

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)		2. REPORT TYPE Final		3. DATES COVERED (from. . . to) May 2002 to November 2002	
4. TITLE AND SUBTITLE Measurement Methods for Human Performance in Command and Control Simulation Experiments				5a. CONTRACT OR GRANT NUMBER	
				5b. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) William R. Sanders (U.S. Army Research Institute for the Behavioral and Social Sciences)				5c. PROJECT NUMBER A790	
				5d. TASK NUMBER 211	
				5e. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: TAPC-ARI-IK 2423 Morande Street Fort Knox, KY 40121-5620				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: TAPC-ARI-IK 5001 Eisenhower Avenue Alexandria, VA 22333-5600				10. MONITOR ACRONYM ARI	
				11. MONITOR REPORT NUMBER Research Note 2003-11	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT (<i>Maximum 200 words</i>): The U.S. Army's proposed Future Combat System of Systems (FCS) will include automated Command and Control (C ²) capabilities that will allow tactical commanders, assisted by a small command group, to effectively lead a future force composed of large numbers of manned and robotic elements. This paper describes research conducted by the U.S. Army Research Institute (ARI) to develop measurement methods to enhance the existing Human Functional Analysis (HFA) approach (Sanders, Lickteig, 2002) for estimating human performance requirements associated with FCS C ² design concepts. Measurement techniques are demonstrated that can address C ² human performance requirements through the evaluation of verbal communications, Human-Computer Interaction (HCI) behavior events, and subjective survey data. Specifically, automated word count, and task-time estimation methods were applied to existing HFA data sets to provide estimates of the frequency and time duration of verbal communications for individual					
15. SUBJECT TERMS Future Combat System Human-Computer Interaction Command and Control Human Functional Analysis					
SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT Unlimited	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number) Dr. James W. Lussier 502-624-3450
16. REPORT Unclassifie	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified			

FOREWORD

Concept exploration and development research for the Army's transformation to the Future Combat System of Systems (FCS) is a key concern of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The Future Battlefield Conditions (FBC) Team of the Armored Forces Research Unit (AFRU) is conducting research to support the development of measures of human performance required for FCS command and control (C²) under work package (211) FUTURETRAIN: Techniques and Tools for Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C⁴ISR) Training of Future Brigade Combat Team Commanders and Staffs. This research also supports the Science & Technology Objective (STO) "Methods and Measures of Commander-Centric Training."

The U.S. Army's proposed FCS will be a networked force horizontally and vertically integrated from strategic to tactical level, to provide dominant situational understanding. The successful development of the FCS requires that a measurement approach be developed for estimating human performance requirements associated with FCS C² design concepts to address issues regarding performance success, workload, task allocation, and training development. The objective of this report is to demonstrate additional measurement techniques that can be applied to the existing Human Functional Analysis (HFA) approach to address C² human performance issues through the evaluation of verbal communications, Human-Computer Interaction (HCI) behavior events, and subjective survey data. Preliminary data gathered in a series of U.S. Army warfighter-in-the-loop battle simulation experiments were reanalyzed for the present research to develop new measurement methods that can be used to address FCS C² human performance issues.

The information provided in this report was developed as part of an effort to support the Defense Advanced Research Projects Agency (DARPA) investigation of FCS C² command group functional requirements. Results of this research are valuable to the U.S. Army and other organizations involved in conducting FCS C² research, and in developing automated training support systems for FCS C² leaders and staffs. The measures developed in this research are applicable to a wide range of current and future FCS C² systems. The results of this research were provided to the Program Manager FCS C², and briefed at the annual conference of the International Military Testing Association (IMTA) on 23 October 2002.

KATHLEEN A. QUINKERT
Acting Technical Director

MEASUREMENT METHODS FOR HUMAN PERFORMANCE IN COMMAND AND CONTROL SIMULATION EXPERIMENTS

EXECUTIVE SUMMARY

Research Requirement:

The U.S. Army's proposed Future Combat System of Systems (FCS) will be a networked force horizontally and vertically integrated from strategic to tactical level, to provide dominant situational understanding. Automated Command and Control (C²) capabilities in FCS units will allow tactical commanders, assisted by a small command group, to effectively lead a future force composed of large numbers of manned and robotic elements. This paper describes research conducted by the U.S. Army Research Institute (ARI) to develop measurement methods to enhance the existing Human Functional Analysis (HFA) approach (Sanders, Lickteig, 2002) for estimating human performance and cognitive workload requirements presented by FCS C² systems. A major shortcoming in current experimentation is that the frequency and time duration of verbal communications and Human-Computer Interaction (HCI) behaviors must be obtained through time consuming analysis and coding of video recordings. Due to time constraints, previous HFA assessments have not been able to provide estimates of the frequency and time duration of verbal communications for individual members of the FCS C² command group, and could not provide time duration estimates for all HCI actions. As a result, the estimation of human performance and cognitive workload requirements was limited to frequency comparisons. The objective of this report is to demonstrate additional measurement techniques that can be applied to the existing HFA approach to address C² human performance requirements through the evaluation of verbal communications, HCI behavior events, and subjective survey data. Specifically, automated word count, and task-time estimation methods were applied to the existing HFA data to provide estimates of the frequency and time duration of verbal communications for individual members of the FCS C² command group, and task time estimates for all HCI actions.

Procedure:

Data gathered in a series of U.S. Army warfighter-in-the-loop battle simulation experiments were reanalyzed for the present research to develop new measurement methods that can be used to address FCS C² human performance requirements. Automated word count, and task-time estimation methods, were applied to existing HFA data to provide estimates of the frequency and time duration of verbal communications for individual members of the FCS C² command group, and task time estimates for all HCI actions. Figures were developed to demonstrate the utility of developing the new estimates of FCS C² system performance requirements. In addition, methods were demonstrated for estimating the reliability and validity of self-report surveys used to estimate individual command group member workload.

Findings:

The present research has identified a number of measurement approaches that can support simulation-based research assessments of human performance requirements for future FCS C² systems. Figures were developed to demonstrate how the use of the new automated word count, and task-time estimation methods can provide estimates of human performance that support decisions regarding workload, task allocation, and training requirements. Methods used to estimate the reliability and validity of self-report surveys of workload provide evidence that these scales are sensitive to changes in task demands, and identify limitations in comparisons of workload across C² command group members.

Utilization of Findings:

Measures developed from this research can be used to ensure that human performance requirements are identified early in the new system design process. The word count and task-time estimation methods can be applied to the existing HFA approach to partially overcome some of the data limitations associated with the lack of automated frequency and time duration measures for verbal communications and HCI actions. By demonstrating the types of estimates that can be provided when verbal communications, and HCI frequency and time duration data are available, the present research has served to promote the development of automated measures of command and control performance that would reduce the laborious process of manual HCI video data reduction.

MEASUREMENT METHODS FOR HUMAN PERFORMANCE IN COMMAND AND CONTROL SIMULATION EXPERIMENTS

CONTENTS

	Page
Introduction.....	1
Overview.....	1
The Human Functional Analysis Approach.....	2
Research Objective	3
Previous Human Functional Analysis Approach and Limitations.....	3
Verbal Communications Measurement.....	4
HCI Performance Measurement	5
Focused Surveys	5
Data Sources Used to Develop New Measurement Methods	6
Results: Additional Measures of C ² Verbal Communication	7
Verbal Communications Skill Proficiency with Practice: Estimating Communications Time Across Battle Runs.....	7
The Changing Nature of Verbal Communications Demands During a Battle: Estimation of Communications Time Within a Battle Run.....	9
Results: Additional Measures of C ² HCI Behavior	12
Estimates of HCI Task Time Requirements for Workload Assessment.....	13
The Changing Nature of HCI Task Demands During a Battle	14
Indicators of HCI Task Training and Proficiency Levels.....	16
Estimated Time as a Common Metric for Verbal Communications and HCI Task Comparisons	17
Results: Additional Measures for Focused Surveys	18
Validating Subjective Ratings of Workload and Performance Success	18
Workload Ratings Comparison Across Participants and Experimental Treatment Conditions.....	20
Summary and Discussion.....	22
References.....	25

CONTENTS (Continued)

	Page
Appendix A. Verbal Communication Rating Codes: Definitions and Examples.....	A-1
Appendix B. Examples of Coded Verbal Communications Transcript Passages	B-1
Appendix C. Human Computer Interaction (HCI) Rating Codes.....	C-1
Appendix D. Example of a Coded HCI Record	D-1
Appendix E. After Run Survey: Task Load Index (TLX)	E-1

List of Tables

Table 1. Data Available from Previous C ² Experiments	6
Table 2. Time Per Word Results from Experiment 1 Runs	8
Table 3. Human Target Recognition and Battle Damage Assessment Task Time Comparison (Seconds).....	16
Table 4. Composite Workload Scale Inter-Item Correlation.....	19
Table 5. Comparison of Workload and Performance Success Ratings to Run Complexity Level.....	20
Table 6. Comparison of Composite Workload Ratings Across Run Complexity, and Duty Position	21
Table 7. Comparison of Command Group Member Workload Ratings Across Three Battle Run Complexity Levels	22

List of Figures

Figure 1. Command group communication times derived from Experiment 1 Run 1 data.....	9
Figure 2. Command group member communication times derived for Experiment 1 battle runs.....	9
Figure 3. Frequency of words per 10-minute time interval, Experiment 2 Run10.....	10
Figure 4. Estimated communication time for Cell Commander across 10-minute time intervals.....	11
Figure 5. Estimated communication time for command group members across 10-minute time intervals	12
Figure 6. Comparison of HCI task frequency and task time workload estimates	13
Figure 7. Estimated HCI task duration time for command group members	14
Figure 8. HCI task load for command group members across 10-minute time intervals	15
Figure 9. Frequency of Commander's HCI task performance across 10-minute time intervals	15
Figure 10. Information Manager HTR task performance across 10-minute time intervals.....	17
Figure 11. Combined verbal communications and HCI task time estimates for command group members.....	18
Figure 12. Average workload ratings by command group member, and run complexity	21

MEASUREMENT METHODS FOR HUMAN PERFORMANCE IN COMMAND AND CONTROL SIMULATION EXPERIMENTS

Introduction

Overview

The U.S. Army's proposed Future Combat System of Systems (FCS) will be a networked force horizontally and vertically integrated from strategic to tactical level, to provide dominant situational understanding. Automated Command and Control (C²) capabilities in FCS units will allow tactical commanders, assisted by a small command group, to effectively lead a future force composed of large numbers of manned and robotic elements. This paper describes research conducted by the U.S. Army Research Institute (ARI) to develop measurement methods to enhance the existing Human Functional Analysis (HFA) approach (Sanders, Lickteig, 2002) for estimating human performance and cognitive workload requirements presented by FCS C² systems. A major shortcoming in previous HFA assessments was that the frequency and time duration of verbal communications and Human-Computer Interaction (HCI) behaviors had to be obtained through time consuming analysis and coding of video recordings. Assessments based on manual video data reduction of command and control performance can only examine a fraction of the data potentially available from each FCS C² experiment, or any future FCS training, testing, and evaluation effort. Previous HFA verbal communications video data reduction required approximately one day of analyst time for each hour of recorded performance. The time demands were even greater for the reduction of HCI data. There were eight separate video screens present in the FCS C² Cell, with data reduction requiring approximately eight days to complete, identifying over 1,000 HCI actions total for the command group. Due to time constraints, previous HFA assessments have not been able to provide estimates of the frequency and time duration of verbal communications for individual members of the FCS C² command group, and could not provide time duration estimates for all HCI actions. As a result, the estimation of human performance and cognitive workload requirements was generally limited to frequency comparisons.

