

Predicting Communications and Workload with a Team of GOMS Models: Lessons Learned

Thomas P. Santoro

Naval Submarine Medical Research Laboratory
Groton, CT
santoro@nsmrl.navy.mil

David E. Kieras

University of Michigan
Electrical Engineering & Computer Science
Department
Ann Arbor, MI 48109-2110
kieras@eecs.umich.edu

ABSTRACT

Lessons learned from modeling the performance of human surface warfare teams with a team of GOMS models are presented. The resulting model teams successfully demonstrated the extension of GOMS technology from fine-detail ‘keystroke’ level tasks to tactical operations shared among cooperating individuals engaged in contact management activities typical of surface, air, or undersea warfare. The predictions of the models served to bracket upper and lower limits of actual team performance well in most cases. However the direct comparison of the ‘best’ model team predictions to actual data produced mixed results for task completion latencies and reasonable matches only for auditory and vocal workload. In some instances the models predicted as much as 30% shorter latencies than the real teams achieved. Data for communications and workload sharing in the critical air track management tasks for this domain were difficult to collect and lacked definition because critical aspects of these team activities were left largely to the prerogative of individual team members. Consequently the model provided clear characterizations of possible team structure designs while the data was ambiguous about what actual design was being implemented by a given team. Alternative models for these interactions predicted significantly different outcomes implying that explicit procedures for how scenario events are acquired and enter the individual and collective decision loops of the team are critical to achieving optimal performance.

INTRODUCTION

GOMS models, introduced by Card, Moran, and Newell (1983) are a way to characterize the

procedural knowledge required to use a system within the time constraints of human sensory-motor and cognitive behaviors. To construct a GOMS model, one determines the user’s Goals, lists what Operators can be executed in the interface, discovers the Methods, which are sequences of operators that will accomplish the goals, and the Selection rules that pick out which method to use to accomplish a goal when more than one applies. Once the initial model is calibrated with actual operator performance data, it is usually very accurate in predicting the effects of design variations on expected performance. Inserting a model into the build – test – build cycle then allows exploration of the entire design space in a more comprehensive fashion with more confidence that, when prototype designs are constructed, they are likely to be basically good.

Since the original Card et al. (1983) proposal, considerable progress has been made in developing this concept into several useful techniques (John & Kieras, 1996a, b), and greater clarity has emerged on the role of how models can be used in interface design, Kieras (2003). More recent work has begun to connect GOMS modeling with the developing computational cognitive architectures such as ACT-R, Anderson & Lebiere (1998) and EPIC, Kieras & Meyer (1997), Byrne (2003); such work will connect the practically-oriented GOMS methodology with fundamental mechanisms of human cognition and performance. A step in this direction is GLEAN (GOMS Language Evaluation and ANalysis), a tool for constructing and running computational GOMS models; it has been under development for several years, Kieras et al. (1995); Wood, (1993), and is currently available for research purposes, Kieras (1999).

GLEAN provides an executable programming language for GOMS models, GOMSL (GOMS Language), that resembles a familiar programming language, making the models relatively easy for system developers or other non-specialists to construct. The GOMSL program is interpreted and executed by a simplified computational cognitive architecture (Figure 1) that incorporates some basic facts and parameters about human performance in addition to the conventional GOMS keystroke-level model. The simulated human on the right-hand side of Figure 1 interacts with a simulated device on the left-hand side. Simplified perceptual processors translate simulated sensory input from the simulated device display to the cognitive processor, and the cognitive processor can command vocal and manual motor processors to produce simulated movements on the device's inputs, such as simulated keystrokes or mouse movements. The device can then change the contents of the simulated display accordingly.

While well known for its capacity to deal with very fine details of the HCI and associated human behaviors (John & Kieras, 1996a, b), GOMS has, however, rarely been applied to model and predict the performance and outcomes of team activities rather than activities of individuals working on their own. This paper presents the lessons learned about

team communications and workload using GLEAN in a modeling project for a user interface under development for new Naval surface combatants (SC-21). Kieras & Santoro (2004) provides further details concerning the development of GLEAN for use in this project. The work demonstrates that the concept of modeling a team of humans with a team of human models can provide a bridge between the psychology of individual humans and the organization and functioning of teams.

