CALIBRATING AND VALIDATING A HUMAN PERFORMANCE MODEL TO SUPPORT PREDICTIONS OF FUTURE MILITARY SYSTEM CAPABILITY

Shelly Scott-Nash
Tom Carolan
Christine Humenick
Christy Lorenzen
Micro Analysis and Design, Inc.
Boulder, Colorado

James Pharmer
Naval Air Warfare Center
Training Systems Division
Orlando, Florida

ABSTRACT

Throughout the entire design process of any future military system, from function analysis through system operation testing, a key human factors question will be, "Will warfighters meet required performance criteria on demanding operational scenarios?" Although experimentation with real warfighters is essential, it may be too expensive to for all possible equipment and team design options. The Integrated Performance Modeling Environment (IPME) focuses on simulation of humans in complex environments and allows us to evaluate system concepts, designs, and team structures with simulation at a far lower cost than with real humans. For the Office of Naval Research Science & Technology Manning Affordability Initiative, the IPME was used to model human processes and human interactions with current consoles, as well as internal and external communications networks. The model is based on a demanding air defense warfare scenario containing in excess of 80 air tracks, 1160 scenario events and 150 human tasks. Modeled processes include air track detection and identification, escort, queries, warnings, and threat evaluation and mitigation. Some measures that can be provided by the model include the time to first identification, wait time associated with various tasks, and crewmember workload parameters. In parallel with model development, experimental data were collected aboard ship from eight intact crews using the same demanding operational scenario. This paper describes a multiple-step process in which the model is validated and calibrated, and discusses progress to date in this area. Then, the paper discusses how the project will modify the timing and function allocation rules in the model to allow experimentation with alternate team designs, automation and alternate human machine interfaces. This cycle of model validation and model-based design evaluation provides a powerful way to integrate human factors engineering into the design of future systems.

Biographical Sketches of Principal Authors:

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 Pharmer is a Research Psychologist in the Science and Technology Division of the Naval Air Warfare Center Training Systems Division (NAWCTSD), Orlando, Florida. Mr. Pharmer holds an M.S. in Engineering Psychology from Florida Institute of Technology, Melbourne, Florida. He is currently a doctoral candidate in Human Factors Psychology at the University of Central Florida. His research interests include team performance, communications and usability.

CALIBRATING AND VALIDATING A HUMAN PERFORMANCE MODEL TO SUPPORT PREDICTIONS OF FUTURE MILITARY SYSTEM CAPABILITY

Shelly Scott-Nash
Tom Carolan
Christine Humenick
Christy Lorenzen
Micro Analysis and Design, Inc.
Boulder, Colorado

James Pharmer
Naval Air Warfare Center
Training Systems Division
Orlando, Florida

INTRODUCTION

Current technology within the Combat Information Center (CIC) has allowed warfighters to maintain high performance on even the most demanding operational scenarios. However, the goals of manning optimization for future surface combatants has provided a unique challenge to the designers of these systems to maintain this high performance with fewer warfighters. The Office of Naval Research Science and Technology Manning Affordability Initiative has taken on the task of demonstrating that this goal may be 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. These techniques are currently being utilized in the design of the Multimodal Watchstation (MMWS) that will demonstrate how advanced display and control technology in conjunction with task and workload management can support high performance with significantly reduced team sizes.

The purpose of this paper is three-fold. First it will describe the role of human performance modeling and its application to the development of the MMWS. Specifically, this paper will focus on the use of the Integrated Performance Modeling Environment to model the tasks performed by the ADW component of a CIC watchstanding team. Second, this paper will describe how these models are calibrated and validated against experimental data collected from intact watchstanding teams aboard ship using current state of the art equipment to perform a complex ADW scenario. Finally, this paper will discuss how these models are being applied to the development of the MMWS to evaluate the impact of design alternatives on team and individual performance and workload.

HUMAN PERFORMANCE MODELING

Due in part to the increase in computational speed and power over the past decade, human performance modeling tools to support the system design process are becoming more widely accessible. Human performance models offer a means of evaluating complex system designs, measuring the impact of design changes on performance, and assessing design effectiveness by simulating operators using the system. Additionally, human performance models can be applied to answer questions about expected changes in task performance times as a function of system design, task allocation, individual capabilities, and performance shaping factors such as stress and workload.

Laughery and Corker (1997) have classified human performance models into two general categories: reductionist models and first principle models. Reductionist models use the task sequence as the primary organizing structure. The modeling process involves task analytic decomposition of human behavior, from larger units to successively smaller elements of behavior, until a level of reduction is reached that can provide reasonable estimates of human performance for these task elements. Task network modeling is an example of a general-purpose reductionist approach based on task analysis methods. In a task network model, the results of a task analysis process are used to construct a task network that defines the hierarchical and sequential relationships among tasks.