The objective of this report is to demonstrate additional measurement techniques that can be applied to the existing HFA approach to address C² human performance requirements through the evaluation of verbal communications, HCI behavior events, and subjective survey data. Specifically, automated word count, and task-time estimation methods were applied to the existing HFA data to provide estimates of the frequency and time duration of verbal communications for individual members of the FCS C² command group, and task-time estimates for all HCI actions. Data gathered in a series of U.S. Army warfighter-in-the-loop battle simulation experiments were reanalyzed for the present research to demonstrate new measurement methods that can be used to address FCS C² human performance requirements. Particular attention was paid to identifying the frequency and time duration of verbal communications and human-computer interactions associated with basic C² functions (e.g., Plan, See, Move, Strike) for the individual members of the C² command group.

The Human Functional Analysis Approach

The HFA approach has been developed by ARI to identify and describe C² functions associated with command group performance. Results of this analysis can be used to support decisions regarding workload and task allocation, assess the effects of changes in automation support on workload, and serve as indicators of training and proficiency levels of the C² command group. The term “functions” generally refers to groups of related actions that contribute to a larger action to achieve a definite goal or purpose. The approach taken to assess human functions (required to accomplish C² tasks) was to classify elements of behavior, namely verbal communications and HCI events, into meaningful command and control functions, providing estimates of the behaviors and workload demands associated with command and control of an FCS C² Cell. For this paper the term “C² Cell” will refer to a co-located command group composed of a Commander, and three Battle Managers. The HFA approach has been used to identify and describe the C² behaviors of the command group for an FCS Unit Cell in an ongoing series of experiments sponsored by Defense Advanced Research Projects Agency (DARPA) and the U.S. Army Communications-Electronics Command (CECOM) as documented in previous ARI test reports (Sanders, Lickteig, Durlach, Rademacher, Holt, Rainey, Finley, and Lussier, 2002, and Lickteig, Sanders, Durlach, Rainey, and Carnahan, 2002). Data for the HFA have been obtained from videotaped records of verbal communications, HCI actions associated with operating the Commanders Support Environment (CSE); and from subjective responses obtained in after action reviews, surveys, and interviews, as follows:

- Verbal analysis of “communications” included transcription from audio recordings of all spoken exchanges by members of the command group with one another, with higher headquarters, and with subordinate personnel. A taxonomy of communications was developed as a structural framework for the Verbal Communications Rating scheme. Verbal analysis identified the source and type of communication, C² function, subject matter, and time duration.
- HCI analysis of player and CSE interactions included iterative review of video recordings of command group performance in the C² vehicle. A taxonomy of HCI tasks was developed as a structural framework for the HCI C² Rating Scale. A related goal in the HCI analysis was to promote the development of automated measures of command and control performance.
- Responses obtained from command group players in after action reviews, surveys, and interviews addressed multiple research issues including: workload, performance success, effectiveness of the CSE prototype, and function allocations among humans and machines.

At the Unit Cell level, the overall function of command group actions was to command and control the Unit Cell and accomplish the assigned mission. To support the assessment of human functions, a candidate set of subordinate command and control functions was developed from a review of the FCS C² experimental design plans, U.S. Army documents addressing FCS C² functions (DARPA, 2001), and the U. S. Army Objective Force Operational and Organizational Plan for Maneuver Unit of Action (TRADOC, 2002). This review suggested that four basic C² functions could be identified (Plan, See, Move, Strike). These four functions have provided a framework for the analysis of C² Cell verbal communications and HCI performance as follows:

- Plan: Develop, assess, and modify a plan including combat instruction sets provided to robotic elements in response to changing events.
- See: Control and interpret input from a heterogeneous set of advanced sensors to mentally construct an accurate picture of the battlefield in terms of METT-TC (mission, enemy, terrain, troops, time, civilians) factors.
- Move: Control the movement and activity of friendly manned and unmanned systems to maintain desired movement rates and formations.
- Strike: Distribute a variety of indirect and direct effects over a set of targets.

Research Objective

The objective of this report is to demonstrate additional measurement techniques that can be applied to the existing HFA approach to address C² human performance requirements through the evaluation of verbal communications, HCI behavior events, and subjective survey data. Specifically, automated word count, and task-time estimation methods were applied to the existing HFA data to provide estimates of the frequency and time duration of verbal communications for individual members of the FCS C² command group, and task time estimates for all HCI actions. The detailed assessment of C² functions and workload requirements can support many important decisions related to manpower, personnel, task allocation, and training requirements. For example, the HCI analysis provides useful estimates on the impact of C² prototype design changes introduced during experimentation on command group performance and workload. Also, the behavior-based HCI measures provide an empirical basis for the development of automated C² performance assessment and feedback tools for training. Measurement problem areas were identified, and ways of reducing the burden of experimental data analysis were suggested. Examples of HFA measures were presented that illustrate the assessment of verbal communications, HCI behavior, and the use of interviews and focused surveys. Specific measurement techniques demonstrated in the present research were as follows:

- Verbal Communications
 - Task skill proficiency with practice.
 - Changing nature of task demands during a battle.
- Human Computer Interaction
 - Estimates of task time requirements for workload assessment.
 - Changing nature of task demands during a battle.
 - Indicators of training and proficiency levels.
- Subjective Measures (Self-Report Survey)
 - Validating subjective ratings of workload and performance success.
 - Workload ratings comparison across participants and experimental treatment conditions.

Previous Human Functional Analysis Approach and Limitations

The development of new C² system human performance measures was based on needs identified from previous research and utilized existing data bases from previous research to develop and test out the measurement approaches. In previous research the HFA approach has been used by ARI to identify and describe the C² behaviors of the command group for a future fighting force, through the analysis of verbal communications, HCI actions, and surveys. Data

gathered in a series of U.S. Army warfighter-in-the-loop battle simulation experiments were analyzed and presented in formal reports describing the human requirements associated with future highly automated command groups (Sanders, Lickteig, Durlach, Rademacher, Holt, Rainey, Finley, & Lussier, 2002), and (Lickteig, Sanders, Shadrick, Lussier, Holt, & Rainey, 2002). Data were collected from a simulated C² Cell environment developed for research purposes composed of a hardware and software system located in a command group C² vehicle. The simulated C² environment included workstations for four key command group members—Commander, Battle Space Manager, Information Manager, and Effects Manager—that allowed them to command and control a large number of robotic airborne and ground vehicle sensors, and other ground vehicles.

Verbal Communications Measurement

Verbal communications were analyzed after each experiment from audio recordings to identify the content, frequency, and time duration of communication. The method used to analyze command group verbal communications basically required the transcription of recordings of all spoken exchanges by members of the command group, and the coding the content of these exchanges using a set of Verbal Communications Rating Codes developed for this purpose. The assessment of verbal communications required the systematic decomposition of Unit Cell verbal communications into functions rating categories (Plan, See, Move, Strike), and several sub-functions categories to include Mission, Enemy, Terrain, Troops, Time, Civilian (METT-TC). Experimental trials were videotaped and transcripts of verbal communications were developed. The text of the transcripts was separated into blocks or “chunks” of dialogue specific enough in their meaning that they did not fall under multiple ratings categories. After the transcript text was separated into chunks, raters would individually assign codes from the Verbal Communication Rating Codes sheet to each chunk of text, yielding a record of the types and frequencies of communications between C² Cell players for the experimental run. The development of the Verbal Communication Rating Codes was an iterative process, involving multiple reviews of transcripts and revision of codes. Several runs were coded independently by multiple raters and the ratings compared to assess inter-rater agreement. The Verbal Communication Rating Codes sheet used to evaluate Experiment 2 data (Lickteig, Sanders, Durlach, Rainey, and Carnahan, 2002) is provided as Appendix A. Examples of transcript passages coded using this rating scheme are provided as Appendix B.

The previous evaluations of verbal communications data have been limited by the fact that communication time was reported for the total command group, and was not broken out individually for each member of the command group. It is important to estimate the amount of time individual command group members devote to communication, and the subject matter, as this provides insights as to the nature of the C² tasks they perform, and cognitive workload they experience. This problem is largely due to the prohibitive amount of manual effort that would be involved in recording start and stop times for each member’s statements, with the command group averaging a total of 222 statements per run in Experiment 1. Without individual level communication time estimates it is not possible to estimate communication requirements for individual command group members, and the contribution of verbal communication to individual workload during the course of battle runs.

HCI Performance Measurement

It is important to identify the frequency and amount of time command group members devote to HCI task actions as this provides an estimate of the functional performance requirements associated with operating a C² system, and the cognitive workload they experience. The method currently used to analyze command group member interactions with C² vehicle computerized systems requires a review of video recordings of command group performance at each workstation in the C² Cell. Written records of the HCI actions performed by the C² Cell Commander and the three Battle Managers were developed from video recordings of a battle run. An HCI C² action scoring scheme was developed to separate observed HCI actions into functions categories. The coding categories were developed from a review of the CSE software operators training materials, reviews and scoring of HCI video recordings, and the iterative revision of scoring categories to ensure coverage of HCI actions, and rater reliability. Estimates of inter-rater agreement were calculated to ensure that the ratings attained a high level of reliability. The primary measure of performance used for assessing HCI C² performance was HCI task frequency. Performance times for 45 of the 50 Experiment 2 HCI tasks were typically less than 5 seconds, and were not recorded due to the great demand this manual task would represent for researchers. Start and stop times were recorded for four tasks that were typically longer than five seconds in duration (Create Ground Route, Create Air Route, Human Target Recognition, and Battle Damage Assessment). The HCI C² Scoring Codes sheet from Experiment 2 (Lickteig, Sanders, Durlach, Rainey, & Carnahan, 2002) is provided as Appendix C. An example of a coded HCI actions record for one battle run is provided as Appendix D.

Previous evaluations of C² HCI data have been limited by the lack of HCI task time data for 45 of the 50 HCI tasks, and by a lack of detail in some analyses. Time estimates are needed for all HCI tasks as the presentation of HCI task frequencies can serve to underestimate the workload demands associated with long duration or difficult tasks. Additional detail in the presentation of HCI findings can serve to address important questions regarding specific task demands placed on individual command group members during the course of a battle run.

Focused Surveys

Subjective measures used to gather information on command group performance requirements for C² FCS have included an In-Place After Action Review (AAR), focused surveys, and interviews. The present research will describe efforts to gain evidence of the validity of measures used to analyze perceived Workload, and perceived Performance Success survey data as these measures have shown some evidence of being sensitive to manipulations of task complexity in previous experiments. Workload and Performance Effectiveness were assessed at the conclusion of each battle run using the NASA Task Load Index (TLX) (NASA, 1986). The TLX scale is a multi-dimensional rating scale that has been shown to be very sensitive to changes in operator workload levels in many different contexts (see Appendix E for TLX rating scale). The TLX survey was administered in part to assess the impact of manipulations in battle run complexity, where run complexity was manipulated by changing the level of friendly force automation support, and by changing the composition of the threat force. These surveys provided an alternative method to address workload issues, which complemented the C² Verbal Communications and HCI actions task frequency and time duration criteria.

Previous survey-based evaluations of workload and performance success have been limited by a lack of information regarding the reliability and validity of the workload survey measures that were employed. Responses to four TLX workload subscales (Mental, Temporal, Effort, Frustration) were averaged to provide the Composite Workload self-report estimate used in previous analyses, but data have not been analyzed to estimate the reliability and validity of this composite measure. Analyses need to be conducted to identify whether the Composite Workload scale, and the TLX Performance Success subscale, are related to other indicators of FCS C² workload, such as the frequency and duration of verbal communications, and HCI task performance events. Previous estimates of the relationship between workload ratings (and Performance Success ratings) and Run Complexity have been limited to a visual comparison of mean ratings data plots, rather than an empirical estimate of the relationship between the two variables. Empirical measures are needed that facilitate comparisons between simulation test participant workload and performance ratings. Measures are also needed to determine whether self-report surveys are sensitive to experimental manipulations, such as battle run complexity.

