The Effects of Multimodal Collaboration Technology on Subjective Workload Profiles of Tactical Air Battle Management Teams

Victor S. Finomore
Benjamin A. Knott
General Dynamics Advanced Information Systems
Dayton OH 45431

W. Todd Nelson
Scott M. Galster
Robert S. Bolia
Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

May 2007

Approved for public release; Distribution is unlimited.

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Collaborative Interfaces Branch
Wright-Patterson AFB OH 45433
NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Wright Site Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TP-2007-0012

THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.

FOR THE DIRECTOR

//signed//
Scott M. Galster
Experimental Research Psychologist
Collaborative Interfaces Branch

//signed//
Daniel G. Goddard
Chief, Warfighter Interfaces Division
Human Effectiveness Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government’s approval or disapproval of its ideas or findings.
The Effects of Multimodal Collaboration Technology on Subjective Workload Profiles of Tactical Air Battle Management Teams

Victor S. Finomore*
Benjamin A. Knott*
W. Todd Nelson**
Scott M. Galster**
Robert S. Bolia**

General Dynamics Advanced Information Systems*
Dayton OH 45431


A tactical air battle management task required a team of two weapons directors, two strike operators, and a tanker operator to communicate with each other in order to coordinate offensive and defensive air attacks, and aerial refueling. This study compared the impact of two types of communication modalities (Voice or Picture Chat) and the number of enemy targets (4 or 6) on team performance and perceived team workload. Three subjective workload scales were evaluated in their ability to characterize task difficulty, communication demands, and demands of the different team roles.
THE EFFECTS OF MULTIMODAL COLLABORATION TECHNOLOGY ON SUBJECTIVE WORKLOAD PROFILES OF TACTICAL AIR BATTLE MANAGEMENT TEAMS

Victor S. Finomore, Benjamin A. Knott
General Dynamics, Advanced Information Systems, Dayton, Ohio

W. Todd Nelson, Scott M. Galster, Robert S. Bola
Air Force Research Laboratory, Dayton, Ohio

A tactical air battle management task required a team of two weapons directors, two strike operators, and a tanker operator to communicate with each other in order to coordinate offensive and defensive air attacks, and aerial refueling. This study compared the impact of two types of communication modalities (Voice or Picture Chat) and the number of enemy targets (4 or 6) on team performance and perceived team workload. Three subjective workload scales were evaluated in their ability to characterize task difficulty, communication demands, and demands of the different team roles. The results are discussed with respect to the descriptive and discriminating abilities of the three team workload scales.

Introduction

Team Workload Assessment

The assessment and measurement of various aspects of team performance is increasingly important to many military and civil aviation domains such as tactical air battle management (ABM), air traffic control, and emergency medical response. Within these domains, modern communication and information networks support collaboration within and between widely distributed teams. As a result, the design and evaluation of collaborative technology (CT) interfaces with respect to their impact on team communication, coordination, and information sharing is receiving a great deal of recent attention among human factors researchers (Cummings, 2004; Bolstad & Endscale, 2005; Knott, Bola, Nelson, and Galster, 2006a).

When exploring the effectiveness of CTs, the evaluation of performance measures is important. However, measurement of mental workload imposed by the task is also critical. Mental workload refers to the demands on cognitive resources experienced during the performance of a task. This concept has been used to identify bottlenecks in systems where task demands exceed operators' supply of cognitive resources (O'Donnell & Eggemeier, 1986). Mental workload has also been used to evaluate the different resource demands placed upon the operator from alternative system designs in order to determine the most optimal interface (Wickens & Hollands, 2000).