First principle models of human behavior involve structures that represent basic principles and processes of human performance. Tools that support first principle modeling provide algorithms that embody some theory-based aspect of human cognition or behavior. First principle and reductionist modeling approaches are not mutually exclusive. Integrating first principle or theory-based models with reductionist approaches, such as task network modeling, provides a powerful tool for modeling the impact of system design and environmental factors on human performance.

Task network models provide a method to generate human performance predictions for complex task sequences under different human-computer interface conditions. Task network modeling, in use since the 1970s, extends the power of task analytic methods by

developing executable computer models that can be predictive of human performance. The use of task analytic methods, combined with human information processing primitives to predict human performance, is well established for computer interface evaluation (e.g., Card, Moran and Newell, 1983; Gray, John and Atwood, 1993; Kieras, 1998).

Human Performance Modeling and Human Experimentation

While computer modeling of human behavior provides another tool to enhance the system design process, it does not replace the need for other analysis methods such as rapid prototyping and human experimentation. The variability involved in human performance makes humans less predictable. There is, therefore, a need to conduct empirical testing with more than just computer models in order to validate and extend the models.

As a complement to human experimentation for achieving design and usability goals, models of human-system interaction can be used to predict human performance under a variety of conditions. Human performance models can be used to leverage the information gathered from experimental studies by extending the usability evaluation to additional features, tasks and conditions. A combination of human performance models and experimental methods provide a practical approach to usability evaluation and prediction of task performance time and accuracy. Experimental studies provide the calibration data to validate the structural and parametric components of the human performance models and increase their predictive power. Once validated, human performance models can be used to help focus the human experimental data collection on those areas where model predictions are not consistent with expectations or observations. This relationship between experimentation with human subjects and experimentation with human performance models is illustrated in Figure 1, adapted from Laughery and Corker (1997).

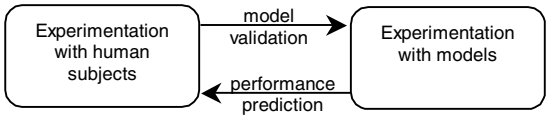

Figure 1. The relationship between experimentation and human performance modeling

The Integrated Performance Modeling Environment (IPME)

The Integrated Performance Modelling Environment (IPME) was used to develop a full scenario model with multiple team members for this research. 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 Model Human Processor); a multiple resource theory (e.g., Wickens, 1992) based 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 crewmembers and use 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.

APPLICATION TO TEAM PERFORMANCE IN THE SHIPBOARD COMBAT INFORMATION CENTER

A key objective of the Manning Affordability Initiative research is to demonstrate the potential reduction in manning requirements that can be achieved, without loss of performance effectiveness, by improving system usability through advanced human-computer interface and workload management approaches. These approaches are currently being developed and evaluated in the design of the MMWS. The MMWS concept has many advanced features such as a multi-modal interface, graphical information displays, advanced attention management features, multiple screens,

synthesized speech, collaboration support, and task management tools, to name a few.

These advanced interface features are designed to enhance and optimize team performance and reduce onboard manpower requirements. As part of this advanced MMWS project, studies are being conducted to gather baseline human performance data from CIC teams using today's workstation. The data being collected in these investigations will be used to evaluate the usability of MMWS advanced interface features and determine the effects of those features on operator performance.

To augment these human studies, the IPME was used to develop a full scenario model of the ADW component of a CIC team including interactions with consoles and internal and external communication networks. The full scenario model is integrated within a Human Centered Design Environment (HCDE). The HCDE contains performance parameters from other modeling tools, a common scenario definition, the advanced workstation functional requirements, as well as output from the full scenario model. This includes the task network model and its output data that has been made available for use by other tools in the HCDE.

The primary IPME modeling objectives are to simulate performance on a target scenario for CIC teams using today's workstation and then to predict the impact on performance that can be expected from MMWS advanced interface and workload management functions. Figure 2 provides a simplified conceptual overview of the role of the IPME model in the HCDE and in the MMWS design and evaluation process.

