VERIFICATION AND VALIDATION OF LATENCY AND WORKLOAD PREDICTIONS FOR A TEAM OF HUMANS BY A TEAM OF COMPUTATIONAL MODELS

Thomas P. Santoro, Naval Submarine Medical Research Laboratory
David E. Kieras, University of Michigan
James A. Pharmer, Naval Air Warfare Center

Accepted for publication, U.S. Navy Journal of Underwater Acoustics, Special Issue on Modeling and Simulation

ABSTRACT

The behavior of a four-member air defense warfare team composed of experienced crews from Aegis guided missile cruisers was simulated by a team of four computational models for the purpose of verification and validation of the modeled output with data from the human teams. The individual models are capable of communication and (hence) collaboration. They provide predictions of task execution times and latencies and estimates of individual task and total operator workload for sensory-motor and cognitive modalities. An event interrupt mechanism facilitates free voice communications as well as both spontaneous and deliberate visual search. Each model team member could perform all its required actions without assistance from the other members when allowed spontaneous capture of all visual events. When models were allowed to capture visual events only during deliberate search periods, they needed assistance via verbal communication from other model team members to achieve coverage of events equivalent to that performed by the human teams. Model predictions of latencies and workload were found to correspond well to certain data from the human teams, although other data were not well matched. Further refinement of both the modeling tool and the empirical measurement techniques are needed to realize the tool's full potential for predicting these complex behaviors.

INTRODUCTION

The GOMS (Goals, Operators, Methods, and Selection Rules) methodology developed by Card, Moran, and Newell, (1983) for describing human sensory-motor and cognitive behavior is among the most common engineering models in use for human-computer interface (HCI) design. While well known for its capacity to deal with very fine details of the HCI and associated human behaviors (John & Kieras, 1996a, b), it has, however, rarely been applied to model and predict the performance and outcomes of team activities rather than activities of individuals working on their own.

Under the SC21 Manning Affordability Initiative, GOMS models of watchstanders in an air defense warfare (ADW) team have been constructed and validated against measured performance of real teams. These models were developed using GLEAN, the GOMS Language Evaluation and Analysis tool, created by Dr. David Kieras of the University of Michigan to support the application of GOMS in simulations of human behavior for the purpose of interface design (Kieras et.al., 1995; Kieras, 1998). The programming of communications and collaborations among model team members was facilitated by a dynamic interrupt mechanism in the GLEAN tool. The sensory modality processors in GLEAN can be programmed to generate asynchronous event-driven interrupts to on-going activity in the cognitive processor in order to insert volatile information from sensory memory into cognitive working memory .

For example, the auditory processor can be primed to constantly listen for keywords in ongoing voice communications over the team's internal voice network. When such a word or stream of words occurs, an interrupt service routine is triggered to load related information from sensory memory into the working memory store where it can be accessed by higher level cognitive processes at a later time. Analogous behavior in the visual system is problematic, however, as vision is used in conjunction with cognitive processes where specific items on a display or in a table are held in focus causing events outside of the field of view to be missed and therefore not available to trigger an interrupt. Thus the spontaneous capture of all critical visual events should be considered as the upper limit to expected visual performance and more realistic models of deliberate visual search should be used to probe for the lower limit in order to bracket expected performance. However, the automatic interception of auditory events via interrupts is a reasonable model of auditory behavior and essential to spontaneous inter-operator voice communications. Since spontaneous communication is critical to team collaboration and workload sharing, this interrupt mechanism is the key to using a team of GOMS models for studying these behaviors in a team of humans.

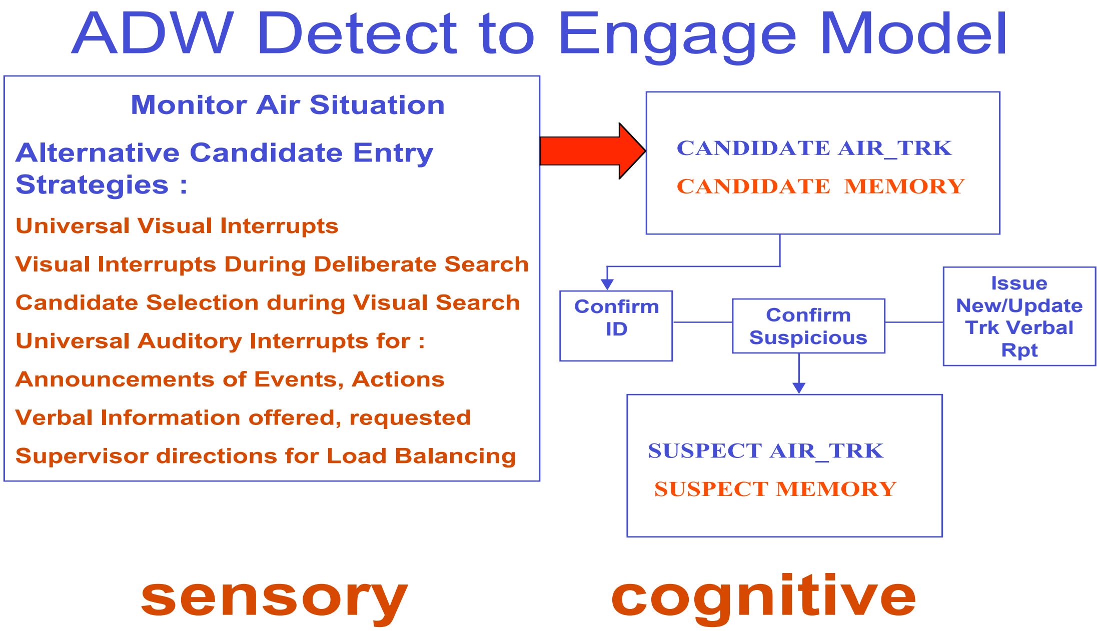
The value of human performance models in human system integration, team design, and human computer interface (HCI) design hinges on their accuracy in predicting actual human performance for system operations under realistic problem scenarios. Once this has been established, models can be used to provide much faster evaluation of initial designs and design changes, team composition and task allocation than the traditional empirical human user testing approach. In comparison to actual user testing, testing a design in a simulation with a human model or team of models facilitates the examination of a much broader range of design parameters in a much shorter time. This advantage can only be realized if the models can be relied upon to accurately represent real human performance. The Verification and Validation process (Pew & Mavor, 1998) is an accepted method to provide this necessary confidence in the use of models in simulations for design and training purposes.

GOMS MODELS FOR AIR DEFENSE WARFARE

The essential ingredient in GOMS modeling is the development of a representation of the tasks the human has to perform. The predictions of the model can only be as good as the quality of the task analysis upon which it is based. The fundamental premise of GOMS is the creation of models for complex tasks from combinations of elementary actions which are well defined and understood. Hence, a GOMS task description must be cast in terms of fine details involving elementary behaviors and requires significant knowledge of the tasks, the HCI, and the task execution strategies.

For this reason, GOMS models may not be appropriate for use in the earliest stages of a design when little is known about the details of system operations other than the constraints on performance for the highest level tasks required to satisfy mission goals. At that point, a modeling tool such as SAINT, Systems Analysis as an Integrated Network of Tasks (Laughery, 1989), may be the only choice to evaluate basic system feasibility. Conversely, as details of different components and their operating procedures become available, GOMS models can be used to improve the SAINT predictions and address more complex, system and HCI-dependent questions on task performance. Access to subject matter experts and system designers with first hand knowledge on the various tasks to be modeled is critical in creating GOMS task descriptions. The task descriptions used for this study were developed from extensive, in-depth analysis of the ADW mission and advanced, multi-modal watchstation designs by investigators at the Space and Naval Warfare Center Systems Division, San Diego, the Naval Surface Warfare Center Dahlgren Division, and Basic Commerce and Industries (BCI).

The top-level tasks for the ADW model are shown in figure 1. In general, these tasks follow the detect to engage process, sometimes referred to as the 'OODA Loop' (Observe, Orient, Decide, Act; Boyd, 1984). According to this sequence, 'Observations' are made by sensory mechanisms interacting with the HCI. Next, in the critical 'Orientation' stage, the meaning of those observations is interpreted in the context of the on-going tactical situation through assessment processes (Santoro and Amerson, 1998) as shown in figure 2. Depending on the results of that 'Orientation' stage, a threat assessment is developed that then leads to a 'Decision' being taken on possible 'Actions.' This sequence of processes is iterative with each Decision and Action stage followed by Observations and Orientations which serve to correct and guide the sequence to an acceptable end result.



