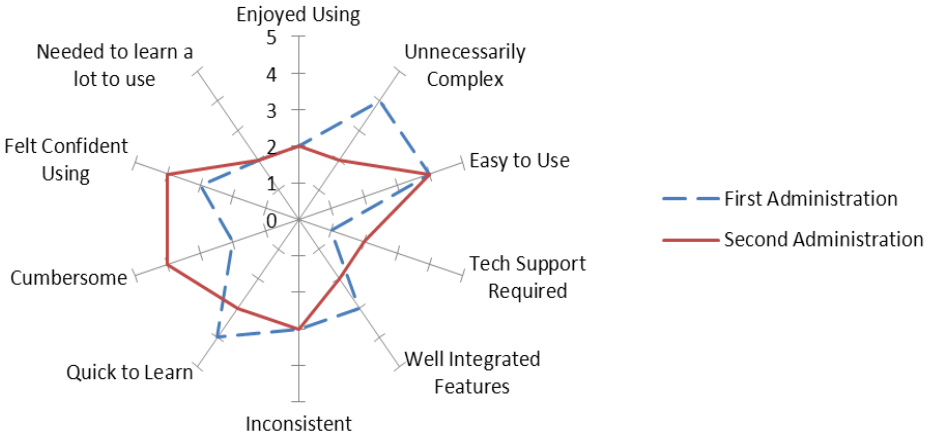

Figure 9. Change in SUS item scores for 'Assess' function between first and second administrations for Participant 3.

3.1.2.2 Changes between first and second administrations in Review item scores

Changes in attitudes towards the Review function also occurred in some areas. Participant 2 increasingly thought the software was unnecessarily complex through the week, and felt that the features of the system were not as well integrated as they had thought earlier. Participant 2 also felt late in the week that they would be more likely to require technical support than they did after initial use. Participant 1 and Participant 3 did not express any large difference in views between the first and second SUS administrations for any of the SUS items.

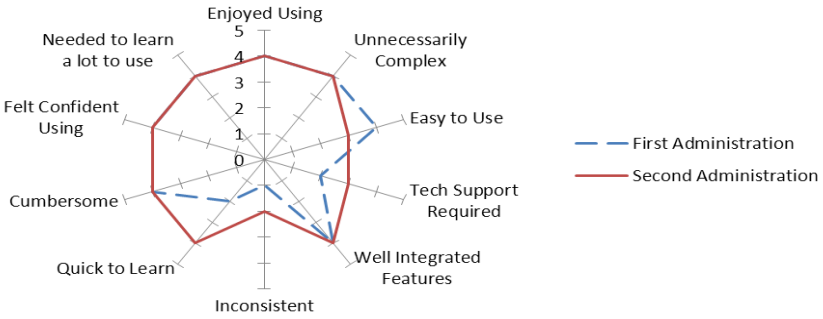

Figure 10. Change in SUS item scores for 'Review' function between first and second administrations for Participant 1.

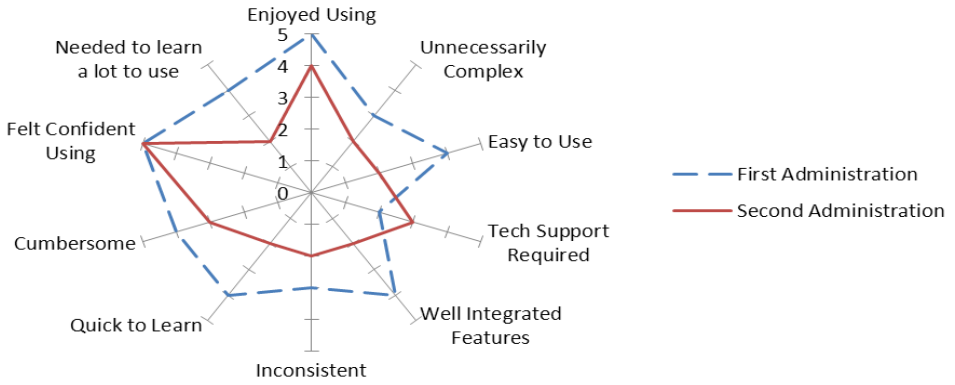

Figure 11. Change in SUS item scores for 'Review' function between first and second administrations for Participant 2.

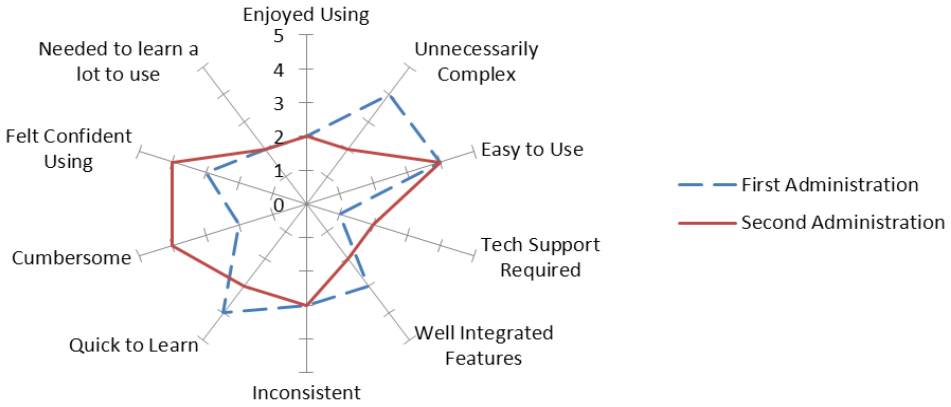

Figure 12. Change in SUS item scores for 'Review' function between first and second administrations for Participant 3.

3.1.3 Top 3/Bottom 3 User Comments

All AWAR-related comments were collated from the 'Top-3, Bottom-3' section of the daily questionnaires. These comments are listed in Table 2 and Table 3.

Table 2. 'Top-3' AWAR Comments.

Participant	Day	Comment
1	4	*"Being shown/given an example of how an AWAR [sic]"*
2	5	*"AWAR usage by assessors"*
2	5	*"AWAR software stability"*

Table 3. 'Bottom-3' AWAR Comments.

Participant	Day	Comment
1	2	*"A lot of information may have been missed by observers as the complex functionality of AWAR Assess drop-down menus takes extra time to find the correct slot to add the info to"*
1	2	*"Constantly altering the grading slider. The slider keeps changing after you click the correct gradient"*
2	2	*"AWAR usage 42WG (lack of)"*
2	2	*"AWAR procedure not refined"*
3	2	*"AWAR. Assessor still needs to interact with TAC display and comms. AWAR, whilst simple to use, is an unnecessary distraction. Early days but unclear of how AWAR is an improvement upon traditional methods. Data entry is not real-time – not achievable."*
3	3	*"AWAR data entry a distraction, particularly during periods of high workload."*
1	4	*"AWAR Assess has too many drop-down menus. Would be good to have less sublevels and potentially a hot-key that aligned to more frequented items (like comms/posture/execution)"*

A general comment was also made outside of the 'Top-3, Bottom-3' questionnaire, written at the bottom of the SUS questionnaire. It was as follows:

"Found AWAR a distraction – assessor MUST interact with display + comms. Replay (timestamp) not linked to my display (DP2A). Have similar functionality through Wedgetail Mission Support Segment. Not a real-time tool – hand written notes remain my preference. Structure needs to be reviewed – individual or team assessment tool? Not difficult to use but number of tech issues during replay (i.e. 2 min time offset, no FF or RW available on selection)."