BACKGROUND

This work was sponsored by the ONR SC-21 Manning Affordability Initiative, a large project that sought to explore how modern computer technology could reduce the size of warship crews. A major thrust of the Manning Affordability Initiative was the development of a watchstation design for use by operators in the Combat Information Center (CIC), the Multi-Modal Watch Station (MMWS), and a new concept of how the CIC jobs associated with Air Defense Warfare (ADW) would be organized. A full description is provided in Osga et al., (2002). The new system was designed and evaluated using conventional human factors and user testing techniques supported by a software rapid prototyping tool that brought new concepts into

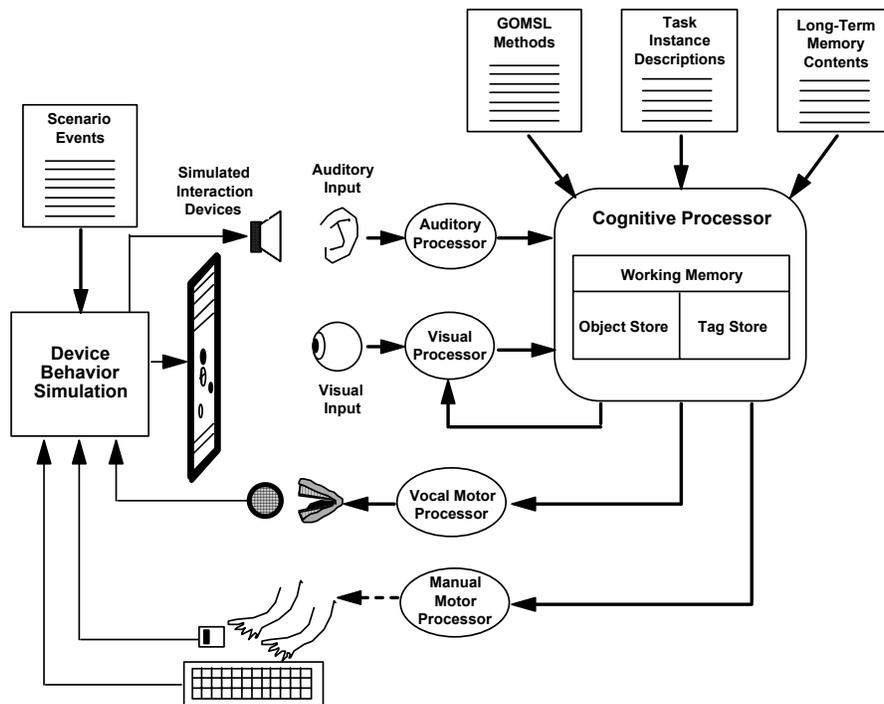


Figure 1. Architecture of the GLEAN system.

simulated operation for evaluation very quickly. The goal for the GOMS/GLEAN tool was to determine whether and how it could contribute to this design process.

THE OODA LOOP MODEL

The top-level tasks for the ADW model are shown in Figures 2 and 3. In general, these tasks follow the detect to engage process, sometimes referred to as the ‘OODA Loop’ (Observe, Orient, Decide, Act; Boyd, 1984; Fadok et al., 1995). According to this sequence, ‘Observations’ are made by sensory

mechanisms interacting with the HCI. Next, in the critical ‘Orientation’ stage, the meaning of those observations is interpreted in the context of the ongoing tactical situation through assessment processes (Santoro and Amerson, 1998) as shown in Figure 3. Depending on the results of that ‘Orientation’ stage, a threat assessment is developed that then leads to a ‘Decision’ being taken on possible ‘Actions.’ This sequence of processes is iterative with each Decision and Action stage followed by Observations and Orientations which serve to correct and guide the sequence to an acceptable end result.