ADW Detect to Engage Model

Monitor Air Situation

Alternative Candidate Entry Strategies :

Universal Visual Interrupts

Visual Interrupts During Deliberate Search

Candidate Selection during Visual Search

Universal Auditory Interrupts for :

Announcements of Events, Actions

Verbal Information offered, requested

Supervisor directions for Load Balancing

CANDIDATE AIR_TRK

CANDIDATE MEMORY

Confirm ID

Confirm Suspicious

Issue New/Update Trk Verbal Rpt

SUSPECT AIR_TRK

SUSPECT MEMORY

sensory

cognitive

Figure 1. Air Defense Warfare Top Level Tasks

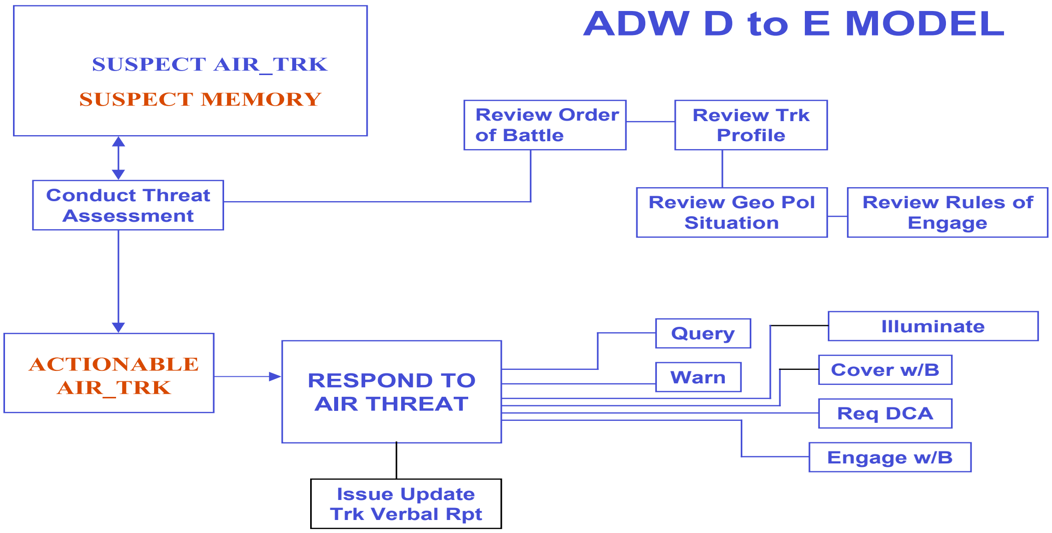


Figure 2. Orientation, Decision, and Action stages of the ADW Model.

ALTERNATIVE STRATEGIES FOR VISUAL SEARCH USING THE HCI

The Observation stage of the OODA loop depends both on the functionality of the HCI and the observer's capability to extract information from it. Critical pieces of information must be easily accessible when available and there must be an efficient strategy for searching the data to acquire this information in a timely fashion. The mechanics of simulating the positions of air track icons on a tactical situation display and dynamically adjusting close control read-out data for track locations with respect to own ship are a straight-forward, but non-trivial, programming problem which is implemented in GLEAN through the C++ language. The representation of visual search for new and changed track icons on a simulated tacsit display is a more difficult simulation problem that can tax the capabilities of much more complex modeling approaches than GOMS.

The philosophy of the GOMS methodology toward such problems is to bracket expected performance by posing hypothetical best case and worst case task execution strategies. This allows the estimation of high and low limits for very complex behaviors like visual search with simple combinations of elementary behaviors that have been well defined. For example, the average time to search a display and determine the presence of an icon that has good contrast and reasonable size has been established by experimental measurement, as have the average times to fixate a detected icon, interrupt on-going cognitive processes, and create a new location for later reference in working memory. These elementary behaviors can be combined to put useful limits on expected visual search performance.

When visual or auditory events are initially inserted into the GLEAN synthetic environment, objects are created for them in sensory memory, which has a nominal lifetime of 0.5 seconds. The objects must find their way into working cognitive memory during that brief lifetime in order to be available for further use in decision-making. For visual objects, this may occur via an automatic interrupt mechanism or, more realistically, by deliberate visual search. The upper limit for performance under condition of best possible capture of critical air track events is obtained when the events are immediately detected as they occur and logged into working memory for future evaluation. This condition, which can be easily implemented with the GLEAN dynamic interrupt facility, affords the model the maximum possible time to evaluate the potential threat and decide on appropriate actions as the exercise evolves. Any search strategy that does less than this universal coverage of critical visual events would be subject to increasing likelihood of missing events or responding to them too late to perform important actions.

More realistic less-than-universal visual search techniques would involve deliberate start and stop of the search process at selected points during or after execution of other tasks where vision is occupied. This would assume that when the operator is engaged in locating or reading visual information for a particular task from another screen window, the visual system is not capable of simultaneously detecting track events from a tacsit display window. A conservative lower limit for worst visual search performance would be the case where only the final range tripwire event would be captured, i.e. the model only notices hostile tracks when they come into such close proximity with ownship that they pose immediate danger. These two assumptions then bracket the range of expected visual search performance.

SENSORY-MOTOR AND COGNITIVE WORKLOAD MEASURES

In characterizing human performance, it is common to describe discrete tasks in terms of their relative distribution of work in sensory-motor and cognitive modalities. For example, a task that involves extensive visual search or monitoring of displays for detection and recognition of new or change events is considered to be a high visual workload task as opposed to one requiring careful selection and evaluation of information leading to fine judgments about tactical situations which would be considered a very cognitive task. In general such task workload descriptions are estimated by individuals familiar with the subject matter area who have performed related tasks. Ultimately, overall workload for a given job description is obtained by adding the work estimates over all modalities for the various tasks assigned to that position.

Since the GLEAN tool simulates the activity of sensory-motor and cognitive processes using time parameters derived from experimental psychology and psychophysics, it has the capability of estimating not only the complete time duration of a given task, but also the portions of that duration when the different modalities are active. For example, statistics are recorded by the GLEAN tool on the total number and duration of visual actions performed on each repetition of each task. The overall totals for any designated time period can also be computed as desired for any individual operator model. These numbers can form the basis for a visual workload estimate. In order to correspond with the observer estimates, these totals were made at 10 minute intervals for each GOMS model over the test scenario. By combining the estimates from the different modalities according to a stepwise multiple regression analysis a prediction equation can be constructed for workload and the correspondence of model workload totals to observer subjective estimates during the actual exercises can be determined.

COGNITIVE DECISION-MAKING AND WORKING MEMORY MANAGEMENT

The Orient – Decide stages of the OODA Loop are implemented with GOMS Methods for confirming air track identifications and assessing air threat actionability as shown in Figure 2. These functions are each supported by a dynamic working memory store that expands and contracts as needed to accommodate current track activity. Events captured during universal or deliberate visual search of the tactical situation displays are first stored in the candidate air track memory for later confirmation. This is because multiple events may be observed during a search period and also because more urgent functions may be pending and they must be completed first. Tracks have a short life span in the candidate air track memory, being deleted upon selection for confirmation.

When candidate air tracks are confirmed as possible threats, they are stored in a suspect air track memory and are periodically assessed to determine whether any action is required to be taken against them. The assessments are made in the order of entry of the tracks into suspect memory. Once a track is assessed as a possible threat, it is placed at the end of the suspect memory list and will be re-assessed continuously in order of its insertion in the list until such time as it becomes actionable or leaves the operating area. The GOMS tool does not place any limit on the number of working memory tags that can be assigned by a given Method. The Methods are provided with tags to store candidate and threat air tracks as needed and tags are deleted as they are vacated to conserve tag space. For the scenario used in this study, the maximum number of air track identifiers held in either of the two memory stores at any given time was nine.

In actuality, there are significant sensory-motor activities as well as decision-making logic involved in the confirmation and threat assessment Methods. This is because information must be acquired as needed in the decision-making process and memory tags are required to accommodate it. Such information is generally provided when a track icon is 'hooked' or pointed at with the cursor and clicked on, a process taking an average of over 2 seconds according to the GOMS model. Hooking a track opens a window known as the Close Control Read Out or CCRO table which displays range, bearing, closest point of approach, transponder and radar emissions and other critical track parameters.