Data Sources Used to Develop New Measurement Methods

The current research effort to identify and apply new measurement methods used available data from a series of three U.S. Army warfighter-in-the-loop battle simulation experiments. For these three experiments four active duty lieutenant colonels served as the command group members. The four-person command group consisted of the Commander, Battlespace Manager, Information Manager, and Effects Manager. Experiments contained from 9 to 11 simulated battle runs lasting approximately 60-90 minutes each, with one or two battles conducted each day over the course of each two week experiment. During each battle the command group would direct the actions of robotic ground and air assets in simulated operations against a Red Force. The analyses provided in previous reports were limited to descriptive statistics. Inferential statistics were considered inappropriate due to limitations in experimental design and execution, and the small size of sample. With regard to experiment execution, one notable constraint on the available data is that command group members changed a great deal across battle runs for Experiment 3, so that combining data across runs was not always appropriate. Also, while the manipulation of Run Complexity level was a key independent variable for analysis, this was likely confounded with run order effects. The data sources from the previous three experiments that were available for the present research are summarized in Table 1.

Table 1

Data Available from Previous C² Experiments

Experiment	Verbal Communication Transcript	Record of HCI Actions	TLX Workload and Performance Success ratings
1	8 Battle Runs	No Data Collected	9 Battle Runs
2	3 Battle Runs	1 Battle Run	9 Battle Runs
3	None at this time	None at this time	11 Battle Runs

Results: Additional Measures of C² Verbal Communication

The present research developed a method for producing verbal communications performance estimates for the individual C² Cell members which have not been available from the manual audio data reduction and coding approach used in previous CSE experiments. The verbal communications performance estimates were developed by reanalyzing available data from previous experiments to group together the verbal statements made by each separate C² Cell member, generating a word count for his statements, and then assigning a time value to his communications based on the average amount of time-per-word for the experimental run. Two estimates of verbal communications were developed to enhance the understanding of C² experiment participant performance. First, estimates of the individual C² Cell member verbal communications duration were developed for a single run, and then across all runs for an experiment. These performance estimates might be valuable in providing an indication of any increasing skill proficiency with practice. The second verbal communications performance estimate was developed by breaking an experimental run into ten minute time intervals, and identifying the individual C² Cell member contribution to communication within each time interval. This detailed description of changing verbal communications time requirements for C² Cell members across a battle run provides valuable information about the possible cognitive workload experienced by C² Cell members at different phases of the battle.

Verbal Communications Skill Proficiency with Practice: Estimating Communications Time Across Battle Runs

Measures are needed that can provide evidence of whether command groups demonstrate increasing skill proficiency with practice in modern C² systems. A related issue is whether verbal communication becomes more efficient, perhaps becoming less frequent, and less time consuming, as the command group becomes familiar with tasks and coordination requirements. To address this question estimates of the frequency and time duration of verbal communications for each member of the command group are needed. While a direct measure of the behavior would be best, in the absence of measurement data methods for generating estimates of verbal communications behavior were required. Estimates of individual command group member verbal communications were developed using an automated word count feature. The number of words within battle run transcripts (recorded in Excel format) were estimated using an automated word count feature (available in MS Word) for seven available Experiment 1 battle runs. Battle run transcript data were not available for Experiment 1 Runs 6 and 8. The frequency of words per battle run were then divided by the time devoted to communications within each battle run (available from previous reports) to produce a time-per-word estimate (see Table 2). Averaging the time-per-word data across the seven battle runs yielded an average time-per-word of approximately one second (1.00 second per word, SD = 0.45). The mean time per word estimates show some variability across Runs within Experiment 1, ranging from a low of 0.69 seconds per word for Runs 2, 3, and 4, to a high of 1.86 seconds per word for the last Run (Run 9). No analyses were conducted to investigate whether the greater time per word value for Run 9 was associated with the command group speaking slower, using longer words, or whether the longer duration was associated with pauses or breaks in speech.

Table 2

Time Per Word Results from Experiment 1 Runs

Run	Communication Time (Seconds)	Word Count	Mean time per word (Seconds)
1	3960	2964	1.34
2	1260	1832	0.69
3	2340	3415	0.69
4	3480	5064	0.69
5	2160	2097	1.03
7	2340	3299	0.71
9	3420	6377	1.86
Mean			1.00

The use of communication time estimates allows for an examination of the verbal communication demands on each simulation experiment command group member, while previous analyses could only provide a total command group communication time per run estimate. Based on the mean word time across the seven battle runs of 1.00 seconds, a value of 1.0 seconds per word was accepted as a reasonable estimate of communication time for Experiment 1 data. It should be noted that new time-per-word estimates would be required for later Experiments as characteristics of command group communications could change with experience. The verbal communication statements for each battle run were sorted by command group member (a four-way sort for each battle run using an automated Excel sort feature). An automated word count program was applied to estimate the frequency of words per command group member, and the word frequencies were multiplied by the 1.0 second time-per-word value to generate an estimate of communication time for each command group member, for each battle run. Figure 1 presents a summary of command group member communication time (in minutes) for Run 1 of Experiment 1. The previous method of grouping all communications together provided only an overall time estimate of 42.91 minutes of communication for the C² Cell members, while the use of word counts, and mean time-per-word values, provides an estimate of communication times for each C² Cell member.

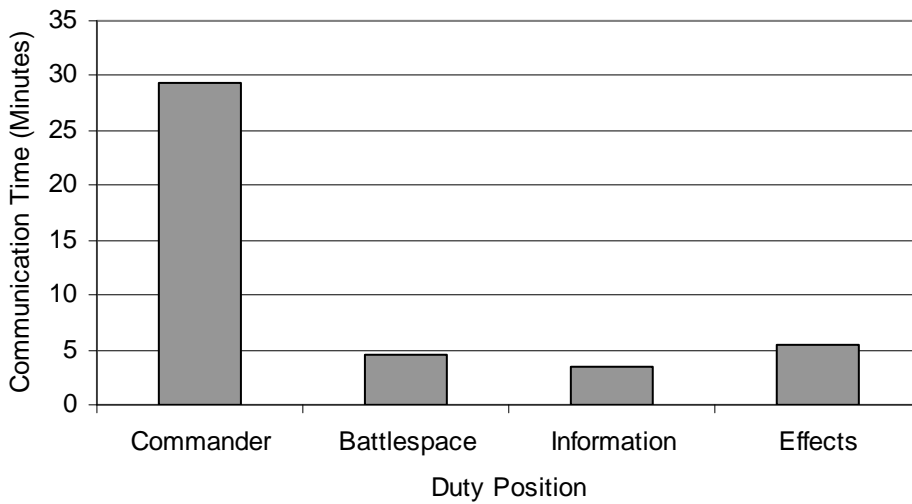


Figure 1. Command group communication times derived from Experiment 1 Run 1 data.

The use of C² Cell member word counts, and mean time per word values, was applied to the full set of Experiment 1 verbal communications transcripts, and the results were organized to present a comparison of communication times across all Experiment 1 runs (see Figure 2). Battle run transcript data were not available for Experiment 1 Runs 6 and 8. The presentation of data in Figure 2 provides a means of identifying patterns or trends for communication over time as C² Cell members become more proficient with the FCS C² tools.

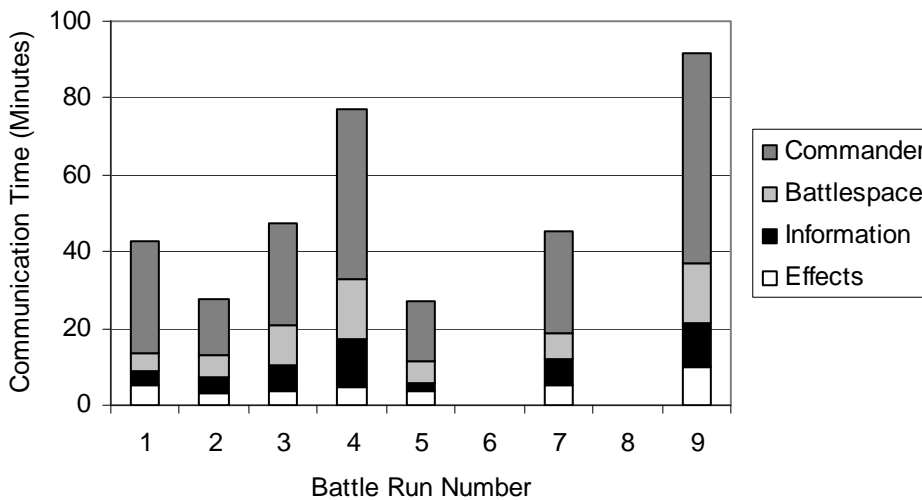


Figure 2. Command group member communication times derived for Experiment 1 battle runs.

The Changing Nature of Verbal Communications Demands During a Battle: Estimation of Communications Time within a Battle Run

The available data were reanalyzed to provide an estimate of how verbal communication demands vary across the performance of a battle run for the individual command group

members. These findings can also suggest whether some tasks might be temporarily reallocated to a cross-trained back-up person during periods of high task load. In order to develop estimates of the communication requirements faced by simulation experiment test participants, the transcripts of verbal communication from Experiment 2 Run 10 was first divided into sequential 10-minute time intervals, and an automated “word count” was conducted to estimate the frequency of words within each ten-minute time interval. 10-minute time intervals were chosen to allow a comparison of word frequency estimates across the performance of the run. Figure 3 presents the frequency of spoken words within the nine ten-minute time intervals for Run 10 (90 minute duration).

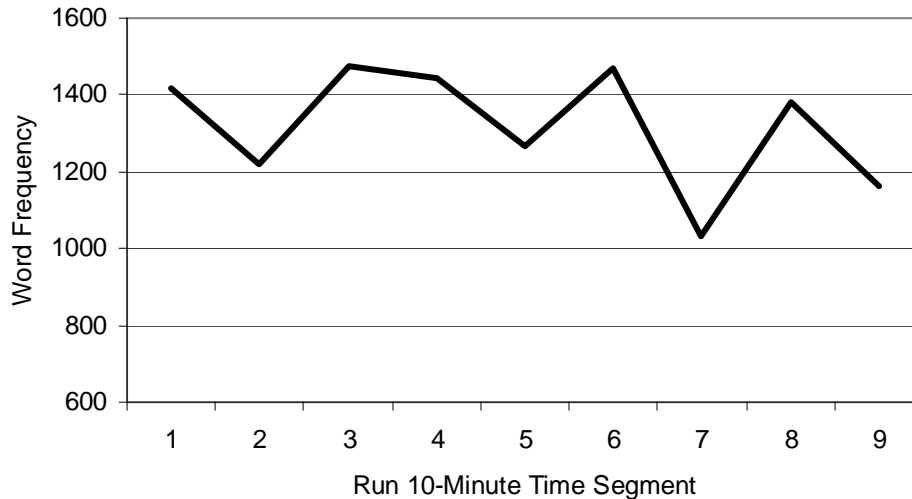


Figure 3. Frequency of words per 10-minute time interval, Experiment 2 Run 10.

The results presented in Figure 3 suggest that there is some variation in the frequency of verbal communications across time intervals during the course of a battle run. The use of word counts, and mean time per word values was applied to the Experiment 2 Run 10 data to provide an estimate of communication times for each C² Cell member across the battle run. The average time per word for each ten-minute time interval for Run 10 was calculated (mean value of 1319 words per 10-minute interval, STD 156.87), from which the estimated time-per-word in minutes (mean value of 131.9 words per minute, STD 15.69) and in seconds (mean value of 0.45 seconds per word, STD 0.26) was derived. The mean word time of 0.45 seconds per word is less than half the 1.0 seconds per word value identified in Experiment 1 data analysis, which suggests that the time-per-word estimates are not generalizable across experiments. While communications appear to be more rapid in Experiment 2, it is not known whether this is due to the command group speaking faster, using shorter words, a reduction in the pauses or breaks occurring in speech, or some other source.

