

**Final Technical Report**  
Office of Naval Research

Project Title:

**Computational Modeling and Eye Tracking of Multitasking Performance  
with Multimodal Auditory and Visual Displays**

Award number: N00014-06-10054

Grant Period: October 1, 2005 to September 30, 2008

This report was completed by Anthony J. Hornof on Sept. 1, 2009

**Personal Information**

PI Last name and first name: Anthony J. Hornof  
Performing Institute: University of Oregon

Mailing Address: Department of Computer and Information Science  
1202 University of Oregon  
Eugene, OR 97403-1202

Telephone: (541)346-1372  
Fax: (541)346-5373  
Email Address: hornof@cs.uoregon.edu  
Your website URL: <http://www.cs.uoregon.edu/~hornof/>  
Contract or Grant Number: N00014-06-10054  
Contract or Grant Title: Computational Modeling and Eye Tracking  
of Multitasking Performance with  
Multimodal Auditory and Visual Displays

Program Officer: Dr. Paul Bello, Ph.D.

**Introduction**

Critical Navy activities such as anti-submarine warfare require enormous automation but ultimately rely heavily on human performance capabilities and limitations. Yet it is impossible to collect human data in a sufficient subset of complex task scenarios to fully evaluate the overall effectiveness of the human-machine system. Reducing the number of crew members required to operate a vessel will ultimately require more automation and multitasked information assessment and decision-making by a smaller crew. However, designers have a limited understanding of human capabilities and limitations in these complex tasks and environments. Multimodal watchstations, which present vast arrays of data in visual, auditory, and perhaps even in the tactile modalities, are being introduced and will need to be evaluated and tested before they are implemented and deployed.

The Navy needs a better understanding of human capabilities and limitations with respect to monitoring and responding to situational and tactical data presented to the human on multiple displays and in multiple modalities. The Navy needs better predictive modeling, based on sound psychological theory, to explain and ultimately predict the complex human perceptual-motor interactions that lead to both success and to breakdowns in the operation of complex Navy equipment. Past and current modeling efforts have provided insight and guidance, but more work is needed.

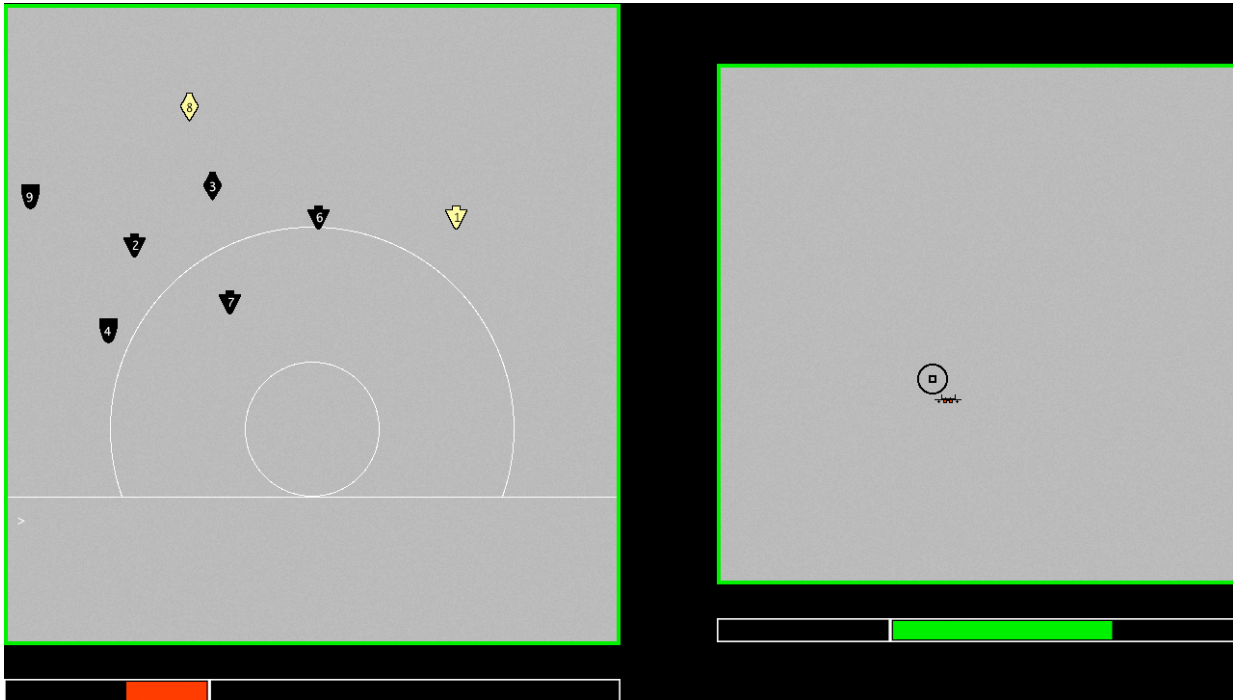
The project builds directly on previous research by multiple investigators (Brock, Ballas, Stroup, & McClimens, 2004; Kieras, Ballas, & Meyer, 2001; Kieras & Meyer, 1997) and advances an understanding of how to present situational and tactical data on multiple auditory and visual displays to be used in a multitasking environment. The project has accomplished this through cognitive modeling, empirical studies with eye tracking, and model refinements based on the eye tracking data.

The project investigates the human performance capabilities and limitations associated with performing multiple decision and motor tasks while monitoring multiple auditory and visual displays. Research findings address the problem in a number of ways. First, new models of human performance have been constructed to predict and explain the human information processing involved in a benchmark multimodal task. The models explain how and why performance limitations are encountered. Second, eye tracking data have been collected to reveal how a human employs their perception in this demanding multitask situation. Third, the models have been evaluated and refined based on eye tracking data. The project informs design opportunities afforded by the multimodal watchstation, and how to utilize the complex environment—and human capability—to the Navy’s advantage.

### ***The Task***

The task is based on the NRL (Naval Research Lab) dual task developed by Dr. Jim Ballas and colleagues at NRL in the early 1990s (Ballas, et al., 1992) to explore aspects of human-machine interaction. The task was used in early models built using the EPIC Cognitive Architecture (such as in Kieras, Ballas, & Meyer, 2001) in part because the task provides a good framework for recording aspects of multimodal dual task performance, which is of great interest to both the Navy and the scientific development of computational cognitive models—computer programs that simulate aspects of human behavior.

Figure 1 shows a screenshot of the NRL dual task as it was adapted to the current project. In short, the participant’s task is to interleave two subtasks: (a) looking at the *tracking* display on the right, keep the reticle on an airplane by moving a joystick with the right hand; (b) looking at the *tactical* display on the left, classify the blips (icons) on the radar screen as “hostile” or “neutral” based on a complex set of rules, and type in the blip number and classification using the left hand. Each experimental trial, or scenario, takes about eight minutes and has about sixty blips.



*Figure 1. A screenshot of the NRL dual task as used in the project.*

The source code for running the experiment was inherited from NRL, and was effectively entirely rewritten for this project. From the experimental perspective, perhaps the two most important enhancements to the software include (a) connecting the software to an eye tracker such that a “gaze contingent” paradigm can be used, in which case only the contents of the currently looked-at task display are visible, during which the other task display cannot be monitored in the visual periphery, and (b) a means of providing a participant with feedback on their speed and accuracy performance from moment-to-moment, as well as at the end of each scenario, including how much extra money they earned in bonuses for good performance.

The experimental software has also been substantially updated to accommodate contemporary peripheral devices and software interfaces, such as a USB joystick and a VRPN (Virtual Reality Peripheral Network) connection to a VR Sonic spatialized sound server, and to compile in a contemporary programming environment (Apple XCode) in the current version of the operating system (Mac OS X). These software engineering details do not directly contribute to the scientific merit of the exercise, but they do make it easier for other researchers, such as our collaborators at NRL, to reuse the software for their own running of the experiment.

The experimental procedure has also been substantially enhanced and refined, compared to previous executions of the experiment, to improve the validity of the experiment with regards to capturing expert behavior. Participants are methodically trained to a performance criteria for every subcomponent of the task before they can start the experiment; if they cannot reach

criteria, they are politely dismissed. Data is collected for each participant over the course of three consecutive days, with no modifications to the task from day to day. A carefully-tuned payoff matrix has been developed and tested to motivate participants to perform as quickly as possible while maintaining roughly 95% accuracy, and to balance their efforts between the two subtasks; monetary bonus feedback is provided at the end of each scenario, and a progress bar below each subtask window (see Figure 1) is continually updated, with the goal of keeping both bars equally in the green. The experimental instrumentation and procedure have been carefully designed to motivate and capture expert behavior at this complex task.

We introduced a number of precise characteristics to the waves of blips: (a) Five distinct wave sizes were used: 1, 2, 4, 6, and 8. There was at least a one second interval between the disappearance of the final blip of a wave, and the appearance of the first blip in the next wave. (b) The designation of blips within wave sizes was not simply random. (c) Some waves contained blips that were, given the task scenario, *preclassifiable*; other waves had blips randomly distributed. (d) Careful consideration was made so that no two blips were ever within 2° of visual angle of each other.

Additional experimental modifications include the following:

- Sounds is used slightly differently than in previous versions of the task. First, auditory signals are now derived from an actual real-world ATC task environment. Second, the signals do not indicate the *type* of blip but instead the *designation*.
- Each participant completed all conditions of the task for each of three consecutive days, thus permitting participants to start to develop some real expertise with the task.
- The experimental design was updated from a simple 4-way experimental design (four sound conditions, or four input conditions) to a 2x2x3x5 design (sound on or off, peripheral information available or not, day, and wave size).

### ***Instrumentation***

The instrumentation for running the experiment and collecting the data was quite complex. For the actual data collection with participants, two technicians were required to initiate procedures and monitor three different computer systems—a Macintosh Dual G5 that presented stimuli and recorded responses, a Windows-based LC Technologies 120 Hz Eyegaze eye tracker, and a Windows-based VR Sonic SoundSim Cube spatialized audio server. The three systems communicated in real-time via TCP/IP ethernet connections and appropriate communication protocols. The software for presenting stimuli and recording participant responses, and for communicating with the eye tracker and spatialized audio server, was written in C++ using Apple XCode. C++ was chosen over a language such as Java because there was no room for system latency, such as in recording or responding to eye movements in real time.

Particularly impressive is the instrumentation that was conducted to analyze the eye movement data, which made the possible the data and analyses presented later in this report. Extensive

analyses were conducted. But before these could be done, a rigorous examination of the error in the eye movement data was conducted using the VizFix software, created in the PI's lab, and post hoc error correction was applied using the RFL (required fixation detection) technique developed by Hornof and Halverson (2002). For this experiment, blips were used as RFLs, and the error correction technique was extended to incorporate both moving RFLs (the blips), and also to incorporate multiple error-correction "signatures" across an eight-minute scenario.