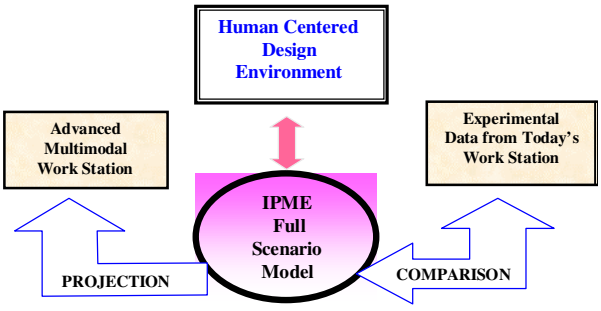

Figure 2. The of role of the IPME model

Overview of the IPME CIC Air Defense Warfare Model

The IPME CIC ADW team model (hereafter referred to as the IPME ADW model) is a highly complex model that includes a scenario that contains in excess of 80 air tracks, 1160 scenario events and 150 human tasks. Modeled processes include air track detection, track identification and re-identification, monitoring of changes in track profile, threat evaluation and engagement. The complexity of the model also involves multiple operators and operator task assignment configurations, workload tracking, internal and external communication network activity, and a complex track and task prioritization scheme. Figure 3 provides an illustration of the model's top-level network and the IPME network development interface.

The IPME ADW model contains many "flags," controlled by scenario events that determine which set of inputs is being processed. These flags allow different HCDE inputs to be evaluated using the same task structure and decision logic. For example, the HCDE could provide timing and operator-to-task allocations that differ from the original input data. By switching one simple "flag," the model could produce a different set of output data based on the inputs from the HCDE. This capability provides one means to modify task performance parameters based on MMWS features or based on input from other modeling tools.

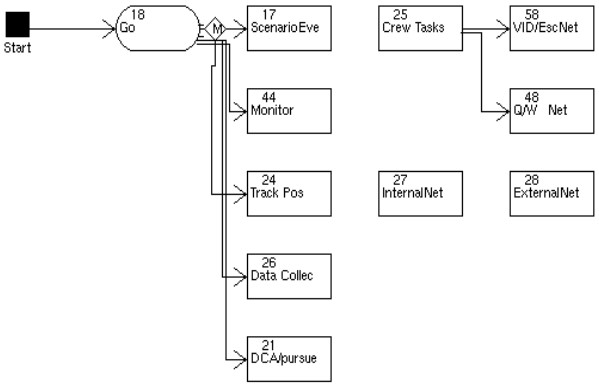

Figure 3. IPME ADW Model Top Level Network

ADW Scenario Model. A challenging ADW scenario was developed for this research by a team of subject matter and tactical training experts. The scenario is designed to have a 45-minute low workload period, followed by a 15-minute coast period, and a subsequent 45-minute high workload period. The operators are tested by several "critical tracks" which require special attention and actions. These critical tracks involve events such as "pop-up" tracks that disappear from radar coverage and later reappear; friendly tracks that cross into enemy airspace; enemy aircraft that masquerade as commercial aircraft; an attack and subsequent engagement; and many other challenges for the operators.

The most complex code contained in this 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 tracks, using a combination of initial bearing and range from ownship. It then updates track positions over time, accounting for course and speed changes, in addition to converting to and from nautical miles and latitude and longitude. The model also calculates the equations to represent the location of coastlines, commercial air routes and other pertinent geographic features. This position information is crucial to the model because it impacts nearly every aspect of the operators' tasks. However, it also presents the challenge of finding the right balance between continuous simulation and discrete event simulation.

ADW Team Task Model. The task model for the Air Defense Warfare component of the Combat Information Center Team is much too complex to provide a complete overview in a page or two. Thus, we will focus on the overall structure and those tasks that are pertinent to our initial validation process described in the next section. An important feature of the ADW team task model is that it is not designed to be scenario specific. Instead, it is designed to process ADW scenarios with common sets of task types and events.

The ADW team task model consists of six of the networks displayed in Figure 3. The "Monitor" network, the "Crew Tasks" network, the "VID/Escort" network, the "Queries and Warnings" network, and the "Internal and External Communications" networks. The VID/escort network involves employing defensive aircraft to visually identify, intercept and escort potential threats. The queries and warnings network involves communicating with potential threats. While these are important task networks that provide performance measures for validation, they are not elaborated in this paper.

The "monitor" network simulates tasks that operators do to maintain situational awareness when they are "not busy" doing their main tasks. These monitor loops are designed to cycle at specific intervals, with specific operators. When it is time for a monitor task to be performed, the operator's workload is checked to determine if they are currently performing another task. If they are, when the task that they are performing is completed, their workload decrements and the monitor task begins.

Internal and external communications are modeled extensively in this model. The model also has a complex communication structure that simulates more than 30 different communications. Moreover, the model does not restrict operators from listening to internal and external incoming messages at the same time. However, it does restrict operators from talking on one net if they are talking on another net. 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 primary task network, the "Crew Tasks" network, is displayed in Figure 4. The "Crew Tasks" network consists of sub-networks (represented by rectangles) that contain tasks, as well as tasks that trigger other networks (represented by triangles). There are actually two distinct series of tasks represented at this level. The top series of tasks represents the operator tasks that happen in response to new track detection. These tasks are organized into the "Track Identification (ID)" network, the "Threat Assessment" network, the "Respond to Air Threats" network, and the "Engage" network. The bottom series of tasks, which extend beyond this graphic, represents operator tasks in response to detection of electronic signals.

