EYE-MOVEMENTS DURING UNIT-TASK EXECUTION IN A COMPLEX PROBLEM-SOLVING SITUATION

Myeong-Ho Sohn, Scott A. Douglass, Mon-Chu Chen, & John R. Anderson
Department of Psychology, Carnegie-Mellon University, Pittsburgh, Pennsylvania

We have studied the performance of subjects as they acquired skill in the Georgia Tech Aegis Simulation Program (GT-ASP) with a particular focus on their eye movements. Our task analysis showed that the GT-ASP breaks down into the selection of unit tasks and the execution of these unit tasks. We focused on the Identification unit-task. Our results showed that most of the practice benefit in Identification came from increasing efficiency during cognitive process, in which people make inferences and decisions on the basis of the currently available information. We also analyzed eye fixations when people perform this unit-task. Participants showed different fixation patterns, depending on what portion of the unit-task was being executed. Fluency in a dynamic complex problem-solving seems to be achieved by efficiency in cognitive as well as perceptual processes.

GT-ASP

The purpose of the current study is to investigate critical aspects of skill acquisition and their relationship with eye movements when people learn complex dynamic tasks such as the Georgia Tech Aegis Simulation Program (GT-ASP, Hodge, Rothrock, Kirlik, Walker, Fisk, Phipps, & Gay, 1995). The GT-ASP is a tactical decision making computer game that simulates tasks of an anti-air warfare coordinator (AAWC) on board a US Navy battleship.

The radar screen (**Figure 1**) an AAWC faces consists of three major areas. First, the radarscope shows various aircraft and surface tracks within the current radius. Vectors emanating from the aircraft indicate speed and course. Second, there is a group of information boxes. The character readout (CRO) box provides information about a hooked aircraft (e.g., speed, altitude, bearing, course, etc.). The character type in (CTI) box is used when the AAWC wants to change speed, altitude, and course of friendly fighter jets. The message box displays information requested by AAWC, which is also presented in an auditory channel. Third, the menu panel shows the currently available function keys and their labels. Because menu has a hierarchical structure, the same key can map to several different functions depending on the context and the depth of the current task.

An AAWC is responsible for monitoring only the air traffic that appears on the radar screen by carrying out the following unit tasks. First, the AAWC is supposed to identify as many unknown aircraft as possible, identification (ID). Identification includes both primary intent (i.e., what are the intents of aircraft: hostile, friendly, etc.) and air type (i.e., what type of aircraft they are, helicopters: strikes, commercial airliner, etc.). Second, to be able to identify, the AAWC needs to collect information about the unknown aircraft as much as possible, information search. Some information is relayed through other officers or through friendly combat air patrols (CAPs). Third, therefore, the AAWC has to change altitude, speed, or course of a CAP to send it close to an unknown aircraft for information, CAP control. Fourth, once an aircraft turns out to be hostile, the AAWC has to issue military action commands (e.g., warnings, assignment or engagement of a missile) depending on the range of the aircraft from the ownship, military action. Fifth, to make all the above actions more efficient, the AAWC is allowed to control the display, such as to change its radius, center, or to put some supplementary range rings, display control.

Figure 1. GT-ASP display. A: information, CRO box, B: information, CTI box, C: information message, D: menu panel, E: radarscope, F: ball-tab.

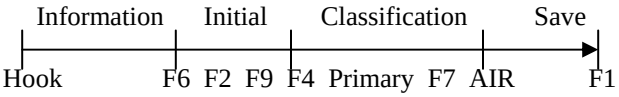
RESEARCH ISSUES

How the Unit-tasks Are Executed

The purpose of GT-ASP is to develop training methods that can be adapted to train AAWCs. Therefore, it is critical to determine what aspects of the task benefit from practice. GT-ASP is a complicated task consisting of multiple unit-tasks. Two issues are critical in GT-ASP as a representative of complex dynamic problem-solving tasks: unit-task selection and unit-task execution. Unit-task selection in GT-ASP involves such topics as aircraft-selection and transition between unit-tasks. Unit-task execution is about how each unit-task is performed in terms of accuracy and latency. Due to space limits in this report, we focused on unit-task execution and explored the relationship between key press results and eye movements.