The next four paragraphs will explain in some detail how the error correction technique was adapted and applied to this experiment. To incorporate the moving RFLs, a series of screen objects are used to represent the trail of each blip. Thus, several objects that representing the same blip could fall into a fixation's duration time. Each object and the fixation location will then form a vector. However, only the shortest vector was used in the extended error-correction algorithm, and we call these vectors error vectors.

To dynamically form multiple error-correction "signatures" in one scenario, a lot of parameter study was conducted. Firstly, the error correction vectors (ECV) should be calculated separately for left and right half of the tactical display. The reason is described in Hornof & Halverson 2002 — the ECV will change for different locations on the display. Secondly, each fixation will have several error vectors, if a number of blips were within 4 degrees of visual angle to the fixation. Higher priority were assigned to the shorter error vectors, in other words, to those blips which were closer to the fixation. Weighted average error-correction vector was calculated each time a new error vector was processed. The weight is the squared duration of the fixation, since we believe the longer a fixation is, the more likely the fixation is landed on a blip.

This weighted average ECV is a temporary error-correction signature, and it will be compared to the next error vector. If the next error vector is about the same as the weighted average ECV, then it is included in this current error-correction group. If the next error vector is far away from the weighted ECV (1 degree of visual angle from the weighted ECV in the algorithm), a new error-correction group will start and hence a new error correction signature will be calculated.

After all of these ECV-related calculations, several error-correction signatures should be outputted for two parts of the tactical display and for different periods in that scenario. The next thing we will do is to check whether these error-correction signatures are all less than 1 degree of visual angle. If they are, we believe there is no need to do the error correction on that scenario file. If not, then those error-correction signatures should be applied to the raw data file, and re-imported to Vizfix.

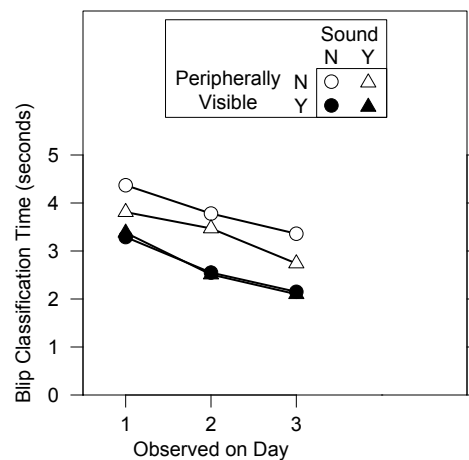
## **Human Performance Data**

Numerous data were collected, primarily classification time for each blip, classification accuracy, tracking error sampled every 83ms, and eye movements.

Statistical analyses of all data were performed using the *R* statistical analysis software. The statistical analyses were quite rigorous. The following series of procedures were conducted for each response variable. First, the distribution of the response variable is examined. If the data does not follow a normal distribution or other common distribution, data transformation is carried out. Usually, the reaction time data is *log* transformed. Then several graphs will be plotted to check if there were any extreme outliers, or if any assumptions of the intended statistics methods were violated. Graphs were also used at this step to investigate the interactions between factors. The third step is to fit models to the data. We use mixed effects model technique, because this advanced statistic method is very robust. It can handle unbalanced, nested data very well. A number of alternative models were compared before the final model is determined. The model comparison was based on several information criteria, and finally, the standard residual — fitted value plots were also examined. After the above procedures, we checked all the significant interaction effects and see if there is anything informative. Note that most of the means reported in this section (with count data one exception) are geometric means of the raw data because geometric means are the best central tendency estimate for a positively skewed distribution.

### *Classification Time*

Figure 2 shows the mean blip classification time across all participants and all days. Important trends in the data are as follows: Peripheral visible conditions (filled plot symbols) are faster, and peripheral not-visible conditions (unfilled plot symbols) are slower. Whether sound is on (triangular plot symbols) or sound is off (circular plot symbols) only makes a difference when peripheral information is *not* visible. Performance improves across the three days. Three of the four conditions show a negatively accelerating downward slope approaching asymptotic performance, whereas one condition (peripheral not visible and sound on) shows a positively accelerating downward slope that is *not* approaching asymptotic performance.



*Figure 2. Blip classification time as a function of day, for each of the four conditions.*

Figure 3 breaks out the data in Figure 2 by wave size. Important trends are as follows: In general, classification time goes up with wave size. Looking at conditions in which peripheral information is *not* available (the unfilled plot symbols) and sound is on (unfilled triangles) or off (unfilled circles), the two data plots are quite close to each other on Day 2, and pull apart from each other on Day 3. On Day 1, there is a steeper slope for classification time in the peripheral visible condition (the filled plot symbols in the first frame) than in all other plots.

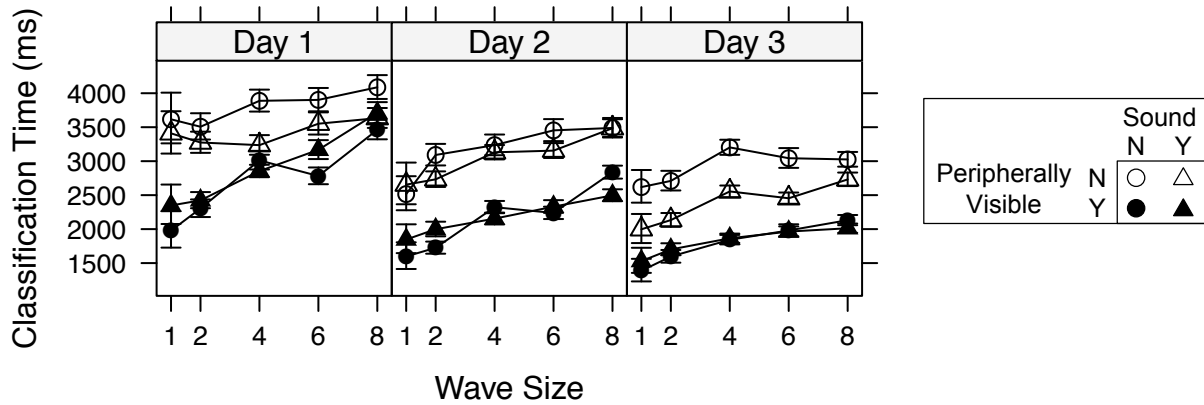


Figure 3. Blip classification time as a function of wave size, for each of the four conditions. The three panels show performance across the three days.

Looking at how blip classification time improves across the three days for waves of different sizes, we see that the day-to-day performance increase for waves with 1 and 2 blips, which we will call *small* waves, improves differently than waves with 4, 6, and 8 blips, which we will call *large* waves. Figure 4 shows the difference in these trends. Most notably, we see that the negatively accelerating slope for no peripheral information and yes sound (unfilled triangles) comes more from the large waves than the small waves.

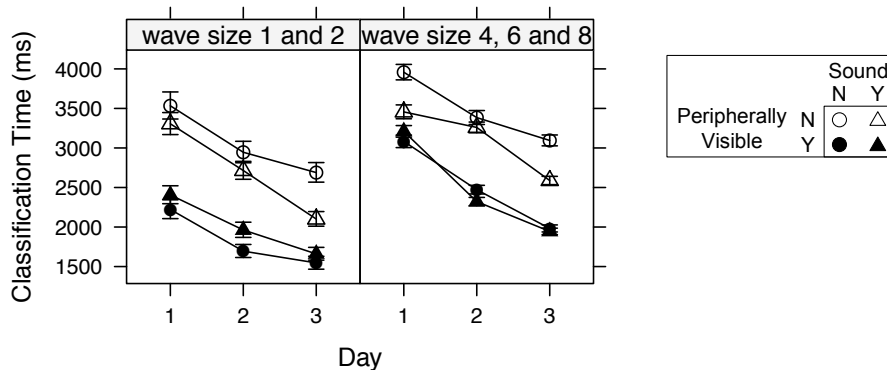


Figure 4. Blip classification time as a function of day, for each of the four conditions. The left panel shows small waves (1 and 2 blips). The right panel shows large waves (4, 6, and 8 blips).

Figure 5 shows blip classification time as a function of blip color. The most important effect is that yellow blips, which require consideration of shape, location, direction, and speed, take 519 ms longer to classify than red or green blips, which can be classified based just on color, and that red blips are classified 130 ms faster than green.

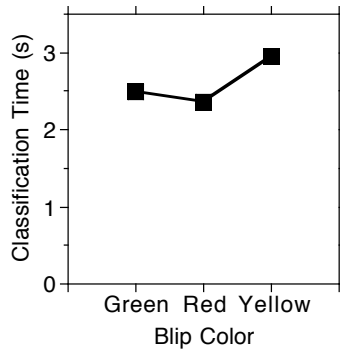


Figure 5. Classification time as a function of blip color.

### Tracking Error

Figure 6 shows the mean tracking error—the average distance between the tracking cursor and the tracking blip. Important trends are as follows: For all conditions, error decreased across the three days. Tracking error was always greater when no peripheral visual information was available, but the difference decreased on Days 2 and 3.

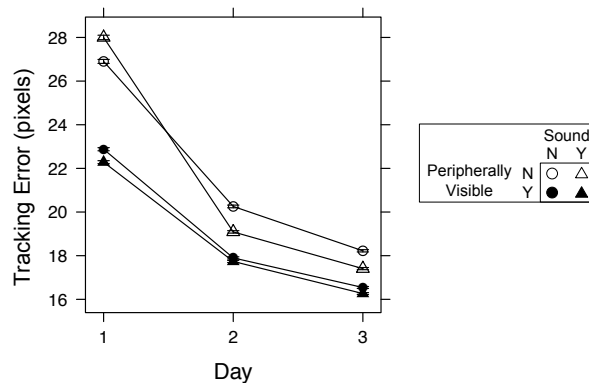


Figure 6. Mean tracking error (in pixels) for each of the four conditions, as a function of day.

## Eye Movement Data to Explore Fundamental Aspects of Dual Task Execution

This section will discuss the extensive eye movement data collected for this experiment and a number of analyses conducted on that data.

### *Fixations on Various Display Regions*

Eye movement data are first presented to explore or confirm some fundamental aspects of how participants did the task. Figure 7 shows the five regions that were used for classifying fixations on the task display. Perhaps note that the regions expand a bit beyond the actual display components to accommodate the error that is inherent in all eye tracking data.