Cognitive processes typically load several of these items into working memory, do comparisons and IF-THEN decisions with them, and then delete them from memory to conserve tag usage. Furthermore, various track properties such as icon color, air space, proximity to a comercial air route, and inbound or outbound with respect to ownship, must be fixated visually or they will not be available for use in the decision process. Hence, a good deal of the task execution time as well as the total memory tag usage for decision-making is actually to support information retrieval involving sensory and motor actions.

HUMAN TEAM CONFIGURATION

Data were collected on six human teams of five experienced ADW operators using warfighter centered design prototype watchstations in a simulated intermediate-to-advanced level ADW scenario. The two hour scenario was segmented into two halves of equal duration. The first half of the scenario was designed to be a low difficulty segment, consisting of a manageable load of low threat tracks, requiring a moderate number of action responses. The second half of the scenario, the high difficulty segment, contained a higher load of more threatening tracks, involving a high number of required actions. Latencies of actions taken on critical air tracks by the teams were recorded and workload for each team member was estimated by observers at ten minute intervals across the scenario.

Typical actions for a threat air track would include the initial new track verbal report, update report(s), possible multiple queries and warnings to the air track, requests to friendly defensive counter-air (DCA) assets for visual intercept, identification, and escort of the track, and defensive measures such as covering and engaging the track with ownship missiles. These actions are governed by pre-defined rules of engagement which can change during the course of the exercise and depend to some extend on information concerning the on-going geo-political situation in the hypothetical scenario. The time latencies of the actions taken by each team were recorded in seconds measured from the time of appearance of each of 25 critical air tracks. Individual workload for each operator was rated on a scale of 1 to 7 based on the observers' subjective judgment of how much overall effort each operator expended during a given interval relative to the other team members and the other measurement intervals in the scenario.

The composition and task allocation for the human team was designed by the scenario authors who based their decisions on their own expert knowledge of the task areas as well as the capabilities of the advanced watchstations the operators were using. Actual teams were composed of a supervisor and four specialist members. Each of the five team members was assigned an external communications circuit that provided contact with certain off-shipboard parties under his task responsibility; one operator communicated with other radar operators on other friendly vessels, another talked to friendly DCA, and so-fourth. Reports and messages on their assigned circuits formed a significant part of each operators tasking. In most cases, only the operator with a given circuit would perform the tasks related to communications on that circuit. Some operators were identified as workload sharers for tasks related to circuits not under their control, but they generally passed information or directions to the primary operator who then performed the report or message. In addition to their off-board circuit, operators all had access to an open internal circuit which was their primary means of communication for the purpose of workload sharing and collaboration.

MODEL TEAM CONFIGURATIONS

The rules of engagement (ROE) governing the scenario were quite specific as to prescribed reports and actions to be taken as a function of the position of a given threat track with respect to own ship. As such, the rules could be easily installed in a GOMS Method using simple IF-THEN decision logic based on observations of track parameters such as range, bearing, and any available radar emission records. In addition, an “Expert Solution” for the scenario was provided by its designers which identified time windows for prescribed actions that should be taken according to the ROE. These items, along with the task analysis information, provided sufficient guidelines for developing and testing different versions of the GOMS models used in the Verification process in order to iteratively “home in” on the best possible team performance estimates.

Initial models involved a set of GOMS Methods that included the cognitive working memory and decision-making logic needed to perform the actions required in the Expert Solution and the sensory-motor behaviors to acquire the necessary information to support decisions using the HCI. In the first team model, the Methods were assigned to three separate operator models corresponding to three members of the five-member human team. The three operators chosen for modeling were the ones with primary responsibility for the critical reports, queries, and warnings, as well as the defensive actions involving threat air tracks in the scenario. These are the Air Warfare Coordinator (AWC) and the two Information Quality Coordinators (IQC1 and IQC2). The team leader, or Tactical Action Officer (TAO), tasks consisted largely of monitoring the performance of those three operators based on his own evaluations of the threat tracks, redirecting their efforts when needed, confirming their statements of intended actions as appropriate, and passing along to them high-level strategic information from an off-board circuit under his control. These tasks were not modeled although a fourth team member was later found to be necessary for workload sharing on certain model teams. No model was build for the fifth real team member, who was dedicated to communicating with friendly DCA for extended periods concerning their surveillance and prosecution of threat air tracks.

Three-station-universal-search model. In the first case, runs against the test scenario with the three-operator model had each operator working independently on their own separate tasks without verbal communications or any collaboration with the other two team members. In addition, visual search was assumed to occur in parallel with all other activities and capture all critical visual events for examination by the cognitive processor at a later time. Under this condition, no track appearance or change event was missed by any of the three model operators, all appropriate actions on critical tracks were taken, and the time latencies of the actions were similar to, and in many cases shorter than, the fastest times produced by the human operators. This model, designated as **TSU**, represents the best expected performance, corresponding to the upper bracket, but it involves an unrealistic ability to reliably detect all of the significant visual changes in the display.

Three-station-deliberate-search. The next case attempted to bracket the lower performance limit by using the same no-communications model with the additional restriction of brief 3-5 second deliberate search time windows occurring between threat-related activities. Only track appearance and change events that happened to fall within these windows would be captured by this model. If such events were not captured, the next opportunity to respond to a threat track would occur when it crossed the last range tripwire and came dangerously close to ownship. Again, the model operators could not collaborate through verbal or any other communication mode in order to alert each other to critical track events. This model resulted in the highest number of missed required actions of all models tested, did not have a good fit of predicted to actual latencies for actions taken, and the relative workload fit was also poor. This model, designated **TSD**, defined the worst-case performance, or lower bracket, model.

MODELS WITH VOICE COMMUNICATIONS

Further models were build to bridge the above bracket cases and make a closer match to actual team performance in the model Verification process through the introduction of various hypothetical inter-operator communications modes. The actual human teams freely communicate over their internal network depending on different individual styles. Many ad-hoc remarks are made about track events and various pieces of information are passed. It is not at all clear from a study of these communications to what extend they are useful to, or used by, their respective recipients. Hence our approach was to hypothesis the passing of certain discrete pieces of information on threat tracks and determine its effect to drive the worst case results towards those of the best case and thereby iteratively approach the actual team performance as closely as possible. In this endeavor, two variations of a three member team with voice comms and two variations of a four member team with voice comms were built and tested as follows.

Three-station-voice(1). Since the assigned tasking to one of the three model operators was clearly less than that for the other two, it was decided to give that model an additional visual search task thereby creating a new team model designated as **TSV(1)**. This operator's primary responsibility was only for responding to detections by ownship's passive electronics surveillance measurement (ESM) systems and reporting them over the off-shipboard network under his control. The deliberate search for new and changed air tracks performed by the other two operators was considered a secondary work-share task for this operator and so it was added to the regular top-level task procedure as follows. The model was programmed to do a periodic 3-5 second visual search and report all appearance and change events captured for threat tracks over the local network which would be heard by the other two model operators. The interrupt rules for the two listener models were set to key on these reports and load the identifying information for the new or changed track into their respective new track working memories as though they had been captured by their own visual search process.

Three-station-voice(2). A variation, designated as **TSV(2)**, of the three-member team model was next created in which the ESM operator reported threat ESM information locally in addition to sending it over the off-board network. Threat ESM information is an important cue for certain defensive actions by the other two operators and is often passed along by the real team ESM personnel. If one of the other operators has missed a threat track, hearing such a report should bring attention to that track. The ESM model operator still had to do the visual search task and make reports on all threat track events captured as well.

Four-station-voice(1). Among his available actions as team leader, the TAO may alert the others to new or changed tracks which they may have missed. Again, it is difficult to determine the extent to which this is done in the different real teams and its effectiveness in improving team performance. To model this function, a fourth member model, with verbal communications, was added to the three member team creating a new team using deliberate visual search and designated as **FSV(1)**. This member's only task was to do brief 3-5 second deliberate searches on the tactical situation display and announce critical events concerning threat air tracks that were captured in the search.. the other members were programmed to listen for these announcements and, once again, store the relevant track information in working memory as thought it had come from a deliberate search of their own.