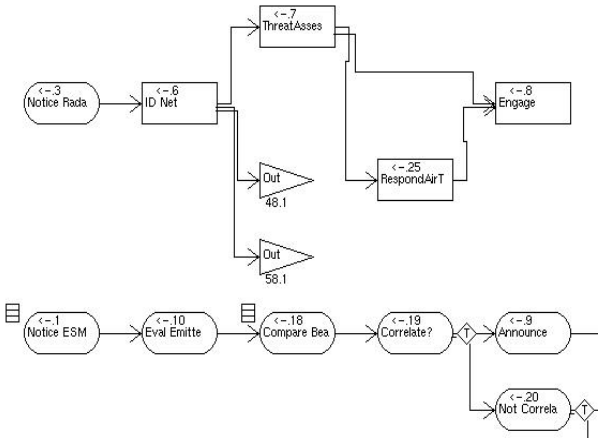

Figure 4. Crew Tasks Network

The "ID Process" network is partially displayed below in Figure 5. The ID Process represents all of the tasks that an operator must complete in order to assign a track into an identification category. An operator must seek out and use all available pieces of available track data in order to arrive at the correct identification. Correct identification at this stage does not necessarily mean ground truth identification. It implies that, based on the knowledge that the operator has obtained, the identification assigned is correct according to the ID matrix. The ID matrix is a set of rules that relate the various sources of identification information and the combination of indicators required for each ID classification. Embedded in the IPME model "ID Process" Network is a complex track prioritization scheme that allows the model to simulate the choices

the ADW team makes about which track to attend to next. The ID process is one of the more critical and complex aspects of ADW team performance. It is, therefore, one of the tasks that advanced workstation technology would be expected to impact by decreasing the time and manpower requirements while maintaining or improving on current performance levels.

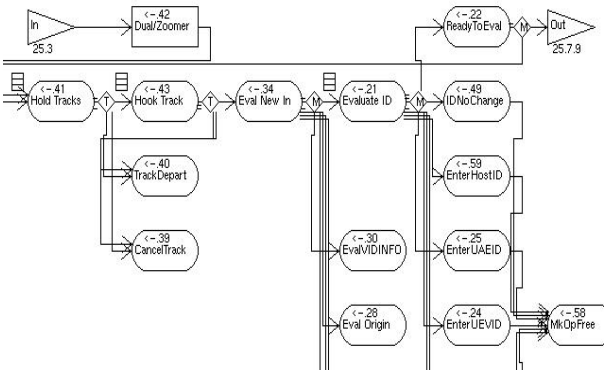

Figure 5." ID Process" Network

Operator Allocation and Workload. The IPME ADW model simulates multiple operators and operator-to-task assignment configurations. A complex method of assigning operators is used. Operators are assigned a probability in the release condition for a given task based on the operator's role in the scenario and whether the task is a primary or secondary responsibility for the operator. An operator is assigned to a task based on availability and a random number generator. Availability is determined by workload estimates defined for each task based on a multiple resource theory framework. For the ADW model, values were assigned to specific tasks for visual, auditory, speech, cognitive, and psychomotor workloads on a relative scale. The model keeps track of workload for each operator and uses these workload values as inputs to the operator-to-task assignment algorithm. Workload estimators are used in the IPME model to monitor the effort required by an operator to perform a task or set of tasks. Additionally, these values for workload can be used to shape the outcome of a task network as operators become too loaded to perform certain tasks.

Figure 6 shows an example of the current workload output for a single operator. In this case, the operator is the Air Intercept Controller (AIC). The x-axis of the chart shows time elapsed though the scenario in seconds. The y-axis of the chart shows a scale representing the sum of all workload estimators at any given point in time. The value for total workload for each operator is gathered every 10 seconds, and represented here as a point on the chart. This chart demonstrates how, given the tasks that the AIC is managing, relative workload increases and decreases over the course of the scenario. Note how the relative workload predicted by the model peaks during the high workload period of the scenario.

