VERIFYING AND VALIDATING THE AEGIS AIR DEFENSE WARFARE HUMAN PERFORMANCE MODEL

Christine H. Brockett
Shelly Scott-Nash
Micro Analysis and Design, Inc.
Boulder, Colorado

James A. Pharmer
Naval Air Systems Center
Training Systems Division
Orlando, Florida

ABSTRACT

In some industries, simulation and modeling techniques are a widely accepted, integral part of system design, while in others these techniques may be perceived as expensive, unreliable, or inconclusive. Within the Manning Affordability Initiative (MAI), which was funded by the Office of Naval Research and managed by the DD-21 Program Office, we have attempted to demonstrate that simulation and modeling techniques can play a significant role during the design of future combatants, especially in light of future Naval goals to optimize shipboard manning. The MAI used a warfighter-centered design approach to developing a prototype air defense warfare (ADW) system, and human-in-the-loop data was collected from the watchstations in use today and from the prototype watchstations. Under a simulation experiment, the Integrated Performance Modeling Environment (IPME), a discrete event simulator, was utilized to represent a demanding ADW scenario, and models were created to simulate performance for ADW teams using today's watchstation and then to predict the impact on performance that can be expected from the prototype. These complex models include multiple operators, dynamic operator task assignment configurations, workload tracking, internal and external communication network activity, and processes such as air track detection, track identification and re-identification, monitoring of changes in track profile, threat evaluation and engagement. This paper discusses the process of calibrating, verifying and validating models of the current and prototype watchstations, and present the conclusions made.

Biographical Sketches of Authors:

Christine H. Brockett is a Systems Engineer for Micro Analysis and Design, Inc. and holds an M.A. in Agricultural Economics from Cornell University. Mrs. Brockett is the lead analyst for MA&D on the SC21 Science & Technology Manning Affordability Initiative sponsored by the Office of Naval Research, which seeks to demonstrate that the goal of manning optimization is achievable within the air defense warfare (ADW) component of the CIC through a human-centered design approach and the use of human performance modeling techniques.

Shelly Scott-Nash has a B.A. in Mathematics with a minor in Computer Science from the University of Colorado at Boulder, 1987. At Micro Analysis & Design, Inc., Ms. Scott-Nash is the team project lead for the SC21 Science & Technology Manning Affordability Initiative sponsored by the Office of Naval Research. For this effort, the MA&D team has developed human performance models and has integrated one of MA&D's products, the Integrated Human Performance Modeling Environment, into a recently developed object-oriented data repository.

James A. Pharmer is a Research Psychologist at the Naval Air Systems Center Training Systems Division (NAWCTSD) in Orlando, Florida. Mr. Pharmer currently supports the efforts of NAWCTSD under the ONR SC-21 Manning Affordability Initiative through implementing research, which compares performance of reduced sized air defense warfare teams using advanced watchstation technology to intact teams using current equipment. In addition, Mr. Pharmer leads the effort to use the human in the loop data collected under Manning Affordability to validate human performance models developed to support the goals of optimal manning on future naval surface combatants. He holds an M.S. in Engineering Psychology from Florida Institute of Technology, and is currently a doctoral candidate at the University of Central Florida in Orlando, Florida.

VERIFYING AND VALIDATING THE AEGIS AIR DEFENSE WARFARE HUMAN PERFORMANCE MODEL

Christine H. Brockett
Shelly Scott-Nash
Micro Analysis and Design, Inc.
Boulder, Colorado

James A. Pharmer
Naval Air Systems Center
Training Systems Division
Orlando, Florida

INTRODUCTION

In some industries, simulation and modeling techniques are a widely accepted, integral part of system design, while in others these techniques may be perceived as expensive, unreliable, or inconclusive. Within the Manning Affordability Initiative (MAI), which was funded by the Office of Naval Research and managed by the DD-21 Program Office, we have attempted to demonstrate that simulation and modeling techniques can play a significant role during the design of future combatants, especially in light of future Naval goals to optimize shipboard manning. The MAI included three major thrusts. The first thrust demonstrated that applying a warfighter-centered design approach to developing a prototype air defense warfare (ADW) system would enable a 50 percent reduction in team size, with equivalent performance and decreased workload. The second thrust focused on defining the processes and providing the environment and tools necessary to support collaborative, warfighter-centered design. The third thrust sought to demonstrate how human performance modeling could be instrumental within the warfighter-centered design process (Scott-Nash, Carolan, Humenick, Lorenzen, & Pharmer, 2000).