Even in unit-task execution, we further narrowed down our current focus to learning of the Identification unit-task only. There are three reasons to focus only on Identification. First, practically, our preliminary analysis revealed that our participants spent great deal of time on Identification. Second, conceptually, the most important task of an AAWC is to identify all the unknown aircraft on the radar screen. Third, analysis of other unit-tasks revealed that characteristics of learning are consistent across different unit-tasks.

The Identification unit-task can be broken into four sub-intervals as shown below. During the Information interval, the



AAWC gathers information about the aircraft by requesting information from other officers or reading from CRO or message boxes. During the Initial interval, the AAWC types in a specific key sequence to initiate identification. During Classification, the AAWC selects among several options to assign a correct identification to the aircraft. During Save, the AAWC finishes the key sequence so that the action is complete and the aircraft's identification is updated.

The motivation for the above break-down of the Identification period is that each interval is differently loaded by cognitive and motoric demands. During the Information and Classification intervals, an AAWC has to make inferences about the relationship between the currently available information and the primary or air type identity of the aircraft. Therefore, these two intervals can be regarded as more cognitively loaded, in the sense that efficiency in higher level cognitive process may be the main source of learning in these intervals. In contrast, during the Initial and Save intervals, the AAWC simply executes a fixed sequence of key presses. Therefore, these two intervals seem to be more motorically loaded in the sense that most of learning in these intervals may be related to familiarity with the key locations. We examined whether learning in these intervals would be different or not.

How Eye-movements Are Related to Performance

One of prominent characteristics of skill learning is that people can learn to distinguish task-relevant information and task-irrelevant information (Haider & Frensch, 1999). Lee and Anderson (2000) showed that, when learning a complex problem-solving task such as air-traffic control, much inefficiency in performance can be attributed to suboptimal scanning of the screen. They looked at participants' eye movements and examined fixation time on several regions of the screen as a function of whether the information available in the region is relevant to performing the task. They found that as participants' performance become more efficient, the proportion of fixations on the relevant regions increased.

In GT-ASP also, not the entire task screen is relevant to task performance. The relative importance of each region may be different depending whether a unit-task is currently performed or not. For example, before a unit-task is selected, a participant may actively search the radarscope to find an aircraft to work on. Our focus in this report is eye movements during execution of the Identification unit-task.

Once the task is initiated, gathering information about the selected aircraft by looking at the aircraft itself or at information boxes may be more important. Also during performance of a unit-task, it is important to make sure the right menu key is selected. We categorized hooked-aircraft, information boxes, and menu panel as on-task regions. The rest of areas on which eye movements are fixated are called off-task regions.

We suggest that learning in GT-ASP should also be reflected in the pattern of attention distribution, which is defined as eye fixation in the current study. That is, people should learn to pay more attention to on-task regions relative to off-task regions with practice.

EXPERIMENT

To investigate the above research issues, we conducted an experiment, in which participants performed the GT-ASP task assuming the role of an AAWC while their eye movements were recorded. Performing the GT-ASP task requires relatively good mastery of the briefing materials which convey extensive amount of information including the purpose of the game, different rules about when to execute military actions, useful strategies, types of aircraft and their characteristics, relationship between specific information and the intents and type of aircraft, etc.. To encourage participants to study the material, two brief quizzes were conducted.

Method

Task and equipment. The GT-ASP task was performed with an IBM compatible computer. The GT-ASP system we adopted was developed by Hodge and his colleagues (Hodge et al., 1995) and modified by CHI system (Zachary, Ryder, & Hicinbothom, 1998).

The eye-tracking equipment was ETL-500 video-based, head-mounted eye-tracking system with magnetic-based head tracker from ISCAN®, Inc. The software for collecting and analyzing eye data consisted of the EPAL (Douglass, 1998) software suite that was internally developed in our lab to facilitate the development of eye-tracking experiments and their analyses.