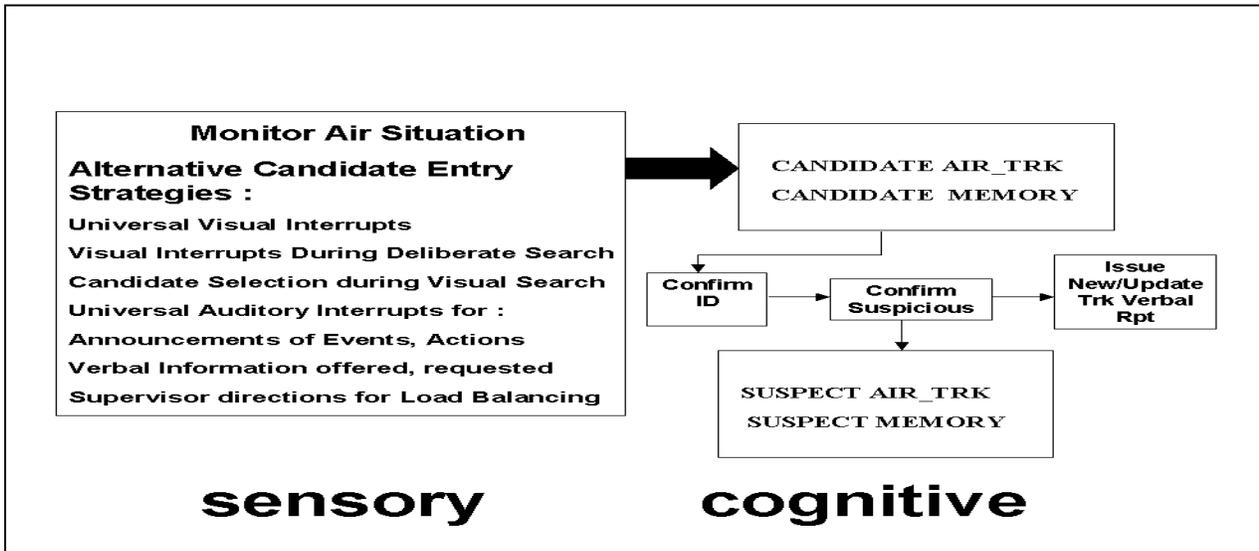


Figure 2. Air Defense Warfare (first half) Observation stage

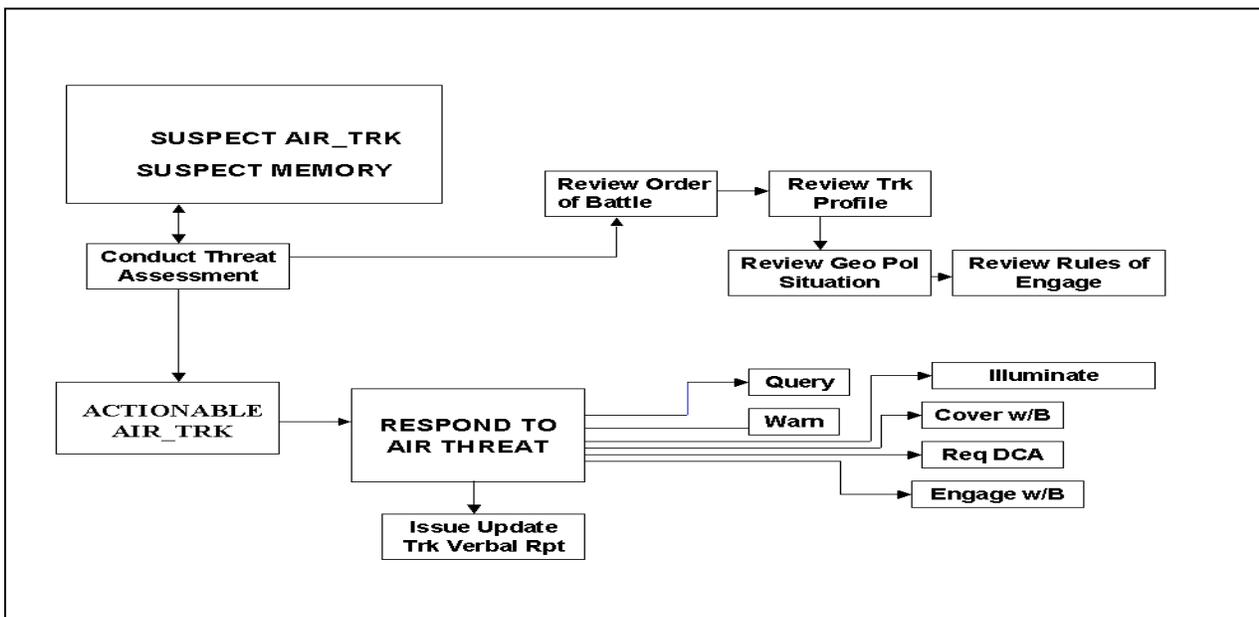


Figure 3. ADW (second half) Orientation, Decision, and Action stages

The task procedures for individual team members were constructed from task analysis by subject matter experts (SMEs) based on an extrapolation from the existing AEGIS watchstation to the new MMWS. A five-member team structure was constructed with primary and secondary task allocations as shown in Fig 4. In particular, the philosophy of one-member to one-circuit for off-board communications was carried over from AEGIS. Also carried over was a certain level of freedom in task allocation and supervision by the team leader. In particular, each team could adapt

their own way of sharing the critical Air Sit Monitor task, a primary task for the leader but a secondary task for all members. This critical task determines the entry point for contact information in each operator's OODA loop. If the contact is acquired visually as a new candidate air track, it must pass through the entire cycle, but if another operator performs some of the OODA stages and announces the results to the team, other members are saved valuable processing time. As will be explained later, variations on this protocol have a significant effect on overall team performance.

Task Allocation						Task
Operator:	Team Ldr	AWC	IQC 2	IQC 1	AIC/RC	
Ckt:	Cmd	AAW	EW	IAD/MAD	Redcwn	
	S	P				Conduct Engagement
	S	P				Order Air Engagement
						Review Engagement Status
		S	S	P		Respond to Air Threat
		S	S	P		Issue Level I Query
	S	P				Issue Level II Warning
	S	P				Illuminate Track
	S	P				Order Cover
	S	P				Order DCA Action
	P	S	S	S	S	Monitor Air Situation
		P	S	S		Monitor Air Situation
		P	S	S		Issue New Track Verbal Report
						Issue Update Track Verbal Report
		S	P	S		Respond to ESM
		S	P	S		Issue ESM Heads Up Report
		S	P	S		Issue ESM Racket Report
		S	P	S		Review an Uncorrelated ESM
					P	AIC Tasks