Extending this concept from individual operators to teams, it is crucial for identifying bottlenecks related to team structure, processes, or system interfaces where demands of the environment exceed team resources. A better understanding of team workload would allow researchers to identify problems with perceptual, physical, and cognitive demands and aid in the development of more efficient CTs and team structures. One goal of the present study was to assess and characterize the workload demands placed upon teams engaged in an ABM task when using different CT interfaces. The experiment evaluated and compared the relative contribution of three subjective workload scales to the understanding of demands placed upon ABM teams. The three workload scales that were compared are:

1. The NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988), which is one of the most effective measures of perceived mental workload currently available (Wickens & Hollands, 2000). It provides a global workload index on a scale of 0 to 100 and identifies the relative contributions of six sources of workload: Mental Demand, Temporal Demand, Physical Demand, Performance Effort, and Frustration. The NASA-TLX has been used in many laboratory and real-world tasks to identify situations that are cognitively demanding (Warm, Dember, & Hancock, 1996).

2. A modified Multiple Resource Questionnaire (MRQ, Boles & Adair, 2001), in which observers are presented with a set of mental priorities based upon a combination of dimensions drawn from Wickens' Multiple Resource Theory (Wickens & Hollands, 2000). The MRQ consists of the 17 resource dimensions listed in Table 1.

Using a modified scale (Finomore et al., 2006) from 0 (no usage) to 100 (extreme usage), observers are asked to rate the extent to which a task they just performed utilized each dimension. Research with the modified MRQ has indicated greater sensitivity without modifying its diagnostic profile (Finomore et al., 2006). This result supports the conclusion that a
The proliferation of collaborative technologies such as instant messaging (chat), data visualization, and digital whiteboards, promise alternative or enhancements to the traditional voice communication. However, commercial-off-the-self CIs are often employed without proper evaluation or consideration of the work domain and the cognitive requirements of its operators (Scott, Cummings, Graeber, Nelson, & Bolia, 2006).

In the present study, teams conducted a simulated ABM scenario in which they communicated in the traditional manner using radio headsets, or by using a custom graphical whiteboard tool, called “Picture Chat” (PC), to augment voice communication. Picture Chat is a collaboration tool that was developed with consideration to the communication-intensive demands of ABM. This tool allows WDSs to send images of a tactical display to teammates, annotated with task-relevant symbols indicating desired actions or directives. This type of visual communication is germane to our evaluation of subjective workload instruments for team research, because it presumably imposes significantly different cognitive demands on the team compared to the baseline voice communication.

Methods
The Distributed Dynamic Decision-Making (DDD) Simulator was employed to create a set of ABM scenarios conveyed to participants through a tactical display that exhibited the movement of entities within a battle space.

The ABM scenario was constructed around a five-member team. Two WDSs coordinate operations by communicating with each other, two Strike Operators (SOC) and a Tanker Operator (TO) to intercept threats and resupply assets as needed. To do this effectively, WDSs must understand the capabilities, limitations, and resources of their operational environment. Within the simulation, three classes of friendly fighter assets (F-15, F-16, & F-18) and two classes of hostile targets (MiGs & Su-27s) were employed. The F-15s and F-16s were equipped with two missiles and could only attack two MiGs. The F-18s were outfitted with four missiles to target two MiGs and two Su-27s. Moreover, each fighter began the mission with different fuel capacities, the F-15s and P16 could refuel at the Air Force tanker and the F-18s at the Navel Tanker.

The WDS’ role was to match friendly fighters with the appropriate enemy targets, schedule fighters for refueling and resupply, and communicate their plan.
of action to the SOSs and TO. As such, WDs had primary decision making and leadership responsibility. The role of the SOSs and TO was to maneuver assets as instructed and to provide pertinent information to WDs concerning the resources of team assets. The number of targets present throughout scenario was deliberately controlled as a manipulation of task difficulty. A more in-depth explanation of the scenario can be found in Knott, Bollia, Nelson, and Gaiser (2006b).

Participants
Five men and five women between the ages of 18 and 25 yrs. (Median = 22) were paid $15 for each hour of participation in the experiment. Participants were combined to create 10 unique teams. Seven of the participants were undergraduate students and three were graduate students.

Procedure
Prior to the experiment, all participants completed two 4-hour training sessions in which they practiced on the PDD platform for all team roles and with the CT software. Team members communicated either through simulated radio headsets or by using Picture Chat (PC). An example of a PC message is shown in Figure 1, in which a WD was requesting two target intercepts and refueling of two fighters.