Four-station-voice(2). As a final team model, the **TSV(2)** and the **FSV(1)** models were combined to create a model with two listener operators and two speakers. The team leader did the same search and announce task as in **FSV(1)** while the ESM operator only announced threat ESMs to the team as opposed to the case of **TSV(2)** where both ESM and captured new and changed threat tracks were announced by the ESM operator. In this model, designated **FSV(2)**, only the team leader did new and changed threat track announcements**.**

STEP 1. VERIFICATION: BEST FIT TO FIRST HALF DATA FROM SIX REAL TEAMS

The available data from human team exercises consisted of latencies on track actions and estimates of individual operator overall workload made by observers at ten minute intervals during the exercise. The workload observers were selected for their expertise in the tasks of a particular watchstander and were instructed to base their estimates on the activity level of that operator relative to their estimate of the individual's maximum possible work output. Latency data was taken for 25 air tracks in the scenario which had been designed to be of critical importance. While there were over 100 air tracks in the complete two hour scenario, omission of required actions on these critical 25 tracks would indicate failure to perform in accordance with given rules of engagement (ROE) for the exercise. Moreover, the timing of events as determined by air track initial bearing, range, course, and speed, and their subsequent maneuvers served to create rather narrow windows for the critical tracks in which actions had to be performed if the ROE were to be satisfied.

Data from the first half of the test scenario, the low difficulty period, was provided for the Verification process. There were 11 of the 25 critical tracks in this period. All required actions on these tracks were taken by the combined six human teams, although errors of omission and commission were made by individual teams. In addition, a set of four workload values covering the period were provided for each of four operators in the five-operator real teams. The Verification process for matching predicted to actual data consisted of doing the iterative adjustments to functionality and composition of the six GOMS model teams previously described, holding fundamental GOMS behavioral parameters and operational parameters of the HCI constant, until an acceptable match was reached. Once the theoretical upper and lower expected performance brackets were defined, it was necessary to select criteria to govern the iteration process to get the best fit between those limits. Otherwise, this process could be quit open-ended without reasonable stopping criteria. Three criteria were chosen to guide the Verification process through the different model iterations as follows.

CRITERION 1: Completion of Required Actions.

All real teams made new track reports on all 11 first half critical tracks. While several teams did make omissions for the other required actions (queries, warnings, etc) on some critical tracks, they were few enough to expect that the model team should perform at or very close to a 100% action completion criterion. Since the ROE are built into the GOMS decision-making process, the models will only take actions when all ROE conditions are satisfied. Hence, 100% completion of required actions by a given model is highly dependent on how effective the visual search strategy is in locating those critical air tracks in a timely fashion before the ROE conditions are no longer valid.

Certain models performed very well in terms of their adherence to the actions prescribed by the ROE at the earliest permissible times when they became valid to be performed. The **TSU** model, since it automatically acquired all track appearances and changes, did complete 100% of the required actions with very short latencies, occasionally even taking actions sooner than the ROE required. The challenge to the models with more realistic deliberate search was to also take the required actions in the appropriate time windows. After several iterations, a team model was built that did so. This was the **FSV(2)** team, the four-member team with two members speaking over the internal net to the other two operators for the purpose of helping in the early acquisition of new air tracks. No three-member team, nor the four-member team with just one internal speaker, could achieve this stopping criteria.

CRITERION 2: Workload Estimation.

A first question about the measured workload data is its reliability; these ratings were produced by a handful of observers, only one per workstation, who were asked to make subjective estimates of a relatively ill-defined concept without any detailed training or instructions. Clearly any attempt to understand or predict such data requires first that it is well-behaved and consistently reflects aspects of the task that were supposed to be present. As a first look, the reliability of the workload ratings was assessed by considering the extent to which the workload ratings over time for a particular job role in a team correlated with the ratings for the same role in a different team. This sort of question could potentially be addressed by some of the standard reliability measures which summarize such intercorrelations, but the structure of this data does not appear to match the structure assumed by these measures. An approach similar in spirit to those measures was to compute all possible correlations between the ratings for each role on each team and that of the other five teams. In addition the ratings for each time period and role were averaged across teams as a summary measure. The correlations of each individual team and role with this mean rating was also determined. Table 1 summarizes the results of this analysis for each workstation/job role.

Job/Role	AWC	IQC1	IQC2	TAO	Mean
Mean intercorrelation	0.582	0.688	0.604	0.555	0.607
Mean correlation with mean	0.807	0.860	0.817	0.790	0.818

Table 1. Reliability of the Workload Ratings

As an example, for the AWC role, the eight time-period workload ratings for an individual watchstander correlated, on the average, at 0.582 with the other AWC watchstanders. On the average, the individual AWC watchstander's ratings correlated at 0.807 with the average of all six AWC watchstanders' ratings. There was considerable variation in the individual intercorrelations, ranging from a low of -.124 to a high of 0.944, but no one team stood out as problematic. The correlations with the mean ratings are fairly high; the lowest observed was 0.492, and the highest was 0.951. Of the 24 possible correlations of individual watchstanders with the mean ratings for the role, 22 were significant at or beyond the .05 level. As shown in Table 1, the mean correlations for AWC, IQC1, and IQC2 are similar in magnitude, while the TAO role had the lowest intercorrelations and correlation with the mean ratings. This is consistent with its essentially flat workload profile show in Figure 3; there was relatively little systematic variation in this role's workload over time. Overall, the intercorrelations are high enough to be considered adequately reliable as an assessment instrument. The correlations with the mean workload ratings are high enough to assume that the mean ratings adequately reflect the individual team ratings.

Effects of Role and Time Period on Workload Ratings. To the extent that the workload ratings are both reliable and meaningful, there should be systematic and reasonable effects of job role and time period. The workload ratings were subjected to a 4 X 8 X 6 repeated measures analysis of variance. The four workstation/role levels were a between-subjects factor, and the eight time periods were the between-subjects factor; there were 6 subjects nested in each workstation/role level. Each cell in the data table was thus occupied by one workload rating for one individual watchstanders at one time period. The main effect of workstation/role was marginally significant (F(3, 20) = 2.36, p = 0.1025). As shown in Figure 4, overall, the level of workload was very similar for all roles except for the TAO (Team Leader) where it was slightly less. Note also that the overall level of workload rating averaged only 2.14.

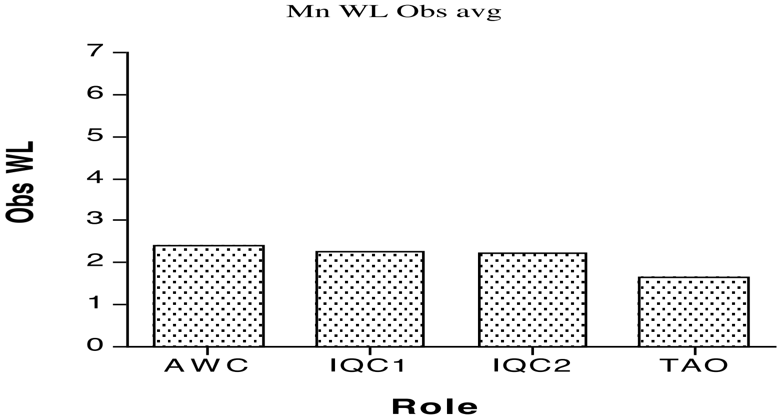


Figure 3. The mean observed workload rating for each workstation (job role) averaged over all eight periods and six teams.

The effect of time period was quite significant (F(7, 140) = 32.30, p < 0.001). As shown in Figure 4, the mean workload was highest in the last period, and overall was higher in the second half of the scenario, consistent with the intended effect of the scenario design. Again note that the excursions in workload are not large, varying from a minimum of about two to a high of about four.