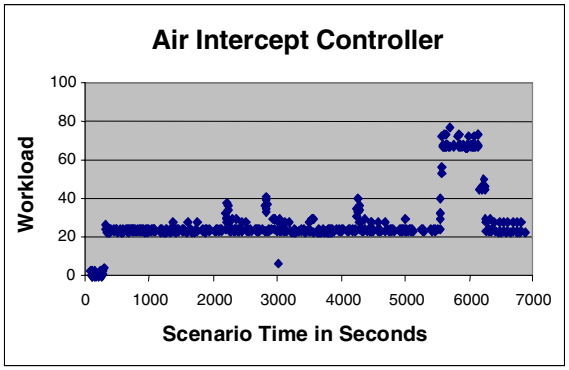

Figure 6. Example of workload output

Model Calibration and Validation

Before a model can be used reliably to support decision making, the model's validity and utility must be established. The underlying modeling assumptions and limitations as well as the type of human performance outputs predicted by the model, are important aspects of establishing the scope of a model's utility. Through the process of building, calibrating and validating specific models, the modeling approach gains validity as a tool that can provide value within a particular application. The comparison of model predictions to empirical data is the essence of model validation, which typically involves one, or more, calibration steps. Calibration is the process by which a baseline model is adjusted to agree with known data. The level and type of validation and calibration required typically depends on the purpose of the model and the decisions it is meant to support.

The purpose of the IPME ADW model is to support various levels of decision-making relative to the goals of the Manning Affordability Initiative advanced workstation demonstration. One purpose is to use the model to support manning optimization decisions. This involves evaluating the impact of different team configurations on performance, workload, task allocations, and task automation. Calibration and validation for this purpose involves assuring that the distribution of outputs on the selected performance measures are consistent with the performance distribution produced by actual teams under the same conditions. A second purpose of the model is to support specific advanced workstation design decisions. In this capacity, the focus is on predicting potential

performance gains to be expected by adding specific advanced MMWS interface functionality, such as information visualization tools, control input modalities and attention management features. Supporting this type of decision usually requires validation of the model structure to assure that it reflects the different methods by which the human-computer interface (HCI) is used to perform tasks. Finally, the model can also be used as a diagnostic tool to help investigate and isolate observed differences in performance between teams or between teams and models.

Three types of validation are being utilized in the IPME ADW model. The first type of validation is to assure that the model operates as intended and performs within expected standards. The primary source of performance data for the IPME ADW model is subject matter expert (SME) estimates of the performance times and variability expected of an intact expert team. While these performance estimates are not necessarily formal performance standards, they are, in the judgment of very experienced subject matter and tactical training experts, acceptable levels of performance for an expert team given the conditions in place for the test scenario. They are, therefore, considered "performance standards" for the purposes of model development and initial evaluation and are referred to as performance standards throughout the validation discussion below. Discrepancies between model performance and performance standards can be attributed to one of two possible sources. Either the model structure or parameters do not accurately reflect the SME defined expert performance, or the performance standards may not be realizable in some situations (e.g., high workload). If appropriate, the model may be calibrated at this point to reflect expected performance. The output of this step is a "baseline" model that performs at the level expected of an expert ADW team on the modeled tasks.

The second level of validation is a comparison to performance data from intact teams using current workstation technology. One purpose of this level is to assure that the model output is consistent with the observed performance of intact expert teams and can, therefore, be used to predict expert level performance when validated modifications are made to reflect different performance conditions. Not all teams are performing at the same level. Different teams are in different stages of development and experience. Some may be bringing new team members up to speed. Another reason for validating the model against a variety of intact teams is to determine the average and range of performance of these teams on the selected set of performance measures.

The IPME model needs to be calibrated to reflect the differences between the observed performance of human teams and the expected performance built into the model. Calibrating the model will improve its ability to predict expected changes in performance due to advanced interface technologies for the average ADW team, not just for the most proficient teams. A number of different calibration methods can be used, depending on the level of validation required and the performance data available. One method is to proportionally adjust the original performance time estimates to fit the mean observed performance time and variance. This approach to calibration provides a good first approximation and may be sufficient to support model-based research on alternative manning configurations.

Another approach to calibration is to take a closer look at the workstation interface and modify task performance time estimates, as required, using the research based micro-task performance times built into IPME or imported from other tools. Micromodel-based task times provide a more fine-tuned estimate of performance times with respect to human performance capabilities and limitations derived from the research literature. This calibration process also involves working with the SMEs and experimenters to determine if there are prioritization heuristics, different from those in the model, that are being used by the test teams to guide decisions. This more detailed and diagnostic approach to validation of model structure can be used to support evaluation of specific design alternatives. It is appropriate where observed performance differences can be localized to specific sets of tasks. Ideally, after calibrating the IPME ADW model to more closely reflect the range of observed team performance, comparing it to another subset of the experimental data then further validates this calibrated model.

Once the model has been calibrated so that it produces valid performance data that reflects the distribution of performance results expected from proficient ADW teams, it is ready to be used as a predictive tool. At this point the third type of validation applies. Changes that are made to the model to simulate changes to the HCI, or to team structure or function allocation, must be validated to assure that the affected model parameters accurately reflect the intended design changes.