3.2 SSTA and User Testing Video Recording Analysis

The results of the User Testing and SSTA analyses of the video footage of the Assess function were found to generally be highly convergent, and have been combined. However, the researcher applying User Testing video analysis felt that there wasn't enough use of the Review toolset to be able to make statements about future development of this aspect of the AWAR tool via video analysis. The video footage was still useful to the researcher applying SSTA analysis, the results of which are presented below.

3.2.1 Combined SSTA & User Testing Analysis of AWAR Assess Function

The results of the SSTA video analysis of the AWAR Assess function are presented in Table 4. The first two columns of the table present an algorithmic summary of the motor and cognitive steps undertaken in the use of the Assess function of AWAR. The third column presents the functional analysis of self-regulation processes in the form of the evaluation of goal conditions the assessor needs to make in order to progress assessment activity though the use of the AWAR tool. The fourth column represents points of the process where features of the tool breakdown or cause a focus-shift.

The nature of these 'breakdown' features is explained in greater detail after the summary in Table 4. The breakdowns are presented in temporal order as per Table 4, from an SSTA and User Testing perspective. Under-utilised AWAR assessment-tool features are presented after the breakdown material.

Table 4. Summary SSAT algorithmic and functional analysis of the AWAR Assess function.

Motor Algorithm	Cognitive Algorithm	Functional Analysis	Breakdown
	1. Monitor Solipsys screen and comms to identify ABM behaviour to assess.	a. Is the current behaviour something worth assessing?	
	2. Decide if whether the event maybe captured via AWAR and orient to AWAR tools.	b. Is the observed behaviour likely to have been included as assessable in the AWAR tool? Is it worth entering the assessment point now, or wait for a break or more significant event?	- (3.2.1.1) AWAR separate tool which takes attention away from air battle; particularly so when user has separate workstation tools. (3.2.1.2) Caused one assessor to revert to pen and paper.
3. Move mouse/cursor over AWAR categories to reveal relevant assessment criteria.	3. Identify a match between AWAR category and observed behavioural event.	c. Do the criteria options make sense in terms of the assessment I want to make?	- (3.2.1.3) Long search times indicate inability to find suitable criteria to explore and populate. Search task difficult, but (3.2.1.4) assessors also assessed some events across several criteria, which indicates a problem of meaningfulness/specificity of criteria.
4. Click left mouse button to open sub-category.	4. Identify a match between AWAR category and observed behavioural event.	e. Do the criteria options make sense in terms of the assessment I want to make?	- As Above
5. Click left mouse button to open sub-subcategory.	5. Identify a match between AWAR category and observed behavioural event.	e. Do the criteria options make sense in terms of the assessment I want to make?	- As Above
6. Move mouse/cursor to data entry window, click left mouse button on "+" (add data) button, to enter rating and comment.	6. Track opening of data entry window.	f. Has the tool responded reasonably?	- (3.2.1.5) Question over whether to force rating or comment first. Comments first may aid rating decision.
7. Move mouse/cursor along rating scale, and click left mouse button to select rating score. [Rating window transforms to an comment window].	7. Select an appropriate rating level to match the observation. Check that the rating level has selected appropriately.	g. Which rating level/descriptor represents my assessment of the observed behaviour? How essential is it to me to rate the observed behaviour accurately? Has the software done something to indicate the rating action has registered so I can enter a comment?	- (3.2.1.6) Rating level sometimes registered differently to the intended rating. - (3.2.1.7) To change rating requires change to time stamp which impacts debrief; may be judged as not worth the effort. - (3.2.1.4) Rating descriptor and number were seen as incongruent, impacts ratings.
8. Type comment in comment field. [Comment saved automatically]	8. Formulate and enter words for meaningful comment.	h. Do these words make sense to me and will they make sense to the training audience? Are they congruent with the selected rating score?	- (3.2.1.5) The act of typing likely concretises a position on the observed behaviour, informing a more reliable rating selection. - (3.2.1.8) Lack of feedback promotes "save session" behaviour.
9. Return to step 1.	9. Orient to new task.	i. What needs to be done next?	

Note: The motor and cognitive algorithm steps are denoted by numbers, with the corresponding functional analysis steps denoted with letters by way of distinction. Breakdown enumeration corresponds with the order discussion in the report below.

3.2.1.1 Physical positioning of AWAR relative to other software applications

One of the participants was provided with a completely separate computer on which to run AWAR, instead of having AWAR run on the same computer on a second monitor. This participant had to use an entirely different mouse and keyboard, and had to physically move his chair to the left to interact with AWAR. This may have contributed to

this participant's negative attitude towards AWAR. It also may have contributed to Participant 3's 'Bottom 3' comment on Day 3 about AWAR being a distraction (Table 3).

3.2.1.2 Persistent preference for pen and paper

The participant with the separate AWAR computer also displayed a persistent preference for using a pen and paper to make comments before entering them into AWAR, several comments at time. Unfortunately, this negated the benefit of time-stamping those comments and made referencing those events more difficult during the AAR.

3.2.1.3 Hierarchy searching

Often, between three and six hierarchy items (Figure 1) were clicked before the correct hierarchy items was found and a rating and comment was entered. Over the course of a mission assessment, a considerable amount of time was spent by the assessors attempting to find the hierarchy item they wished to make a comment and rating on. Participant 1 noted this issue on Day 2 and Day 4 in their 'Bottom 3' comments (Table 3).

3.2.1.4 Validity of scoring

There were two issues that arose from the review of video data that related to scoring of performance. These were: (1) the assessors were observed scoring and commenting on the same event across multiple criteria, due to uncertainty about which of hierarchy criteria best encapsulated the issue they wished to assess (this is related to the point in 3.2.1.3 above); and (2) the design of the assessment process and rating scale appeared to encourage a somewhat bimodal preference in scoring, with '2' and '4' being preferred over other rating options, as demonstrated in Figure 13.

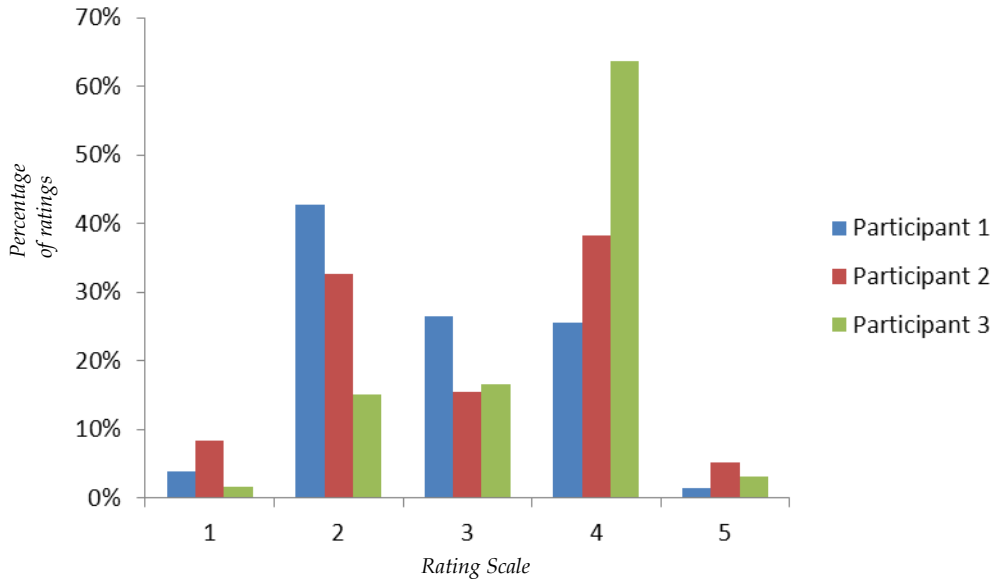

Figure 13. Bimodal preference in AWAR scoring.