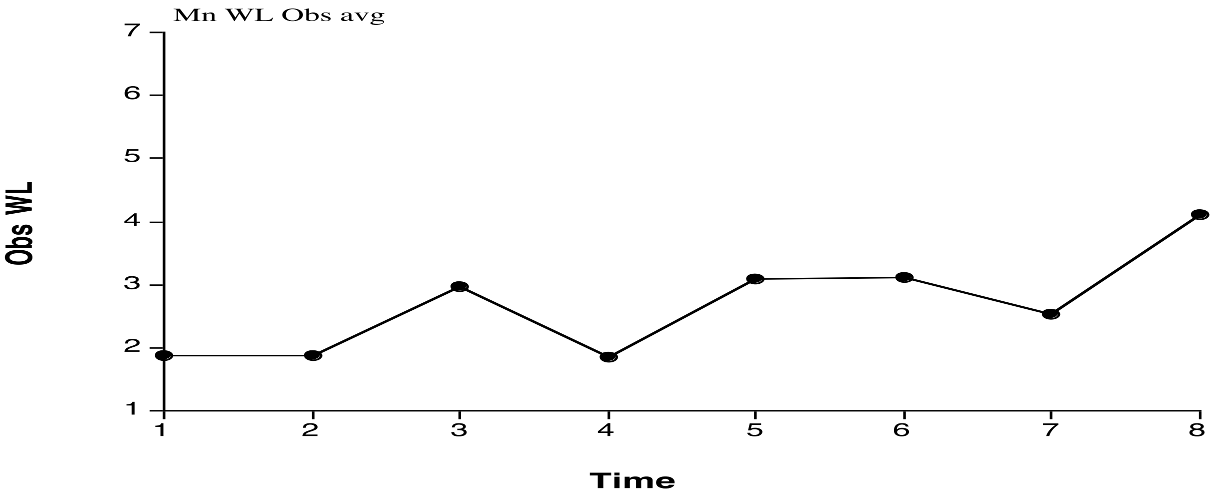


Figure 4. The mean observed workload rating for each time period, averaged over all workstations (job roles) and teams.

Of most interest for this project, the pattern of workload variation over time was different for the different workstation/job roles. This is shown by the significant interaction of workstation with time period (F(21, 140) =3.47, p < .001), shown in Figure 5. Clearly the watchstations were affected differently by the distribution of task activities over time.

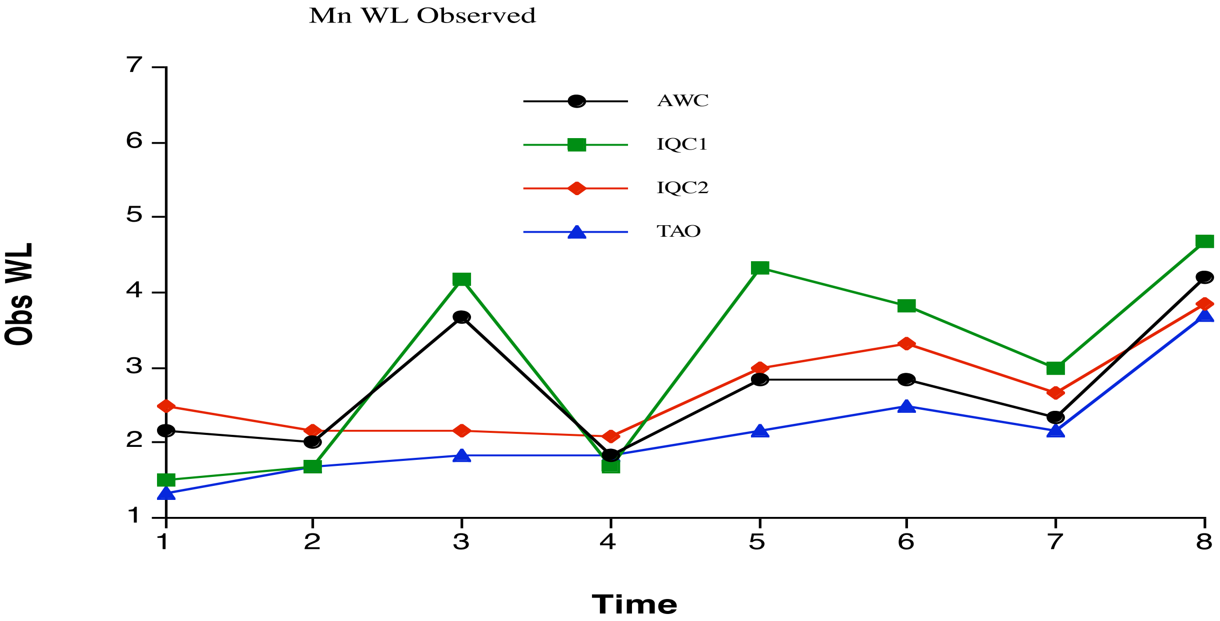


Figure 5. The mean observed workload rating for each time period as a function of job role, averaged over all teams.

All job roles had their peak workload in the last period,. But the TAO role had a basically flat rated workload through the first seven periods . The IQC2 job averaged somewhat higher than TAO, but was also fairly flat across time. AWC workload was similarly uniform, with the one exception of the third time period, where it spiked to almost as much as the last time period. IQC1 was generally the most affected by the scenario, with high workload in the third, fifth, and sixth periods, as well as the last.

To provide additional information for the split-half validation procedure, a similar analysis of variance was performed using the data from only the first four time periods. The same pattern of effects at similar significance levels was obtained, showing that the pattern for the first four time periods in Figures 4 and 5 was also statistically reliable.

Together with the correlational assessment of reliability, these results show that the workload ratings are a well behaved measure of how the level of task activity changed over the time periods and changed differently for the different job roles. However, these analyses do not clarify what these ratings were actually measuring; it merely says that whatever was being measured is reasonably reliable and shows effects consistent with how the scenario was designed. It should be kept in mind that the mean intercorrelation was only about 0.6, and thus it should not be expected to be able to predict the mean data substantially better.

Workload Regression Analysis. The **FSV(2)** model was used to try to predict the mean workload ratings for each job role at the eight time periods. These are the data plotted in Figure 5. Following the split-half validation logic, the first step was to fit the model to the workload ratings from the first half of the scenario (time periods 1-4). This subset of the data thus had 16 data points, 4 for each team role X 4 time periods. A stepwise multiple-regression analysis was performed in which all of the vf6 model's workload predictors were allowed to compete for entry into the prediction equation, along with dummy-coded variables for the roles, in case there was a systematic difference in average observed workload value for the roles.

Only two predictors entered the equation. One was the vocal activity, but in addition, the auditory activity entered also. The resulting prediction equation was:

$$WL = 1.323 + 0.08795 * \text{auditory} + 0.01589 * \text{vocal} \qquad \text{(Eq. 1)}$$

Eq. 1 accounted for 54.7% of the variance (adjusted R2 = .477). Both the vocal predictor and auditory predictors were significant ($p < .01$). Fig. 6 shows a scatter plot of these results.

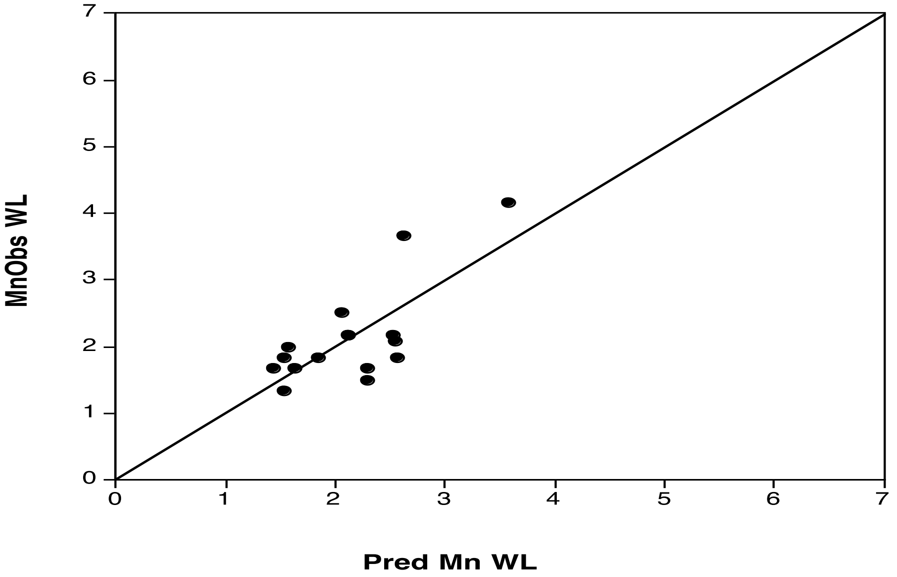


Fig. 6. Scatterplot showing predicted and observed workload values for the first half of the data using Eq. 1 to predict the values.

CRITERION 3: Action Latency.

The time latency from initial air track appearance on the tactical situation displays to each of the operators' actions for the 25 critical air tracks over the entire two hour scenario was recorded for each of the six test teams and their medians were calculated. Since certain GLEAN Operators for sensory processes are probabilistic and depend on a random number starting point, model runs on the test scenario for different starting points can result in different time latency predictions for required air track actions. Hence, medians of predicted latencies over four different model runs were calculated for the 25 tracks.