The mean word time of 0.45 seconds per word was used to estimate verbal communications time. A reanalysis of Experiment 2, Run 10 data was conducted to provide an estimate of changing communication task demands for each command group member across nine ten-minute long time intervals in Run 10 battle execution phase. Run 10 verbal communication statements (recorded as Excel format transcripts) were first sorted into nine sequential 10-minute time intervals, and sorted into command group member categories (using an Excel sort feature).

An automated word count program was then applied to estimate the frequency of words in each command group member's transcript statements. The resulting word count frequencies were multiplied by the 0.45 second per word time value to generate an estimate of communication time for each command group member, within each of the nine time intervals for the battle run. Figure 4 presents the communications time estimate for the C² Cell Commander across Run 10 time intervals based on the word count and time-per-word estimation method.

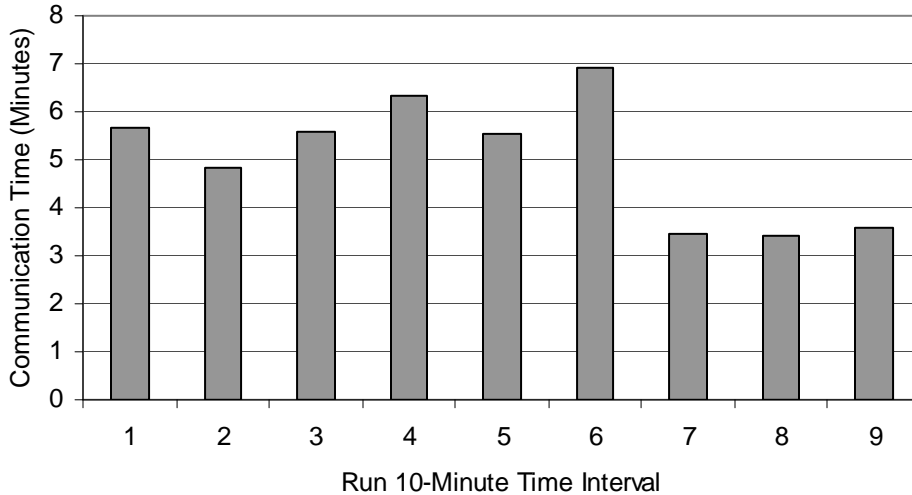


Figure 4. Estimated communication time for Cell Commander across 10-minute time intervals.

Variations in communications demands across the course of a battle might provide evidence of the relative level of cognitive demands placed on C² Cell members during the course of a battle. The word count and time-per-word estimation method can provide estimates identifying times during the battle during which each member of the command group is devoting more or less time to verbal communications tasks, and the absolute time duration of the task requirement. This analysis might help to identify times during a battle when command group members having a relatively light load of tasks might assume some of the task duties of a member who is more heavily loaded. Figure 5 presents a summary of command group member communication time (in minutes) for each of the Experiment 2, Run 10, 10-minute time intervals. It should be noted that the actual time for intervals generally ranged from 8.8 to 10.5 minutes in length, as the original data source did not contain precise 10 minute time tags. In particular, the final time interval, Interval 9, was only 6.5 minutes in length when Run 10 ended.

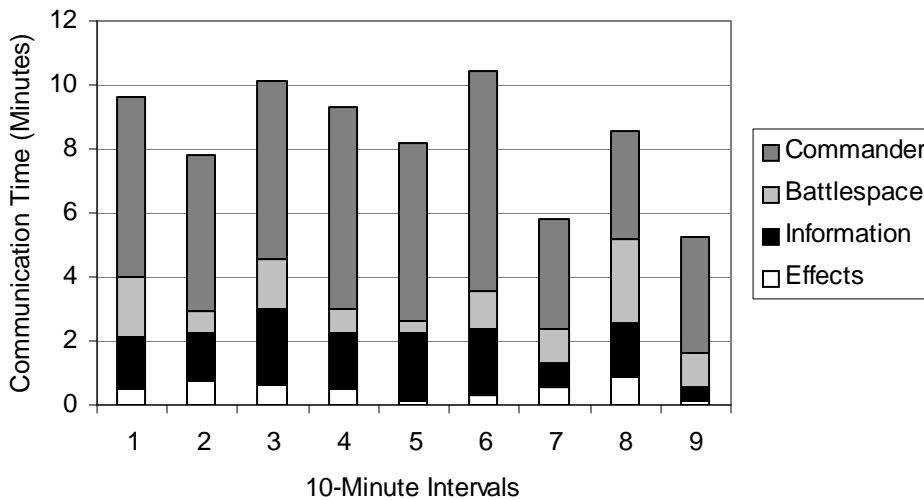


Figure 5. Estimated communication time for command group members across 10-minute time intervals.

The estimation of communications time within a battle run could be useful in identifying the impact of changes to the command group membership, providing evidence of any reallocation of tasks across participants. As a training tool, this estimation of communications time might serve to provide a baseline performance estimate for comparison across groups. This analysis could identify command group trainees who fail to coordinate with others, or unusual patterns of communication.

Results: Additional Measures of C² HCI Behavior

The present research addressed the need for HCI task time estimates by reanalyzing available data to apply a default time value to un-timed HCI tasks, and then combining these task time estimates with data from timed long-duration tasks for presentation. Figures were developed to provide a comparison of HCI task frequency, and HCI task time duration representations of workload demands. Four methods of analyzing HCI task data were developed to enhance the understanding of C² experiment participant performance. First, a measurement method was identified that could be applied to existing data to provide HCI task performance time estimates for short-duration (previously un-timed) tasks. This method was then applied to data from previous experiments to demonstrate how the time duration estimates provide a better estimate of HCI task load compared to simple HCI task frequency measures. The second HCI measurement effort demonstrates how a detailed description of changing HCI task requirements across a battle run can provide valuable information that goes beyond a simple HCI task frequency count. The third measurement approach provides an example of how HCI task performance time data can be used as estimates of training and proficiency levels, and serve to identify high priority training tasks. The fourth measurement effort provides an illustration of how verbal communications and HCI task time duration estimates can be combined to provide a combined estimate of task load for individual command group members.

Estimates of HCI Task Time Requirements for Workload Assessment

Previous estimates of HCI workload have relied primarily on reporting task frequency data. However, the presentation of HCI task frequency data can provide an underestimate of the workload associated with long duration tasks, when compared to the presentation of HCI task time duration data. Five HCI tasks consistently required more than five seconds to complete (Create Ground Routes, Create Air Routes, Human Target Recognition [HTR], Battle Damage Assessment [BDA], and Computer Reboot). When a command group member must perform a number of these long-duration tasks, a simple frequency count that treats long-duration and short-duration tasks equally could underestimate workload. The HCI task data from Experiment 2 was reanalyzed to provide an illustration of HCI task frequency versus HCI task time duration workload estimates.

For Experiment 1 HTR and BDA target Imagery Analysis tasks were performed by higher headquarters, however for Experiment 2 these long duration tasks were allocated to the C² Cell command group. This change suggested the need to assess changes in task allocation (similar to a change in automation support) on workload. Previously HCI task frequency data was presented to estimate the effect of changes in task on workload. Using the Information Manager data as an example, the decision to allocate Imagery Analysis tasks to the C² Cell command group for Experiment 2 added a total of 68 long-duration HTR and BDA tasks to the Information Manager, for a combined total of 400 tasks, representing a 17% (68/400) increase in task frequency. An HCI task time estimate was developed for Information Manager tasks by assigning the actual time duration values to the long-duration tasks, and a default value of 5 seconds to short-duration tasks. The HCI task time approach estimates that the Information Manager target imagery tasks required 976 seconds (16.27 minutes) to complete, out of a total HCI task time of 2986 seconds (49.77 minutes), so that the Imagery Analysis tasks represent a 32.69% (976/2986 seconds) increase in tasks. Figure 6 provides a comparison of the task frequency and task time workload estimates for the Information Manager to illustrate how the use of task time criteria might provide a better estimate of the impact of allocating specific tasks (imagery analysis) to a command group member's duties.

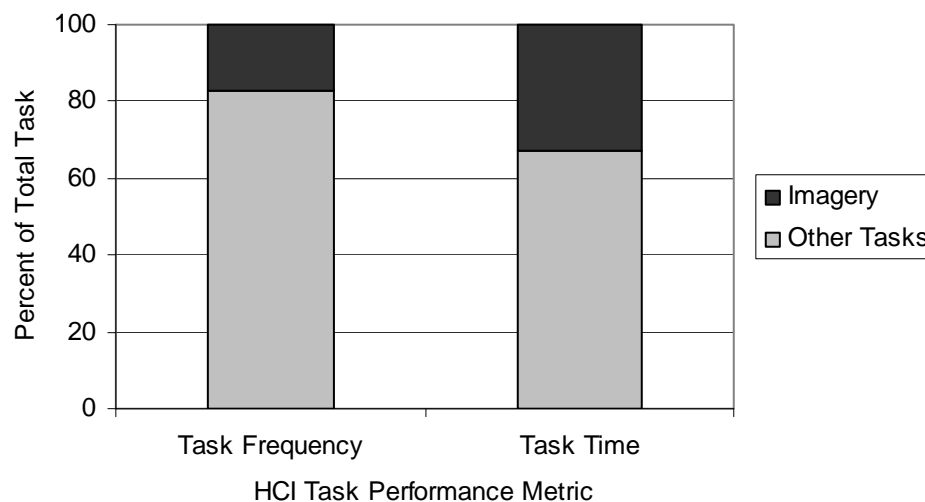


Figure 6. Comparison of HCI task frequency and task time workload estimates.

Figure 7 illustrates how the workload demands for specific tasks, in this case the target Imagery Analysis tasks, can be examined across all members of the command group using task time estimates. Data from Experiment 2 Run 10 were analyzed to illustrate how the Imagery Analysis tasks were allocated across all members of the command group, rather than being duty position specific. While the size of the task load estimate is greatly impacted by the choice of a default time value, the HCI task time estimation approach appears to be a valuable method for combining available measured time data with reasonable estimates of time for tasks which do not have measured time data available. The value of task time duration comparisons should also reinforce the argument that automated measures that can capture time data for all HCI tasks are a high priority requirement to support C² system simulation experiments.

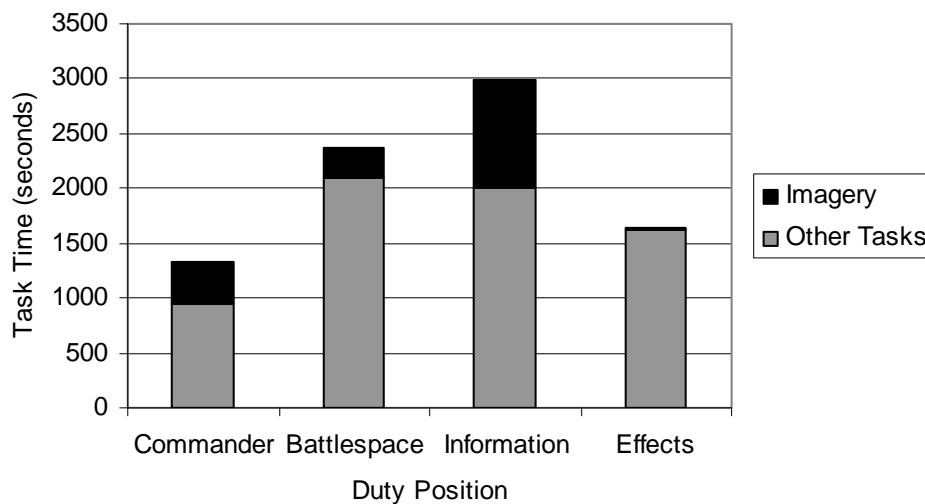


Figure 7. Estimated HCI task duration time for command group members.