Under the umbrella of this program, an experiment was devised to empirically evaluate the claim that modeling can aid the design process with the following steps:

1. Model performance and workload of ADW teams using current watchstation technology on a two-hour intermediate-level ADW scenario.
2. Using the same scenario, collect human in the loop data from intact teams aboard ship using this current technology
3. Validate the 'baseline' model with the experimentally collected data, and, if necessary, calibrate the baseline model.
4. Modify the model to reflect warfighter-centered design changes to the watchstations and manning reductions achieved under MAI comparison study.
5. Again using the same ADW scenario, collect performance data from these smaller teams using the warfighter-centered design watchstations.
6. Using the modified model, predict changes in human performance and workload trends with the new watchstation and reduced team size.
7. Validate these predictions against the collected data.

Not only did this simulation experiment provide the sought evidence, but also the availability of this data for comparison has provided a unique opportunity to evaluate how well the systems were modeled, and how well the verification and validation process worked. Moreover, it ultimately demonstrated the value that this modeling technique in general, and this model in particular, could add to the design process.

ADW SCENARIO AND DATA COLLECTION

The ADW scenario utilized for this effort was derived from an unclassified, intermediate to advanced level scenario developed by SMEs for fleet ADW training. The scenario consisted of a 1-hour 'low difficulty' segment followed by a 1-hour 'high difficulty' segment. This level of difficulty was manipulated by the number, ambiguity, and threat-level of tracks within the scenario.

Data collection for the baseline condition was conducted aboard ship, pier-side for 8 teams, each consisting of 8 team members. Participants performed the simulated ADW scenario utilizing the same equipment that they currently use operationally to perform their tasks.

Latency and accuracy data were collected on the performance of tasks within the detect-to-engage sequence for 25 selected air contacts of interest. This 'event-based' measurement strategy has been used successfully in previous military research (Johnston, Cannon-Bowers & Smith-Jentsch, 1995). ADW SMEs collected these data throughout the scenario using the Air Warfare Team Performance Index (Dwyer, 1992).

Workload was measured in three ways. First, team members rated themselves at the end of each difficulty segment using a modified version of the NASA-TLX (Hart & Staveland, 1988). Second, team leaders rated the effort of each team member at the end of each segment. Finally, SMEs rated workload for each team member at ten-minute intervals.

Data were later collected, using the same measures on four-man teams using the new watchstations in a similar manner using the same measures. Crewmen were selected for the new team based on similarities between old and new team positions. Additionally, the team received two hours of training on using the new system, prior to the data collection effort.

THE APPLIED HUMAN PERFORMANCE MODEL

The Integrated Performance Modelling Environment (IPME)

The Integrated Performance Modelling Environment (IPME) is a discrete event simulation tool, and was used to develop the ADW model for a typical 8-member "air alley" within an AEGIS combat information center. The IPME utilizes Micro Saint as the core simulation engine. Micro Saint is a general purpose, task network-based, simulation and modeling tool that has been widely used to support military simulation (Pew & Mavor, 1998) and is well suited for supporting task-centered design.

The IPME is a performance modeling and analysis tool that augments the task network modeling capabilities of Micro Saint in a number of ways. It includes a database of human performance micro-models derived from the research literature (e.g. Card, Moran & Newell's (1986) Model Human Processor); a multiple resource theory based (e.g., Wickens, 1992) workload analysis tool; and a capability to model the influence of performance shaping factors such as stressors, training, and aptitude.

As a task analysis and modeling tool, the IPME supports the hierarchical decomposition of missions into functions and tasks. Tasks are assigned to human resources, as well as to the physical interfaces of the workspace. Features include micro models of basic perceptual, motor, and cognitive operations to support accurate performance time estimates.

As a performance analysis tool, the IPME provides users with a method to assign workload estimates to tasks that are performed by team members and uses those workload estimates to dynamically model the impact on task and system performance. A primary feature distinguishing the IPME from other workload modeling tools is the implementation of workload management strategies. These strategies can be used to evaluate how humans will dynamically change their tasks in an attempt to manage workload overload.