Figure 1. An example of a Picture Chat message. Dottiles are communicated to teamates in a visual manner by connecting assets and targets with lines and task-relevant symbols.

The trainer informed participants that the purpose of the study was to evaluate how teams used communication technology to work together to meet the ABM scenario’s objectives. In addition, participants were trained on and practiced communication brevity, such that voice communication mimicked the highly structured communication used in ABM. Participants were also trained on the specific objectives and rules of the mission, and were instructed that the performance of the team would be measured for each trial based on how well they met their objectives.

The experimental session consisted of four 10-minute experimental trials per team. After each trial, participants completed the three subjective mental workload scales (NASA-TLX, MRQ, and TWS). All major simulation events (e.g., the occurrence and outcome of attacks, refueling events, etc.) were recorded in data logs for later analysis. In addition, all Voice and PC communications were recorded.

Experimental Design
There were two levels of collaboration technology (CT) (voice-only, and voice & PC), and two levels of task difficulty (4 or 6 targets). Three independent variables were combined factorially, yielding a 2 x 2 within-subjects design. In the Voice & PC condition, WDs were instructed to use the PC tool for all movement-related communication and could use voice to check asset status. The order of conditions was counterbalanced across trials.

Results
Team Performance
A single measure of team performance was calculated for each trial by averaging the percentage of (1) targets that penetrated friendly airspace, (2) high value assets destroyed (the air base, infantry units, and tanker aircraft), and (3) fighter assets lost. The average was then subtracted from 100, resulting in a team performance score in which 100 indicated optimal team performance in accordance with the mission objectives.

A target x collaboration technology repeated measures analysis of variance (ANOVA) was performed on the team score, which revealed a main effect for number of targets, F (1, 9) = 73.78, p < .001. In this and all subsequent ANOVA’s, Bonferroni’s correction was employed to correct for violations of the sphericity assumption. Performance scores were significantly higher for the 4 target condition (M = 82.7, SE = 2.6) compared to the 6 target condition (M = 69.1, SE = 2.9). All other sources of variance in this analysis lacked significance.

Subjective Workload Scales
Subscale ratings for each of the three subjective workload instruments were submitted to a Subscale x Team Role (WD/SO) x Targets (46) x CT (voice/voice & PC) mixed-design ANOVA with team role as a between-subjects factor. A series of analyses were conducted to determine how each of the instrument’s subscales varied as a function of 1) changes in task difficulty, 2) the demands of different team roles, and 3) the demands of different
collaborative interfaces. When appropriate, a Bonferroni correction was used to test post-hoc comparisons. Correlations between subscales and the team performance measures were also computed to determine how each of the workload subscales could serve as predictors of team performance. Results are reported for coefficients that are significant at the p < .05 level.

NASA-TLX

Weighted NASA-TLX ratings were submitted to a 3 x 5 (Subscale) x 2 (Team Role) x 2 (Targets) x 2 (CT) mixed-design ANOVA. The main effects of targets and CT indicated that the NASA-TLX instrument was indeed sensitive to a) the primary task difficulty manipulation, and b) the differential demands of the two CT interfaces. The 4-target condition (M = 133.8, SE = 6.8) yielded higher average scores than the 4-target condition (M = 130.5, SE = 7.1).

Additionally, a significant interaction emerged in higher ratings (M = 133.9, SE = 3) than voice communication alone (M = 130.5, SE = 3).

Post hoc comparisons of the significant targets x subscale interaction (F (5, 90) = 2.8, p < .05) indicated that the sensitivity of the NASA-TLX to targets was due to the reliable difference in the temporal demand and performance subscales for 6 and 4 targets (Figure 2). The significant CT x subscale interaction (F (5, 90) = 2.8, p < .05) indicated that the sensitivity of the NASA-TLX to CT conditions was due to the reliable difference between the voice only and PC conditions for frustration and performance subscales (Figure 3).