For the first half Verification process, the medians on the 11 critical tracks in that period were averaged to form the average predicted latency for each of the major required actions: new track reports, queries, warnings, and ESM racket reports. Other actions such as update reports, requests for visual ID or escort, and engagement actions were too infrequent or otherwise complicated to analyze. The absolute percent error of predicted to actual average overall latencies for the selected actions on the 11 tracks was calculated for each of the six model teams and is given in table 2. It is generally the case that the actual team latencies were longer than those predicted by the models. For example, as shown in Table 3., the 32.67% difference in average query latency for model 6 is caused by the predicted value, 3355 seconds,

	TSU	TSD	TSV(1)	TSV(2)	FSV(1)	FSV(2)
NTVR	6.692585	242.193	113.6374	105.5206	51.78423	13.04699
QUERY	34.58329	72.81135	22.1072	30.12349	38.54178	32.67472
WARN	43.27371	32.54104	44.27836	16.63808	12.15388	9.556481
ESM RPT	13.95349	13.95349	12.65451	12.71737	13.70207	13.89063

Table 2. Average Absolute Percent Error of Latency Predictions

	TSU			FSV(2)		
	real team	model	abs%err	real team	model	abs%err
ntvr	151.8	141.7	6.6	151.8	132	13
query	528.6	345.8	34.5	528.6	355.8	32.6
warn	680.1	385.8	43.2	680.1	615.1	9.5
esm	298.3	256.6	13.9	298.3	256.8	13.8

Table 3. Average real and predicted latencies for models TSU and FSV(2).

being some 173 seconds less than the actual value, 528 seconds. Occasionally, actions were even taken before they were valid according to the pre-defined ROE time windows. In some of these cases discrepancies between the GLEAN track motion simulation and that used in the actual team testing may be responsible.

For the model with universal visual event capture, **TSU**, most queries and warnings are done at the earliest possible valid (or near valid) ROE time explaining their short latencies compared with the real teams, as also shown in Table 2. In addition, the shorter predicted latencies in the models must to some extent be due to other activities, both sensory-motor and cognitive, that are taking place in the real teams that have not been fully accounted for in the models thereby delaying the real teams' activity with respect to the models. This implies that gaps exist in the task analysis and further refinement of the GOMS models would be required to shore them up.

As occurred for the first two stopping criteria, the **FSV(2)** team once again did better in this criterion than the other teams. This model thus satisfied the 100% critical action criterion and had the closest fit to the observed workload trend as well as the measured action latency values. Iterations were therefore brought to a stop with this model and it was selected as the model team of choice for use in the following Validation process.

STEP 2. VALIDATION: BEST FIT TO SECOND HALF DATA

Once a team model, the **FSV(2)** team, was built that performed acceptably according to the three Verification criteria as applied to the data from the 11 air tracks in the first half of the test scenario, that model was used to predict the data on the remaining 14 critical air tracks from the second (high difficulty) half of the same exercises for the Validation process following the split-half validation process. As shown in the following sections, the **FSV(2)** team model did the best of the models in completing required tasks on the 14 critical second half air tracks, in several cases the general fit for latency predictions with **FSV(2)** was within ten percent, a common rule of thumb for engineering design purposes, and the predictions of overall workload on the Validation section of the exercise was a reasonable match to the trend of the observed workload averages over that period.

CRITERION 1: Completion of Required Actions.

The second half of the test scenario was designed to be more challenging to the operators and required an advanced skill level on their part to negotiate it successfully. In comparison to the first half, there were more threatening air space incursions by known hostile air tracks requiring quick decisions and defensive actions, including defending a missile attack on ownship. Of the six models constructed for the Verification process, only the team with universal interrupts completed 100% of required actions on the 14 critical tracks in the second half. The **FSV(2)** team, the best of the non-universal interrupt teams, missed one required action concerning a non-threat commercial air track that should have received a new track report. The team did achieve 100% performance on all threat air tracks and all query and warning actions. The four other model teams did not do this well, as they missed required query and warnings on certain threat tracks.

CRITERION 2: Workload Estimation.

Workload Regression Analysis. The prediction equation developed from first half only data was used to predict the second half workload for the human teams. Fig. 7 shows these results and the regression fit of these predictions to the second half data.

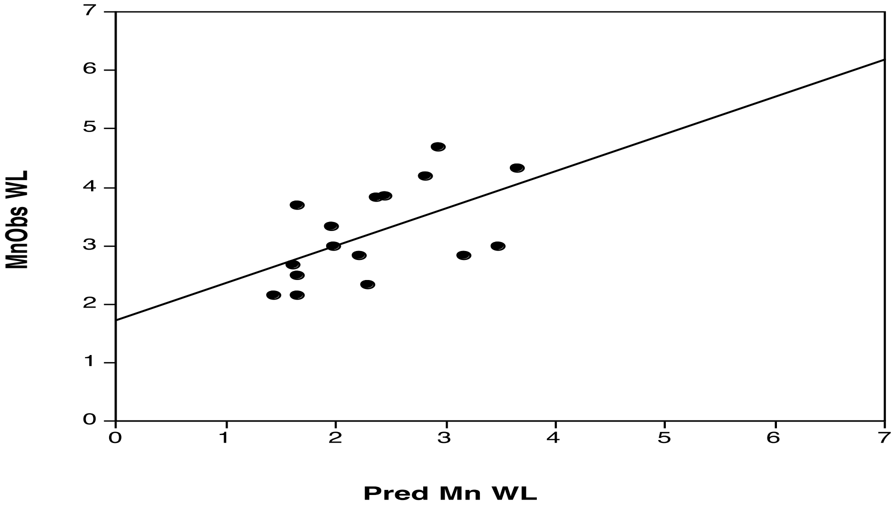


Figure 7. Observed and predicted workloads for the second half, using the same prediction equation that was derived from the first half data (Eq. 1). R2 = .31.

The R2 for predicting the second half workload ratings with the regression equation from the first half is .31. The prediction equation also provides a scaling from the predictor variables to the observed values in a way that facilitates comparing predicted and observed.

Figure 8 shows the Observed and Predicted workloads for AWC and TAO across all 8 time periods. The Predicted values are those from Eq. 1. For these two, the predicted and observed values track quite close to each other, but the fit is better in the first half (1-4) than in the second (5-8). Note that for both roles, the model under predicts the workload substantially at time 8.

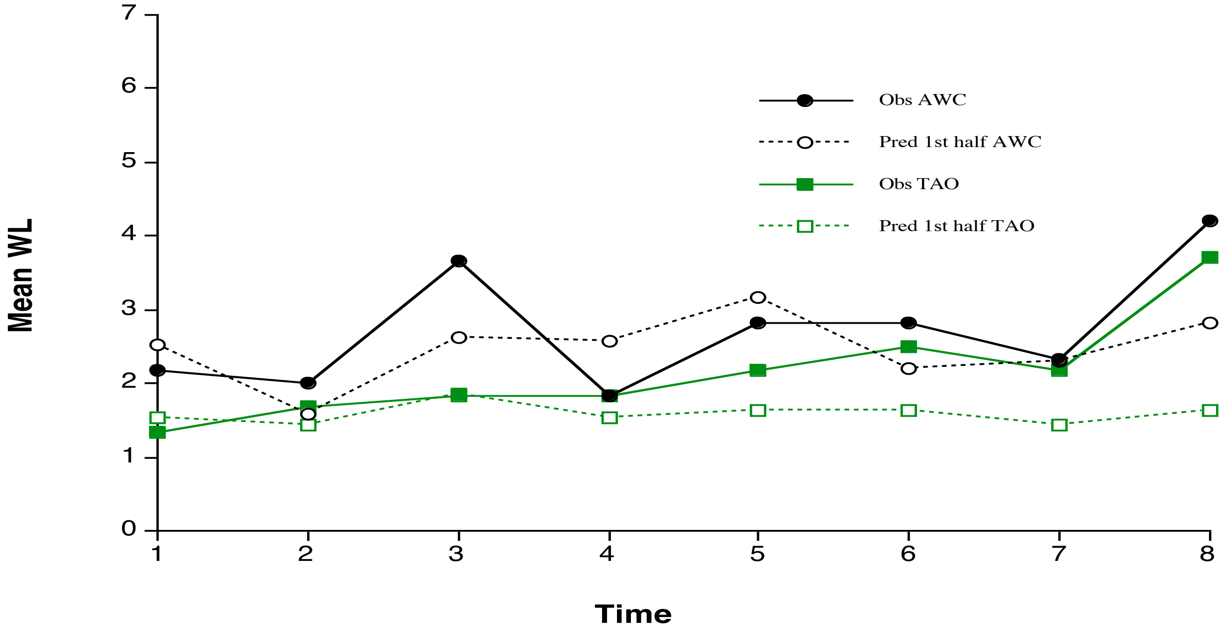


Figure 8. Observed and predicted workloads for AWC and TAO for each time period; the predicted values are from Eq. 1.