					P	Clear Aircraft to Proceed on Mission / Hand Off
					P	Clear Aircraft to RTF
					P	Manage DCA Intercept
					P	Manage DCA Cover
					P	Manage DCA Engagement
		S			P	Respond to Hot Dog Violations
	P	S				Ownership Situational Awareness
	P	S				Review Doctrine
	S	S	S	S	P	Review Readiness
	P	S				Review ATO
						Monitor Team Workload/Performance

Figure 4. ADW Task Allocation for 5 Member MMWS Team

EXTENDING GOMS TO A TEAM

Initial models were based on SME analysis of how the task should be conducted in order to follow rules of engagement, etc. The GOMS models were programmed to carry out the stated tasks on the new MMWS interface according to the SME procedures. The resulting models could predict observed task execution times reasonably well, 10% or better in some cases. This would be expected

given the long history of using GOMS models for this purpose (John & Kieras, 1996a, 1996b.).

The analysis of team structure in system design normally tries to subsume team activity under a more general resource allocation framework in which the boundaries between team members and their supporting devices tend to be blurred, and the details of the interactions between team members are lost. In this project we demonstrated a rather straightforward approach: if one can simulate an

individual user acceptably well, then one can model a team of such users by setting up a model of each user and having the models interact with each other according to specified team procedures or team strategies. These are simply part of each individual's methods.

For example, the GOMS model for the operator who services Electronic Sensor Measures (ESM) events may specify that when the operator notices a new ESM event on the workstation display, the operator will announce it by speech over the internal net. The methods for another team member could specify that upon hearing this announcement, the air contact in question should be sent a query. In this way, the first operator has saved the second the work of that decision-making process.