The ADW Model

An ADW team task model was created using the IPME tool for the purpose of providing a level of fidelity to the domain sufficient to evaluate potential design changes to the AEGIS ADW capabilities, such as task automation and user interface improvements. The model development team worked closely with SMEs to decompose the major ADW activities into networks of tasks to a level of granularity that both met the purpose of the model and allowed the SMEs to reasonably estimate duration, standard deviations and workload for each task.

In addition to individual tasks, the model captured the relationships between tasks, the flow of information through the system and between crewmembers, and some decision-making by crewmembers, particularly as it pertained to the prioritization of tasks. The model was designed to reflect an expert team in the sense that errors due, for example, to inexperience or fatigue were not explicitly captured. Lastly, crew-to-task assignments were defined. In some cases, only one crewmember would ever perform a particular task. In others, as many as six different crewmembers might potentially perform a particular task, depending on timing and availability.

The Task Networks. The resultant model, shown in Figure 1, is composed of six networks representing major ADW activities, and four others that drive events internal to the model. The primary task network, shown in Figure 2, is the "Crew Tasks" network.

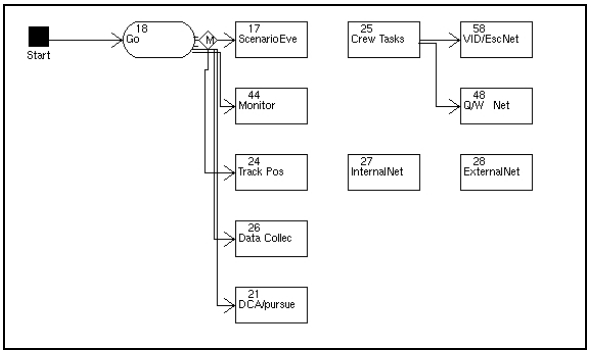

Figure 1. IPME ADW Model Top Level Network

This network includes crew activities such air track identification and re-identification, ESM monitoring and correlation, threat assessment and engagement The "Crew Tasks" network manages this range of activity with sub-networks, represented by rectangles. For example, the "ID Net" sub-network contains all of the tasks involved in identifying and re-identifying air tracks, and even uses lower-level sub-networks to group tasks such as those performed to confirm the identification of a track. Some tasks trigger events and actions in other networks; this is represented on the task diagram with rectangles.

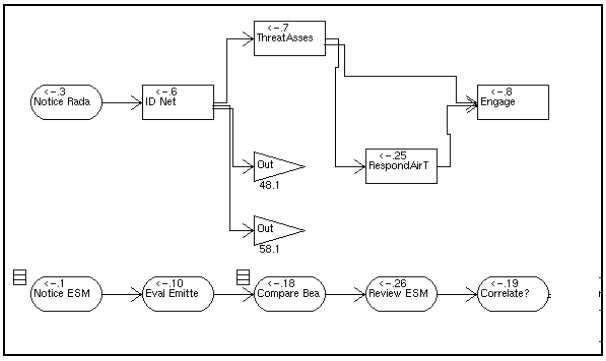

Figure 2. "Crew Tasks" Network

The "ID Process" is shown in Figure 3, and it represents one of the more critical and complex aspects of ADW team performance. In this process, an operator (a simulated crewmember) must seek out and use all available pieces of air track information to derive a correct identification. Correct identification at this point does not necessarily mean ground truth identification, but rather that the crewmember has assigned an ID classification according to rules relating the various sources of identification information and the particular combination of indicators. The "ID Net" also includes a complex track prioritization scheme that allows the model to simulate the choices the ADW team makes about which track to attend to next.

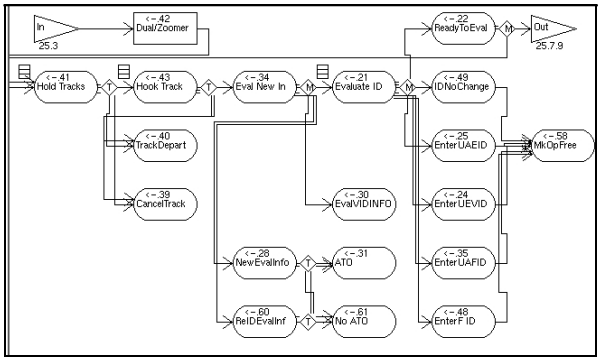

Figure 3. "ID Process" Network

Internal and external communications are modeled extensively in this model in the major task networks, "Internal Communications" and "External Communications", and include over 35 different categories of communications. The model restricts operators from talking on one net if already talking on another. However, it does not restrict operators from listening to internal and external incoming messages at the same time. The time needed to broadcast a communication is computed by inputting the number of words contained in that particular message to an embedded speech micromodel.