Procedure. The experiment was conducted individually and consisted of five days. On Day 1, a participant studied the briefing materials with guidance of an experimenter. After studying the materials, the experimenter demonstrated the GT-ASP task for 30 minutes. On Day 2, the participant took the first quiz, and the experimenter went over the answers given and corrected the wrong answers. After the quiz, the participant practiced two sessions of part tasks, which were specifically designed to have people practice key press sequences for unit-tasks. On Day 3, the second quiz was conducted and feedback was given. After the quiz, the participant performed a short demonstration scenario for 10 minutes, followed by Scenario 1 for 20 minutes, and Scenario 2 for 30 minutes. On Day 4, Scenarios 3 and 4 were performed, and Scenarios 5 and 6 on Day 5. After each

scenario including the demonstration scenario, participants were given feedback about their timing and accuracy. The experiment on each day did not exceed more than two hours.

Participants. Fourteen undergraduate students, graduate students, and staff members of Carnegie-Mellon University were recruited with a monetary compensation of $20.00 a day.

Results and discussion

Task execution. Due to the composition of GT-ASP scenarios we used, true status of more than half of unknown aircraft was commercial airliner and the rest are various kinds of military aircraft. We examined task execution of each type of identification separately from Day 3, Day 4, and Day 5 performance. Each day, two scenarios were performed. For task execution, both accuracy and interval latency were analyzed. Accuracy was calculated as proportion of correct identifications among attempted identifications and was subjected to a one-way ANOVA with practice day (Day 3, Day 4, and Day 5) as a factor.

For latency, only the correct identifications were subjected to a two-way ANOVA with practice day and interval (Information, Initial, Classification, and Save) as variables. Preliminary three-way ANOVA with identification type (commercial vs. military), practice day, and interval revealed significant three-way interaction. As will be elaborated later, this interaction occurred because practice effects in intervals were different depending on the type identification. Therefore, we report two-way ANOVA with practice day and interval for each type of identification.

Mean accuracy of identification of commercial airliner was .99 and there was no main effect of practice, $p > .4$. Figure 2 shows the latency data for commercial airliner. As for latency, both main effects of interval and practice day were significant, $F(3,39) = 167.58$, $MSE = 1.78$, $p < .0001$, and $F(2,26) = 38.95$, $MSE = 1.04$, $p < .0001$, respectively. More important, the interaction between interval and practice day was significant, $F(6,78) = 14.90$, $MSE = .77$, $p < .0001$. Simple effect analysis revealed that practice effect was significant with every interval, $F(2,26) = 20.58$, $MSE = 2.93$, $p < .0001$, , $F(2,26) = 3.46$, $MSE = .18$, $p < .05$, , $F(2,26) = 63.1$, $MSE = .22$, $p < .0001$, and , $F(2,26) = 7.59$, $MSE = .032$, $p < .01$, for Information, Initial, Classification, and Save, respectively, although it appears that practice effect is greater with Information and Classification intervals.

The pattern of results with military aircraft identification is generally consistent with commercial airliner but somewhat different. Mean accuracy was .94 and there was no main effect of practice, $p > .2$. Figure 3 shows latency for military aircraft. For latency, both main effects of interval and practice day were significant, $F(3,39) = 117.90$, $MSE = 5.83$, $p < .0001$, and $F(2,26) = 7.23$, $MSE = 1.93$, $p < .01$, respectively. More important, the interaction between interval and practice day was significant, $F(6,78) = 5.08$, $MSE = 2.08$, $p < .01$.

Figure 3 is quite similar to Figure 2, except that the Identification interval increased in Day 5 compared to Day 4. This increase is due to the fact that, in Day 5, different kinds of military aircraft were introduced and they required quite different information request procedures than before. Among individual intervals, only the Initial interval did not show significant reduction with practice, $p > .9$. The other intervals showed significant changes with practice, $F(2.26) = 5.64$, $MSE = 7.32$, $p < .01$, , $F(2,26) = 5.59$, $MSE = .69$, $p < .01$, and $F(2,26) = 7.62$, $MSE = .06$, $p < .01$, for Information, Classification, and Save, respectively.