This team modeling approach can be expected to make good predictions for actual team behavior only if the actual team is also following the model's communications and task sharing protocol. However, in the case of the critical Air Sit Monitor task, the guidance for real human teams, both with the legacy AEGIS system and the new MMWS, is vague in this area and, as will be seen in the lessons presented below, teams were free to follow previous experience and personnel style in both communications and secondary task sharing procedures. The simulated humans, on the other hand, were programmed to communicate and select tasks in very specific ways. They complied strictly to the given protocol for communications on the internal team network as well as the off-board circuits. When their primary tasks were not active, the models diligently performed the Air Sit Monitoring task and shared their products with other team members.

Using this approach, we explored alternative team designs that focused on how the team members cooperated on the critical Air Sit Monitor task. Different team organizations were simply represented in the individual GOMS methods that specified when announcements would be made over the intercom, and what actions would be taken in response. This modeling approach was effective in that different team procedures were found to produce clear differences in overall team performance, thereby providing a rational for optimization. This demonstrates that the powerful and useful capability of GOMS for comparing alternative operating procedures does indeed easily

extend from the traditional individual task model to tasks that are shared among model operators. This approach to modeling team behaviors may also work with other types of cognitive-architectural models of human cognition and performance.

MODEL VALIDATION

Research on GOMS and other modeling methodologies has usually tested the validity of the model by comparing its predictions to empirical data collected with actual human users in the same tasks and interface. In the case of GOMS and related methods, there is enough of a record of validation success to accept that the GOMS model methodology is basically valid (John & Kieras, 1996a, b). But a GOMS model is based on a task analysis that identifies the user's goals and procedures, and can be wildly inaccurate if the task analysis does not accurately represent them.

In the human exercises for this study we were unable to discern a set pattern for the Air Sit Monitor task from observations, both direct and recorded, of team interaction and communication. It is a U.S. Navy tradition that each team can evolve its own structure and procedures of operation at a certain level, under the direction of its leaders. Other than the "Control by Negation" directive and certain "no sooner than..." and "no later than..." rules of thumb, teams seemed to be free to execute scenario tasks according to their own personnel "style." In monitoring air tracks, that style seemed to vary from "anyone call out a contact anytime" to "only the team leader can designate tracks and order actions" with intervening variations of team member participation. This made the results difficult to interpret without tediously viewing video and audio records to determine which operator announced which contact, what information was passed and who did what with it when it was passed.

Without a uniform protocol to model, the various model designs for representing team interaction on the critical monitoring task met with mixed results. For example, estimates for task latency times (the time from when a given air track is first actionable to when the particular action is performed) were in certain cases significantly shorter than the latency actual teams achieved. Predictions in some cases matched the data within 10% but others were better than 30% shorter. Presumably, if the human team had followed the procedure programmed for the

model team, their latencies might also have been as short, but there is no way to really know. The lesson is that empirical data in which the humans are doing something different from the task analysis can neither validate nor invalidate a model based on that analysis. Given the difficulty of data collection in these complex tasks, any simple concept of validation against data is unworkable when strict adherence to protocol is not maintained.

WORKLOAD VALIDATION

If the goal of balanced team workload is to be achieved, a reliable measure that determines when a team member is under-loaded, over-loaded, or balanced must be obtained. In the exercises reported, SME observers monitored each operator at ten minute intervals over the scenario and generated an overall workload figure in the form of a number from 1 to 7 that represented their subjective measure of the work performed by the operator in that period. The workload observers were selected for their expertise in the tasks of a particular watchstander and were instructed to base their estimates on the activity level of that operator relative to their estimate of the individual's maximum possible work output.

As a first look, the reliability of the empirical workload ratings was assessed by considering the extent to which the workload ratings over time for a particular job role in a team correlated with the ratings for the same role in a different team. Overall, the intercorrelations are high enough to be considered adequately reliable as an assessment instrument. The correlations with the mean workload ratings were high enough to assume that the mean ratings adequately reflect the individual team ratings.