The other three networks representing ADW activities are the "Monitor" network, the "VID/Escort" network and the "Queries and Warnings" network. The "Monitor" network simulates tasks that crewmembers perform to maintain situational awareness when they are "not busy" with their main tasks. The "VID/Escort" network involves employing defensive aircraft to visually identify, intercept and escort potential threats. The "Queries and Warnings" network simulates communicating with potentially threatening aircraft through a protocol of issuing Level 1 queries to those within a certain radius of the ship, and Level 2 warnings to those within yet a smaller radius.

The Scenario Model The same ADW scenario used to collect human-in-the-loop data from the crews was translated into a scenario model, represented by the "Scenario Event" network within the ADW model. This network plays the critical role of driving the team task networks described above, simulating the vast array of information that appears on the crewmembers' watchstations, as well as other information inputs such as communications from sources external to the crew. This particular scenario employs over 80 air tracks, and includes 1160 scenario events that take place over a simulated two-hour time frame.

The most complex code contained in the scenario model updates the geographical locations of all the air tracks relative to ownship, as well as in latitude/longitude coordinates. This model keeps track of all the positions of all of the air tracks tracks, using a combination of initial bearing and range from ownship, and then updates track positions over time as a result of course and speed changes. The model also calculates the equations to represent the location of coastlines, commercial air routes, and other pertinent geographical features. This position information is crucial to the model because it impacts nearly every aspect of the operators' tasks.

Model Outputs. The inputs into the ADW model, the individual tasks, task durations, standard deviations and workload, were all estimated and/or defined in isolation. Through the use of the team task model, the scenario model that simulates flow of information and triggers activity, and operator-to-task assignment algorithms, IPME can use those inputs to answer more complex questions, such as: how long does it take to perform an activity comprised of multiple tasks and decisions? How often are crewmembers able to perform an operation or specific task within the scenario? And lastly, as a crewmember juggles and prioritizes competing tasks, what is the effect on his or her workload?

"Baseline" versus "modified" ADW models The ADW model was developed to represent today's AEGIS watchstations and today's ADW crew, and this model is referred to in this paper as the "baseline" model. The baseline model was constructed in such a way to allow for the relatively easy comparison of multiple design options that incrementally modify the existing AEGIS watchstation, as opposed to the design of a completely new system. In a similar manner, the baseline model will accept modified inputs, such as new operator-to-task assignments for the purpose of comparing different team configurations.

For this experiment, the baseline model was changed to capture the warfighter-centered design changes to the AEGIS watchstation implemented under (MAI). These changes included the automation of certain tasks, and user interface design changes. A flag employed within the ADW model allows the model user to run the model for either version. In this paper, this second version is referred to as the "modified" model.

VERIFICATION & VALIDATION OF THE MODELS

The intent of the V&V effort for the ADW models was to demonstrate that the models had a level of fidelity sufficient to predict the effects on performance and workload of potential design changes, such as task automation and user interface improvements, to the AEGIS watchstations.

V&V Criteria

Five key measurements representing major ADW activities were selected as the criteria for validating the models. These included:

- Average Time to First ID
- Average Time to New Track Report
- Number of Tracks Queried
- Number of Tracks Warned
- Number of Tracks Identified

The goals of this project were to calibrate the baseline model to within a 10% difference from the human-in-the-loop data for these key measurements and to further predict performance with the modified model within a 10% difference.

"Time to First ID" is measured from the time that a specific air track appears on the radar, until the time that a crewmember assigns an ID to it. As described in previous sections, track identification is both a critical and complex process as it is performed with today's watchstation. As such, this was an important criterion for validating the baseline model. For this same reason, track identification was a key area targeted by the MAI for human-centered design changes that reduced manpower requirements while maintaining or improving on current performance levels. Consequently, track identification is automated by the new watchstation. Although crewmembers still must mentally verify track IDs, this process is negligible with the new watchstation, and so this measurement is not used for validating the modified model.