IPME ADW Model Validation Process – Initial Steps

At the time of this writing, the IPME ADW model has been completed and is being validated against the SME defined performance standards. In addition, some very preliminary comparisons have been made to the

experimental data collected from intact ADW teams. These data will be used to determine how the model performs relative to an observed performance distribution and to start to consider the type of model calibration that may be needed. While not enough progress has been made to present detailed model validation results, some preliminary findings are provided in the following sections to illustrate the validation process.

Comparison to expected performance standards. The first step in validating a model is to compare the model results against expected results (i.e., Does the model behave as it was initially designed?). The first performance measures to be evaluated to determine if the model is behaving as designed are related to the track identification process. Two performance measures "Time to First ID" and "Time to New Track Report" are used. "Time to First ID" is a team performance measure which includes the amount of time it takes any of four possible operators to notice a new track on their tactical display following detection by an air search radar and to correctly identify it by kinematic and other data. "Time to New Track Report", measures the time from the point of radar sensor detection, to the time when the appropriate operator, the Anti-Air Warfare Coordinator (AAWC) begins to report the new track on the external communications circuit. This measure is dependent upon the time it takes the team to notice and correctly identify the track, but the identification is typically completed by a different operator, the Identification Supervisor (IDS). Although timely identification may increase the likelihood of a correct and timely voice report, this is not always the case, as the voice report is dependent upon the situational awareness and workload of the AAWC. The voice report is also limited by the amount of voice traffic over the communication circuit, since only one person can speak on that circuit at a time.

SME's provided latency estimates of expected performance for the two performance measures described above. In both cases, multiple tasks are performed within the task network to arrive at these latencies. For example, to reach "Time to First ID", an operator must first observe that a new track appears on the radar screen. An operator then must make the decision to 'hook' that track (physically click on the track to request information from the system regarding that track), evaluate the information, determine a classification for that track and input that classification into the system. Similarly, to arrive at "Time to New Track Report" latency, a second operator (AAWC) decides to issue the report, confirms the identification of the track, and then waits for a moment when the appropriate external communication circuit is free.

Running the scenario through the IPME model ten times, yielded average times for "Time to First ID" that were about 36% slower than the SME-generated values, although the median value was only slightly slower (12%). As noted above, multiple tasks and decisions were involved in modeling the latencies, as opposed to assigning a single task an average duration specified by the SME's. Additionally, the SME-provided task times are nominal and independent of situational workload, and thus these results are encouraging in terms of validating the model. Likewise, both the average and median "Times to New Track Report" for the model are close by similar intervals to the SME-provided estimate.

Although these results demonstrate the model's ability to generate valid data when compared against expert estimated performance, they do suggest that some model calibration may be required. However, these two measurements are only two out of approximately eight that we are using to validate the IPME model. Validation will include comparisons such as whether queries and warnings were issued appropriately, how many verbal update reports were issued, and wait times for communication circuits.

The next step in validating the IPME ADW model is to compare the model performance to observed human performance.

Comparison to Human Performance Data. Data were collected from eight active-duty intact ADW teams, using the same two-hour scenario used in the IPME model development. This scenario consists of about forty-five minutes of 'low workload' activity and forty-five minutes of 'high workload' activity, with each period bounded by short periods of little activity. Additionally, the tracks that appeared through the scenario are a mixture of different levels of criticality. For example, a commercial airline would have a low level of criticality, whereas an aircraft originating from enemy territory and heading straight for ownship at high speed would have a high level of criticality. The "high workload" period in the scenario has a large percentage of high criticality tracks, many of which are moving at fast speeds and thus demand short reaction time by the operators.

Although the entire scenario includes seventy-nine air tracks, human performance data were collected on only twenty-five. To control for the differences in the characteristics of each track, we compared our data to only those twenty-five tracks. Since the model performance parameters were selected to reflect highly proficient performance, the model data were also compared to one of the eight teams tested, which was

identified by a SME to perform the best on these measures. For the purpose of this paper, we will refer to that team as the "best performing" or ''best" team. It is expected that model performance will be closer to the best performance than to the average performance.

The following sections provide some preliminary discussion of the IPME ADW model's "Time to First ID" performance as compared to the human experimental data collected. The discussion illustrates the process to be used for further validating and calibrating the IPME model for all performance measures.

Comparison of Time to First ID. Table 1 below contains a comparison of the results of the "Time to First ID" between the IPME ADW model, the eight ADW teams, and the "best performing" team for the same twenty-five tracks. These values express the percentage model results differ from the experimental data. For example, the mean IPME model "Time to First ID" was 55.4% lower than the "Time to First ID" for all of the experimental teams. However, the median model "Time to First ID" was 66.7% higher than that for the best performing team.