A bimodal pattern in rating scores is understandable when the structure of the AWAR rating task is considered. The assessor is likely to be drawn to salient performance that is either better or worse than expected performance. However, instances of very poor and perfect performance are also likely to be rare. The current design of the software also requires a rating of performance so that a comment on performance can be entered against a hierarchy criterion. Therefore, in order to increase the sensitivity of the rating scale data, some consideration is required to assess whether there is a benefit in increasing the number of scale points.

There is also a risk of the assessor blindly entering a rating score in order to be able to enter a comment about an observed event. In conducting the assessment of performance, the assessor is also using AWAR to start the process of developing the material for the AAR. There may be times where an assessor would like a time-stamped comment to refer to during the AAR, that does not require or benefit from a rating score. An option to not record a score with every comment may also be beneficial to the quality of score data.

A comment made by one assessor also suggested that they felt rating scale descriptors were slightly incongruent with the numbers to which they corresponded. This may have also impacted the approach to rating.

3.2.1.5 Entering comments before ratings

AWAR forces the user to add a rating before a comment can be entered, and on a number of occasions this caused clear frustration for the user. It was observed that an assessor, under time pressure and having just seen an event to comment on, would click the 'add new' button and immediately start typing their comment. The user would then realise that they had to enter a rating first. The user had to decide what rating to give (though in these cases they had probably not yet considered a rating) and then retype the comment they had already entered but which had not been registered by the system. This kind of inflexibility can be very frustrating to users.

It is also likely that as an assessor types a comment, the process of putting their observations into word concretises their thinking about the situation (Flower & Hayes, 1981). Entering a rating either before or after entering a comment may impact the way in which the goal of entering the comment is developed. If the rating is entered first, a goal of entering the comment may be to maintain congruence relative to the rating. If the comment is developed first, the rating may be selected to be congruent with the comment. This is another argument for giving the user a choice as to whether they rate or comment first.

3.2.1.6 Rating slider malfunction

On a number of occasions, a user would select a rating (Appendix B. Figure B1.3), and when the comment box appeared (Appendix B. Figure B1.4), the incorrect rating had been recorded, and shown in the rating display on the right hand side of the comment box. It was observed that this occurred due to a slight lag between clicking the rating and the comment box appearing. During this lag, if the user moved their mouse cursor over another rating, the rating on which the cursor finally rested would be saved, not the rating that was clicked. In these cases, it was up to the assessor to notice that the incorrect rating

had been entered. Participant 1 noted their frustration with this problem in their 'Bottom 3' comment on Day 2 (Table 3).

3.2.1.7 Recoverability from error

The software did make it relatively easy to recover from error. For example, when a rating was to be changed, or the rating slider registered the incorrect rating, the user was able to recover quickly by simple clicking the rating number on the left side of the comment box and reselecting the rating they desired. The only issue with this process is that reselecting a rating also automatically updates the time-stamp for that rating, which may not be what the assessor desires depending on when the event that is being rated occurred.

3.2.1.8 Lack of feedback

A key feature that exists in many software applications that require user input is feedback of some sort that recognises that the user has completed an action. One of Nielsen's (1993) heuristics refers to system status visibility. Nielsen notes that the user should always be made aware, via feedback from the system, of what the system is doing (for example waiting for user input, processing a calculation). This may be in the form of a dialogue box confirming the action, or a status bar, for example.

AWAR lacks this feature, particularly in the comment/rating adding process (See Appendix B. Figures B1.1-1.4). The rating and comment is saved without the user having to confirm their input in any way. While this may seem like a positive feature (less action is required by the user to complete the action), it was observed to be confusing. Conventionally, programs require a user to click an 'add' or 'save' button to register input. One user was compelled to click the 'save session' button (Appendix B. Figure B2) each time they added a rating, and another user, repeatedly clicked the green 'add new' (Appendix B. Figure B1.2) button thinking that it would add their comment to the list of previously made comments.

3.2.1.9 Unutilised and underutilised functionality

A number of features were either unutilised or underutilised. The first was the 'Notes' section (Appendix B. Figure B3.1) that appears at the bottom of the comment editing panel. This feature exists so that assessors may create a drawing as a comment instead of describing the situation in words, and was implemented as a time saving device. However, this feature was not used once during the entire exercise. It is suspected that it would not be as time-saving a feature as was intended due to the difficulty of drawing on a computer using a standard mouse.

Another feature that was underutilised was the 'Left/Right' switch (Appendix B. Figure B3.2) which allows users to swap the hierarchy and comment panels. The only time this switch was used was when a user was exploring the software just prior to a mission. The swap in this case was not saved and the user switched it back to the original configuration immediately. This is a feature that may well be useful for some users, but it does not need to be on the main Assess screen. The position that it is currently in may contribute to screen clutter.

3.2.2 SSTA Analysis of AWAR Review function

The results of the SSTA video analysis of the AWAR Review function are presented in Table 5. The first two columns of the table present an algorithmic summary of the motor and cognitive steps undertaken in the use of the Review function of AWAR. The third column presents the functional analysis of self-regulation processes in the form of the evaluation of goal conditions the assessor needs to make in order to progress the AAR activity though the use of the AWAR tool. The fourth column represents points of the process where features of the tool breakdown or cause a focus-shift. The nature of these 'breakdown' features are also explained in temporal order from an SSTA perspective.

Table 5. Summary SSAT algorithmic and functional analysis of the AWAR Review function.