The Changing Nature of HCI Task Demands During a Battle

The HCI data provides an emerging empirical basis for reallocation of tasks among command group members, and can identify high priority areas for task automation. Previous analyses have presented HCI task frequencies data showing periods of low and high task performance frequency during a battle run. While the absolute frequency of HCI task performance provides a general indication of task demands, a more detailed level of analysis employing estimates of task performance time is needed to identify the changing nature of task demands during the course of a battle run. Figure 8 was presented in the ARI Experiment 2 report to illustrate the frequency of HCI task performance for each of the C² Cell members during 9 10-minute intervals of battle execution (Lickteig, Sanders, Durlach, Rainey, & Carnahan, 2002). Considering the HCI workload of the Commander as one example, he performed 219 HCI tasks during the 90-minute long battle simulation. His lowest HCI workload occurred during the 30-40 minute time interval (Interval 4), while his greatest HCI workload occurred during the 50-60 minute time interval (Interval 6). While this figure was valuable in suggesting overall task performance levels, questions arose as to what specific tasks the Commander was performing during his low and peak performance intervals of the battle.

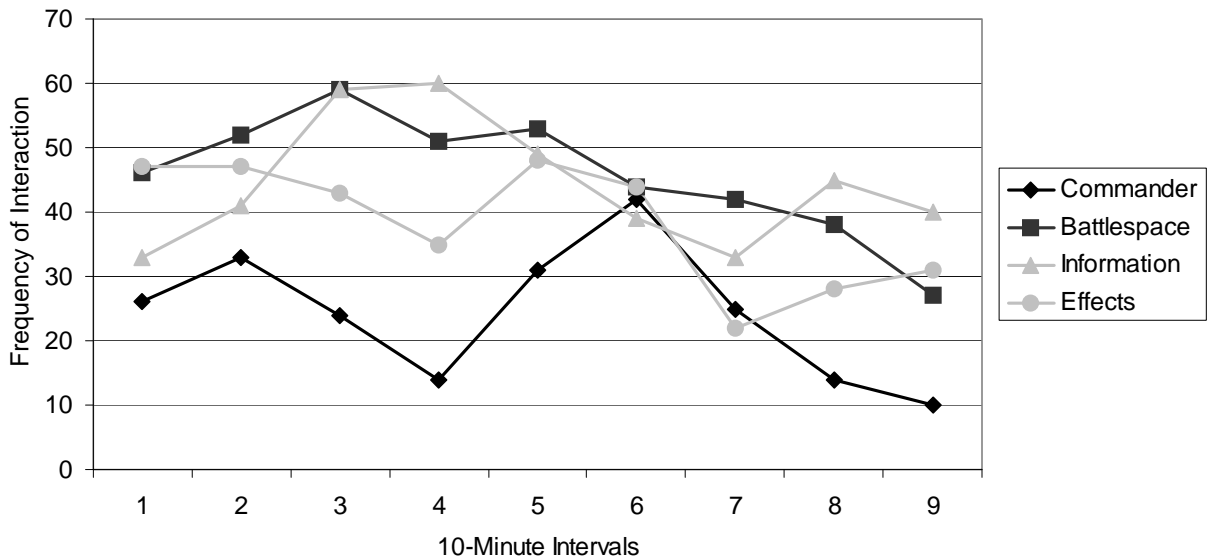


Figure 8. HCI task load for command group members across 10-minute time intervals.

The HCI task frequency data from Experiment 2 used to create Figure 8 were reanalyzed using the task time method that assigns the actual time duration values to the long-duration HCI tasks, and a default value of 5 seconds to short-duration HCI tasks. Figure 9 demonstrates this method of presenting a more detailed time duration-based depiction of the Commander's HCI task performance across a battle run. This assessment reveals that during Interval 4 when HCI task time requirements were lowest, the Commander's HCI actions were split between Target Imagery Analysis (HTR and BDA) and text-based Sensor Data review (to include accessing information on Friendly and Enemy assets). In comparison, the task time estimates for Interval 6, when HCI task time requirements were high, suggest that the Commander's HCI actions have increased in duration, but have not changed in type, as he devoted more time to both Target Imagery Analysis, and Sensor Data review.

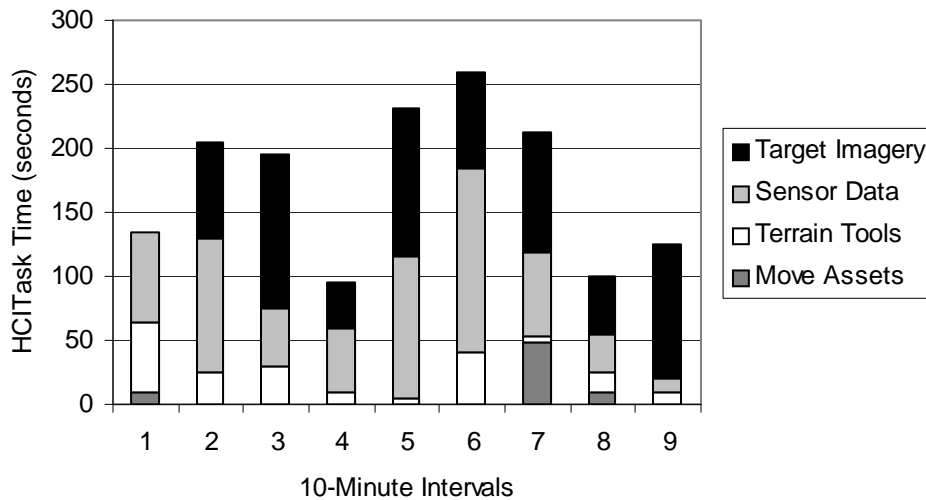


Figure 9. Frequency of Commander's HCI task performance across 10-minute time intervals.

Indicators of HCI Task Training and Proficiency Levels

Measurement techniques are needed that can provide estimates of training and proficiency levels for task performance, which would support the identification of high priority training requirements. One approach to accomplishing this is to identify tasks that show high time demands, and those that show a large degree of variation during performance. A reanalysis of Experiment 2 HCI task performance time duration data for long-duration tasks was conducted to examine performance variations between command group members, and across battle run time intervals. The long-duration HTR and BDA tasks were selected for evaluation as their long duration suggests a relatively high task demand, and also the potential for high payoff if training could serve to reduce this high task time requirement.

Previous reports presented the HTR and BDA task performance times only as mean values for the command group. The data from Experiment 2 were reanalyzed and separate task time estimates were developed for the individual members of the command group (see Table 3). The large standard deviations for HTR and BDA times for each command group member suggest that there was a great deal of variability in each member’s task performance times. The comparison of mean performance times across command group members suggests that performance differences might exist between members. A One-way ANOVA comparison of HTR and BDA task performance times across the command group members for Experiment 2 Run 10 was conducted to identify whether some members performed tasks faster than others. Differences in task performance speed might imply that this task skill is variable, and might be a high priority for training development. However, the results of the ANOVA analysis on this small sample showed no significant difference between command group members for task performance time for both HTR ($F = .805, r = .451$), and BDA ($F = 1.491, r = .250$).

Table 3

Human Target Recognition And Battle Damage Assessment Task Time Comparison (Seconds)

HCI Task	Commander Mean/SD	Effects Mean/SD	Information Mean/SD	Battlespace Mean/SD	Combined Mean/SD
HTR	15.0/11.4	6.0/one case	20.0/19.9	14.8/8.9	17.9/16.7
BDA	17.3/11.3	NA	19.3/12.3	9.0/5.1	16.0/10.9

A descriptive analysis of command group member performance can be useful in identifying whether task performance improves with practice across a series of battle runs, or within a single battle run. Data for HCI performance was only available for Experiment 2 Run 10, so that only the within-run comparison of task performance time could be conducted. The data were analyzed to provide the depiction presented in Figure 10 of the Information Manager HTR task performance time requirement across battle run 10-minute time intervals. This simple plot of task time data can be valuable identifying extreme “outlier” data values that would be lost when simply summarizing performance as mean values. For the present example, the long

duration of task performance during the first 20 minutes of the battle run might reflect the situation where the Information Manager had few targets available for assessment. The more rapid HTR task performance during the 20 to 40 minute time intervals may reflect the detection of large numbers of targets and the demand to rapidly identify them so they could be engaged. The relatively infrequent and long duration HTR tasks beginning at the 40 minute point in the battle might reflect fewer potential targets, and the availability of time for the Information Manager to evaluate target imagery, or to simply leave target imagery windows open in the absence of competing task requirements.

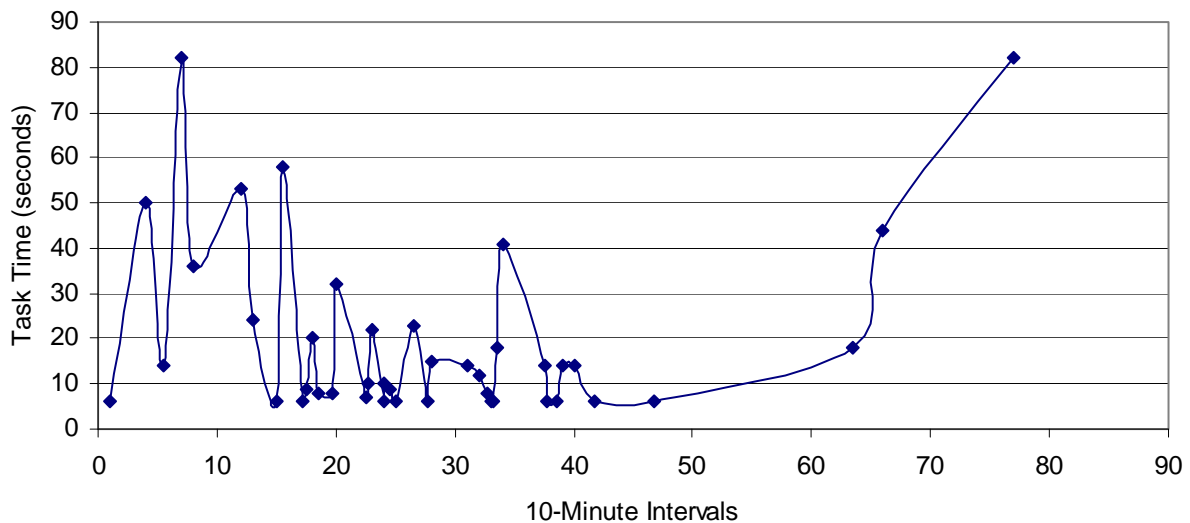


Figure 10. Information Manager HTR task performance across 10-minute time intervals.

The plotting and visual examination of HCI task performance data can be very useful in identifying “outlier” performance values that might not reflect the task-focused behavior desired when developing performance estimates. Through visual analysis the researcher might want to identify stable periods of performance, and focus performance estimates on these periods to eliminate outliers that might not represent periods of highly motivated task oriented performance. As one example, the Figure 10 plot of data suggests that particularly late during the battle, command group members may open task windows and slowly review target imagery information that had already been examined earlier. As automated data collection is introduced, and the requirement for human data entry declines, there will be a growing need to screen the data for outlier values that are not representative. The inclusion of outlier performance data can impact summary estimates of performance. For the Figure 10 example, trimming the last three values has the impact of lowering the HTR target imagery task time estimate from a mean of 20.0 seconds (STD = 19.9) to a mean of 17.9 seconds (STD = 17.5).