Although statistics do not exactly correspond, patterns of Figure 2 and Figure 3 are quite consistent. In both cases, practice effect seems to be concentrated more on Information and Classification intervals, in which cognitive processes of making inferences and decisions are emphasized. Instead, during Initial and Save intervals, which require typing of fixed key sequences, performance quickly reached asymptotic level and did not benefit from practice as much as other intervals.

Eye movements. We broke the screen (Figure 1) into six regions of interest for the purposes of analyzing where people are gazing while executing the Identification unit-task. The regions are information boxes, menu panels, radarscope, hooked-aircraft, middle-of-nowhere (MON), and off-screen. The hooked-aircraft is the aircraft that is currently selected for identification. MON is on-screen areas that are not classified into any meaningful regions. Eye movements can be directed to off-screen areas when participants looked at keyboard to press appropriate keys.

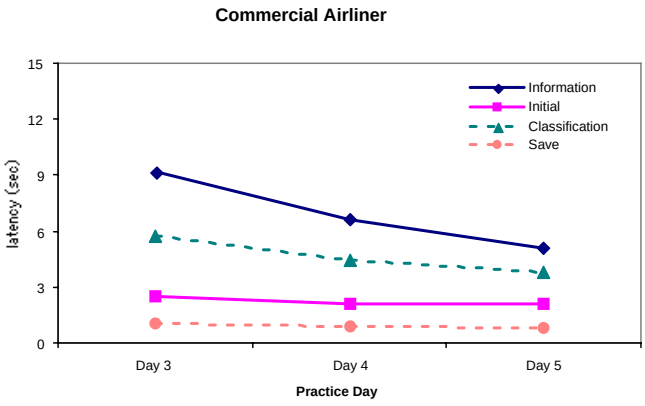


Figure 2. Mean interval latency for commercial airliner identification.

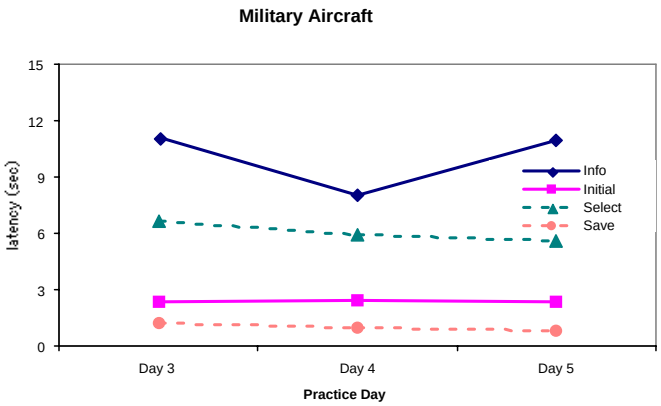


Figure 3. Mean interval latency for military aircraft identification

As noted earlier, we categorized the above regions into on-task and off-task regions. On-task regions are hooked-aircraft, menu panels, and information boxes. They are the regions an AAWC has to pay attention to at one point in time during the execution of the identification unit-task. Off-task regions are radarscope, MON, and off-screen. These are the regions the AAWC does not necessarily have to pay attention to while performing the unit task.

Fixation times on either on-task or off-task regions were subjected to three-way ANOVA with identification type, practice day, and region (on-task and off-task) as variables. Figure 4 shows mean fixation time on each region as a function of practice. Because three-way interaction involving identification type was not significant, we only report effects related to practice day and fixation regions. People fixated more on off-task than on on-task regions, $F(1,13) = 90.22$, $MSE = 11.63$, $p < .0001$, undoubtedly because they are a much greater portion of the visual field. Fixation time decreased with practice, $F(2,26) = 18.92$, $MSE = 4.89$, $p < .0001$. The interaction between practice and region was significant, $F(2,26) = 4.54$, $MSE = 10.28$, $p < .05$. Further analysis revealed that fixation time on off-task regions was significantly reduced with practice, $F(2,26) = 12.12$, $MSE = 5.11$, $p < .01$. However, fixation time on on-task region differed with practice only marginally, $p > .07$.