	IPME vs. All Teams	IPME vs. Best Team
Mean	-55.4%	-51.0%
Median	-29.7%	+66.7%
Std Dev	-88.9%	-86.7%

Table 1. "Time to First ID", percent change of IPME from ADW team data for 25 tracks

As might be expected, the mean values for the IPME ADW model versus the ADW teams are very different, and there is clearly more variation in the human performance than in the performance of the IPME model. A key part of the model validation and calibration process is to determine valid estimates of the levels of performance variability to be expected and calibrate the model accordingly. The observed variability in "Time to First ID" performance is likely due to a number of factors that are not currently considered in the IPME ADW model. These factors might include level of experience and amount of training. The differences between the mean and standard deviations of the ADW model and the ADW teams suggests that calibrating the model to reflect observed ADW team performance on the track identification task might be achieved by modifying mean and standard deviations across ID process subtasks, or including random modifiers such as experience level or amount of training to increase variability.

However, some of the variability in observed "Time to First ID" performance might also be due to factors induced by the human-machine system, by human information processing limitations or by performance heuristics used by the ADW teams that have not been captured in the ADW model. These are just the kind of performance shaping factors that the advanced technology workstation is designed to address. Performance variability for "Time to First ID" might also to be significantly reduced as a result of MMWS capabilities. The IPME ADW model will provide a more accurate basis for predicting expected performance changes to the extent that it has been calibrated to capture specific sources of variability.

The comparison of median data in Table 1 suggests that some of the observed "Time to First ID" performance variability may be due to more specific factors than experience or training. The median values are much closer in terms of absolute values. The distributions of the individual measurements for the IPME model, all of the ADW teams and the "best performing" ADW team are similar to each other. However the ADW teams, and especially the best performing team, clearly distinguish between some tracks that are identified very quickly and others that are not identified for a substantial period of time.

In the next section, the ADW team "Time to First ID" data are compared to the IPME ADW model data in terms of the two previously mentioned factors: the low versus high workload period of the scenario, and the criticality level of individual tracks. The purpose of this analysis is to further define the distribution of observed "Time to First ID" performance so that the IPME ADW model can be calibrated to capture the performance shaping factors.

Scenario Workload Levels and Track Criticality. We compared the percent differences in "Time to First ID" for the tracks occurring during the low workload period versus the high workload period for IPME, all of the teams and the "best performing" team. The median values for each were very similar during the low workload period. However, the median values were much different during the high workload period, with the most notable difference being between the "best performing" team and all of the teams together. In fact, based on the median values, the best team seems to be better than the model at determining those tracks that should be attended to very quickly and those that are not as critical.

For example, the difference in mean and median times for the IPME model between the low and high workload periods is slight. From the low to high workload period, the mean time increased 2%. The median time decreased from the low to high workload period by 6%. The "best" ADW team shows a similar pattern, although the percent changes in times between low and high workload periods are much greater. While the mean time increased 14% from low to high workload, the median time decreased 45% from the low workload period to the high workload period. Conversely, both the mean and median times for all teams increased from the low to high workload period: 23% and 19% respectively.

To further explore differences in mean and median times for the ADW teams versus the IPME ADW model, we compared data for tracks with low or medium levels of criticality with those high criticality tracks during the high workload period. Again, the percentage change in times for the IPME model for the less critical tracks versus the highly critical tracks is small, with both the mean and median values falling by 4% and 7%, respectively, for the highly critical tracks. This time, however, this pattern is shared both by all of the teams and by the "best" team. Furthermore, the mean and median differences in "Time to First ID" for less critical versus highly critical tracks, is much larger for all teams, as well as the "best" ADW team. For example, the median value for all teams fell 64% from the less critical tracks to the highly critical tracks, whereas it fell 94% for the expert team.

To summarize, the IPME ADW model appears to perform in a way most similar to the "best" team, which is to be expected at this point. During the high workload period, the ADW teams are clearly prioritizing tracks by those they think are more critical. The difference is even more pronounced with the best performing ADW team. The mean time of the "best" team for highly critical tracks during the high workload period matches the original SME-provided estimate for the average "Time to First ID". The model differs from the experimental teams in that, although it operates using well defined prioritization rules that give priority to the most critical tracks, it does attend to less critical tracks much earlier than the human teams. This suggests that the human teams may be using additional prioritization rules that are not captured in the model or that the HCI and information processing limitations may limit attention to the most critical tracks and tasks. The model can be used to help determine which of these factors may be operating by testing the various candidate hypotheses. By calibrating the model to capture these sources of performance variability, the model will be in a better position to provide valid performance predictions for future workstation functionality. The calibration process will allow the development of versions of the model that reflect expert and average team performance.