Temporal and performance demand subscales varied with task difficulty, such that higher task difficulty resulted in higher ratings on these scales. Frustration and performance varied with CT in that the PC condition resulted in higher ratings on these subscales. Notably the NASA-TLX was not sensitive to differences in team role.

Figure 2. Mean NASA-TLX subscale scores for number of targets.

![Mean NASA-TLX subscale scores for number of targets](image)

Figure 3. Mean NASA-TLX subscale scores for communication modality.

MRQ

Ratings were submitted to a 17 (Subscale) x 2 (Team Role) x 2 (Targets) x 2 (CT) mixed-ANOVA, yielding main effects for subscale, F(16, 288) = 55.0, p < .05, and a two-way interaction of CT x Subscale, F(16, 288) = 4.8, p < .05. The significant three-way interaction, F(16, 288) = 1.74, p < .05, indicated a different CT x subscale profile for the two team roles (Figure 4). For WDs, the voice only condition lead to greater demands on vocal processes, but the PC condition led to greater spatial emergent processes demands. In addition, there were marginal differences between CT conditions on SC, SE, and SP subscales, though these differences did not reach significance in post hoc comparisons. For SOs, only vocal process differed across CT conditions, and the trends for the spatial processing dimensions were not present. MRQ ratings were sensitive to CT and team role differences as evidenced by these profiles.

However, the MRQ was not sensitive to the primary task difficulty manipulation of targets.

None of the subscale ratings for SOs were correlated with performance. However, Auditory Linguistic subscale ratings were positively related to team performance for WDs (r = .44). One possible explanation is that teams performed better when they were more actively listening to and comprehending messages from their teammates.
The purpose of this study was to evaluate and compare the relative contribution of three subjective workload scales to our understanding of the demands of collaboration tools and team tasks. Ten teams were trained on an ABM scenario and tested under two communication modality conditions (Voice only and PC), and two levels of task difficulty (6 or 4 enemy targets). The latter factor was reflected in team performance scores.

Reliable workload measures are crucial to the evaluation of team tools and the technologies employed for collaboration. Diagnostic workload instruments may be used to understand how to restructure team responsibilities, or design interfaces consistent with the communication demands of a task. A useful instrument must be sensitive to the demands of the task, team members' responsibilities, and technology interface manipulations.

All three workload instruments were sensitive to the overall demands of the CT, however only the NASA-TLX was sensitive to the task difficulty manipulation, showing that more targets imposed greater temporal demands on the team. While the NASA-TLX was sensitive to the demands of the CT interfaces overall, it was not diagnostic in detecting differences between the team roles.

The TWS was sensitive to team roles, and indeed provided a useful profile that differentiated team process demands for the WD and SO responsibilities. Perhaps the MRQ was the most diagnostic with respect to CT and teammate responsibilities, in that the three way interaction provided descriptive profiles for both of these factors. In particular, it measured the high demands on vocal processes for WDs using voice only communication compared to the PC. The PC served to reduce vocal processing demands but resulted in higher spatial emergent processing and did not reduce workload overall.
While there was no evidence that high vocal processing workload was related to poor performance, the descriptive nature of these results may still suggest a path forward for future CT development. For instance, although the PC relieved vocal process demands, the STM demands were quite high in both CT conditions. Future CT development should focus on strategies for reducing this source of workload.

The findings indicate that not one of the workload scales is sufficient in themselves at capturing the complete profile of team workload. Each scale has its own strengths and limitations in their ability to accurately discriminate or describe task demands. However, together these scales paint a descriptive picture of the demands placed upon the team.

Further research into the development of a team workload scale is necessary. A reliable and valid team workload battery will be critical for the exploration of a range of collaboration technologies and design of team environments such as that used in command and control situations.

Acknowledgements

The authors gratefully acknowledge the contributions to this research of James Hysen, from Miller and Allen Duku, without whose innovative and efficient software development, troubleshooting, and lab support, the experiment would not have been completed. Special thanks are also extended to Becky Brown and Dan Schwartz for their role in developing the CBT and data collection.

References