Eye fixation data show that our participants were sensitive to different types of regions on the screen. It turned out that people fixated on off-task region longer than on-task region, mainly because off-task region occupies greater proportion of the visual field. We also suspect that this is because of sub-optimal characteristics of our participants' performance. For example, the GT-ASP task heavily requires using function keys. However, even for a quite skillful typist, it is not easy to fully memorize the function key locations, which may have caused relatively long off-screen fixations.

More important, fixation on off-task regions reduced more dramatically with practice than on on-task regions. This interaction suggests that people are not only sensitive to different types of region but they quickly learn not to pay too much attention to irrelevant regions, consistent with previous studies (Haider & Frensch, 1999; Lee & Anderson, 2000).

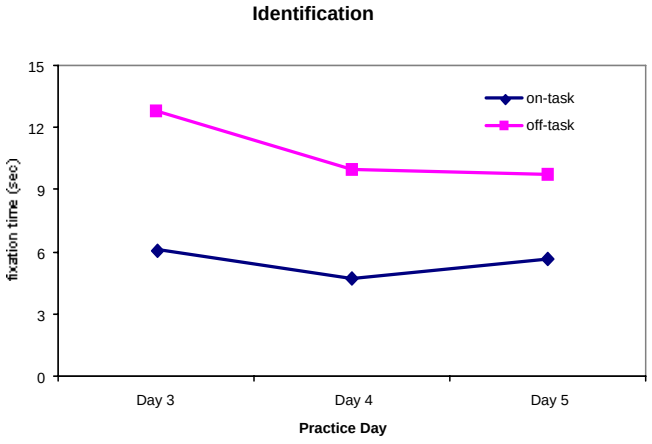


Figure 4. Mean fixation time on different regions during execution of an identification unit-task.

GENERAL DISCUSSION

The purpose of the current study is to examine critical aspects of skill acquisition in a dynamic, complex, problem-solving situation and to related them to eye movements. The results showed that, at least for some unit-tasks, speed up in task execution mostly came from increasing efficiency in cognitive processes to make inferences and decisions on the basis of the currently available information. Eye movement data showed that people are quite sensitive to where they should and should not look on screen. Moreover, they looked at irrelevant regions less and less as they practiced more.

The current results strongly suggest that it is critical to incorporate how people "think", where they are "looking at", and what they are "typing in", to fully understand the dynamic processes involved in GT-ASP. To this end, our next goal is to build a model that captures the cognitive, perceptual, and motoric demands of the task. We will use an extension of the ACT-R framework called ACT-R/PM (Byrne & Anderson, 1998) that is concerned with simulating the coordination of cognition, perception, and action.

ACKNOWLEDGEMENTS

This research is supported by a grant N00014-99-1-0097 from the office of Naval Research and is performed in collaboration with the Training Systems Division of the Naval Air Warfare Center.

REFERENCES

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson and C. Lebiere (Eds.), *The atomic component of thought*. Mahwah, NJ: Erlbaum.

Douglass, S. A. (1998). EPAL: Data collection and analysis software for eye-tracking experiments. Pittsburgh, PA: Carnegie-Mellon University.

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 172-190.

Hodge, K. A., Rothrock, L., Kirlik, A. C., Walker, N., Fisk, A. D., Phipps, D. A., & Gay, P. E. (1995). Trainings for tactical decision making under stress: Towards automatization of component skills. (HAPL-9501). Atlanta, GA: Georgia Institute of Technology, School of Psychology, Human Attention and Performance Laboratory.

Lee, F. J., & Anderson, J. R. (2000). Does learning of a complex task have to be complex? A case study in learning decomposition. Manuscript submitted for publication.

Zachary, W. W., Ryder, J. M., & Hicinbothom, J. H. (1998). Cognitive task analysis and modeling of decision making in complex environments. In J. A. Cannon-Bowers & E. Salas (Eds.), *Decision making under stress: Implications for individual and team training*. Washington DC: APA.