"Time to New Track Report" is measured from the time that a specific air track appears on the radar, until the time that a crewmember begins to issue a verbal report announcing the track. This measurement does not include the time required to actually broadcast the report, so that it is not influenced by variations in wording. "Number of Tracks Queried" measures the number of air tracks that received at least one Level 1 query during the scenario, and "Number of Tracks Warned" similarly captures the number receiving at least one Level 2 warning. Lastly, the "Number of Tracks Identified" is simply a measure of the number of tracks identified by crewmembers. Like the first measurement, "Time to First ID", this measurement was not used for validating the modified model because this task was automated.

In addition to the key measurements, the project compared IPME calculated trends in workload to the collected workload data. Because of its complex nature, workload is often a difficult concept to quantify (Tsang & Wilson, 1997). Consequently, a multi-dimensional approach was taken using three different measurements collected from the crews to describe the workload experienced during the ADW scenario. Trends, rather than actual numbers, were compared between these three measures and the IPME predicted workload because of the slight differences in what was being measured.

The first collected workload measurement is a modified version of the NASA Task Load Index (TLX), where participants were asked to rate their workload across 10 dimensions, including 6 taskwork dimensions (e.g., Mental Demand, Physical Demand) and an additional 4 teamwork dimensions (e.g., Communications Demand, Coordination Demand). Participants rated themselves at the end of each half of the scenario. At these times, the team leaders were also asked to rate the percentage of capacity of effort exerted by each member of the team. Lastly, SME evaluators assigned to observe each crew member during the scenario, rated the participants' workload on a scale of 1 to 7 every 10 minutes during the active portions of the scenario.

Verifying the Baseline Model

In addition to using standard software testing techniques, two methods were employed to verify the baseline model.

Firstly, results were compared to an "expert solution'' for the ADW scenario, provided by its developer. Because the ADW scenario drives all of the simulated crew actions in the model, it was important to verify that the models were correctly interpreting and then processing information from the scenario. The expert solution identified correct actions crew members should take for specific air tracks, the correct conclusions that crew members should reach regarding track identification, and the earliest times crew members could possibly perform specific detect to engage tasks, given the times that events occurred in the scenario. For example, the expert solution might state that, for track X at time 00:31:48 in the scenario, information becomes available that allows the crew to identify track X as "Unknown Assumed Friend". If the model predicts an earlier identification time for track X than the answer key, then the development team knew to examine this portion of model.

The second method employed for model verification was to review the model and its results carefully with the SME. This step gave the SME another opportunity to ensure that the development team correctly interpreted the requirements. Additionally, the SME provided input as to whether the results appeared 'reasonable' or not.

Validating and Calibrating the Baseline Model

After completing the verification phase for the baseline model, the development team was given the human-in-the-loop data for validating the model's predictions. If the model did not come within 10% of the collected data for the key measurements, the team began to calibrate the model.

After looking at the human-in-loop data, median time was selected as a calibration point rather than average time for the first two key measurements for two reasons. First was to compensate for the fact that the model was intended to represent the behavior of an expert team, whereas data were collected from representative teams that were not necessarily expert. Second, outlying data points resulting from the greater variation in less-than-expert performance and small sample sizes too easily influence average times causing the results to be negatively skewed. During the calibration process, it was further determined that the initial 10% goal was not appropriate for all of the key measurements. Final calibration results for the baseline model were as follows in Table 1.

Measurement	**% difference from data**
Median Time to First ID	-2%
Median Time to New Track Report	-3%
Number of Tracks Queried	+44%
Number of Tracks Warned	+30%
Number of Tracks Identified	+26%

Table 1. Baseline model, calibration results for key measurements

To achieve the results in Table 1, the development team only made two changes to the logic of the model. Firstly, the external communication sub-network was modified to provide prioritization between the queued Level 1 Queries and Level 2 Warnings that was not identified in the original requirements. Secondly, the duration and standard deviation of two tasks involved in the identification of air tracks were changed as a result of examining the human-in-the-loop data and reviewing with the SME. This second change affected both the "Time to First ID" measurement and the "Time to New Track Report" measurement.

Next, the team found much agreement between the workload predicted by the baseline model and the workload measurements collected for the intact crews. The baseline model predicted that the summed workload of the entire crew would increase by 16% from the 'low difficulty' half of the scenario to the 'high difficulty' half. The collected workload measurements also demonstrated an increase in workload between scenario halves, as shown in Table 2,

which was of a similar magnitude for two out of the three measurements.