Since the GLEAN tool simulates the activity of sensory-motor and cognitive processes using time parameters derived from experimental psychology and psychophysics, it offers the possibility of estimating the portions of that duration when the different visual, auditory, cognitive, and psychomotor (VACP) modalities are active. For example, statistics are recorded by the GLEAN tool on the total number and duration of visual actions performed on each repetition of each task. The overall totals for any designated time period can also be computed as desired for any individual operator

model. These numbers can form the basis for a visual workload estimate.

In order to validate them against SME observer estimates, the numbers were totaled at 10 minute intervals for each GOMS model over the test scenario. A post-hoc prediction equation was constructed using stepwise multiple regression of the GOMS numbers against the subjective estimates. Only two predictors entered the equation. One was the vocal activity, but in addition, the auditory activity entered also. The resulting prediction equation was:

$$WL = 1.940 + 0.01108 * \text{Vocal} + 0.09737 * \text{auditory}$$

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

The fact that vocal and auditory activity are the most important predictors suggests that the observers were basing their ratings heavily on the operators' reporting and other verbal communication tasks. While the tasks in the ADW scenario do involve significant levels of these activities, there is no question that the other modalities, especially visual and cognitive, also play major roles in an operator's performance. Other workload measures such as eye movements or EEG, should be explored to validate the GOMS predictions for these modalities.

DESIGN EVALUATION

In contrast to the uncertainties in the user testing data, the modeling exploration of different team organizations produced very clear conclusions. As described in Santoro et al., (in press), following the general logic suggested in Kieras & Meyer (2000), we first constructed models that bracketed the required performance. The first model made

unrealistic assumptions that every team member noticed and acted upon all critical events and worked completely independently. This superhuman non-team performed the task very well. A second model made more realistic assumptions about attention to events, but still lacked any cooperation. This model performed quite poorly - too many events were missed while the team members worked on their individual tasks. With the next models, we systematically attempted to find the amount and kind of cooperation that would result in fewer missed events. In the better models, if a simulated operator was not busy with a specific task, it would announce critical events to the rest of the team; other team members could hear this announcement and remember it long enough to act on it after completing a task in progress. The model teams that performed well thus define a good team strategy, and the evaluation of the interface then becomes very clear: For each specified set of team organization and procedures, a particular interface design will result in a predicted level of performance.

This experience suggests that in a complex task where user data is hard to collect and user strategies are hard to determine, it will be easier to evaluate the interface with modeling than with user testing. To put it differently, having clear hypothetical accounts of how a task could be done is a more useful state of affairs than having unclear empirical data. Of course, only a fool would skip all user testing in the design of such a critical system, but our experience suggests that a model analysis may be the most practical way to arrive at detailed design decisions for complex team tasks.

CONCLUSION

This application of GOMS modeling to the interactions of operators in an Air Defense Warfare team has illustrated the importance of a well-designed communications and load-sharing protocol on overall team performance. There are significantly different ways to implement a protocol that seeks to maintain team situation awareness and balanced workload. A comprehensive task analysis must be done to identify the most effective one. Team member personalities and individual preferences and capabilities are not necessarily the best determinant of these aspects of the team design structure. The difficulty of validation for communications and

workload sharing from human testing data alone makes a strong argument for the use of a team model that has the level of detail available in the GOMS technology for the design and evaluation of these critical procedures.

Certain results of the workload predictions are encouraging for the prospects of further development with a GOMS tool. The models treated each operator as an individual who had primary tasks assigned and then would perform a secondary task, Air Sit Monitoring as a backup for the team. They provided a valuable insight to the extent to which two or more individuals might collaborate to complete a given task together. With further development of the workload measures, models that account for this kind of collaboration could characterize the flow of workload from one operator to another and provide guidance for workload dynamic balancing across the team.

ACKNOWLEDGEMENT

This work was conducted in support of the SC21 Manning Affordability Initiative under the Naval Submarine Medical Research Laboratory Work Unit 62233N-R3322-50214, entitled "Human performance modeling of the MMWS Build 1 Watchstation." The opinions or assertions contained herein are the private ones of the authors and are not to be construed as official or reflecting the views of the Department of the Navy, the Department of Defense, or the United States Government.