Motor Algorithm	Cognitive Algorithm	Functional Analysis	Breakdown
	1. Use notes for reminder as to which observation is to be discussed next.	a. Is the point reviewable via AWAR?	- (3.2.2.1) Some debrief points not seen as requiring video/audio playback.
	2. Search "review" overview screen for required debrief point.	b. Which is the debrief point to talk about?	- (3.2.2.2) Histogram view had few cues to ID debrief points because time synching was inaccurate and labelling was opaque.
3. Touch the identified debrief point on the AWAR "Review" screen.	3. Interpret outcome.	c. Did it open as expected? Is this the right debriefing point? Should access be replanned or the debrief point be abandoned on AWAR?	- (3.2.2.3) Button touches often failed on smartboard.
4. Touch the timestamp button to open the controls for the replay	4. Interpret outcome.	d. Did the controls open as expected? Should access be replanned or the debrief point be abandoned on AWAR?	- (3.2.2.3) Button touches often failed on smartboard.
(If discard, click "close button" and go back to step 2.)	5. Check the AWAR "replay" screens to ascertain whether the appropriate bookmarked material has been opened and is at the right point in time. If yes, go to step 8, if no examine whether the video needs to moved back or forward (step 6), or discarded.	e. Is the picture on the viewing screen a match for the mission point I'm looking to talk about? Does it need to go forward, back or do I need to go back and find another possible bookmark?	- (3.2.2.4) Bookmark times on AWAR/mission systems not synched, made it very difficult to use non-bookmarking methods to locate video material for replay. Negatively impacted on attitude to tool as a whole.
6. Press forward/backward buttons until video replay will start at the required point in the mission.	6. Interpret outcome.	f. Is the video moving towards what I remember of the mission scenario?	- (3.2.2.3) Button touches often failed on smartboard.
	7. Check the screen covers what is required. If not, go to step 8. If it does, continue to step 9.	g. Is everything of concern going to fit on the screen the way it is set up?	- (3.2.2.3) Button touches often failed on smartboard.
8. Use touchscreen or mouse to move the required screen into position or sizing.	8. Interpret outcome.	h. Is everything that is required here, and will the audience be able to see it?	- (3.2.2.3) Button touches often failed on smartboard.
9. Press play button on AWAR "review" screen to start video.	9. Interpret outcome.	i. Did it open as expected?	- (3.2.2.3) Button touches often failed on smartboard.
10. Introduce debrief point.	10. Introduce debrief point.	j. Does the training audience have enough contextual detail to make sense of this information?	- (3.2.2.5) Comms replay seen as too loud. Assessors attempted to talk over it, and then turned it off in frustration.
11. Watch, listen and discuss debrief material.	11. Watch, listen and discuss debrief material.	k. Is the presentation adequately demonstrating the debrief point?	- (3.2.2.6) Smartboard pens didn't work when required. - (3.2.2.5) Comms replay seen as too loud. Assessors attempted to talk over it, and then turned it off in frustration.
12. Press "close" button on AWAR review screen to close comment box.	12. Interpret outcome.	l. Has the material cleared so the next debrief point can be started or another task can commence?	- (3.2.2.3) Button touches often failed on smartboard.

Note: The motor and cognitive algorithm steps are denoted by numbers, with the corresponding functional analysis steps denoted with letters by way of distinction. Breakdown enumeration corresponds with the order discussion in the report below.

3.2.2.1 *Assessment of appropriateness of video review*

Not all the debrief points that assessors wanted to make were seen as worth reviewing via the AWAR review tools. The assessors decided that talking without reference to the replay could adequately cover many points. For these points the reviewers made an assessment about the value and importance of the possible replay relative to their own description to be transmitted to the training audience. These assessments were made where the goal was to communicate information that was assessed as fairly straightforward in nature.

3.2.2.2 *Opacity of bookmark search from histogram view*

For the assessors undertaking the review function of the AAR, the histogram view was often opened for them to refer to when selecting debrief points for discussion, as shown in Figure 14. The histogram view coupled to the Smartboard touchscreen functionality allowed the assessors to use their fingers to open a debrief point to replay and discuss. However, the histogram view had few cues to help the identification of the briefing points that the assessor wanted to talk about.

Figure 14. Histogram view of AWAR hierarchy criteria grouped by their associated rating score.

Assessors tended to make a judgment based on hierarchy label, rating and temporal order to select the appropriate debrief point. The assessors made paper notes to help them work out which debrief points to address, however the description they gave themselves to refer to was poorly supported by AWAR. Assessors tended to want to refer to time points for the selection of debrief points.

3.2.2.3 *Failure of Smartboard button touches*

An important breakdown with the use of the Smartboards was the fact that assessors often had trouble with button touches. Oftentimes when assessors went to select a button on the

Smartboard they would have to press it repeatedly before it would work. In future some calibration of the Smartboards would be in order to make them more reliable, if possible. When the assessor had trouble using the Smartboard they would have to make decisions about whether to persist in pressing the button to see if they could make it work, decide whether to just speak to their notes, or ask the assistant to use the mouse to click the button they required. Smartboard button touch failures are important to the self-regulation of the action because the success of the button touch indicates that the next step can begin, and eliminates a potential focus-shift.

3.2.2.4 Lack of synchronization of timing between AWAR & Wedgetail TAARDIS

The most important issue with regard to the review function of the AWAR review tools was the lack of synchronicity between the times that were bookmarked and the times that were instantiated on the Wedgetail TAARDIS recording system.

When the Wedgetail assessor opened a debrief point from the histogram display, he would have to refer to the replay display to try to confirm whether the point he had opened was the point he intended to open, and if so whether the video was at an appropriate point in time to play in order to deliver feedback.

When the video opened at a point that was not consistent with the assessor's feedback the assessor was forced to use the step-forward or step-back buttons to step the video in 30 second increments in order to find where video should have started from. The failure to have synchronised bookmark times negatively impacted on the strategy the Wedgetail assessor brought to utilizing the AWAR Review function, and prompted him to abandon AWAR and resort to paper and pencil to support his debrief presentation.

An informative and selectable timeline view of the summary page is recommended to help make use of temporal and ordinal cues to mission events.

3.2.2.5 Verbal channel clash: Assessor speech & comms replay

Once the assessors had started the video replay, the corresponding audio of the communications replay would start. Oftentimes the comms replay was far too loud to make sense of for the training audience, and was distracting for the assessor. The assessors tried to talk over but then turned it off in frustration. This is an example of competing verbal information in the auditory channel.

3.2.2.6 Failure of Smartboard pens

Another related issue to the Smartboard touch failure was the failure of the Smartboard pens to work when required. Functioning Smartboard pens would have been a very useful tool for the assessors to use while trying to illustrate spatial points directly on the workspace that the training audience use.

4. Discussion

This study aimed to identify the key usability issues that existed in the EBS12 version of AWAR software tool, as well as identify options for addressing these issues. As the AWAR is a test-bed for concepts related to assessment and feedback toolset design, the usability issues and design options presented in this report speak to our understanding of what characteristics are important in an effective toolset.

The usability issues were identified through two methods with different emphases. For the most part, the two methods hit upon the same types of issues with regard to instances where the assessors had trouble with the screen entities. The User Testing approach particularly emphasised issues like the placement of screen elements, and design principles such as the use of convention. The SSTA approach emphasised the how the design of the AWAR toolset, coupled with the goals and expectations the assessors brought to the AWAR assessment and review activities shaped their processes of self-regulation.

There are a few limitations to the study. The AWAR usability study was necessarily limited in the ability to conduct a large questionnaire-based quantitative study because of the targeted nature of the software tool. A less limited study may have allowed quantitative analysis comparing groups, sessions and experience levels via the SUS questionnaire instrument.

Another limitation was the relatively infrequent use of the AWAR review function, relative to EBS10. The developer-assessor participant in EBS12 was the assessor carrying out the review function in EBS10. The differential in willingness and ability to engage effectively with the AWAR toolset between EBS10 and EBS12 is at least partly explained by a lack of familiarity. The differential in familiarity makes for an interesting usability evaluation. The circumstances of the EBS12 evaluation are interesting because it allowed the observation of two novices to the tool, but not the domain. Such a study highlights how actions can be thwarted by a tool that is unfamiliar and relatively unintuitive, and how these difficulties in turn impact on user acceptance.