Measurement	% increase between scenario halves per team
IPME baseline model	16%
Task Load Index	11%
Team Leader estimates	12%
SME estimates	27%

Table 2. Baseline model, percent increase in summed team workload between scenario halves

Of the three human-in-the-loop workload measurements, only the SME estimates were taken more frequently than once per scenario half. Figure 4 below shows a comparison in summed and averaged team workload across eight 10-minute intervals. The general rise and fall of the lines suggest a similar trend, although it should be noted that these measurements were scaled for the purpose of depicting them in the same graph.

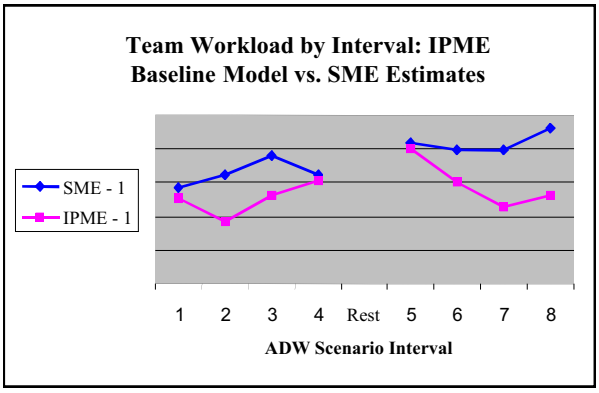

Figure 4. Team Workload by Interval: IPME baseline model versus SME estimates

There was also agreement between the model predictions and the collected data that the extent to which workload increased between scenario halves varied by crew member, where for at least one crew member there was no significant workload increase between scenario halves. Beyond a few obvious trends, however, few of the results agreed to the extent to which the workload increased between scenario halves for a given crewmember. This may indicate that the measurements were not describing the exact same workload dimensions.

At what point should the development team stop calibrating and call the baseline model done? The first two key measurements, "Median Time to First ID" and "Median Time to New Track Report", fit well within the original calibration goal of no more than a 10% difference between predicted and collected values. The baseline model team workload predictions and human-in-the-loop team workload data demonstrated similar trends. However, for the other three key measurements, "Number of Tracks Queried", "Warned" and "Identified", the baseline model predicted numbers more than 10% higher than found in the experimentally collected data. Talking this over with the model's SME, it was concluded that the discrepancies were easily explained by two factors.

One factor is that the model was intended to model an expert team, and data were collected from representative teams. So, for example, for an activity like identifying tracks where a goal of the team is to identify as many tracks as possible, it appears reasonable that the expert model will identify more tracks. A second factor is that for the purpose of this project, the model was designed to reflect crew behavior only to the level that the model would serve as a good tool for evaluating design options. Whereas the model did include a number of complex prioritization schemes to capture how crew members decide what task he or she might perform next, it did not attempt to enter into the arena of tactical decision making. The development team realized that the activities of querying and warning tracks involved this kind of decision-making.

Rather than change the nature of the model to bring the last three key measurements within the 10% goal, the development team decided that the validation and calibration of the baseline model was finished with the following conclusions. The baseline model is a valid representation of an expert team as it performed the major activities of an ADW crew, but the results for the model must be interpreted with the understanding that the model was intended to reflect little tactical decision-making. The baseline model also provides a valid estimate of workload incurred by the ADW crew.

Modified Model Validation Results

The baseline model was modified to reflect the human-centered design changes to the watchstations, as well as the reduction of the crew from a 9-man to a 5-man team. These changes included the automation of certain tasks, as well as modifications that captured improvements in the user interface. Two of the key measurements for validation of the baseline model, "Time to First ID" and "Number of Tracks Identified",

were based on the track identification task that was automated with the new watchstations.

The results for the next key measurement, "Average Time to New Track Report", are as follows:

- The IPME modified model predicted that the Average Time to New Track Report would decrease by 40%.
- The experimentally collected data showed that Average Time to New Track Report decreased by 30%.

As discussed in the previous section, the data collected on this measurement were heavily influenced by the small sample size, and by the variation in levels of crew expertise. As a result, the average times predicted by the baseline and modified models were consistently lower than the collected average times: 55% and 62% respectively.

For the next two key measurements, Number of Tracks Queried and Number of Tracks Warned, the differences between the modified model results and the human-in-the-loop data are as follows:

- Number of Tracks Queried: +16%
- Number of Tracks Warned: +17%

Previously, the baseline model had predicted results that were much higher than the collected data, 44% and 30%, respectively, and it was decided that this reflected both the differences between the expert model and the representative crews, and the lack of tactical decision-making in the model. Consequently, it is interesting to see that these results are much closer to the 10% validation goal. This improved result can be explained by the fact that the new watchstations included user interface improvements that made it easier for crew members to become aware of when tracks should be queried or warned, enabling the crews to operate more like experts.