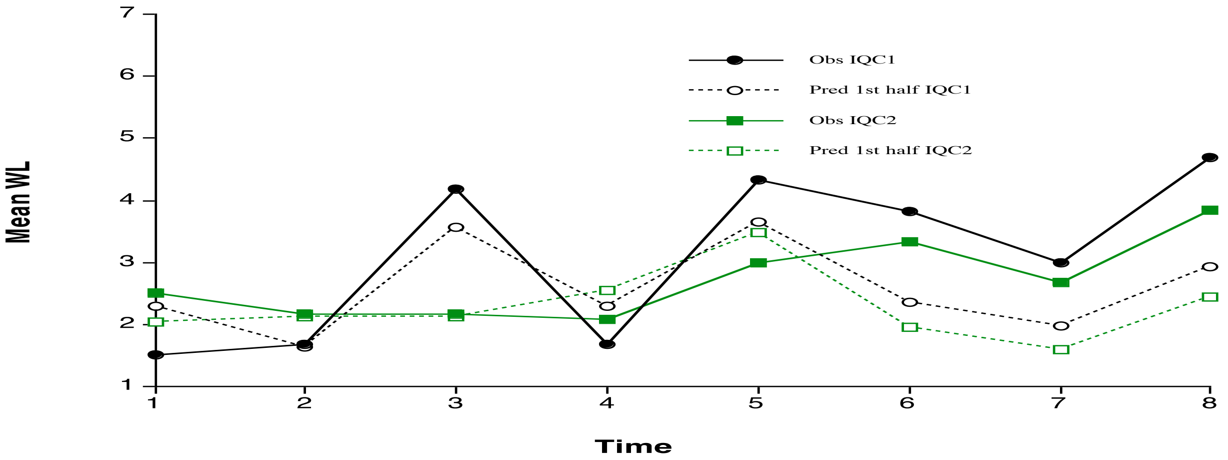


Figure 9. Observed and predicted workloads for IQC1 and IQC2 for each time period; the predicted values are from Eq. 1.

The match isn't as good for the two IQC stations, especially for IQC1, as shown in figure 9. Again, the predicted and observed values track pretty closely in the first half, but in the second half, IQC1's workload increases a lot, as does IQC2's to a lesser extent, but the predicted values for both do not and in fact are nearly identical.

The conclusion is that the model is not predicting the second-half workload very well, and this could be true for a variety of reasons. However, the range of workload variation in the first half was rather limited, and this might have led to poor estimates of how to scale the predicted values to the observed.

Post-hoc Evaluation of Workload Predictions. To evaluate this possibility, a second multiple regression analysis was done in which the data from both halves of the exercise was used. The same two predictor variables again were the only significant predictors, and yielded the following prediction equation:

WL =1.940 + 0.01108 * Vocal + 0.09737 * auditory (Eq. 2)

This accounted for 39.1% of the variance (34.9% adjusted). Both variables were significant ($p < .05$). Figure 10 shows the scatter plot for all of the data:

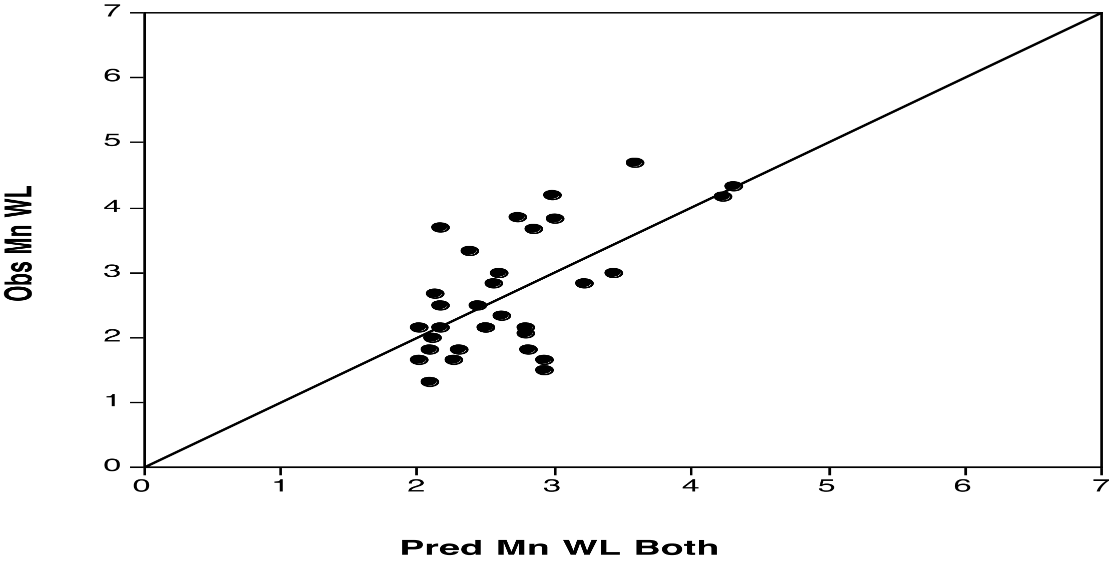


Figure 10. Predicted vs observed values, where the predicted values are based on both halves of the data, using Eq. 2.

The AWC and TAO plots for prediction Eq. 2 were very similar to those for Eq. 1. However, the IQC1 and IQC2 predicted values, shown in figure 11, are much improved for the second half, especially for IQC1, which benefits from auditory activity involved in listening to responses from queried and warned air tracks. These responses are significantly longer than the simple “aye-aye” answers the AWC and IQC2 hear in response to their reports. Unfortunately, in the current GLEAN tool, no workload records are taken for the interrupt-driven auditory work performed by the AWC or IQC1 in listening to asynchronous remarks from the TAO and IQC2 over the model team’s internal voice network.. These messages are more frequent than the query and warning activity and perhaps could offset the increased workload prediction for IQC1 over AWC by the regression equation.

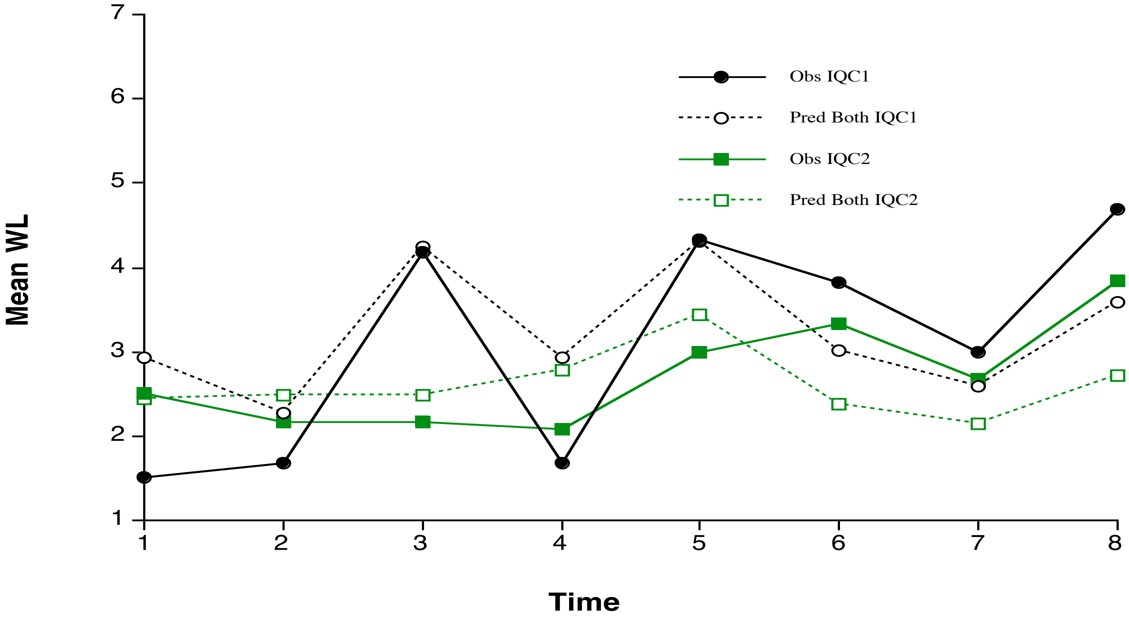


Figure 11. Observed and predicted workloads for IQC1 and IQC2 for each time period; the predicted values are from Eq. 2.