Estimated Time as a Common Metric for Verbal Communications and HCI Task Comparisons

As with the Verbal Communications measures, the HCI measures provide one perspective on task workload, which may not reflect all aspects of workload present. As example, while the Commander may have a low HCI task load during some intervals of the battle, this might simply reflect the fact that he is required to perform high priority verbal

communications at the same time. By developing a common set of task time measures for both verbal communication and HCI task performance their combined cognitive workload demands can be estimated. The data from Experiment 2 were reanalyzed using word count time estimates for verbal communications tasks, and using a combination of long-duration HCI task time measures, and short-duration HCI task time estimates to present an estimate of cumulative verbal and HCI task time requirements for each member of the command group (see Figure 11).

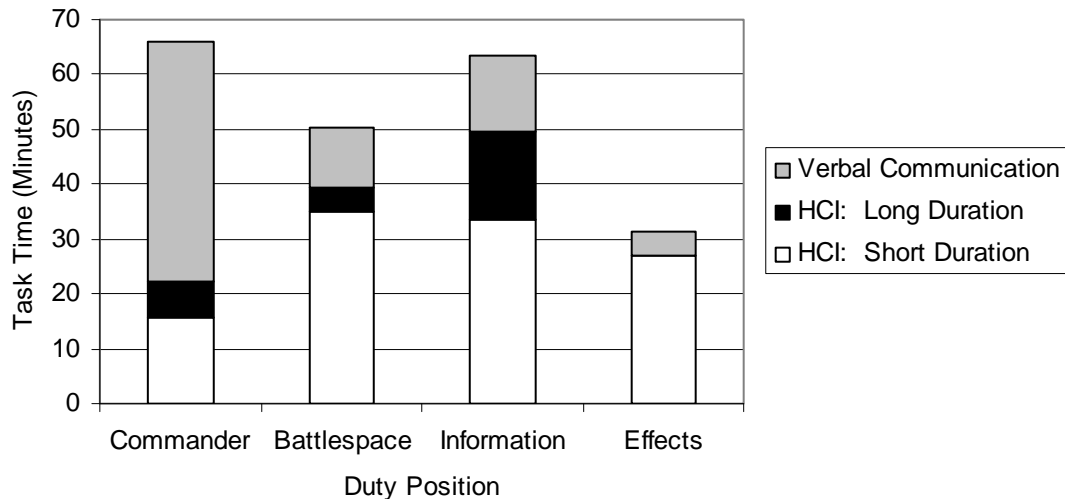


Figure 11. Combined verbal communications and HCI task time estimates for command group members.

Results: Additional Measures for Focused Surveys

Validating Subjective Ratings of Workload and Performance Success

Surveys were conducted at the conclusion of each battle run which allowed researchers to gather self-report estimates of workload and performance success, and to address other issues such as equipment design changes. These surveys provided an alternative method to address workload issues, which complemented the C² Verbal Communications and HCI actions frequency and time duration data. For the present research effort measures of survey scale reliability and validity were applied to survey ratings of workload and performance success from Experiments 1, 2, and 3. First, an inter-item correlation approach suggested by Chronbach was employed to estimate scale internal reliability. Second, scale scores were compared to behavioral examples of performance to establish a validity coefficient as an estimate of the relationship between the self-report scales and an external criteria (battle run complexity).

When reporting Workload results ARI has combined four TLX workload subscales (Mental, Temporal, Effort, Frustration) into a composite average workload score. It was not known whether the mean of the four scale scores combined as the Composite Workload score show high internal consistency. Also, it was not known whether the TLX Composite Workload scale and the TLX Performance Success scale were related to other indicators of FCS C² workload, such as Run Complexity, and behavioral events such as verbal communications, and HCI task performance events. A reanalysis of the data from Experiments 1, 2, and 3 was

conducted to investigate whether the TLX Composite Workload and Performance Success scales were sensitive to changes in experiment run complexity manipulations (Medium, High, Too High, and whether the TLX subjective ratings subscales were related to behavioral measures of workload such as frequency of verbal communication, and HCI task frequency.

The Composite Workload scale reliability was assessed through the generation of an inter-item correlation matrix, and the estimation of scale internal consistency based on the average inter-item correlation (Chronbach's Alpha) for TLX data from Experiments 1, 2, and 3. As an example, Table 4 presents the Composite Workload scale inter-item correlation matrix for Composite Workload data collected in Experiment 2. The calculated average inter-item correlation (Chronbach's Alpha) for the Composite Workload scale was fairly low for Experiment 1 data, the Alpha = .06, n = 36, suggesting low internal consistency across the four TLX scales in ratings of task workload. The internal consistency of the four TLX scale ratings for Experiments 2 and 3 were relatively high, Alpha = .86, n = 48, and Alpha = .79, n = 43 respectively.

Table 4

Composite Workload Scale Inter-Item Correlation

TLX Scale	Mental	Temporal	Effort	Frustration
Mental	1			
Temporal	.6837	1		
Effort	.7578	.7755	1	
Frustration	.3856	.4469	.4888	1

Note: Experiment 2 data set, 11 post-run survey administrations, 4 experiment participants.

One method of validating the subjective TLX workload and performance success ratings is to show evidence of a relationship between the ratings and related behavioral measurements. If workload ratings are correlated with behavioral indicators of workload, then this provides evidence of scale validity. The Composite Workload and Performance Success ratings from Experiments 1, 2, and 3 were compared to their corresponding Run Complexity levels (Medium, High, Too High) using a Pearson Correlation Two-Tailed significance test. Table 5 presents a summary of the comparisons. Two of the three Composite Workload estimates, and all three Performance Success estimates, showed a significant relationship with Run Complexity, while the Workload and Performance Success estimates derived from the combined three-experiment data set also showed a significant relationship to Run Complexity, providing evidence of the validity of both scales.

Table 5

Comparison of Workload and Performance Success Ratings to Run Complexity Level

Experiment	Composite Workload Scale Ratings	Performance Success Ratings
1	$r = .129$, NS $n = 36$	$r = .332$, $p < .05$, $n = 36$
2	$r = .320$, $p < .05$, $n = 48$	$r = .424$, $p < .01$, $n = 48$
3	$r = .371$, $p < .05$, $n = 43$	$r = .482$, $p < .01$, $n = 43$
Combined Data Set	$r = .285$, $p < .001$, $n = 127$	$r = .209$, $p < .05$, $n = 127$

Note: NS = Not Significant. Data from Experiments 1,2, and 3.

Data from Experiment 1 were used to examine whether Composite Workload and Performance Success ratings were related to experience across the nine battle runs (operationalized as the sequential run number, from 1 through 9), and frequency of verbal communication statements. No significant relationship was found, using Pearson Correlation (Two-Tailed) between Composite Workload scale ratings and experience in experimental trials ($r = .274$, $n = 36$), or between Composite Workload and total estimated number of verbal statements made by the command group member ($r = .326$, $n = 32$). Likewise, no significant relationship was found between Performance Success ratings and experience in experimental trials ($r = .218$, $n = 36$), or between Performance Success and the total estimated number of verbal statements made by the command group member ($r = .178$, $n = 32$).

Workload Ratings Comparison Across Participants and Experimental Treatment Conditions

The goal of this research is to identify measures and methods that can be applied to best depict the actions and perceptions of future C² system simulation test participants. A repeated measures analysis was conducted on the Combined Workload ratings obtained after each battle run in Experiments 1, 2, and 3 (Table 6). This analysis provided evidence to identify whether command group members ratings of workload differed based on the run complexity level manipulation, and whether ratings of workload differed across the command group duty positions. The repeated measures analysis of Experiment 1 Composite Workload ratings failed to find a significant difference in workload across run complexity levels, or across command group duty positions. In contrast, for Experiments 2 and 3 the Composite Workload ratings did differ significantly across both run complexity levels, and command group duty positions. The Composite Workload ratings suggest that the manipulation of run complexity levels did have an impact of the command group members perceived workload, and that perceived levels of workload differ significantly between the individual command group members.

Table 6

Comparison of Composite Workload Ratings Across Run Complexity, and Duty Position

Experiment	(Differences Across Run Complexity) Within-Subjects Contrasts	(Differences Across Duty Position) Between-Subjects Effects
1	F = 0.91, p < .368	F = 4.84, p < .033
2	F = 23.85, p < .000	F = 15.80, p < .000
3	F = 78.59, p < .023	F = 93.72, p < .005

Previous reports presented only descriptive results when comparing TLX Workload and Performance Success ratings to battle run complexity levels. Figure 12 provides an example of this type of descriptive results from Experiment 2 data. This figure shows that TLX Performance Success self-ratings generally fell at or above the scale value of 50, which marks the center point on the “Very Low to Very High Workload” 0 to 100 rating scale. A visual analysis of the data suggests that workload ratings differ across both member positions and battle run complexity. For all but the Effects Manager, there appears to be an increase in command group member estimates of Workload at the Too High level of battle run complexity. This result matched the intent of the experimental design for the Too High run which was to challenge the C² command group, identifying an upper performance ceiling for command and control performance.

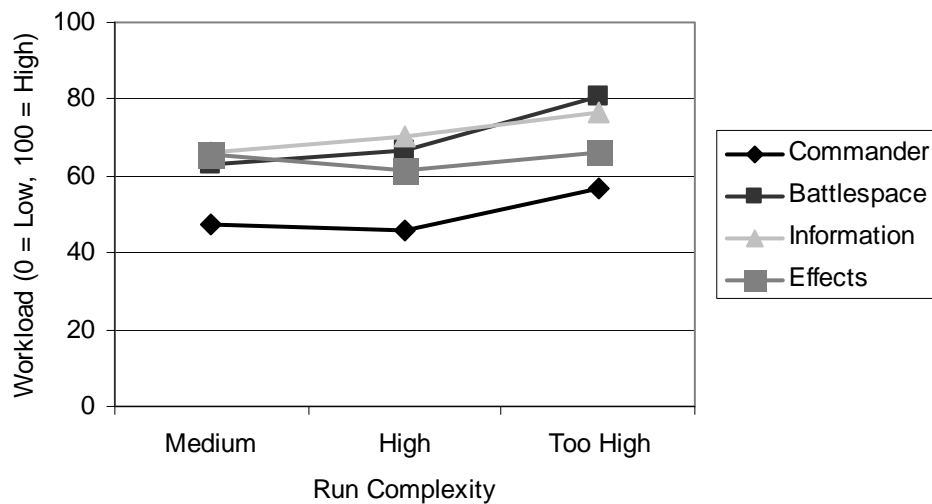


Figure 12. Average workload ratings by command group member, and run complexity.

A One-way ANOVA comparison was conducted to demonstrate an additional measurement method that goes beyond the visual comparison of Figure 10 mean workload ratings plots. A One-way ANOVA of command group member Composite Workload ratings across the three battle run complexity levels for Experiment 2 was conducted with the results presented in Table 7. All tabled comparisons are significant at the .05 level unless otherwise indicated. These Post Hoc multiple comparisons across command group members are based on Tukey’s Honestly Significant Difference (HSD) test statistic. The findings summarized in Table

7 provide support for the conclusion that the trend lines presented in Figure 10 should not be interpreted as indicating real (statistically significant) differences between command group members, that in general the data only support the conclusion that the Commander reported lower levels of workload than the other command group members at the High, and Too High Complexity levels.

Table 7

Comparison of Command Group Member Workload Ratings Across Three Battle Run Complexity Levels

Member	Medium Complexity	High Complexity	Too High Complexity
1 Commander	LT all others	LT Battlespace and Information	LT Battlespace
2 Battlespace	HT Commander	HT Commander	HT Commander
3 Information	HT Commander	HT Commander	ND
4 Effects	HT Commander	No difference	ND

Note: Abbreviations are LT = Lower Than, HT = Higher Than, ND = No Difference

Summary and Discussion

The present research has identified a number of measurement approaches that can support simulation-based research assessments of human performance requirements for future FCS C² systems. Figures were developed to demonstrate how the use of the new automated word count, and task-time estimation methods can provide estimates of human performance that support decisions regarding workload, task allocation, and training requirements. Methods used to estimate the reliability and validity of self-report surveys of workload provide evidence that these scales are sensitive to changes in task demands, and identify limitations in comparisons of workload across C² command group members.