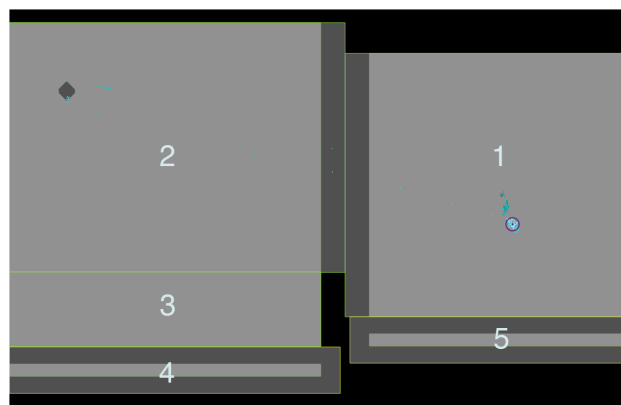


Figure 7. Regions used for classifying fixations in the visual display.

Table 1 shows the percent of fixations observed in each of these five regions, as well as the percent of fixations that fell outside of these regions (Other). The most important observation here is that nearly all fixations fell within either the tracking or tactical display.

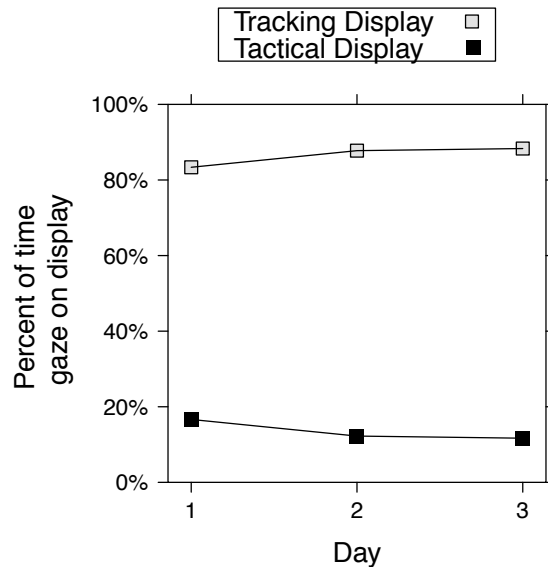
**Table 1. Percent of fixations observed in each screen region.**

Visual Region	Fixations in that Region
1. Tracking	49.6%
2. Tactical	48.7%
3. Text output	0.5%
4. Tactical status bar	0.6%
5. Tracking status bar	0.3%
6. Other	0.3%

The tactical blips (and the moving region of interest around each blip) are perhaps the most important destinations of eye movements. On the tactical display, the eyes can move to black blips, colored blips, and to locations not clearly associated with any blip.

### *Fixations on Tactical Versus Tracking Displays*

Figure 8 shows the percentage of time that participants' gaze was on the tracking versus the tactical display. Important trends include that participants kept their gaze on the tracking display for a *huge* majority of overall task time, and that the trend did not change across the three days.



*Figure 8. Percentage of total task time that participants' gaze was on the tracking versus tactical task display as a function of day.*

### **Eye Movement Data To Explore Tactical Task Strategies**

The analysis of eye movement data is something of an art form, especially if the data are to be used to reveal human strategies and thus inform cognitive modeling. There are few if any eye tracking measures that are as well-defined and reliable as reaction time and accuracy, the two classic human performance measures. In almost all situations, eye tracking measurements are crafted with an understanding of the task, and human performance questions that need to be answered. Simply measuring where the eyes go, for example, is a surprisingly elusive and difficult piece of data to record. It requires, for example, an analyst to define regions of interest for aggregating fixations that are near each other, and these regions will generally need to be defined with an understanding of task parameters. Rigorous eye movement data analysis also requires a thorough consideration of the error that is inherent in any eye tracking device (which generally exceeds a manufacturer's advertised mean error), and ideally a means of preventing and coping with this error when drawing conclusions from data.

What follows are a number of statistically significant results that were identified through extensive data exploration that was conducted with a specific interest in data that would help to reveal the *strategies* that participants used to conduct the task; that is, how participants recruited

the perceptual, cognitive, and motor processes to perform the task well. This section presents the data. Discussion of how the data reveals aspects of strategies will follow. The data are organized based on when the eyes move and where the gaze moves to, and with consideration of what information is likely to be picked up with each fixation. From here on, the eye movement analyses are almost exclusively with regards to events and locations for the tactical task because this subtask represents a complex *secondary* task that is designed to be executed somewhat in the background to a primary task, and thus presents an opportunity for exploring how perceptual, cognitive, and motor processes can be recruited and coordinated for a secondary task in a dual task situation.

### ***When the eyes move***

The eyes move to pick up data that are required to do the task, primarily to look at the blips that need to be classified on the tactical display, and to look at the tracking cursor on the tracking display. In general, each blip on the tactical display requires a fixation in order to determine its number and designation. Additional fixations are also likely required when there is no peripheral visual information in order to make tactical blips appear on the display, so that it can be determined when blips have appeared that will need additional subsequent fixations for classification.

The important events that occur around which eye movement times can be evaluated are as follows: A (black) blip appears on the tactical display (though the eyes must be on that display for it to actually be visible); a black blip changes color to red, green, or yellow and is ready to be classified; the gaze moves from tracking to tactical; the gaze moves from tactical to tracking; and the first of the two keystrokes required to classify a blip is recorded.

Figure 9 shows the time interval from when a blip changes color (and is ready for classification) to when the eyes move from the tracking to the tactical display, for the 80% of the blip color changes that occurred when the eyes were on the tracking display. Important trends are as follows: With peripheral visual information (filled plot symbols), the eyes moved much more quickly than with no peripheral visual information (unfilled plot symbols). With peripheral visual information, sound on or off made little or no difference. With no peripheral visual information (unfilled symbols), whether sound was on or off made a big difference. The time to move the eyes to the tactical display decreased across the three days, but these decreases are much greater when there is no peripheral visual information.

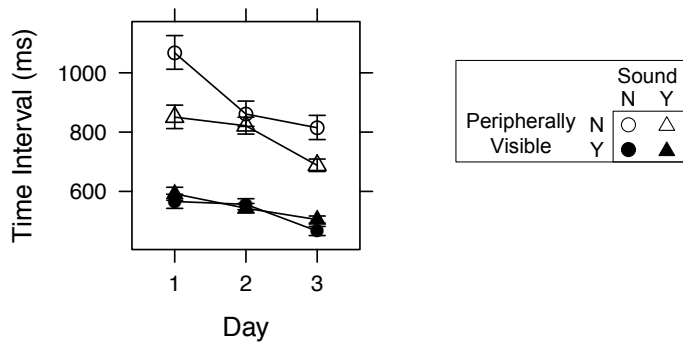


Figure 9. The time interval from when a blip changed color (and became ready to classify) to when the eyes moved from the tracking to the tactical display, for each of the four conditions, as a function of day.

Figure 10 shows the time interval from when the eyes arrived on the tactical display to when the eyes land on the blip that changed color, for each of the four conditions, as a function of the day, and separated by wave size. Important trends are as follows: The time to get the eyes to the blip that just changed color is faster for small waves than it is for large waves. The time is relatively constant across each condition for small waves, but improves across the days for large waves. The time is faster when peripheral visual information is available, but when it is not available, sound seems to help quite a bit; this can be seen in how the unfilled circular plots is clearly separated from the other three plots.

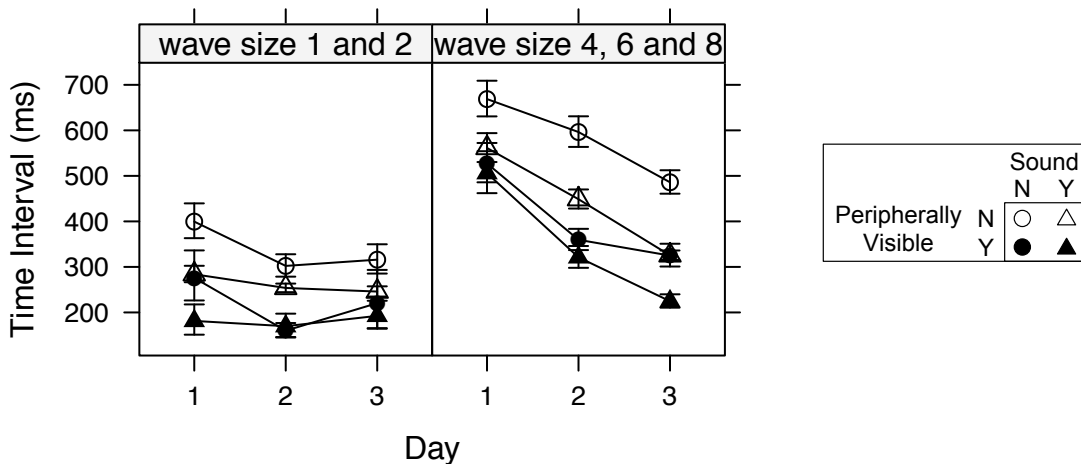
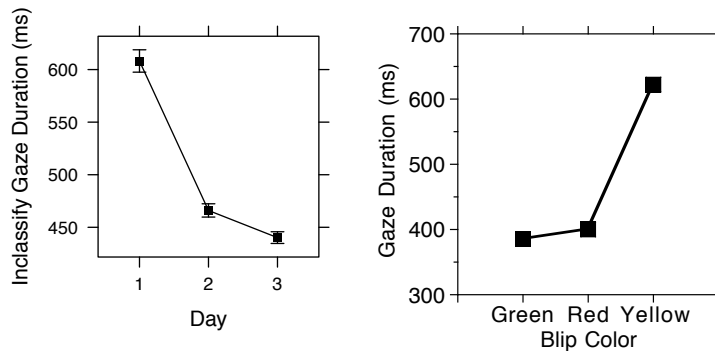


Figure 10. The time interval from when the eyes landed on the tactical display after a blip changed color, to when the eyes land on that blip, for each of the four conditions, as a function of day. The left frame shows small waves; the right, large.

Figure 11 (left) shows how much time was spent looking at a blip while it was in its colored state, ready to be classified. The time will have accumulated across one or more fixations, and is included for the 91% of the blips that received at least one fixation. The other 9% may have been classified without a direct fixation in the blip's colored state. The relevant trend here is that the time interval decreases from Day 1 to 2, at which point a minimum time appears to have been reached. Figure 11 (right frame) shows the same time interval but broken out by blip color and type, and for the same 91% of all blips. Important trends include that yellow blips are examined for 228 ms longer than green or red blips, and red blips are examined for the same amount time (just 14 ms longer) as green blips.