Nevertheless, the observers would not necessarily note that AWC or IQC1 were listening to the TAO and IQC2 comments, the real operators might have given no indication of the fact that they heard the transmission and stored important information away in working memory. On the other hand, the fact that both auditory and vocal activity are again the only significant predictors for Eq.2, suggests the observers, particularly for the second half of the exercise, were basing their ratings heavily on two-way conversations and so the listening portion of the query and warning tasks done by IQC1 could have thus received due attention as a significant part of the workload.

The predictor equation based on both halfs of the data accounts for around 40% of the variance in the observers' workload measures. this is an excellent result given that the estimates are not based on a very well-defined scale as compared to the predictors used to match it. The modeling predictors are measures of the amount of activity going on in various modalities (e.g. counts of vocal actions) and do not directly relate to a 1-7 rating scale like that used by the workload observers. By using a regression analysis, we obtain not only the scaling function to relate these activity metrics to workload ratings, but also information about which activities in the model are most closely related to the observer's ratings.

The fact that the vocal activity is the most important predictor suggests that the observers were basing their ratings heavily on how much talking the operator was doing. IQC1 is strongly affected by the additional activity in second half. In the vF6 model, this appears to be mainly due to how IQC1 is the only team member that has to process speech information supplied by the other team members. In effect, the auditory actions predictor variable is supplying IQC1 an "extra" amount of workload that is related to the overall level of activity in the task, showing that in some sense, IQC1 is working harder than anybody else, at least according to these models. The prediction equation based on the both halves is able to use this "extra" factor to improve the fit for IQC1, but it makes less of a difference if we predict based on just the first half. The role of the auditory predictor should be kept in mind when interpreting these results.

In particular, the fact that the model's vocal actions were a good predictor of these ratings makes suggests that simply counting the different kinds of actions performed by human users would be a better choice than workload ratings, both being more objective, and easily collected by simple instrumentation of the task environment.

CRITERION 3: Action Latency.

Absolute average error for action latencies in the second half scenario was lower than 10% for warnings and ESM report actions for the **FSV(2)** team model . However, it grew to 33% for new track reports and over 45% for query actions. As in the first half scenario, the general trend was for the model operators to perform these actions sooner than the actual teams. Most early actions were the consequence of the track being picked up earlier by the models, given new track reports earlier (by 87 seconds on average), and queried earlier (on average some 182 seconds sooner), on occasion even earlier than the ROE window of validity was supposed to be open. Given that the workload for the IQC1 operator, who has primary responsibility for queries and warnings, was underestimated by the model, this result lends further support to the possibility that this operator was doing additional tasks not treated by this model. Another possibility is that, while the number of required queries in the second half is not dramatically greater than that in the first, the nature of the threats may have necessitated more repetitions of the query actions or related activities than in the first half. Moreover, the AWC workload estimate was correspondingly overestimated in the second half and , together with the fact that the IQC1 had a lot of conversation with the AWC in the real teams, this may indicate additional workload sharing was going on that was not accounted for by the model.

SUMMARY AND CONCLUSIONS

The results of the workload predictions are encouraging for the prospects for further development of workload prediction with a GOMS tool. The change in workload distribution between team members for the low and high difficulty halves of the scenario presents another area for further study and model development. The models developed to this point addressed each operator as an individual who had primary tasks assigned and then might perform a secondary task for another team member as a backup. However, only in the case of visual search, have the models taken account of the extent to which two or more individuals might complete a given task together. Models that account for more of this kind of collaboration could produce better predictions for the dynamic changes in relative operator activities that were observed in these exercises.

The poor correspondence of visual and cognitive workload predictions to overall workload observations also requires further study. Various memory management parameters were tried in the effort to obtain a good predictor for cognitive workload. The memory tag creation, change, and deletion statistics, for example, were tried and did not follow either the across subject or time-wise trends in the subject matter expert observations. Some success was found in efforts to balance deliberate visual search and cognitive processing episodes in the scenario. This was based on the assumption that search would be likely to occur during periods of low cognitive demand as represented by lack of threat track activity. Further work must be done before some type of activity-dependent visual search strategy can be incorporated into the models.

The fact that team voice communications were valuable mainly to help with the visual detection problem suggests that team performance could be improved, possibly to the point of dedicating three rather than four team members to ADW, by improvements in the workstation design. The team communication channel would then be freed for other, more complex, team activities. Previous modeling work for a notional 'next track' button explored the potential for directing operator attention to the air track most likely to require action at any given time in the scenario. When multiple tracks require servicing, such an intelligent decision aid could assist in identifying the most important one to do or allocating different ones to different operators depending on their current workload. The key role of voice communications in operator collaboration as well as the good correspondence of verbal workload predictions to overall workload observations suggests that further development of more detailed, realistic models for the GOMS verbal processor in GLEAN would also be valuable.

Finally, this work demonstrates that the concept of modeling a team of humans with a team of models of individual humans capable of spontaneous communication is a viable approach to bridging the gap between the psychology of individual humans and the organization and functioning of teams. The assumptions and mechanisms supporting the models for sensory memory and information transfer from sensory to cognitive memory are key in implementing this kind of interactive team member model. Further development of task descriptions for dynamic workload sharing by two or more team members is needed to improve the overall workload prediction capabilities.

REFERENCES

Boyd, J.R. (1984). Organic design for command and control. Unpublished briefing paper, pp. 5, 32-35.

Card, S., T. Moran, et al. (1983). The psychology of human-computer interaction. Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.

John, B. E., & Kieras, D. E. (1996a). Using GOMS for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction*, 3, 287-319.

John, B. E., & Kieras, D. E. (1996b). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction*, 3, 320-351.

Kieras, D.E. (1998). *A guide to GOMS model usability evaluation using GOMSL and GLEAN3.* (Technical Report No. 38, TR-98/ARPA-2). Ann Arbor, University of Michigan, Electrical Engineering and Computer Science Department. January 2, 1998. Current version available online at ftp://www.eecs.umich.edu/people /kieras /GOMS/GOMSL_Guide.pdf.

Kieras, D. E., & Meyer, D. E. (2000). The role of cognitive task analysis in the application of predictive models of human performance. In J. M. C. Schraagen, S. E. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis*. Mahwah, NJ: Lawrence Erlbaum, 2000.

Kieras, D.E., Wood, S.D., Abotel, K., & Hornof, A. (1995). GLEAN: A Computer-Based Tool for Rapid GOMS Model Usability Evaluation of User Interface Designs. In Proceeding of UIST, 1995, Pittsburg, PA, USA. November 14-17, 1995. New York: ACM. pp. 91-100.

Laughery, K.R. (1989). MicroSaint – A tool for modeling human performance in systems. In McMIllan, G.R., Beevis, D., Salas, B.E., Strub, M.H., Sutton, R., and Van Breda, L. (eds.) *Applications of human performance models to system design*. New York: Plenum Press.

Pew, R. W. & Mavor, A. S. (Eds.)(1998) Modeling human and organizational behavior: Application to military simulations. National Research Council Commission on Behavioral and Social Sciences and Education Panel on Modeling and Command Decision Making: Representations for Military Simulations. Washington, DC: National Academy Press.

Pharmer, J.A., Freeman, J.T., Kieras, D.E., Santoro, T.P., Brockett, C. (2001). Complementary methods of modeling team performance. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting,* pp 1723-1727, Minniapolis, MN. Oct, 2001.

Santoro, T.P., Kieras, D.E., and Campbell, G.E. (2000). GOMS modeling application to watchstation design using the GLEAN tool. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*, pp. 964-973, Orlando, FL. Nov, 2000.

Santoro, T.P., and Amerson, T.L. Definition and measurement of situation awareness in the submarine attack center. *Proceedings of the Command & Control Research & Technology Symposium*, pp. 379-388, Monterey, June, 1998.