4.1 AWAR design recommendations

The Assess and Review functions of the AWAR have a number of design features that adversely impact on usability and thereby the acceptability of the toolset. The SSTA analysis suggests that the most critical adverse design features happen to be the ones that are key to beginning the task of either documenting an assessment of training audience behaviour, or utilising the video playback tool in supporting a discussion about an event in the mission under review. Where these early actions became difficult, the assessor often tried either a different method to achieve the goal of the action, which often meant not using the AWAR tool, or abandoned the action or task altogether.

4.1.1 Reconsider hierarchy structure

There is a considerable problem with the current hierarchy configuration in AWAR. It is an issue that needs to be addressed as a matter of priority. To be truly proficient with the AWAR Assess functionality – to be able to find a specific hierarchy item with a minimum of clicks – a user must be intimately familiar with the hierarchy that is being used for performance assessment. Users were not familiar with the hierarchy and had to engage in manual searching – clicking on several different categories and sub-categories - to find specific items. The EBS12 hierarchy required an assessor to remember the contents and location of 78 hierarchy items – a difficult feat even for a very experienced user.

Extensive hierarchy searching was also observed in Black Skies 2008 (Tracey, et al, 2009). A deliberate design choice was made during the development of the EBS08 version of AWAR to prevent users from expanding more than one hierarchy category at once. This decision was made in an effort to reduce distraction from other hierarchy items when entering scores and comments. However, this design increases the number of clicks required to find a hierarchy item, and may result in searching a sub-category multiple times. Ultimately, it may prevent users from finding an item efficiently, and may compound time pressure in a high workload situation.

There was also some confusion about whether the hierarchy was an individual or team assessment tool. The assessors who were new to AWAR both commented that it took some time to adjust their assessment mind-set from assessing individuals to assessing a team.

Revising the design approach to the development of the hierarchy may allow the broad domain of ABM/Wedgetail operations to be summarised more succinctly while maintaining or improving the meaningfulness of the descriptions. This revision may place greater emphasis on cueing the instructor to assess team rather than individual performance.

The use of a temporally-based structure of the hierarchy is likely to make the search process easier than the more abstract 'means-ends' based approach of the EBS12 hierarchy. The structure of the hierarchy would benefit from being flattened, so that there is a minimum of levels to be searched.

Options for the future design approach of the hierarchy include: (1) a reformulation of the current content into a more temporally-structured version with fewer levels visible to the assessor; (2) the introduction of another approach to the development of the hierarchy, such as the Mission Essential Competencies (MECs; Symons, France, Bell & Bennett; 2006) or a hierarchy based on a SSTA analysis of mission phases (Bedny & Karwowski, 2007).

The MECs approach has been applied with some success in the United States Air Force (USAF) collective training program. However, it is a large proprietary undertaking that likely entails considerable cost.

Applying a SSTA analysis would represent a middle-ground approach. A SSTA analysis could likely make use of the current content to help begin the development of the new

content. This type of application of SSTA in the air combat command and control domain appears to be novel in terms of published work, but entirely within the capability of the approach.

4.1.2 Time synchronisation & identifying relevant bookmarks

In traditional AARs large sections of video, if not the entire mission, are often played through. This provides a great deal of context to cue the assessor about what is important to talk about. Assessors usually have handwritten notes with time-points that are seen as being particularly important to review.

Assessors trying to use the AWAR review toolset often had difficulty with finding the point they wanted to speak to.

Assessors attempting to use the EBS12 version of the AWAR Review toolset had two strategies for finding the snippets of video they wished to talk about. The first was to use the summary histogram page to select a feedback point. The histogram view had few contextual cues for selecting particular points, so assessors attempted to revert to the time-points they noted while assessing. However, this second strategy was problematic for the Wedgetail assessor due to the Wedgetail version of TAARDIS not being synchronised with the rest of the system. This lack of synchronisation caused Wedgetail video playback to be substantially out of alignment with the assessor discussion topic.

A key suggestion for future implementation of the AWAR tool is to: (1) synchronise the timing between all elements of the AWAR system, and (2) make use of temporal cues by employing a timeline depiction of the mission with assessor comments marked along it. This will allow assessors who want to use the context of the mission to more easily navigate and communicate the context of their commentary.

4.1.3 Add flexibility in time-stamping

The AWAR Assess function should be designed to help the assessor prepare their debrief as they assess. However, the confined structure of the bookmarking process did not foresee how assessors would structure their rating/comment workflow.

At times, it was noticed that users would add comments when there was a lull in activity, or in relation to an event that had not just occurred. As time-stamping is critical for the AAR as it was envisioned for EBS12, it may be useful to give users the option of time-stamping their ratings and comments further in the past than the default two minutes. A time-stamping addition to the proposed comment and rating dialogue box is shown in Appendix C. Figure C1.2.

4.1.4 Address Smartboard function problems

The Smartboard affords intuitive functionality to the assessors in conducting the AAR. However, the screen button presses, attempts to select and move objects as well as the attempted use of Smartboard pen marking replay of ABM/Wedgetail operator screens

often failed and caused the assessor to rely largely on verbal description, and sometimes air-gesticulation to make their point. Failed screen button presses insert focus-shifts from describing a debrief point, to trying to make the display work. The failure of the Smartboard pens also robs the assessor of a planned strategy to illustrate an AAR point, forcing them to reassess how to best present the information given the conditions afforded to them.

Future use of Smartboard technology should, wherever possible, ensure that these pieces of technology are calibrated and activated to work as intended.

4.1.5 Allow comments without ratings and allow user to add comment before rating

As mentioned in Section 3.2.1.5, forcing a user to enter a rating before a comment was a source of frustration. The assessor may observe an event and wish to timestamp it and take note of what happened quickly before they forget or another event that requires their attention occurs. The system currently prevents commenting before rating, though there is no reason why a rating for the event could not be given once the assessor has had more time to consider it. If comments were permitted before rating, the ratings that would be given would be more considered and therefore more useful as a metric for performance. A sample layout for a dialogue box that might be used as an alternative for entering a rating and comment is shown in Appendix C. Figure C1.1.

4.1.6 Maximise the practical integration of AWAR Assess function

Participant 3 was forced to use AWAR on a separate computer. Users must have AWAR on the same PC as the rest of their displays. The physical demands of having to move between two computers, keyboards and mouses are too great, particularly when assessors are expected to respond to up to 78 hierarchy items. This may have negatively impacted this participant's ability and willingness to use AWAR as intended, and in turn their overall perception of AWAR's benefit.

However, the concept of integration could be taken further with future iterations of AWAR. An idea presented by another participant was to design AWAR as an overlay on the Solipsys screen, so that visual attention did not have to be redirected entirely to the second screen. This would require a major redesign of AWAR and maybe a goal to pursue in the longer term.

4.1.7 Give the user the option of scoring with finer granularity

As noted in Section 3.2.1.4, there was a clear preference for assessors to award ratings of '2' or '4'. It seemed that the scores of '1' and '5' were only awarded for fatal errors and perfect performance. It is not clear at this stage why this was the case. So that assessors may choose to capture more detailed information about performance, the AWAR design should consider using a finer rating scale, for example 1 to 7, with appropriate anchors. To compare scores across missions where different scales were used, scores may be subjected to an appropriate normalization procedure.