When the team workload measurements predicted by the baseline and modified measurements across the scenario are summed and averaged by crewmember, the modified model predicts that overall team average workload is 15% lower with the new watchstations, despite the reduction in crew size from nine to five.

The collected workload measurements show similar trends in Table 3. Although there is not necessarily agreement between the three collected measurements on the degree to which workload decreased with the new watchstation, the trends are all in the same direction.

Measurement	Difference in team average workload between watchstations
IPME models	-15%
Task Load Index	-12%
Team Leader estimates	-2%
SME estimates	-32%

Table 3. Change in team average workload from today's watchstation to the new watchstations

Similar to the first watchstation results, the modified model and all of the collected workload data measurements demonstrated a difference in workload between the first and second halves of the scenario, as shown below in Table 4. However, all of the measurements agree that not only is there a difference in workload between the 'low difficulty' and 'high difficulty' halves of the scenario, but that the difference is much larger, by as much as a factor of 2.5 times.

Measurement	Baseline model / today's watchstation	Modified model / new watchstation	Factor
IPME models	16%	40%	X2.4
Task Load Index	11%	23%	X2.4
Team Leader estimates	12%	28%	X2.5
SME estimates	27%	44%	X1.6

Table 4. Changes in team average workload measurements between scenario halves for each watchstation

Given the dramatic reduction in crew size, it would be possible that even though performance times improved with the new watchstation, that improvement might come at the cost of increased stress for the crew, particularly during times of heavy air activity as experienced during the second half of the scenario.

Table 5 below shows the percent change in workload for each scenario half between the baseline model and modified model, and between each workload measurement collected from the current watchstation and the new watchstation. The numbers conclusively show that workload experienced during the first half of the scenario was lower with the new watchstation. Although there is less agreement between the collected workload measurements, it appears that workload may be lower during the second half of the scenario with the new watchstation as well.

Measurement	% Change in workload	
	1st scenario half	2nd scenario half
IPME models	-23.2%	-7.9%
Task Load Index	-17.3%	-7.3%
Team Leader estimates	-9.7%	4.2%
SME estimates	-36.1%	-27.8%

Table 5. Changes in team average workload measurements between watchstations for each scenario half

Lastly, the graph in Figure 5 compares workload trends between the IPME model predictions and the SME workload estimates, which were collected for each 10-minute interval. The first two lines in Figure 5 depict the same data shown in Figure 4 above, comparing workload trends between the baseline model and the SME estimates collected with the current watchstation ("IPME-1" and "SME-1" in the legend). The next two lines depict the workload trends for the modified mode. Again, the general rise and fall of the lines suggest a similar trend, but note that each measurement is scaled for the purposes of depicting them in the same graph.

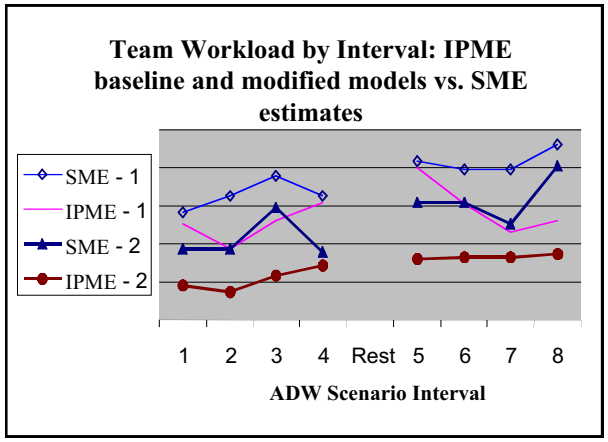

Figure 5. Team Workload by Interval: IPME baseline and modified models versus SME estimates

The similarity in workload trends and the relative success in validating the modified model against key measurements support the conclusion that the IPME ADW model is a useful model for evaluating the impact of incremental design changes to the AEGIS watchstations and changes in crew configuration.