*Figure 11. The time spent looking at a blip while it was in its colored state (and thus ready for classification) as a function of day (in the left frame) and as a function of blip color (in the right frame).*

Figure 12 shows a pseudo-timeline of the eye movements and first keystroke for classifying a blip: First, the eyes move from tracking to tactical. Second, the eyes move to the target blip. Third, there are two overlapped subtasks, one to move the eyes back to tracking and the other to enter the first keystroke. Important trends are as follows: Intervals are shorter when peripheral visual information is available, and the speed increase is constant for the first two fixation events. The eyes return to the tracking task roughly a second before the first keystroke is entered; this can be seen in the vertical distance between the plot symbols for eyes-to-tracking and, directly above, for the keypress; this vertical distance represents an overlap between the tactical and tracking tasks.

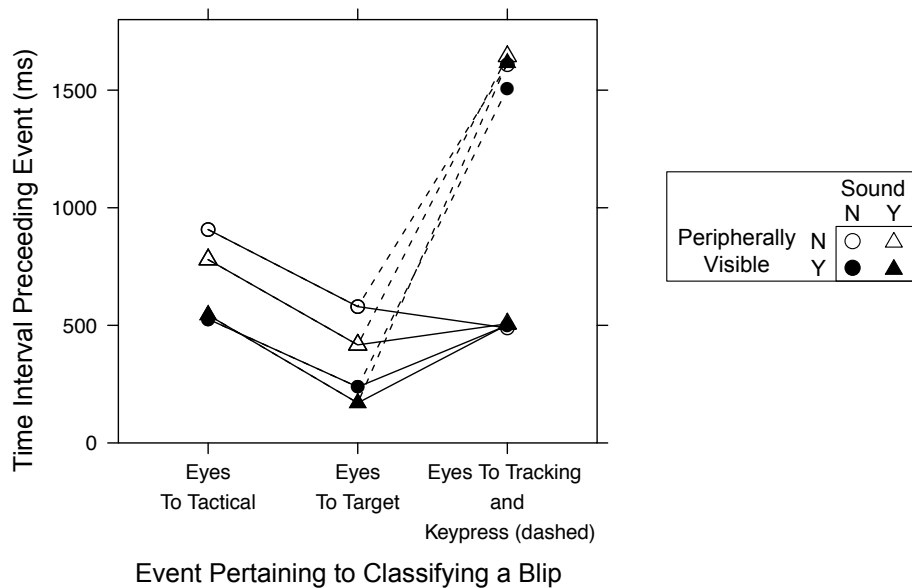


Figure 12. Mean performance times across the lifetime of a colored blip, from when a blip changes from black to colored, to when the eyes leave the blip and it is classified. The horizontal axis shows the subtasks: Get the eyes from the tracking to the tactical display; get the eyes onto the target blip to classify; and then in parallel, get the eyes back to the tracking display (solid lines) and also make the first keypress for classifying that blip (dashed lines).

### Where the eyes move to

The destinations of eye movements can reveal a great deal about the strategies that participants used to do the task. We will next consider fixation locations.

Figure 13 shows the mean number of fixations on each blip while it was black. Important trends are as follows: Participants tended to look at each blip one time while it was black; the one exception is for small waves in the no-peripheral no-sound condition, in which each blip tended to receive two fixations. Participants did not substantially modify their number of fixations on black blips across the three days.

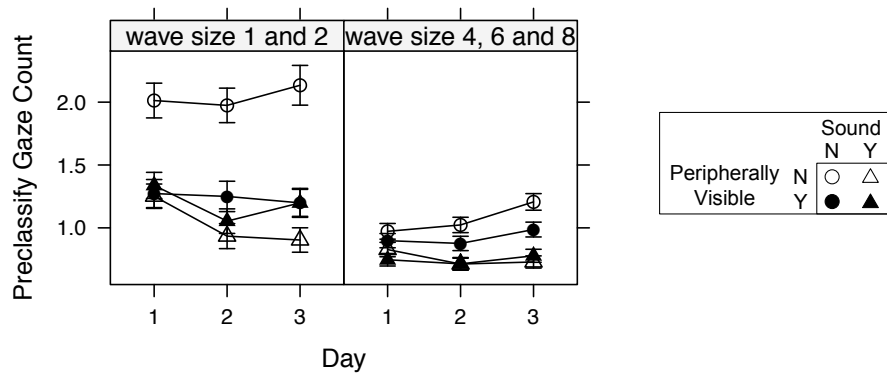


Figure 13. The mean number of fixations on black blips, for each of the four conditions, as a function of day. The left frame shows small waves; the right, large.

Figure 14. shows the average *time* spent looking at blips while they were black, for the 49% of the black blips that were fixated. The other 51% of the blips were not fixated while they were black. Important trends are as follows: When comparing black blip fixations across small and large waves (the left and right frames), the amount of time spent looking at black blips is relatively constant across day and condition for large waves (the right frame). It is in the small waves that distinct differences emerge. For small waves (in the left frame), the amount of time spent on black blips is highest for one no-peripheral-information condition (unfilled circles), and lowest for the other (unfilled triangles). The time increases across the three days for no-periphery no-sound (unfilled circles) and for yes-periphery yes-sound (filled triangles) but stays constant across the three days for the other two conditions.

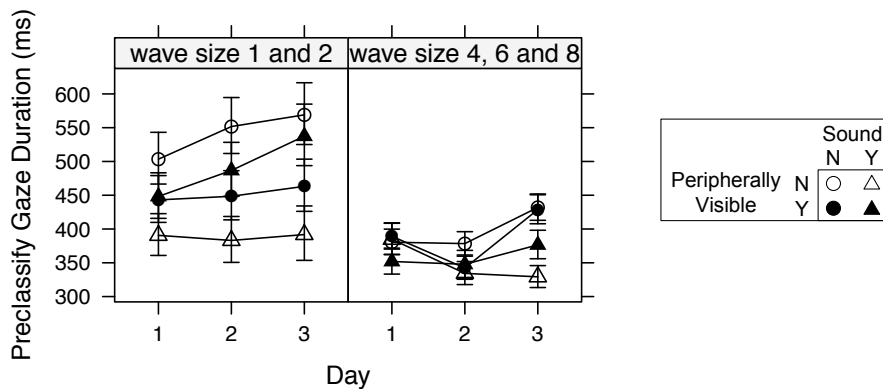


Figure 14. The mean time spent looking at black blips for the 49% of the blips that were fixated while they were black, for each of the four conditions, as a function of day. The left frame shows small waves; the right, large.

Colored blips (ready for classification) are another important eye movement destination. Overall *half* of the eye movements from the tracking to the tactical land directly on the blip that just changed color.

## **Discussion of Human Performance Data**

The human performance data provide insight into how people interleave two perceptual-motor tasks when there is substantial motivation to perform both tasks in parallel, quickly, and accurately.

Important general questions include:

1. How does a person fit two tasks together? What kind of general strategies or approaches appear to be followed? Such as, are the two tasks alternated, interleaved, or something else?
2. Can auditory information be used to enhance a visual display such that one task can be performed more effectively in parallel with another?
3. Can this data support or refute previous theories and models of dual task performance?
4. What have we learned for predicting human performance in future dual task scenarios?
5. What are the implications for designing effective dual task workstations?
6. What are the implications for training?

Questions specific to this experiment, the answers to which may shed some light on these more general questions, include:

1. How did people interleave the two tasks?
2. When engaged in the primary (tracking) task, how did participants use the auditory and visual information from the secondary (tactical) task to assist with that task?
3. How did performance improve across the three days? Did participants just get a little better with each component, or were there any fundamental shifts in how they did the task?

How did people do the task? The goal is to use summary stats to probe to reveal task strategies.

### ***Classification Time***

Classification performance improved across the three days in all four of the gaze/sound conditions. In three of the four conditions, performance appears to be approaching an asymptote, as is observed as people learn a task over time. In one condition, however, the peripheral-not-visible and sound-on condition (the unfilled triangles in Figure 2), performance does *not* appear to be approaching an asymptote. This trend suggests that, in this condition, participants did not just make small improvements to their strategy execution from Day 1 to Day 3, but rather that along the way they adopted a fundamentally different strategy. It appears as if participants may have learned how to use the sound to better monitor what was happening in tactical display when it was not visible.

It appears as if, based on the overall classification times shown in Figures 2 and 3, when peripheral visual information is available, sound did not help with the task and was essentially ignored. In fact, as shown in Figure 4, for small wave sizes, sound even hurt task performance a bit if peripheral visual information was available; this can be seen in the cases in which the solid triangles are well above the solid circles. The negatively accelerating slope in the no-peripheral yes-sound condition (unfilled triangles in Figure 2) is particularly accentuated in large waves (right panel of Figure 4), suggesting that participants especially learned how to use sound to improve task performance for large waves.

It is important to note that, overall, performance improved across the three days. We might conclude that Day 1 performance is that of a novice, and Day 3 performance is starting to approach that of an expert. It seems reasonable to assume, given that participants were given feedback and financial motivation to improve performance, and they did improve across the three days, that participants were using the feedback and motivation to maximize the potential performance based on how the task scenarios interacted with fundamental human perceptual, cognitive, and motor constraints. Hence, it is reasonable to conclude that the micro-level perceptual-motor decisions (such as eye movements) made on Day 3 are in general superior decisions to those made on Day 1, and that optimal macro and micro strategies are emerging.

Blip classification time as a function of blip color, shown in Figure 5, indicates that the yellow blips required 519 ms longer to classify than red or green blips. This time difference likely corresponds directly to the additional perceptual and cognitive time required to convert shape, location, direction, and speed into a hostile or neutral designation. Red blips are classified very slightly (130 ms) faster than green blips; this might result from red blips being slightly more conspicuous in the periphery, or to an implicit task demand to identify enemies before friends (though this was never discussed in the training).

### ***Tracking Error***

Tracking performance, shown in Figure 6, demonstrates that the experimental design, performance feedback, and payoff matrix all worked well together to motivate participants to (a) work hard to improve their performance for *both* the tracking tactical tasks across all three days and (b) find an appropriate balance between expending the information processing resources for the tracking and tactical tasks—this is demonstrated by participants performing *both* the tactical *and* tracking tasks slightly worse when peripheral visual information was not available. The degraded performance in the tactical task when peripheral information is not available cannot be explained away in that, in these conditions, participants simply spent more time and resources on the tracking task. Clearly both tasks were harder and performance suffered in both.