With regard to verbal communications, a measurement approach was presented that described an automated word count approach to estimating the quantity of verbal communications for each command group member. Approaches for estimating verbal communications time requirements for each member of the command group based on time-per-word empirical estimates were presented which allow for the comparison of communications time requirements across battle runs, and also within a single battle run during sequential time intervals. The development of communications frequency and time duration estimates for individual command group members appears to offer a distinct advantage over the presentation of summary estimates of total command group communications. The individual estimates allow for comparisons of communication frequency and time duration between command group members, and across time intervals during a battle run, providing estimates of the cognitive

workload experienced by the individual members. The development of verbal communications time duration estimates is also very useful in providing a common metric that allows HCI and verbal communications task loads to be combined into a single cognitive workload estimate. The primary limitation of the automated word count and task-time estimation approach is that it basically involves replacing word frequencies with a mean time-per-word estimate. There is evidence that the mean time-per-word values will differ across runs within an experiment, and these values might also differ across ten-minute time intervals within a single run. Variations in the time-per-word values could limit the ability to compare results across test events.

With regard to HCI task performance requirements, a measurement approach was presented that provides an estimate of HCI task time requirements, based on the application of task time estimates to un-timed HCI actions. Using this HCI task time estimate method examples were developed to demonstrate how task performance time requirements can provide a useful alternative to the presentation of HCI task frequencies alone when estimating HCI task demands. Examples were provided for comparison of the HCI task time and HCI task frequency estimates for the addition of a target Imagery Analysis tasks to the command group, and for the presentation of detailed HCI task performance times across battle run intervals. The importance of screening raw data for outlier values was demonstrated. The application of task time estimates to un-timed HCI actions appears to have utility in estimating the cognitive effort, or workload associated with FCS C² performance requirements. Compared to a simple HCI task frequency comparison, the HCI task time requirement makes provision for the fact that some tasks, such as target imagery analysis, might be more difficult than others, as evidenced by the greater amount of time required for performance. One obvious limitation of assigning times to un-timed HCI tasks is that this can lead to overestimates of workload, if the assigned time value is too large, and underestimation of workload if the assigned time value is too high. In the present research a single value of 5 seconds was assigned to un-timed HCI tasks. In future efforts multiple time values (such as 1, 3, and 5 seconds) might be identified and assigned to short duration tasks, to reduce the likelihood of over or under-estimating HCI performance time requirements.

A number of measures were identified that should be applied to establish the reliability and validity of survey instruments. Through a reanalysis of the available workload data, measures of ratings scale internal consistency were presented, and methods for estimating the relationship of scale scores to external related criteria were provided. Finally, a repeated measures approach was presented to compare Composite Workload ratings between command group members across run complexity experimental treatment conditions.

Measures developed from this research can be used to ensure that human performance requirements are identified early in the new system design process. The word count and task-time estimation methods that can be applied to the existing HFA approach to partially overcome some of the data limitations associated with the lack of automated frequency and time duration measures for verbal communications and HCI actions. By demonstrating the types of estimates that can be provided when verbal communications, and HCI frequency and time duration data are available, the present research has served to promote the development of automated measures of command and control performance that would reduce the laborious process of manual HCI video data reduction.

With regard to future directions for HFA assessment development, an automated data capture capability is essential for future research efforts, and could provide the basis for an automated performance assessment capability supporting training, evaluation, and C² system design. A major shortcoming experienced in previous HFA assessments was that estimates of the frequency and time duration of verbal communications and HCI behaviors had to be obtained through time consuming analysis and coding of video recordings. Given this constraint, previous HFA assessments have not been able to provide estimates of the frequency and time duration of verbal communications for individual members of the FCS C² command group, and could not provide time duration estimates for all HCI actions. As a result, the estimation of human performance and cognitive workload requirements was generally limited to frequency comparisons. As a first step toward the goal of automated data capture, the Start and Stop actions associated with key C² HCI behaviors have been identified, which could allow software to be developed to pull this information from a data logger file.

While automated data capture might provide a fairly straight forward approach to log the Start and Stop actions, and time duration associated with menu item selections, the automated capture of verbal communications content and time duration presents a much more complicated task. Observations and transcription of FCS C² command group communications reveal that a great deal of communication may occur as short incomplete sentences, or the exchange of a few key words. Much of communication content is FCS C² system-specific jargon, or acronyms, which would exceed the translation capabilities of current voice recognition software.

Future C² HFA efforts might place a greater emphasis on developing decision making and teamwork process measures, which would include the activities, strategies, responses, and behaviors employed in task accomplishment. Cannon-Bowers and Salas (1997) state that this focus on process measures is essential in training development efforts, as outcome measures are usually not diagnostic, and do not indicate the underlying causes of performance necessary to provide constructive feedback. The HFA approach, with its emphasis on individual command group member performance requirements may be well suited to investigations of the processes supporting collective decision making and teamwork.

References

- Cannon-Bowers, J. A. and Salas, E. (1997). A framework for developing team performance measures in training. In Brannick, M. T., Salas, E., & Prince, C.(Eds.), *Team Performance Assessment and Measurement: Theory, Methods, and Applications* (pp 45-62). Mahwah, NJ: Erlbaum.
- Defense Advanced Research Projects Agency. (2001). *Future Combat Systems: CSE Functions Manual (draft)*. FCS Unit Cell C² Study Technical Team, Fort Monmouth, NJ.
- Lickteig, C., Sanders, W., Shadrick, S., Lussier, J., Holt, B., Rainey, S. (2002). *Human-system integration for future command and control: identifying research issues and approaches* (ARI Research Report 1792). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Lickteig, C., Sanders, W., Durlach, P., Rainey, S., Carnahan, T. (2002). *Future Combat Systems Command and Control (FCS C²) Human Functions Assessment*. Available from the Program Manager (PM) FCS C².
- NASA-Ames Research Center, Human Performance Group. (1986). *Collecting NASA workload Ratings: A paper and pencil package* (Version 2.1). NASA-Ames Research Center. Moffet Field, CA.
- Sanders, W., Lickteig, C., Durlach, P., Rademacher, W., (MAJ), Holt, B., Rainey, S., Finley, D. & Lussier, J., (2002). *Future Combat Systems Command and Control (FCS C²) Human Functions Assessment*. Available from the Program Manager (PM) FCS C².
- Sanders, W., Lickteig, C. (2002, October). *Human Command and Control Functional Requirements for Future Systems*. Paper presented at the International Military Testing Association, Ottawa, Canada.
- U.S. Army Training and Doctrine Command, (2002). The United States Army objective force operational and organizational plan for maneuver unit of action: TRADOC Pamphlet 525-3-90. Fort Monroe, VA: Author

Appendix A

Verbal Communication Rating Codes: Definitions and Examples

For each chunk select

SOURCE (for each chunk select one and only one code)

1	Within Cell (Black)	Cell = 4 CSE operators
2	Cell <-> Blue (Team)	
3	Cell <-> White (Higher)	
4	Cell<->Subordinate	Subordinate includes C ² Vehicle gunner & driver
5	Blue<-> White	
6	More than 2-way (e.g., Cell<->White<->Blue)	Only to be used in cases where more than 2 elements involved in SAME conversation
7	Other	e.g., to technical support people

FUNCTION (for each chunk select one and only one code)

1	See	Detect or identify enemy or friendly positions, or significant terrain aspects. (not BDA)
2	Plan	Interpret data, predict enemy COA, generate own COA
3	Move	Manage/Monitor/Control asset movement
4	Strike	Manage/Monitor/Control lethal/nonlethal effects
5	BDA	See for purposes of BDA
6	other	None of the above

TYPE (for each chunk select one and only one code)

1	Share. Announcement, telling what is <u>seen or known</u> .
2	Action. Announcement, telling what <u>speaker</u> is doing at the moment--verbalization <u>accompanying</u> action such as fire or move. <u>Not</u> the decision process. <u>Not</u> actions such as I see, monitor, track, etc. <u>Not</u> describing someone else's actions.
3	Direction. Order, command, delegation of responsibility.
4	Ask. Interaction <u>begins</u> with request for information, confirmation, assistance, or assets and <u>ends</u> either with <u>informational answer</u> or <u>no response</u> , with little or no discussion. <u>Not</u> rhetorical questions.
5	Process. Infer, synthesize, fuse, understand, turn data into information <u>without</u> consequent decision or direction. <u>Can</u> start with Share, Action, or Ask.
6	Decide. Like Process, <u>but in addition</u> , includes a verbalized decision or plan.
7	Other.

Verbal Communication Rating Codes: Definitions and Examples
(continued)

FACTOR (for each chunk select one and only one code)

MISSION	
1	Original Plan: Concerning mission goals and plans prior to execute phase.
2	Dynamic Planning: Tactical re-planning during the execute phase in response to changing events and available assets. <u>Must have stated COA (course of action).</u> Changes from Original Plan.
3	Situational Understanding. Integration/summary of current situation involving multiple factors; <u>but without stated COA.</u>
ENEMY: Concerning enemy situation including	
4	Location: Sensor hit(s) – locate enemy positions.
5	Identification: Identify targets – identify nature of enemy target.
6	Disposition: Probable enemy COA, strategy, or tactics.
7	BDA: Battle Damage Assessment – cell seeks/discusses feedback on damage they inflict on enemy.
TERRAIN	
8	When terrain is the prime focus (e.g., can we travel over that kind of terrain?, we should go this way because it will provide cover). Example: “Moving to low ground.” <u>Not simply map locations (e.g., not, sensor hit north of the wall).</u>
TROOPS and Assets (Soldiers, Equipment, Vehicles) FRIENDLY ONLY	
9	Location Status: Position report/assessment
10	Movement Status: Mobility report/assessment (includes fuel)
11	See Status: Sensor report/assessment
12	Strike Status: Fire power report/assessment (includes # of remaining missiles)
13	Communications/network functionality (radio, internet, or other; cell to outside cell, including semi-autonomous sensors).
14	Information management systems: CSE user interface tools
15	Survivability Concern: Asset in danger.
16	Survivability Move: defensive move to remove asset from immediate danger.
17	Loss/Casualty: Asset destroyed (catastrophic hit).
18	Move Action: Move/Manage/Maneuver [Active, <u>Not</u> position report] Excluding Survivability Move; Also See Terrain.
19	Strike Action Lethal: Launch/fire/deploy with intent to destroy (includes LAMs)
20	Strike Action Nonlethal: Launch/fire/deploy (could include unarmed sensors, propaganda, smoke, jamming of enemy, etc.).
21	Training (soldier training, mission rehearsal)
22	Other-- having to do with troops or assets but none of the above
TIME	
23	When time is the prime focus (e.g., how much time something will take, how much time is available, order of priority, synchronization of actions).

Verbal Communication Rating Codes: Definitions and Examples
(continued)

CIVILIANS	
24	Any issues regarding how to deal with civilians: avoiding, provisioning, protecting, etc. <u>Not</u> mere sensor hits of civilians, <u>unless</u> first time mentioned.
Other	
25	Other (e.g., humor, personal, leadership, morale)

SYSTEMS (for each chunk, select as many systems as apply).