CONCLUSIONS

Capturing the information that describes the activities of and interactions within the ADW component of a CIC and creating a model accurate enough for design evaluation was a non-trivial task. An important conclusion of this effort was that it is possible to model such large and complex processes. One advantage of creating such a model is that now a valid ADW model exists that can be used again and again for various purposes. For example, further analysis on possible design or crew changes can be explored at very little additional cost. Another modeling tool developed under MAI finds optimal crew-to-task assignments, taking into consideration the number of crewmembers, the amount or type of training necessary, best performance and distribution of workload. The resulting crew-to-task assignments can be imported into the IPME ADW model and their impact on performance evaluated. In addition to the MAI, other naval projects may have a use for an ADW model; inquiries into this possibility are already being explored.

A number of conclusions and lessons learned can be drawn from the results of this modeling experiment. During the calibration of the baseline model, very few changes were made to bring the model outputs to an acceptable state. This validates the process used to solicit model requirements from the SMEs, and the estimated task durations and standard deviations that were produced. This process included an initial two-day face-to-face requirements session, followed up by three more face-to-face meetings, as well as much communication via phone and e-mail, with SMEs to review model development and results to ensure that items communicated were correctly interpreted in both directions. It is interesting to note that the need to change the two task durations and standard deviations during the calibration phase resulted from a misunderstanding between a SME and the development team.

Modeling theory clearly maintains that the more effort and time spent at the front end of the model development cycle, the better, and the results of this experiment bear that out with the minimal changes required for calibration. However, a more important conclusion can be drawn from this result, which is that is possible to build a valid model for a complex process such as the CIC ADW component with only estimated data from SMEs. Using estimated data can provide a large financial advantage over collecting human-in-the-loop data either in the field or experimentally, not only because of the initial costs of each effort, but also because the model can be modified and reused multiple times in this case.

In addition to the conclusions drawn from the experiment, a number of lessons were learned as well. Regarding the data collected for this project, one lesson learned related to the measurement of workload. For the purposes of human-in-the-loop testing, it was necessary to use a metric that was unobtrusive to the participants. Consequently, the data collection team chose to use questionnaire-based measurement tools that could be administered during the inactive periods during the scenario. However, the nature of the modeling approach described here would have been better served with multiple measures taken across the scenario. A compromise was reached between the two conflicting needs by utilizing the SME ratings, which were collected at eight intervals during the active portions of the scenario.

An additional lesson learned involved the setting of validation goals for the baseline and modified models. As stated before, the purpose of the model was to provide a tool useful for evaluating design changes to a relatively complex system, and thus the nature of the model was not "critical", requiring independent V&V, or even stringent V&V. However, finding validation goals appropriate to the intended use of the model turned out to be just as complicated a process as for a model of a more critical nature. Problems with the validation goals encountered in this project have been discussed above, and these were resolved through the evaluation of the results with a SME, the reinterpretation of some validation goals, and lastly with the acceptance of the models by the client.

REFERENCES

Card, S., Moran, T. & Newell, A. (1986) The model human processor. In K. Boff, L. Kaufman, & J. Thomas (Eds.). Handbook of perception and human performance (vol. 2). New York: Wiley

Dwyer, D. J. (1992) An index for measuring naval team performance. Proceedings of the Human Factors Society 36th Annual Meeting, 2, Santa Monica, CA; Human Factors and Ergonomics Society, 1356-1360.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P. A. Hancock & N. Meshkati (eds.), Human Mental Workload. Amsterdam: North Holland, pp.139-183.

Johnston, J. H., Cannon-Bowers, J. A., Smith-Jentsch, K. A. (1995). Event-based performance measurement system for shipboard command teams. Proceedings of the First International Symposium on Command and Control Research and Technology (pp. 274-276). Washington, DC: The Center for Advanced Command and Technology.

Pew, R & Mavor, A. (Eds.) (1998). Modeling Human and Organizational Behavior: Applications to Military Simulations. Washington: National Academy Press

Scott-Nash, S., Carolan, T., Humenick, C., Lorenzen, C., and Pharmer, J. (2000) The application of a validated human performance model to predict future military system capabilities. Paper presented at the 2000 Interservice Industry Training Simulation and Education Conference (I/TSEC).

Tsang, P. & Wilson, G. F. Mental workload. In G. Salvendy (Ed.), The Handbook of Human Factors, (pp.417-451). New York: John Wiley & Sons.

Wickens, C.D. (1992). Engineering Psychology and Human Performance (2nd ed.). New York: Harper Collins.