### **Eye Movement Data to Explore Fundamental Aspects of Dual Task Execution**

The eye movement data reveal many aspects of how participants accomplished the task. The first-pass data presented in Figure 7 and Table 1 suggest that participants were successfully

motivated to work very hard at the tactical and tracking tasks both individually and in parallel. That only 0.3% of all fixations were recorded as *not* falling on some component in the task display provides evidence that we have a good set of eye tracking data. That only 0.9% of the fixations are recorded as falling on one of the two status bars used to provide ongoing performance feedback suggest that participants were not unduly distracted by the status bar, that if participants monitored the bars they did so in the periphery, and that the slight increase in the complexity to the overall task resulting from the addition of the status bars is offset by the clear efforts that participants made to use this ongoing performance feedback to explore and optimize strategic decisions and perceptual-motor performance

The percentage of time spent with the gaze on the tracking versus the tactical task, as shown in Figure 8, vividly demonstrates that participants clearly perceived and treated blip classification as a *secondary* task, and that participants were clearly motivated to perform the primary task of tracking very accurately. The instructions, performance feedback, and payoff matrix worked as intended, and participants treated the tracking task as importantly as they would a mission critical subtask such as steering a vehicle.

## **Eye Movement Data To Explore Tactical Task Strategies**

### ***When the eyes move***

The time interval from when blips change color to when the eyes move to the tactical display, shown in Figure 9, reflects most of the same trends of the classification time, including a negatively accelerating downward slope for the no-peripheral yes-sound condition. In this condition in particular, it appears as if participants are learning to use sound to more quickly get their eyes to the tactical display when the contents of that display are not peripherally visible. These trends point to implications for predictive modeling and design of workstations that are intended to support multiple tasks in parallel, and in the training to use such workstations. First, it appears as if auditory alerts are *not* useful for enhancing visual alerts in a situation in which the visual information is readily available in the nearby visual periphery. Second, the extensive performance feedback tied to financial incentives across three days of training helped participants to explore a range of different task strategies and arrive at one on Day 3 that appears quite different from that used on Day 1 or Day 2. It appears as if, from Day 2 to 3, participants shifted to a deliberate decision to use sound to motivate eye movements to the tactical display.

The time from when the gaze arrives on the tactical display to when it gets to the target blip, shown in Figure 3, is relatively constant for the small waves across the three days. There are at important implications here: The revisions to the original NRL dual task appear to have succeeded in providing new independent measures (in this case wave size) that shed light on how people perform multiple tasks in parallel. Handling just one or two blips on the tactical display at a time requires a different strategy or strategy execution than when there are four or more blips. For small waves, it appears as if participants can pretty much just apply the well-trained *single*-task strategy that they are trained in before starting the dual task data collection. But for

larger waves, they need to develop skill over time for *which* blip they need to get their eyes to for classification. An speed improvement of 200 ms from Day 1 to Day 3 shows that, on average, participants learned how to cut remove an entire eye movement from the task execution. Practical implications include that workstation operators may be able to perform well with somewhat complex visual displays, but that they need to be perhaps routinely trained with those displays in their most complex arrangements. Implications for predictive modeling include novice performance should perhaps be modeled with more eye movements and perceptual consideration of extraneous visual information, whereas expert performance should perhaps be modeled with more direct and focused consideration of relevant visual information. Looking again at the data in Figure 10, it is also interesting that the auditory information appears to be somewhat useful for getting the eyes to the target. This may be a situation in which the spatialized component of the audio helped participants to anticipate the target location on the tactical display before it was visible; it could also be that the participant used the color information, which is effectively encoded into the alarm sound, to move the gaze specifically to the blip that just changed color.

The time spent looking at blips after they changed color (and were thus ready for classification) reveals some interesting details of how participants interleaved perceptual, cognitive, and motor processing. As shown in Figure 11 (left frame), participants substantially decreased the amount of time spent looking at a colored blip from Day 1 to 2, but then arrived at a somewhat optimal amount of time. The amount of time spent looking red or green blips as opposed to yellow blips (shown in Figure 11, right frame) can be compared to the data in Figure 5—blip classification time as a function of blip color. Blip classification time for yellow blips was 520 ms longer for yellow blips, when compared to red or green, but the eyes only spent 228 ms longer on the yellow blips. This suggests that participants required an additional 292 ms to translate the shape, location, speed, and direction of the yellow blips into a hostile or neutral designation, but that they did this perceptual-cognitive processing *after* they moved their eyes off of the blip, usually moving their eyes back to the tracking task. A predictive model of a complex visual task should account for holding the gaze on a visual object long enough to perceive object features, and the possibility of additional time spent processing those features after the eyes have moved on. With regards to the same amount of time spent looking at red and green blips, this *might* suggest that red “hostile” blips are classified 130 ms faster than green because of implied task demands.

The pseudo-timeline of the eye movements and keystroke associated with a blip while it is in its classifiable state, shown in Figure 15, summarize some aspects of how people did the task. We see that the overall performance improvement gained with sound and with peripheral visual information can be traced to the time spent getting the eyes first to the tactical display, and then to the target blip. After that, the time spent looking at a blip and getting the eyes back to the tracking task is pretty much the same across the four conditions. It is quite striking how long it took participants to start keying in a blip’s classification after the eyes left the blips to return to the tracking display—upwards of a second. This represents one second for every blip in which cognitive and manual motor processing is still underway for the tactical task while, for the same time interval, perceptual and cognitive processing is also underway for the tracking task. The

human perceptual and motor processing clearly introduce bottlenecks to the multitasking opportunity, but cognitive decisions to manage these resources are perhaps made with greater parallelism. People can do a good job with more than one task at a time provided that the same perceptual and motor processor is not required by more than one task at a time.

### ***Where the eyes move to***

We will next consider the locations where people looked, and how these fixation locations reveal task strategies and other aspects of human performance.

The number of fixations on black blips, shown in Figure 13, and the time spent on black blips, in Figure 14, suggest that participants actively tried to maintain some sort of situational awareness of what was happening on the tactical display, even for conditions in which the tactical display was visible in the periphery, and even while they spent 80% of their time on the tracking task. Twice the number of fixations on black blips for small waves in the no-peripheral no-sound condition is easy to understand—this situation alone would capture the fixations that participants had to make to the tactical display just to make the blips appear so that the participant could see if there were any blips necessary to classify. There are not additional fixations on black blips for large waves in this situation because, after an initial eye movement to make the blips appear, a large wave is much more likely to have a colored blip, and the best performance would be had by moving the eyes to a blip ready to classify rather than a black blip.

That the first keypress for classifying a blip occurs a little over a second after the eyes leave that blip is perhaps the most vivid and striking illustration of how the participants interleave perceptual, cognitive, and motor processing across the tactical and tracking tasks. The eyes routinely leave a colored blip and are back on the tracking task for an entire second before the first keystroke is pressed to classify that blip. Predictive cognitive models of complex dual task scenarios, and designs of multimodal workstations, need to account for expert performance in which visual processing for one task can be coordinated with manual motor task on another.

### **Modeling**

The goals of the project are to advance an understanding of (a) human performance when engaged in multitasking behavior that requires interaction with complex multimodal auditory and visual displays and (b) how to simulate and ultimately predict human performance in such situations by means of computational cognitive modeling. This section discusses the second major effort of the project—building computational cognitive that (a) explain *how* and *why* certain complex auditory and visual display arrangements improve performance while other arrangements do not and (b) contribute to the development of cognitive modeling principles and frameworks that will predict human performance with multimodal auditory-visual displays and in doing so inform the design of those displays.

## *The EPIC Cognitive Architecture*

Computational cognitive models were built using the EPIC (Executive Process-Interactive Control) cognitive architecture developed by Kieras and Meyer (1997) at the University of Michigan. Dr. Kieras continues to pour substantial research effort into improving both the computational power and theoretical veridicality of the architecture (such as evidenced by Kieras, 2009a; 2009b), and provided assistance with numerous technical aspects of this modeling effort.

The EPIC (Executive Process-Interactive Control) cognitive architecture (Kieras and Meyer, 1997) represents the fundamental human information processing—perception, cognition, motor, and memory—by encoding them into data structures and algorithms in the C++ computer programming language. The analyst—the person using the architecture to build a cognitive model—starts with the computer code for the architecture, writes some additional code, combines the two sets of code, and runs the model. The model generates a prediction of human performance.

Figure 15 shows an overview of the EPIC cognitive architecture, with all of its processors, memories, and the flow of control and data among the processors and memories. The diagram also shows the simulated task environment. The components that must be added by the analyst for each model are as follows.

- The cognitive strategy for accomplishing a task.
- The availability of visual features in visual zones, to represent a human's increased acuity vision near the point of gaze.
- Details of the task environment, such as when and where objects appear, and how the user interface responds to mouseclicks and keystrokes.

Once the analyst adds each of these components to the modeling framework and runs the program, EPIC generates as output:

- A prediction of the time required to execute the task.
- The mouseclicks and keystrokes in the task environment, to represent the human task execution.
- The simulated visual layout, including where the eyes are fixated during the task execution.
- A trace of the flow of information and control among EPIC's processors and memories.