1	C ² Vehicle – vehicle, sensors, weapons, and IT systems, crew
2	Future Warrior Vehicle (FWV or IFV)- vehicles, sensors (DVO, IR, and laser range finder), weapons (Javelin), mounted Future warriors
3	Future Warriors-dismounted
4	Roboscout – platform, sensors (DVO, IR, Laser designated range finder, GSR)
5	Line of Sight (LOS) Vehicle, weapons (MP-ERM, and smart cargo)
6	NLOS/BLOS Vehicle, weapons (LAM, PAM, netfires), sensor deployment (UGS).
7	A160 UAV (Unmanned aerial vehicle) platform, sensors (Elint, Comint, CEP, radio hits/links).
8	Shadow UAV platform, sensors (DVO, IR, MTI, SAR)
9	Micro Ducted Fan UAV platform (MUAV), sensors (IR)
10	Internettet UGS deployed in the field, sensors (acoustic, seismic, magnetic)
11	Nonorganic Assets: Team, Unit of Action or Higher
12	Voice radio communications
13	Network communications (among digital information management systems)
14	Logistics (e.g., refuelers, ammo supply vehicles, maintenance).
15	Other
16	Unspecified (relevant but can't tell which system)
17	Not applicable (not relevant)

Appendix B

Examples of Coded Verbal Communications Transcript Passages (Chunks 170-180 from Experiment 2 Run 6)

CO=Cell Commander
 BS=Battlespace Manager
 IN=Information Manager
 EF=Effects Manager

Speaker	Transcript	Chunk	Source	Function	Type	Factor	System
CO	Okay, Brooks you have a target entering phase- line yellow, radio link 21, you need to engage him.	170	1	4	3	19	16
BS	Where is he, in the North?	171	1	1	4	4	7
CO	In the north, heads up.						
BS	Radio link 21.						
CO	Radio link 21, got it.						
IN	The guy in the South has turned NE.	172	1	1	1	4	16
CO	Which guy Ted?	173	1	1	4	5	16
IN	The one brooks engaged earlier, he turned NE. Heavy track.						
IN	Jack, he is still defending forward is all I can tell you.	174	1	2	5	6	17
CO	Yeah, he is still trying to envelop us in the South, is what he is going to try to do, at least with part of his force.						
CO	He is moving faster than us...he seems to be able to move his forces faster than we can move our forces. I know we are moving pretty good because I can see our speed, but somehow he is...maybe because he is using the open ground.	175	1	1	5	6	16

Appendix C

Human Computer Interaction (HCI) Rating Codes

- 100 PLAN (Describe Tactical Situation, Concerns, And Future Activities, Request Information.)
 - 110. Create/Update a Mission and COA
 - 111. Create Overlay Graphics and Map Annotations
 - 112. Place platforms on the map (friendly and threat template)
 - 113. Rehearse the Plan
 - 114. Execute the Plan
 - 115. Indicate an Area/Point on Map Using Cursor
- 200 MOVE (Manage/Monitor Control Asset Movement)
 - 210. Move Ground Assets (Start = First blue line appear, Stop = Click OK)
 - 211. Create routes (clicking map locations to create blue route line)
 - 212. Start, Halt or Resume a platform
 - 213. Edit an existing route
 - 214. Delete all tasks (from execution window)
 - 215. Fire UGS
 - 220. Move Air Assets (Start = First blue line appear, Stop = Click OK)
 - 221. Create Routes (either by creating a direct route or by selecting targets to recon)
 - 222. Delete all tasks (from execution window)
 - 223. Edit an existing route
- 300 SEE (Manage Map and Sensor Data Display)
 - 310. Map Display
 - 311. Zoom Map (either arrow or magnification tools)
 - 312. Scroll Map
 - 320. Use Visualization Aids
 - 321. Range Fans
 - 322. Intervisibility Plotting
 - 323. Measuring Distance
 - 324. Head Up Display
 - 325. Select/Change Windows, State View, or window area for display (increasing/decreasing a window area such as Asset or Alert window for better viewing)
 - 326. Change GCM Settings (Declutter Map – Includes Hide Impacted Missiles)
 - 327. Move Visual Reference Points (red cross used by higher. We hope to eventually include in analysis all who manipulate information during runs on the CSE/C² Prototype)

Human Computer Interaction (HCI) Rating Codes
(continued)

- 330. Sensor Data Display
 - 331. Create Alerts/Automated Filters (includes auto fires)
 - 332. Display Target Catalog (open this spreadsheet window, is it normally open?)
 - 333. Target query (cursor over enemy icons to read properties information)
 - 334. Friendly query (cursor over friendly icons to read properties information)
 - 335. Area query (cursor over area to read properties information)
 - 336. Change Sensor (UAV, Shadow, Roboscout etc.)
 - 337. Toggle sensor fans
 - 338. Alert Window Confirmation

- 340. Human Target Recognition (Start = HTR window open, Stop = Close HTR window)
 - 341. Display target images (through Alert Window, clicking the picture icon on map, select window, etc.)
 - 342. Use tools to refine image (zoom, pan, brightness, etc.)
 - 343. Change Map Icons to reflect target status (i.e. Garm, Draega, Bus, etc.)
 - 344. Remove templated targets (State View selection)
 - 345. Select recon target by clicking icon
 - 346. Select recon target by select window

- 350. Battle Damage Assessment (Start = HTR window open, Stop = Close HTR window)
 - 351. Display target images (through Alert window, clicking the picture icon on map, select window, etc.)
 - 352. Use tools to refine image (zoom, pan, brightness, etc.)
 - 352. Change Map Icons to reflect target status (i.e., targeted, suspected, dead, etc.)

- 400 STRIKE (Distribute Indirect and Direct Effects Over a Set of Targets)
 - 410. Pre-Plan Fires/Execute Pre-Plan Fires

 - 420. Fire A Weapon System
 - 422. NETFIRES LAM
 - 423. LAM Final Attack Command (right click on LAM icon and reassign to attack)
 - 424. NETFIRES PAM
 - 425. LOS
 - 426. C²Vehicle (Gun and Javelin)
 - 427. FW CARRIER
 - 428. DISMOUNT JAVELIN

 - 430. Target Designation (how the cell player chooses what target to be fired upon)
 - 431. Icon Click (on map)
 - 432. Menu Select (in NETFIRES window or LOS window)
 - 433. "Select" Window (right click on cluttered icons on map)

Human Computer Interaction (HCI) Rating Codes
(continued)

440. Monitor Fires Execution

441. LAM query (cursor over LAM icons to read properties information)

442. PAM query (cursor over PAM icons to read properties information)

500 OTHER MANUAL ACTS

510. General

511. Reboot system (Start = Fatal Error, Stop = CSE/C² Prototype full restart)

Appendix D

Example of Coded HCI Record (Battlespace Manager's Left Screen Experiment 2 Run 10)

Duty Position: Battlespace Screen: 1 Left Run: 10

Run Time (minutes)		Code	Description
Start Time	Stop Time		Typical configuration of screen = 50% map, 25 % asset window, and 25 % execution window.
0 00 00			
0 00 28		311	Zoom map
0 00 33		311	Zoom map
0 02 47		334	Friendly Query (FQ)
0 02 50		213	Retask Roboscout Move (Error unable to retask)
0 03 17	0 03 46	511	Crash Reboot
0 04 02		425	Task LOS Fire (FIRE)
0 04 29		213	Retask Transport Move
0 04 38	0 05 08	511	Crash REBOOT (may have crashed due to route over flying icon)
0 05 09		334	Friendly Query
0 05 17		213	Retask roboscout move
0 05 28	0 06 03	511	Crash Reboot
0 06 38	0 07 25	511	Crash Reboot (Unattended crash)
0 14 42		325	Minimize execution window
0 20 35	0 23 43	511	Crash Reboot (Unattended crash)
0 28 15	0 30 46	511	Crash Reboot (Unattended crash)
0 36 09		333	Target Query (TQ)
0 39 00		311	Zoom
0 39 10		312	Scroll map (South)
0 39 11		334	Friendly Query
0 43 14		312	Scroll map
0 43 16		441	LAM Query (LQ)
0 43 19		334	Friendly Query
0 43 24		431	Select target by clicking icon
0 43 28		425	FIRE LOS
0 43 32		312	Scroll map (north) (top 12 km)
0 43 49		431	Select target by clicking icon
0 43 54		425	FIRE LOS
0 43 58		333	Target Query
0 44 18		431	Select target by clicking icon
0 44 24		333	Target Query
0 44 28		425	FIRE LOS
0 48 36		325	Minimize execution window
0 48 39		333	Target Query

Example of Coded HCI Record
(Battlespace Manager's Left Screen Run 10)
(continued)

Run Time (minutes)		Code	Description
Start Time	Stop Time		
0 48 55		431	Select target from select window
0 48 59		425	FIRE LOS
0 52 56		431	Select target by clicking icon
0 53 00		431	Select target by clicking icon NO FIRE
0 53 32		431	Select target by clicking icon
0 53 34		425	FIRE LOS
0 55 41		334	Friendly Query
0 55 45		333	Target Query
0 55 46		333	Target Query
0 55 50		334	Friendly Query
0 56 04		431	Select target by clicking icon
0 56 08		425	FIRE LOS
0 56 17		334	Friendly Query
0 56 18		333	Target Query
0 56 29		333	Target Query
0 57 42		312	Scroll map
0 57 48		334	Friendly Query
0 58 05		325	Enlarge execution window
0 58 13		213	Delete tasks for roboscout (Cancel)
0 58 33		211	Create route
0 58 49		334	Friendly Query
0 59 03		311	Zoom
1 01 59		325	Minimize execution window
1 25 57	1 27 23	511	Crash Reboot
1 28 37			End Exercise

Appendix E

After Run Survey: Task Load Index (TLX)

Part 1. Run Workload


Duty Position _____ Date _____ Run # _____

Task Load Index Rating Scales

Task or Mission Segment: _____

Please rate the task or mission segment by putting a mark on each of the six scales at the point which matches your experience.

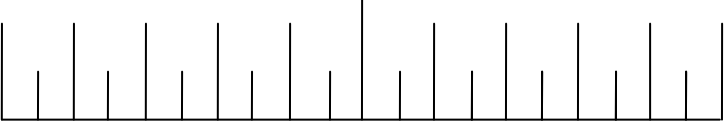
Mental Demand



Very Low Very High

(HOW MENTALLY DEMANDING WAS THE TASK?)

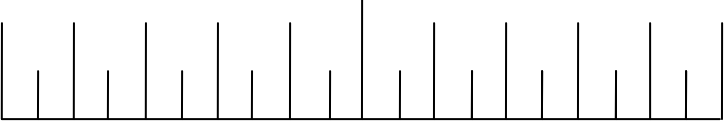
Physical Demand



Very Low Very High

(HOW PHYSICALLY DEMANDING WAS THE TASK?)

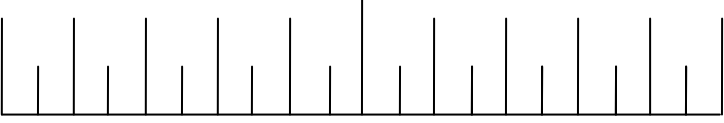
Temporal Demand



Very Low Very High

(HOW HURRIED OR RUSHED WAS THE PACE OF THE TASK?)

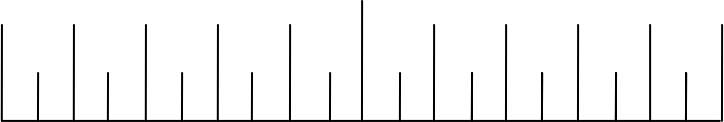
Performance



Failure Perfect

(HOW SUCCESSFUL WERE YOU IN ACCOMPLISHING WHAT YOU WERE ASKED TO DO?)

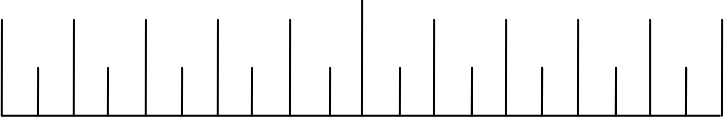
Effort



Very Low Very High

(HOW HARD DID YOU HAVE TO WORK TO ACCOMPLISH YOUR LEVEL OF PERFORMANCE?)

Frustration



Very Low Very High

(HOW DISCOURAGED, IRRITATED OR ANNOYED WERE YOU)