REFERENCES

1. Anderson, J.R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
2. Boyd, J.R. (1984). Organic design for command and control. Unpublished briefing paper, pp. 5, 32-35.
3. Byrne, M.D. (2003). Cognitive architecture. In J. Jacko & A. Sears (Eds), *Human-Computer Interaction Handbook*. Mahwah, N.J.: Lawrence Erlbaum Associates.
4. Card, S., Moran, T. & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, New Jersey: Erlbaum.
5. Fadok D.S., Boyd, J.R., and Warden, J. (1995). *Airpower's Quest for Strategic Paralysis*. Air University Press, Maxwell Air Force Base, AL.

6. John, B.E., & Kieras, D.E. (1996). Using GOMS for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction*, **3**, 287-319
7. John, B.E., & Kieras, D.E. (1996). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction*, **3**, 320-351.
8. Kieras, D.E. (1999). *A Guide to GOMS Model Usability Evaluation using GOMSL and GLEAN3*. Document and software available via anonymous ftp at <ftp://www.eecs.umich.edu/people/kieras>
9. Kieras, D.E. Model-based evaluation (2003). In J. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook*. Mahwah, New Jersey: Lawrence Erlbaum Associates. 1139-1151.
10. Kieras, D.E. & Meyer, D.E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction.*, **12**, 391-438.
11. Kieras, D.E., & Meyer, D.E. (2000). The role of cognitive task analysis in the application of predictive models of human performance. In J. M. C. Schraagen, S. E. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis*. Mahwah, NJ: Lawrence Erlbaum, 2000.
12. Kieras, D.E., Wood, S.D., Abotel, K., & Hornof, A. (1995). GLEAN: A Computer-Based Tool for Rapid GOMS Model Usability Evaluation of User Interface Designs. In *Proceeding of UIST*, 1995, Pittsburgh, PA, USA, November 14-17, 1995. New York: ACM. pp. 91-100.
13. Kieras, D.E. and Santoro, T.P. (2004). Computational GOMS modeling of a complex team task: lessons learned. In *Proceeding of ACM SIGCHI*, Conference on Human Factors in Computer Systems, April 24-29, 2004, Vienna.
14. Osga, G., Van Orden K., Campbell, N., Kellmeyer, D., and Lulue D. Design and Evaluation of Warfighter Task Support Methods in a Multi-Modal Watchstation. Space & Naval Warfare Center, San Diego, Tech Report 1874, 2002.
15. Santoro, T.P., Kieras, D.E., and Pharmer, J. (in press). Verification and validation of latency and workload predictions for a team of humans by a team of GOMS models. *US Navy Journal of Underwater Acoustics, Special Issue on Modeling and Simulation.*
16. Santoro, T.P. and Amerson, T.L. (1998). Definition and Measurement of Situation Awareness in the Submarine Attack Center. in *Proceedings of the Command & Control Research & Technology Symposium*, Monterey, CA, pp. 379-388.
17. Wood, S. (1993). Issues in the Implementation of a GOMS-model design tool. Unpublished report, University of Michigan

AUTHORS

Thomas P. Santoro is a Research Scientist at the Naval Submarine Medical Research Laboratory (NSMRL), Groton, Connecticut. He holds a BSEE from the University of Rhode Island, an MSEE, and a Ph.D. in Engineering Science from the California Institute of Technology. His research interests include measurement of human sensory perception in the visual and auditory pathways, and the modeling of sensory-motor and cognitive performance in human-computer interactions.

David E. Kieras is Professor in the Electrical Engineering and Computer Science Department and the Psychology Department at the University of Michigan. He received a BA in Psychology from Rice University in 1969 and a PhD in Experimental Psychology from the University of Michigan in 1974. His general research field is applied and theoretical cognitive psychology, with specific interests in human-computer interaction, cognitive simulation modeling, human performance, complex human learning, and natural language processing.