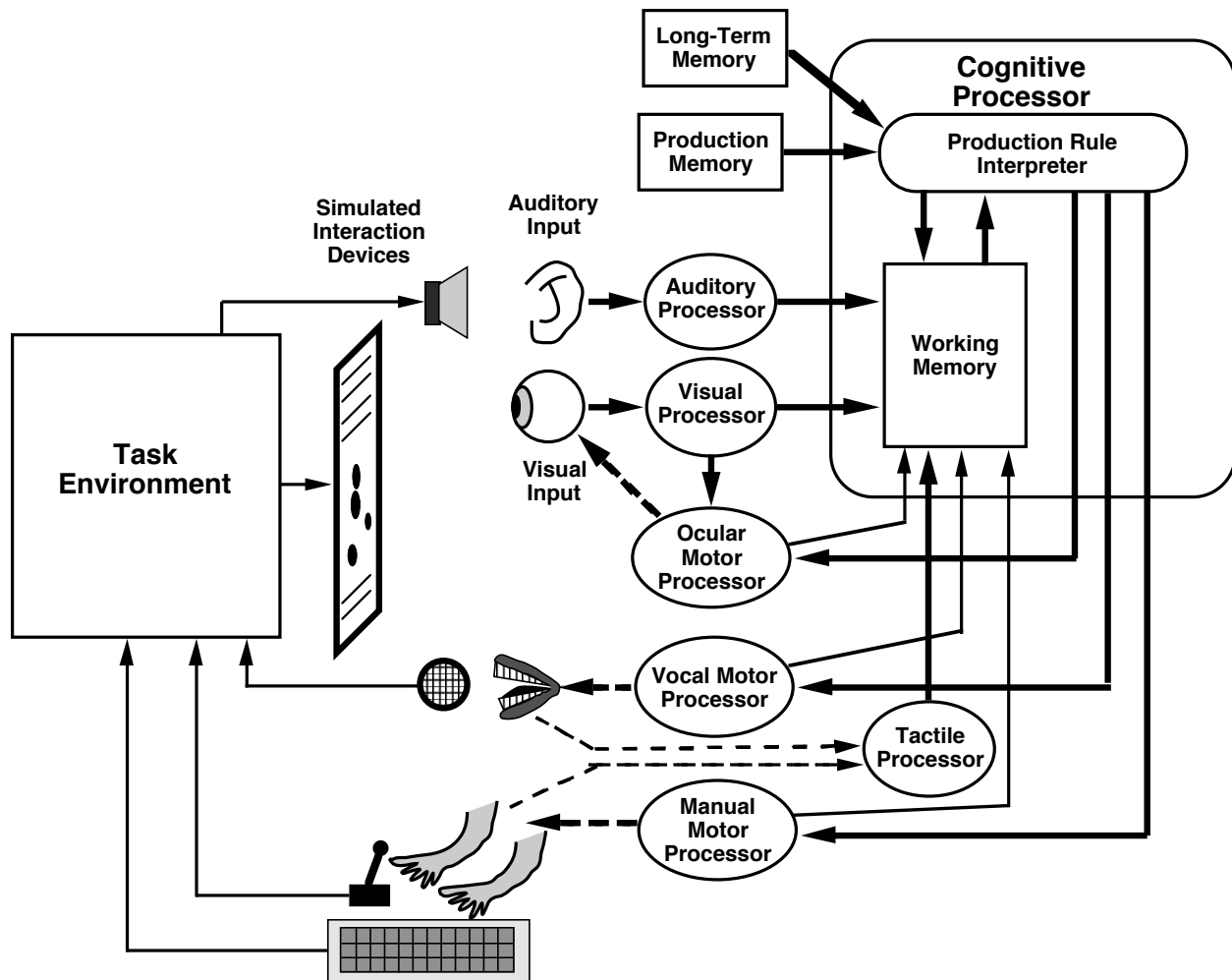


Figure 15. An overview of the EPIC cognitive architecture by Kieras and Meyer (1997). On the left, the simulated task environment, including the simulated input and output devices, and the flow of data among the devices. On the right, the various simulated sensory and motor organs, processors (ovals), memories (rectangles), and the flow of information among the various components.

EPIC is uniquely positioned, when compared to a number of other cognitive architectures, to simulate the perceptual, cognitive, and motor processing involved in complex multitasking task scenarios. Important aspects to the EPIC style and philosophy of cognitive modeling include: (a) a special emphasis on veridical representation of the directly-observable constraints involved in fundamental human information processing, such as the fact that the eyes move in order to bring detailed visual information into view, and the fact that visual features become increasingly available as they move closer to the point of gaze and (b) an underlying assumption that people develop and maintain rich and complex cognitive strategies for the gathering of perceptual information, and the coordinated use of this information with information already in working memory and built-in approaches to utilizing motor processing, to make decisions on how to

physically move body parts (including the eyes) to accomplish observable real-world tasks quickly and accurately.

As part of the EPIC philosophy and technical approach to cognitive modeling, information from all perceptual processors shown in Figure 15 (such as visual and auditory) are deposited into working memory as soon as the information becomes available, even if at the same moment; and commands can be sent to different motor processors (such as ocular motor and manual) entirely in parallel. The EPIC approach to predicting and explaining human performance is to hypothesize, build, and explore different sets of production rule strategies, and much of the theory for a particular task execution is embedded in the production rules. EPIC is uniquely positioned for this sort of theoretical exploration in that its cognitive processor, based on PPS (Parsimonious Production System), imposes very few constraints on the cyclical evaluation of a set of production rules, most importantly that *EPIC permits two production rules to fire in the same cycle*.

Being able to fire two rules in the same cycle is a particularly important theoretical foundation for exploring a wide range of possible strategies for simulating dual task performance. For example, this flexibility would permit a strategy in which one thread of a set of production rules is dedicated entirely to deciding where to point the eyes, and then pointing them there, with no consideration of what the hands are doing. Another thread could be dedicated to just moving the hands. The two threads could both proceed in parallel, communicating by depositing and checking information in working memory, with no performance bottleneck resulting from the cognitive processor only being able to serve one thread at a time. If an analyst would like to explore a model with such a bottleneck, EPIC is perfectly suited to do so in the writing of the production rules. The ability to fire two rules in parallel permits a wide range of models exploring threaded cognition to be built using the core architecture.

### ***Modeling Instrumentation***

In general, a fair amount of instrumentation is needed in the form of computer programming before an analyst can conduct the theoretical exploration afforded by a cognitive architecture, the first of which is often to build a simulation of the physical device and its characteristics within the simulation environment. This task was made easier by starting with code developed by Brock and McClimens at NRL for a previous simulation of the task developed at NRL for a previous version of EPIC, to which the PI added another 2,000 lines of code. These chores are not particularly important details for building theory, but an often-necessary part of the cognitive modeling enterprise.

Much more important theoretically for the application of EPIC to this project was that a temporal processor was built and added to the core EPIC architecture. The processor was built to address the problem that there is one task condition (no peripheral visual information or sound) in which the simulated human will need to move their eyes over to the the tactical display with no external task stimuli to motivate that eye movement. The decision would seem to need to rise up

spontaneously from within the simulated human routinely during the execution of the model. Scanning the modeling literature, the temporal processor built into the ACT-R cognitive architecture and described in Taatgen, Van Rijn, and Anderson (2007) appeared as a reasonably validated architectural component for generating periodic decisions from within the simulated human.

The Taatgen et al. ACT-R temporal processor was adapted to EPIC with as little modification as possible other than to translate the Lisp functions into C++ and adapt the processor commands to a syntax consistent with that used by EPIC. The temporal processor is used to maintain a single working memory item of the number of ticks that have elapsed since the last time that the processor was reset. The counter is reset to 0 and increments over time based on the equations and default parameters discussed in Taatgen, Van Rijn, and Anderson (2007), the net effect of which is that the counter initially increments quite quickly and then more slowly over time, with some noise and hence fluctuation in the pace of incrementing.

### ***New Models of the NRL Dual Task***

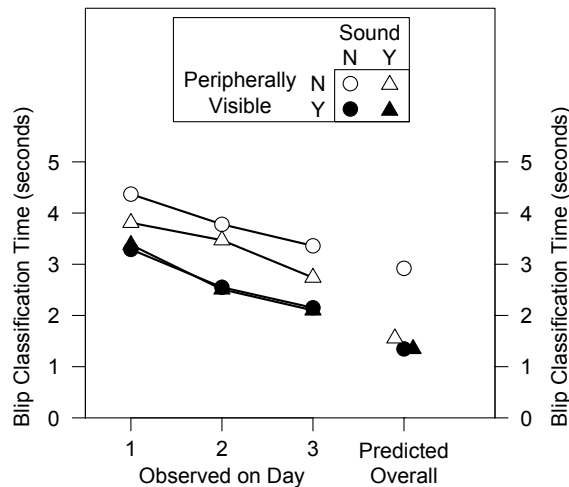
Two models have thus far been developed to explore the strategies that participants developed to decide how to best apply fundamental human information processing constraints to optimally perform the NRL dual task. These models were developed after analyzing the classification time data and in parallel with but somewhat separately from the analysis of the eye movement data. The models represent initial parsimonious baseline strategies for how a person might accomplish the task. The goal is to create two models that “bracket” the data, one model being a fastest-reasonable model—a lower bracket—and the other being a slowest-reasonable—an upper bracket. Following this bracketing heuristic (Gray & Boehm-Davis, 2000; Kieras & Meyer, 2000), it is generally expected that human performance will fall in between the two brackets, and that the observed performance will vary based on motivation and task demands. Following this approach, a final good-fitting model can then be developed by integrating aspects of the slowest- and fastest-reasonable models. Bracketing allows the analyst to examine the plausibility of model parameters early in the process, and provides guidance for closing in on the observed human data. The two models developed thus far are the Maximum-Interleave Model and the No-Interleave Model.

### ***Maximum-Interleave Model***

The Maximum-Interleave Model works very much as described above a few paragraphs earlier—one thread of production rules determines when and how to move the eyes between the two displays, and a second thread of production rules decides when to use the manual motor processor to either move the joystick or key-in a tactical classification. The production rules are intended to provide a lowest-possible bracket and be as fast as possible in every regard, including that they send the eyes straight to every blip as soon as a luminance change is detected. The strategy uses the sound to move the eyes to the tactical display when a blip color-change signal is heard, but not to help the eyes go to the blip. At all times and in all conditions, the change of a

blip from black to colored takes a very high priority over all other visual activities. The parallel threads in the production rules maximize overlapping of all ocular and manual motor activity.

Figures 16 and 17 show graphs that are similar to Figures X and X earlier, but this time the Maximum-Interleave Model's predictions are overlaid alongside or on top of the observed data. Figure X shows overall blip classification time for each of the four conditions, as a function of day, and then a single predicted classification time for each of the four conditions. Note that neither model is designed to predict the learning effects and strategy development that are observed across the three days, in part because EPIC is not well-suited for modeling learning. The models are intended, however, to explore strategic components that may be used across all three days, or perhaps novice strategies (as on Day 1) or more expert strategies (as on Day 3). Based on the predictions shown in Figure X, the model appears to do a decent job of predicting what might be the optimal classification times that participants might exhibit if they extended their practice out for more days. Hence, at first comparison, this appears to be a reasonable candidate for a fastest-possible lower-bracket model.



*Figure 16. Blip classification time observed across the three days, for each of the four conditions, and predicted by the Maximum-Interleave Model.*

Figure 17 shows the predicted and observed blip classification time as a function of wave size for each of the four conditions. We are now conduct a deeper, more rigorous, and more thorough comparison between our observed data and our model using the wider range of probe points that we established through our careful redesign of the NRL dual task. And we start to see that this may not be a reasonable candidate for a fastest-possible lower-bracket model.

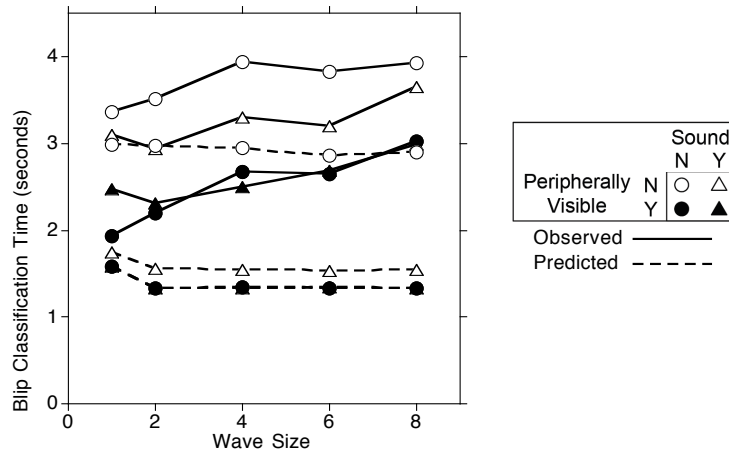


Figure 17. Blip classification time as a function of wave size, for each of the four conditions, observed (solid lines) and predicted (dashed lines) by the Maximum-Interleave Model.