Model Prediction: Application to MMWS

The process of comparing the IPME ADW model predictions to empirical data and calibrating the model(s), as required, to reflect both expert level performance and the performance variability observed in the intact teams, will result in a validated model. Once validation is completed, the validated model can be used as a basis for predicting the individual and cumulative impact of design modifications on ADW team performance and workload. The validated model can then be modified to simulate the alternate team designs, task automation functions, advanced human machine interfaces features, and other advanced technologies proposed for the MMWS design. The same set of dependent variables used as the basis of comparison to the human experimental data during the model validation stage will serve as the primary measures for evaluating the potential impact of the advanced workstation functions on performance. The design of the IPME model allows for modifications to the timing, workload and function allocation rules within the model. This model design feature will facilitate many of these modifications required to evaluate the impact of MMWS technologies on performance. We expect to use the IPME ADW model to provide performance predictions in four areas: automation of task performance, advanced display and control technology, task and workload management, and team configuration.

Automation of Task Performance. One set of tasks that is expected to be automated in future tactical workstations include selected voice communications, such as those required for standard reports, queries and warnings. The system will automatically generate standard reports and queries that are triggered by scenario situations, for the operator's review and approval. Once approved, the reports and queries can be transmitted using synthesized voice. Automation features of the IPME modeling environment will make automated reports a relatively easy modification. The changes to the interface, and the requirement for reviewing and approving communications, will also be modeled. The model will then be run and compared to the baseline version without automated reporting.

Advanced Display and Control Technology. The IPME ADW model can be used to predict changes in performance expected as a result of adding specific advanced workstation design features. During the early

MMWS design and evaluation process, human performance modeling was used in conjunction with rapid prototyping and human usability studies to evaluate the effectiveness of alternative and multi-modal data input (e.g., voice, stylus, keyboard) characteristics for task performance. Supporting this type of decision required validation to assure that the model structure reflected the different methods by which the HCI was used to perform tasks. Where this detailed level of model analysis is required to evaluate specific features, more detailed modifications to the model will also be required. One area where this level of analysis may be required is in modeling the advanced workstation graphical HCI features that support the information gathering and integration required for decision-making (e.g., track prioritization, ID information, and threat assessment). Advanced HCI features might also be expected to impact the structure of information gathering tasks (different button presses, looking in different places), the speed of processing (ease of locating and extracting information from displays), and the accuracy of processing (identifying the information that defines task priorities).

Task and Workload Management. Other MMWS capabilities to be modeled are task and workload management. The "task manager" is a performance support capability that helps individuals and teams prioritize and manage work activities. It is designed to perform activities such as keeping track of active tasks, prioritizing tasks, and alerting operators when task deadlines are approaching. Task management functions are expected to impact performance across most tasks and to reduce search and decision making times, thereby reducing the cognitive task time between the output-oriented tasks. Model modifications will involve performance times for existing cognitive tasks and the structure of information gathering and attention management tasks.

Team Configuration. One objective of this prediction phase is to use the IPME model to predict how reduction in ADW team size, in conjunction with advanced technology features, will impact performance. The IPME ADW model was designed with this key requirement in mind. Once the model is modified to simulate team performance using the MMWS advanced interface and automation functions, the MMWS design team members and tactical training SMEs will validate the model structure. The resulting model will have gone through the three validations steps described earlier and will be expected to produce reliable estimates of team performance under different manning configurations.

CONCLUSION

The IPME ADW model is currently being calibrated against performance standards to assure that it is reliably producing performance data as intended. Validation and calibration using data collected from human experiments performed with the current workstation is also just beginning. The initial pattern of performance data generated by the IPME ADW model demonstrates the potential utility of using human performance modeling as a tool to support HCI usability studies for advanced workstation design. The cycle of model validation and model-based design evaluation described here provides a powerful way to integrate human factors engineering into the design of future systems.

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.

Gray, W.D., John, B.E., and Atwood, M.E. (1993). Project Ernestine: A validation of GOMS for prediction and explanation of real-world task performance. Human-Computer-Interaction, 8, 3, pp. 209-237.

Kieras, D. (1998). Towards a practical GOMS model methodology for user-interface design. In M. Helander, (Ed.) Handbook of Human Computer Interaction. Amsterdam: North-Holland Elsevier.

Laughery, K.R., and Corker, K., (1997). Computer Modeling and Simulation of Human/System Performance. In G. Salvendy (Ed.), The Handbook of Human Factors, (pp.1375-1408). New York: John Wiley & Sons.

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

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