4.1.8 Design in recoverability by providing more exits

Currently, AWAR only allows a user to delete a rating after the rating has been specified, once the rating scale has changed into the comment editing box (refer to the transition shown in Appendix B. Figures B1.3 and B1.4. The lack of a 'delete' button [a small red circle with a cross] can be seen in Figure B1.4). This is poorly placed in terms of the workflow sequence. A user who has accidentally selected an incorrect hierarchy item, will be unable to delete that rating until they have actually specified a rating. Instead of clicking a single 'delete' button (one click), the user must select a rating that is meaningless, click it, wait for the comment box to appear, find the delete button on the comment box, and click it (two clicks plus reaction and waiting time). This defies Nielsen's (1993) principle of always providing the user with clearly marked exits and the ability to recover from error. Aside from being frustrating to the assessor, there are other implications. In a high workload scenario, this may be a stressor and distraction to the user, and wastes time.

4.1.9 Reposition the 'Add New' button closer to the hierarchy

The position of the 'add new' button is inconvenient at present. When adding a rating, the assessor must click on the hierarchy item on the left panel, and then move their mouse to the far right of the right panel. This process is the same every time a rating and comment are added. It would make more sense to position the 'add new' button closer to the hierarchy to increase efficiency.

4.1.10 Provide structured training before an exercise starts

Ensure appropriate levels of training are provided prior to an exercise. Attempting to learn a new system with the addition of time pressure will increase the likelihood that some features of the software are never discovered, explored or used and that the user will default to their old 'pen-and-paper' method of recording comments. Instruction is equally important with the review function, so that the assessors can get some experience in the incorporating the visual and auditory elements effectively.

4.1.11 Adopt conformity with convention where possible

One of Nielsen's (1993) heuristics suggests that developers design systems according to the standards and conventions of their platforms. This ensures that users can navigate the system and perform standard actions (such as saving and creating new files) more easily than if they were learning a new system from the beginning. AWAR was designed for use on the Windows operating system, and as such, design choices should have been made according to the conventions of Windows. An example of such a convention is the presence of drop-down menus at the top left of the program. In most Windows programs, the menus will generally read 'File' and 'Edit', with software-specific menus following to the right, for example 'View', 'Insert' and 'Format' as seen in Microsoft Word. Having these menus means that users may always have a familiar location to refer to when they wish to perform actions, even if shortcut buttons are available. At this stage AWAR does not have these menus, instead only providing users with shortcut buttons.

Menus also provide a location for infrequently used software features that should not really be available on the main screen, as discussed in below.

4.1.12 Hide less-utilised features in menus

A standard menu is an appropriate to place features of the software that do not need to be on the main screen, such as those features that are rarely used. For example, the switch that allows users to swap the hierarchy and comment panels (Appendix B. Figure B3.2), is rarely used, but an important feature, and would be best placed in an 'Options' menu or similar. Only information that is required at the time should be displayed. The most frequently used items should be displayed in the main screen, while lesser-used features should be relegated to secondary screens or menus. Hiding infrequently used features conforms to Nielsen's (1993) heuristic of interface simplification (or using the "Less is More" principle) to reduce learning requirements, the likelihood of misunderstanding, and time spent searching for desired functionality.

4.1.13 Explore design features for reducing the typing burden

Developers cannot make assumption of typing proficiency, particularly when the software is intended to be used under time pressure. In a time-constrained environment, it may be impossible to capture sufficiently detailed feedback in a typed format. Examples of solutions that may be explored to overcome limitations in typing skill include speech transcription, voice recording, and general statement summary buttons, and the use of meaningful icons.

5. Acknowledgments

The authors would like to thank FLTLT Joshua Chalmers (Surveillance & Control Training Unit) for his subject matter expertise and assistance in coming to grips with the specialised material, and Dr. Christopher Best for his comment on earlier drafts of this report.

6. References

Bangor, A., Kortum, P.T. & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale, *International Journal of Human-Computer Interaction, 24(6),* 574-594.

Bedny. G .Z. & Karwowski, W. (2007). *A Systemic-Structural Theory of Activity: Applications to Human Performance and Work Design.* Boca Raton, Fl: CRC Press.

Bedny, G. Z., & Meister, D. (1997). *The Russian Theory of Activity: Current Applications to Design and Learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bedny, G. Z., Seglin, M. H., & Meister, D. (2000). Activity theory: history, research and application. *Theoretical Issues in Ergonomics Science, 1*(2), 168-206.

Bødker, S. (1996). Applying Activity Theory to Video Analysis: How to make Sense of Video Data in Human-Computer Interaction. In B. M. Nardi (Ed.), *Context and Consciousness: Activity Theory and Human-Computer Interaction* (pp. 147-174). Cambridge, MA: MIT Press.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London: Taylor & Francis.

Chin, J. P., Diehl, V. A. & Norman, K. L. (1988). "Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 213-218.

Flower, L. & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition & Communication,* Vol. 32(4), pp. 365-387.

Goldberg, S.L. and Meliza, L.L. (1993). Assessing unit performance in distributive interactive simulations: The Unit Performance Assessment System (UPAS). *Proceedings of NATO Defence Research Group Meeting, Panel 8 (Defence Applications of Human and Bio-Medical Sciences). Training Strategies for Networked Simulation and Gaming). Technical Proceedings AC/243 (Panel 8) TN/5*, 173-182.

Harris, S. R. (2004). Systemic-Structural Activity Analysis of HCI Video Data. In O. W. Bertelsen & M. Korpela & A. Mursu (Eds.), *Proceedings of ATIT04: 1st International Workshop on Activity Theory Based Practical Methods for IT-Design* (pp. 48- 63), September 2-3, 2004, Copenhagen, Denmark. Aarhus: DAIMI.

Hasenbosch, S. & Best, C. (2007). *A Hierarchical Analysis of Air Battle Management Team Goals in the Defensive Counter Air Mission.* DSTO-TN-0781. Melbourne: Defence Science and Technology Organisation.

Kaptelinin, V. & Nardi, B. A. (2006). *Acting with Technology: Activity Theory and Interaction Design.* Cambridge, MA: MIT Press.

Kaptelinin, V., Nardi, B. A., & Macaulay, C. (1999). The Activity Checklist: A Tool for Representing the "Space" of Context. *Interactions, 6*(4), 27-39.

Lewis, J. R., (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7(1),* 57-58.

Nielsen, J. (1993). *Usability Engineering.* San Diego, CA: Academic Press,.

Pressman, R. S. (2005). *Software Engineering: A Practitioner's Approach.* New York: McGraw-Hill.

Symons, S., France, M., Bell, J. & Bennett, W. (2006). *Linking knowledge and skills to mission essential competency-based syllabus development for distributed mission operations* (AFRL-HE-AZ-TR-2006-0041). Mesa, AZ: US Air Force Research Laboratory, Human Effectiveness Directorate.

Tracey, E., Hasenbosch, S., Vince, J., Pope, D., Stott, A., Best, C., Shanahan, C., & Finch, M. (2009). Exercise Black Skies 2008: Enhancing Live Training Through Virtual Preparation – Part Two: An Evaluation of Tools and Techniques. DSTO-TR-2305. DSTO: Melbourne.

Vralic, L. (2003). Evaluating Usability in Context. In H. Hasan, E. Gould & I. Verenikina (Eds.) *Information systems and Activity Theory: Volume 3, Expanding the Horizon* (pp. 171-192). Wollongong: University of Wollongong Press.