In Figure 17, we see that the model does not do a good predicting many of aspects of the observed data. The most glaring problem is perhaps that, for many wave sizes and conditions, the model predicts classification times that are twice as fast as what was observed. Another problem is that the model does not account for the fact that blip classification time increases with wave size. Some of the base prediction times for waves of size 1 (1-blip waves) are within half a second of the observed, but as the waves get larger, the predictions stay constant whereas the observed increase with wave size. The model does not evidently take into consideration, or respond to, the increased overlapping of perceptual, cognitive, and motor processing required for larger waves.

It is perhaps interesting to note that the model predicts higher classification times for Wave Size 1. This is because all of the blips in these waves were yellow, whereas roughly half were yellow in all of the other wave sizes. So there should be roughly an additional 300 ms required to classify blips in Wave Size 1. Note that this trend can be seen in the human data, but only when sound is on, and not when sound is off. This suggests that, in the sound-on conditions, participants did not start determining a blip's classification until after the color-change sound played for that blip; whereas, in the sound-off condition, preclassification was conducted. In the sound-off conditions, the straight slope from Wave Size 1 to 2 to 4 suggests that a similar preclassification was conducted across these three wave sizes. The dip in the line for Wave Size 6 suggests that some kind of processing limit was reached, or a strategy shift occurred, between Wave Size 4 and 6.

We continue our model evaluation with Figure 18, which shows the number of visits per wave observed per day for each condition, and an overall Maximum-Interleave Model prediction, for each of the four conditions. The model overall makes substantially fewer glances over at the

tactical display. The model makes an average of one visit per blip whereas people clearly make more. It appears as if even a lower-bound fastest-possible model might need to account for some efforts to maintain some situational awareness of what is happening on the tactical display across all four conditions.

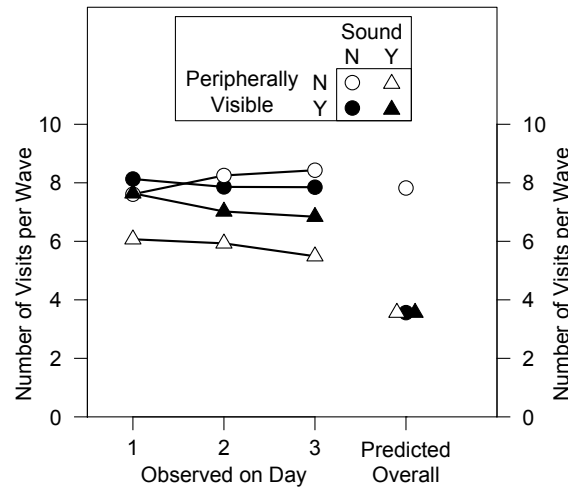


Figure 18. The mean number of visits per wave observed on each day, and predicted overall by the Maximum-Interleave Model, for each of the four conditions.

The Maximum-Interleave Model makes more glances per wave in the not-peripherally-visible and no-sound condition (the unfilled circle) because of all the extra glances made over to the tactical to see if anything needs classified. Though the model does make a pretty good predicting the observed value, this is largely due to the somewhat arbitrarily set delay in terms of how long the model waits between glances at the tactical display.

Figure 19 shows the average amount of time that participants spent on the tactical display every time that the eyes went over to that display. Clearly, participants moved their eyes back to the tracking task more quickly than the supposedly fastest-possible model. The model predicts substantially longer visits than are observed. The model holds the eyes on a blip up until its designation is determined, for roughly 0.5 s for green and red, and roughly 1 s for yellow. As seen in the observed eye movement data discussed earlier, it appears as if participants collect the visual features but continue determining a blip’s designation after moving the eyes back to the tracking task.

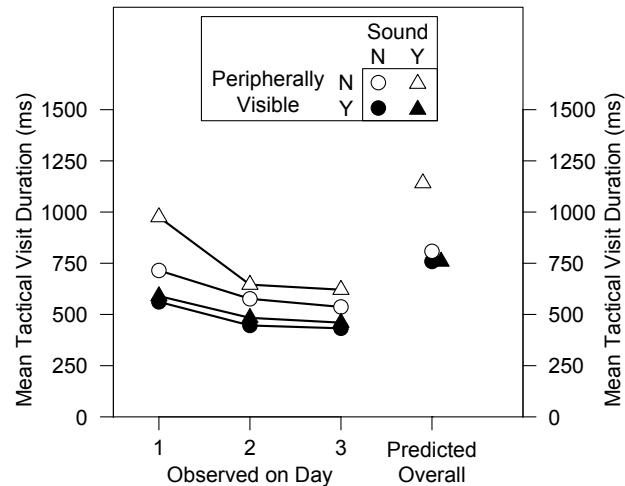


Figure 19. The mean duration of each gaze visit made to the tactical display observed on each day, and predicted overall by the Maximum-Interleave Model, for each of the four conditions.

The no-peripheral-information and no-sound prediction (the open triangle) is substantially higher than the two peripheral-on predictions because in this condition the model needs to make *two* fixations to classify each blip—the first to get the eyes over to the tactical display, and the second to get to the blip that changed color. It would seem quite odd, then, that the sound-off / peripheral-off model (the open circle) does not *also* have a higher value, as it would need two fixations to classify each blip as well. Well, it does, but that model *also* makes quite a few 450ms eye movements over to the tactical display just to see if anything needs classified, and nothing will need classified, and so the eyes go right back to tracking. And these short visits will bring the average down. It’s just a coincidence that the number of visits is so close to that of the two peripheral-on (filled symbol) predictions.

All told, the Maximum-Interleave Model does not appear to be a very good lower-bound model for the task. The model does a reasonable job providing a lower-bound estimation of overall classification time, but fails to account for trends observed across a number of different measures. It appears as if a good model of the task will need to account for a bit more of the complexity in the task and, very likely, the strategies used to do the task.

### ***No-Interleave Model***

The No-Interleave Model was developed as a candidate slowest-reasonable upper-bound model. The model was developed by conducting a hierarchical analysis of the task using the GOMS modeling language, which enforces a single-thread task execution, and then translating this model into the corresponding production rules as needed by EPIC. This exercise turns out to be a bit more difficult than an experienced modeler might expect (for example, every “If” decision

that is made in a GOMS model has an implied “Else go to the next step” that has to be explicitly represented with its own production rule) but the translation was accomplished. Figure 20 shows the GOMS model on which the No-Interleave Model is based.

```

MFG: Do dual task.
// (Steps 2 & 3 are considered in parallel in the PRs.)
Step 1. AG: Determine if a blip
        is ready to classify.
Step 2. Decide: If a blip is ready to classify,
        then AG: Do classification.
Step 3. Decide: If no blips are ready to classify,
        then AG: Do tracking.
Step 4. Go to Step 1.

Selection rule set for goal: Determine if a blip is
        ready to classify.
If tactical display is always visible
        and a blip has changed to high luminance,
        then store that blip as ready to classify
        and RGA.
If tactical display is sometimes visible
        and yes visible now (the gaze is on it)
        and a blip has changed to high luminance,
        then store that blip as ready to classify
        and RGA.
If tactical display is sometimes visible,
        and not visible now (the gaze is not on it)
        and sounds are available
        and a change-color sound has played,
        then there is a blip is ready to classify.
        AG: Check tactical. And stay in the SRFG.
If tactical display is sometimes visible,
        and not visible now (the gaze is not on it)
        and sounds are not available
        and you haven't checked tactical in a while,
        then AG: Check tactical. Stay in the SRFG.

MFG: Do tracking.
Step 1. Decide: If tactical display is sometimes
        visible and the gaze is not currently on
        tracking display,
        then look at tracking display.
Step 2. Decide: If eyes are not on tracking blip,
        then look at tracking blip.
Step 3. Decide: If tracking blip is red,
        then adjust joystick.
Step 4. RGA.

MFG: Check tactical.
// Should only be used in gaze contingent condition
// when the gaze is on tracking.
Step 1. Look at the tactical display.
Step 2. Wait for the blips to appear.
Step 3. RGA.

// Something is ready to classify.
MFG: Do tactical.
// Blips should always be visible when this is called.
Step 1. Look at a randomly selected blip.
Step 2. Look at a blip that is ready to classify.
Step 3. AG: Classify a blip.
Step 4. RGA.

MFG: Classify a blip.
Step 1. AG: Determine the blip's designation.
Step 2. AG: Key in the blip's classification.
// Step 3. Confirm blip changed to gray
// (and that the reward sound played?).
Step 4. RGA.

MFG: Determine a blip's designation.
Step 1. Decide: If blip is red, then blip is hostile.
Step 2. Decide: If blip is green, then blip is neutral.
Step 3. Decide: If blip is yellow,
        then study the blip to determine hostility.
Step 4. Store the blip's hostility value.
Step 5. Store the blip's number.
Step 6. RGA.

MFG: Key in a blip's classification.
Step 1. Key-in blip number.
Step 2. Key-in blip designation.
Step 3. RGA.

Key
MFG  Method for Goal
AG    Accomplish Goal
RGA   Return with Goal Accomplished
//    Comment

```

Figure 20. The GOMS model used to create the No-Interleave Model.

Figure 21 shows overall blip classification time for each of the four conditions, as a function of day, and then the No-Interleave Model's predicted classification time for each of the four conditions. The model does a very good job predicting the Day 3 clip classification time. Hence, at first comparison, out intended slowest-reasonable model might actually be a better candidate fastest-possible model.

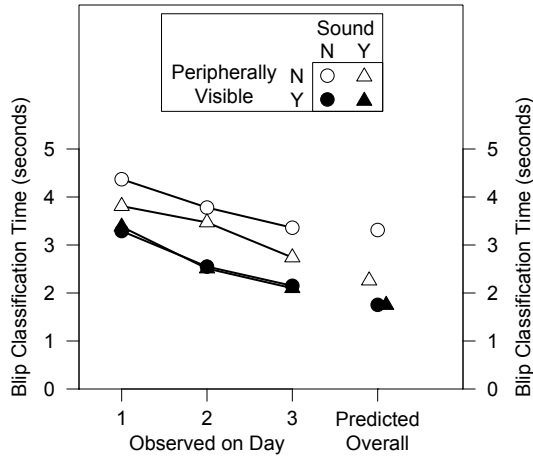


Figure 21. Blip classification time observed across the three days, for each of the four conditions, and predicted by the No-Interleave Model.

Figure 22 shows the predicted and observed blip classification time as a function of wave size for each of the four conditions. This provides a more rigorous comparison between the observed data than a single data point per condition, and we see that the No-Interleave Model does a better job than the Maximum-Interleave Model in this measure, especially for small waves, but it is certainly not a slowest reasonable model. The model also again fails to predict that blip classification time increases with wave size, though in other regards, the predictions are reasonably close to Day 3 performance. There is a small problem in that the spike for Wave Size 1, also seen in the Maximum-Interleave Model is even more pronounced in the No-Interleave Model though only slightly visible in the observed data. Perhaps this wave size in particular needs to incorporate a strategic decision to look at black blips and plan ahead for their eventual classification.

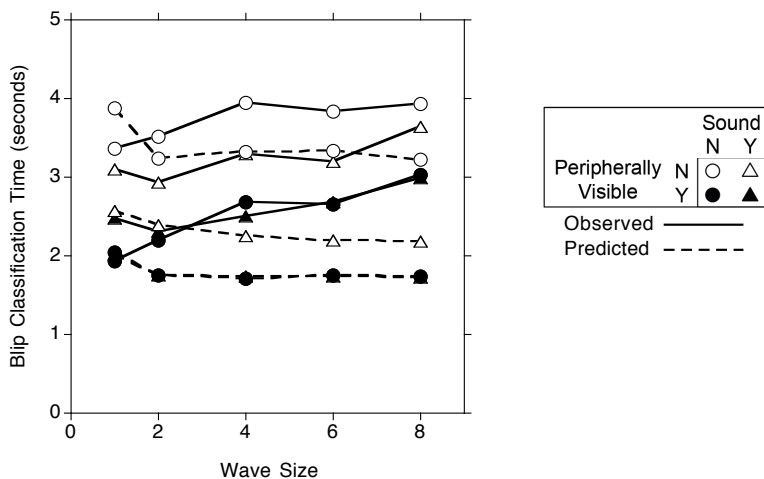


Figure 22. Blip classification time as a function of wave size, for each of the four conditions, observed (solid lines) and predicted (dashed lines) by the No-Interleave Model.

Figure 23 (left) shows the observed number of visits per wave per day, and an overall model prediction. As with the fastest-possible, there are still too few glances at the tactical. The model needs extra glances to the tactical. Figure X (right) shows the average amount of time that participants spent on the tactical display every time that the eyes go over to that display, by day, and an overall model prediction. The model is *way* too slow. Perhaps people use a two-phase perceptual process: First, a quick glance to get the features. Second, a slower compilation of the features into a designation, and only one blip can be compiled at a time; this compilation could even be done with PRs.

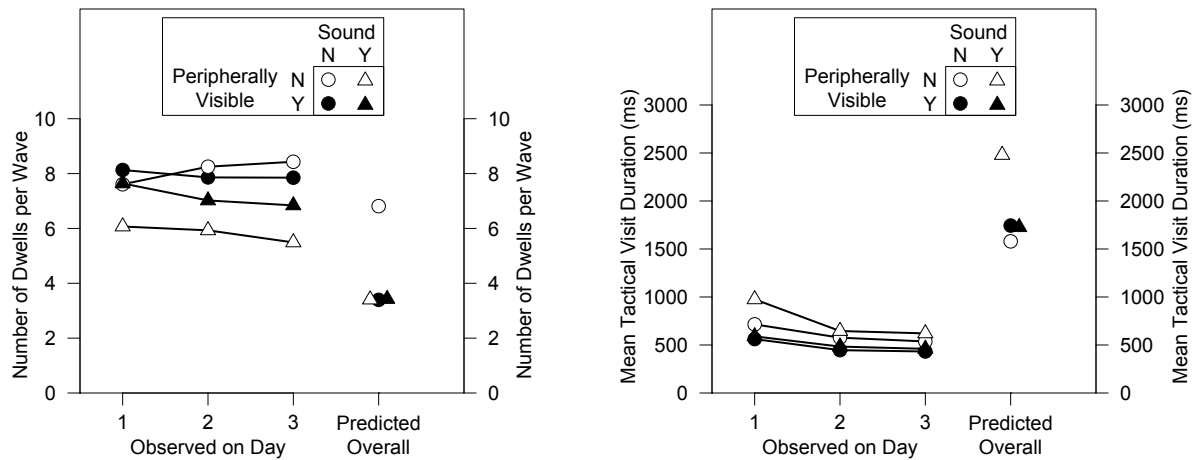


Figure 23. On the left, the mean number of visits per wave observed on each day, and predicted overall by the No-Interleave Model. On the right, the mean duration of each gaze visit to the tactical display observed on each day, and predicted overall by the No-Interleave Model.

## Modeling Discussion

It appears as if, overall, neither the Maximum-Interleave Model nor the No-Interleave Model accurately capture some important aspects of how people conduct this dual task, and how task performance is determined in part by the interaction between task parameters and fundamental human-performance constraints. This is perhaps most noticeable when there are trends that can be seen in the observed data that are not present in the predictions, such as how blip classification times rise slightly with wave size.

It does appear as if some aspects of each of the two models may be correct. The incredibly high tactical visit duration seen in the No-Interleave Model (Figure 23, right) suggests that there are clearly some processes occurring in parallel, such as getting the eyes off of the tactical display long (in terms of human cognition) before a blip's classification is fully determined. Some interleaving is clearly occurring. But it does appear as if some aspects of the the No-Interleave

model may be correct. This can be seen if we compare the No-Interleave Model predictions as a function of wave size against the observed classification times just from Day 3, as shown in Figure 24. Clearly, the No-Interleave Model needs to reduce the spike for waves with just one blip, and to account for how larger waves require increased perceptual-motor coordination. We are starting to approach a baseline model and can see strategic components that should likely be included and excluded from an accurate model, and from an explanation of how people produce the skill that they develop across the three days.

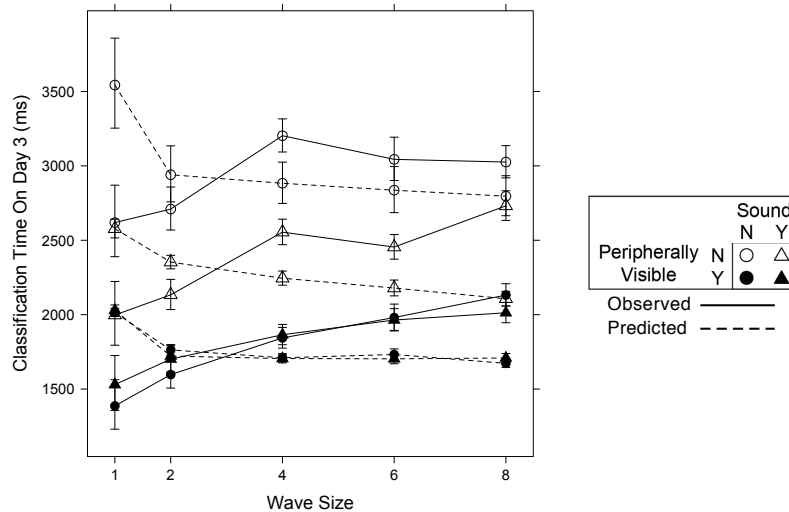


Figure 24. Blip classification time as a function of wave size, for each of the four conditions, observed on Day 3 only (solid lines) and predicted (dashed lines) by the No-Interleave Model.

## Conclusion

The NRL Dual Task as re-designed and re-instrumented in this project has created an opportunity to explore the skill that humans develop for executing complex multitask scenarios. The data, collected with experimental participants highly motivated to perform both tasks quickly and accurately, show a complex interleaving of eye movements with manual motor activity. The human data demonstrate a situation in which dual task workstation operators have ready access to task information in their near visual periphery and in which auditory alerts of the same information is not useful, but when that visual information is in effect moved further out of the visual periphery, the audio alerts become very useful for fast and accurate task performance. But it appears as if the cognitive strategies used to accomplish the task do not simply have a <perceptual cue> parameter in which a visual cue is immediately replaced with an auditory cue, but instead that people take some time and practice to evolve an effective strategy.

The modeling of the task reveals components that will need to be included in predictive dual task models, including the interleaving and overlapping of at least some visual perceptual and manual motor processing, even across competing subtasks. Parsimonious task strategies (that include

details such as only looking at visual cues when they are needed) may do an okay job at predicting overall task performance, but may not be able to account for performance degradation with increased task complexity, when the going gets rough, which for mission-critical Navy operations are probably the part of the task or interface. The important next step in this ongoing research project is to further develop the models to account for more complex interleaving of tasks as demonstrated by the human eye tracking data, and to better account for the situational awareness, or advance planning of blip classification, that participants appear to exhibit as evidenced by trends such as blip classification time as a function of wave size.

The project has successfully advanced an understanding of human performance when engaged in multitasking behavior that requires interaction with complex multimodal auditory and visual displays and how to simulate and ultimately predict human performance in such situations by means of computational cognitive modeling

## References

Ballas, J. A., Heitmeyer, C. L., & Perez, M. A. (1992). Evaluating two aspects of direct manipulation in advanced cockpits. *Proceedings of CHI'92: ACM Conference on Human Factors in Computing Systems*.

Brock, D., Ballas, J. A., Stroup, J. L., & McClimens, B. (2004). The design of mixed-use, virtual auditory displays: Recent findings with a dual-task paradigm. *Proceedings of ICAD 04, The Tenth Meeting of the International Conference on Auditory Display*, Sydney, Australia, July 6-9, 2004.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied*, 6(4), 322-335.

Hornof, A. J., & Halverson, T. (2002). Cleaning up systematic error in eye tracking data by using required fixation locations. *Behavior Research Methods, Instruments, and Computers*, 34(4), 592-604.

Kieras, D. (2009a). The persistent visual store as the locus of fixation memory in visual search tasks. Paper presented at the *Proceedings of ICCM 2009, The Ninth International Conference on Cognitive Modeling*, Manchester, UK.

Kieras, D. (2009b). Why EPIC was wrong about motor feature programming. Paper presented at the *Proceedings of ICCM 2009, The Ninth International Conference on Cognitive Modeling*, Manchester, UK.

Kieras, D. E., Ballas, J., & Meyer, D. E. (2001). *Computational Models for the Effects of Localized Sound Cuing in a Complex Dual Task. (EPIC Report No. 13)*. Ann Arbor, Michigan: University of Michigan, Department of Electrical Engineering and Computer Science.

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(4), 391-438.

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.

Taatgen, N., Van Rijn, H., Anderson, J. R. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review*, 114(3), 577-598.