This page is intentionally blank

Appendix A: System Usability Scale

A.1. SUS AWAR Assess Variant

AWAR Assess Function Usability Scale

You used the AWAR 'Assess' function during the VULs to assess and rate the teams. Below are a number of statements about the AWAR Assess function and its usability. Please mark on the scales the extent to which you agree or disagree with each statement.

		Strongly Disagree	Moderately Disagree	Neither Agree nor Disagree	Moderately Agree	Strongly Agree
1	I think that I would like to use the AWAR Assess function frequently					
2	I found the AWAR Assess function unnecessarily complex					
3	I thought the AWAR Assess function was easy to use					
4	I think that I would need the support of a technical person to be able to use the AWAR Assess function					
5	I found that the various functions in the AWAR Assess function were well integrated					
6	I thought there was too much inconsistency in the AWAR Assess function					
7	I would imagine that most people would learn to use the AWAR Assess function very quickly					
8	I found the AWAR Assess function very cumbersome to use					

9	I felt very confident using the AWAR Assess function				
10	I needed to learn a lot of things before I could get going with the AWAR Assess function				

Figure A1. SUS AWAR Assess Variant.

A.2. SUS AWAR Review Variant

AWAR Review Function Usability Scale

You used the AWAR 'Review' function to conduct AARs. Below are a number of statements about the AWAR Review function and its usability. Please mark on the scales the extent to which you agree or disagree with each statement.

		Strongly Disagree	Moderately Disagree	Neither Agree nor Disagree	Moderately Agree	Strongly Agree
1	I think that I would like to use the AWAR Review function frequently					
2	I found the AWAR Review function unnecessarily complex					
3	I thought the AWAR Review function was easy to use					
4	I think that I would need the support of a technical person to be able to use the AWAR Review function					
5	I found that the various functions in the AWAR Review function were well integrated					
6	I thought there was too much inconsistency in the AWAR Review function					
7	I would imagine that most people would learn to use the AWAR Review function very quickly					
8	I found the AWAR Review function very cumbersome to use					
9	I felt very confident using the AWAR Review function					

10	I needed to learn a lot of things before I could get going with the AWAR Review function				

Figure A2. SUS Review Variant.

Appendix B: AWAR Screenshots

B.1. AWAR Add rating and comment process

The following shows the steps that must be taken to add a rating and comment in the AWAR Assess function.

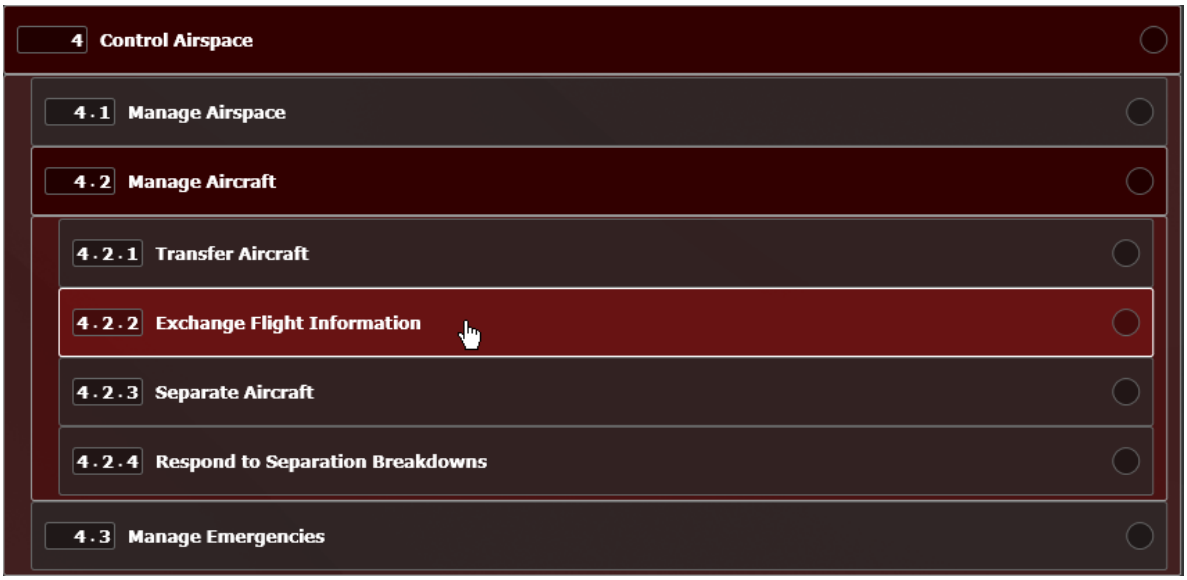

Figure B1.1. Step 1: Find and select criteria in hierarchy

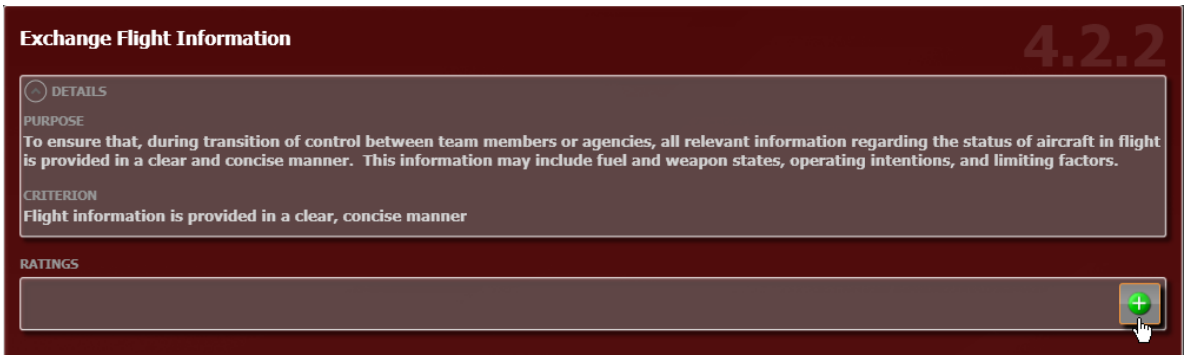

Figure B1.2. Step 2: Click the green 'add new' button. This criterion currently has no comments or ratings

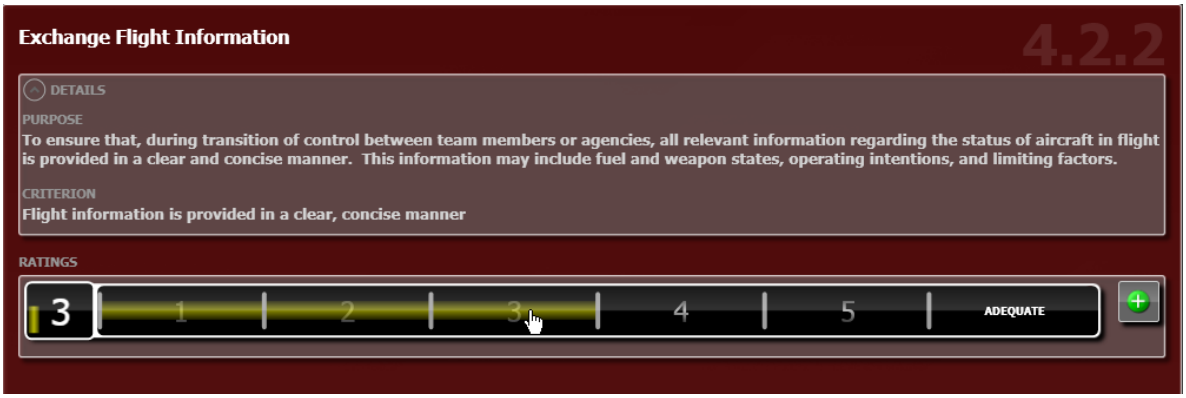

Figure B1.3. Step 3: The rating slider will appear. Select the rating for that criterion

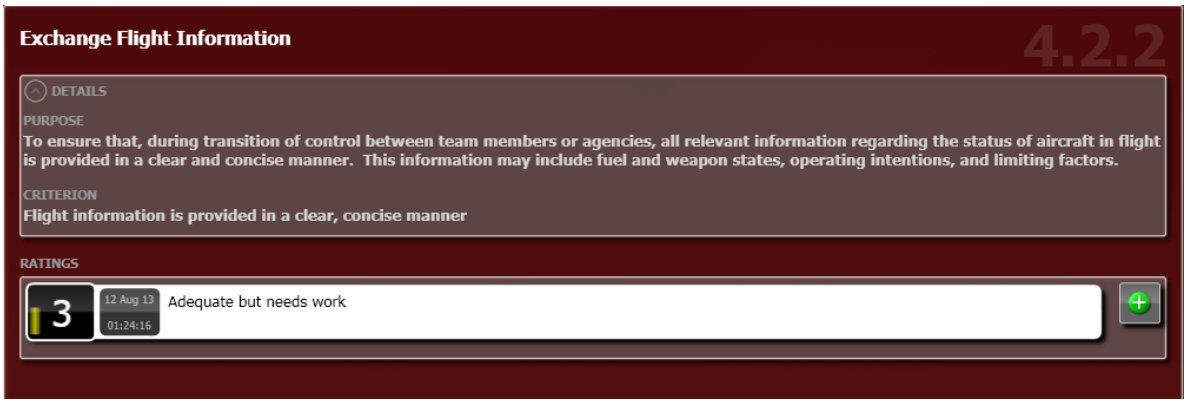

Figure B1.4. Step 4: The comment box will appear. Type your comment into the box

B.2. Save Session process

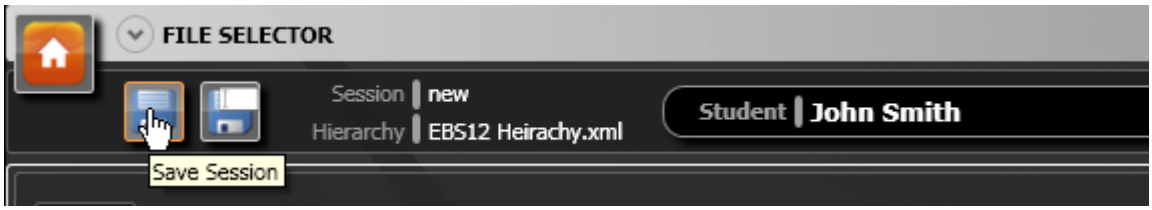

Figure B2. Save session button.

B.3. Infrequently utilised functionality

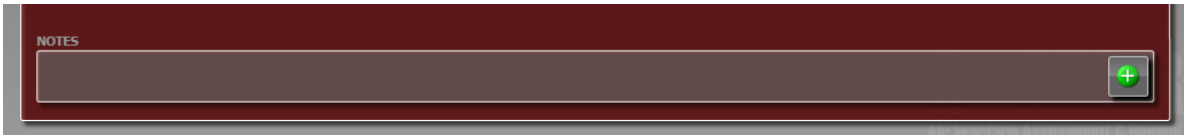

Figure B3.1. 'Notes' capability.

Figure B3.2. Left/Right pane orientation switch.

Appendix C: Suggested Modifications

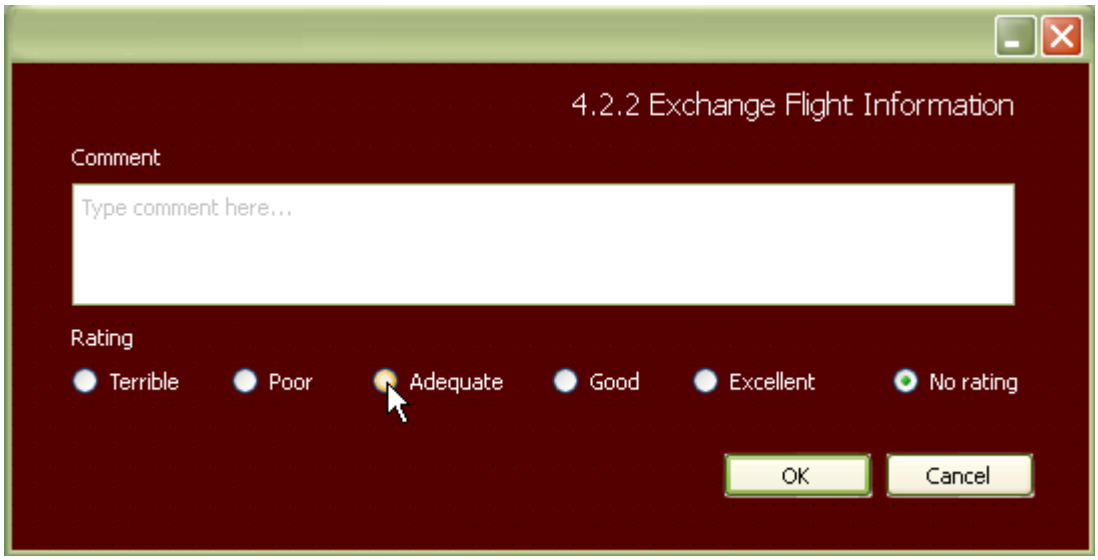

Figure C1.1. Suggested rating/comment dialogue.

Figure C1.2.1 Suggested rating/comment dialogue with time stamp selector.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. DLM/CAVEAT (OF DOCUMENT)

2. TITLE

Usability Evaluation of Air Warfare Assessment & Review Toolset in Exercise Black Skies 2012

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)

Julian Vince and Jessica Parker

5. CORPORATE AUTHOR

DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend Victoria 3207 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TR-2923	AR-015-817	Technical Report	December 2013

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2013/1163049/1	07/327	DGSP-AF	44	19

13. DSTO Publications Repository

http://dspace.dsto.defence.gov.au/dspace/

14. RELEASE AUTHORITY

Chief, Aerospace 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

After-Action Review, Training, Feedback, Collective Synthetic Training, User Testing, Activity Theory, Usability.

19. ABSTRACT

A usability evalution of a suite of after action review tools was undertaken as part of a synthetic collective training research exercise (Exercise Black Skies 2012). The suite of tools represent a test-bed for enquiry into what qualities and features are useful in after action review tools used in the collective synthetic training context. This report provides a description of the usability design issues observed in the use of the tool suite and proposes opportunities for future development. The study makes use of two complementary approaches to usability: User Testing and the Systemic-Structural Theory of Activity.