What the Eyes Reveal: Measuring the Cognitive Workload of Teams

Sandra P. Marshall

Department of Psychology, San Diego State University, San Diego, CA 92182 and
EyeTracking, Inc., 6475 Alvarado Rd., Suite 132, San Diego, CA 92120
smarshall@eyetracking.com

Abstract. This paper describes the measurement of cognitive workload using the Networked Evaluation System (NES). NES is a unique network of coordinated eye-tracking systems that allows monitoring of groups of decision makers working together in a single environment. Two implementations are described. The first is a military application with teams of officers working together on a simulated joint relief mission, and the second is a fatigue study with teams of individuals working together in a simulated lunar search and recovery mission.

Keywords: eye tracking, pupil dilation, cognitive workload, team assessment.

1 Introduction

Many activities require teams of individuals to work together productively over a sustained period of time. Sports teams exemplify this, with players relying on each other to maintain vigilance and alertness to changing circumstances of the game. Other types of teams also require vigilance and alertness to detail and often do so under life-threatening circumstances, such as medical teams, SWAT Teams, or First Responder Teams. Each team depends upon the good performance of all its members, and weaknesses in any one of them will change the way the team performs. For instance, sometimes one team member is overloaded and cannot perform his or her duties quickly enough so the entire team slows down; sometimes a team member loses sight of the situation and makes an error so the entire team needs to compensate; and sometimes the team member is fatigued and cannot function effectively so the other members need to assume more responsibility.

It is not always immediately evident when a team member is experiencing difficulty. All too often, the first indication is a major error that occurs when the team member reaches the critical point of being seriously impaired (either overloaded or fatigued). Early indication of such problems is clearly desirable but difficult to achieve.

This paper describes a networked system for evaluating cognitive workload and/or fatigue in team members as they perform their tasks. The system uses eyetracking data to create a non-intrusive method of workload evaluation. The paper has three parts: the first describes the system itself and how data are collected, the second describes assessing cognitive workload in teams of military officers as they determine

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2009, LNCS 5620, pp. 265–274, 2009.

how to share resources, and the third describes evaluation of performance and fatigue in a NASA study.

2 The Networked Evaluation System (NES)

The networked evaluation system, hereafter called NES, is a unique network of coordinated eye-tracking systems that allows monitoring of groups of decision makers working together in a single environment. Two versions have been developed and tested. One uses lightweight head-mounted optics and the other uses unobtrusive, remote eye-tracking cameras to monitor each individual's eyes. Each system then synthesizes data from all subjects in real time to enable the comparison of attention level and cognitive workload of all team members. The end product is a functional state-of-the-art eye-tracking network that can produce information in real time about all the team members collectively as well as individually.

The head-mounted NES utilizes the SR Research EyeLink II, which is a binocular eye tracking device that samples at 250 Hz. The remote NES utilizes the Tobii X120, which also is a binocular eye tracking device with a sampling rate of 120 Hz. Both eye trackers provide excellent data for eye position (horizontal and vertical pixels) and pupil size. In both configurations, each eyetracker is controlled by GazeTrace™ software from EyeTracking, Inc. which in turn produces the workload measure before feeding it to the central CWAD server (also produced by EyeTracking, Inc.) software for data synchronization and integration [4].

Both NES systems capture the same data: the location of each eye in terms of horizontal and vertical location on the display and the size of each pupil. A primary difference between the two systems is that the head-mounted system records data every 4 msec while the remote system records data every 8.33 msec. The data are transformed by the central processing unit of the NES into more conventional eyetracking metrics such as blinks, fixations, and measures of vergence. The pupil data also are transformed uniquely in the Index of Cognitive Activity (ICA), a patented metric which assess the level of cognitive workload experienced by an individual [2, 3]. Altogether, these metrics may then be combined to provide estimates of cognitive state [1, 4]. In particular, they are useful for examining whether an individual is overloaded, fatigued, or functioning normally.

All eyetracking systems in either NES are interconnected by a private computer network. GazeTrace software controls the eyetrackers, instructing them first to calibrate and then to start collecting data. In real time, the GazeTrace software computes the ICA workload measure and sends it to the CWAD server where it is synchronized into a database with eye and workload data from the other eyetrackers in the session.

3 Assessing Cognitive Workload Level

The research reported here was conducted under the Adaptive Architectures for Command and Control (A2C2) Research Program sponsored by the Office of Naval

Research. It was conducted at the Naval Postgraduate School in Monterey, CA. Researchers from San Diego State University and the Naval Postgraduate School collaborated to carry out the study. The primary purpose of the study was to examine how team members work together to overcome limitations, changes, or problems that arise during a mission. Three-person teams worked together in scenarios created within the Distributed Dynamic Decision-Making Simulation (DDD), a simulation system that allows multiple computers to interface and display coordinated screens for a defined environment.

The general focus of the study was the Expeditionary Strike Group (ESG), a relatively new military concept of organization that unites several different commands into a single unit that can move rapidly in response to problem situations. In the ESG simulation, the decision makers are given a mission, a set of predefined mission requirements, and information about assets that they control individually. Working together, they formulate plans of action and execute those plans to accomplish the overall mission objective. Examples of simulations in the DDD environment involve humanitarian assistance, disaster relief and maritime interdiction.

The simulation was designed to foster interactions among three specific positions in the ESG: Sea Combat Commander (SCC), Marine Expeditionary Unit (MEU), and Intelligence, Surveillance and Reconnaissance Commander or Coordinator (ISR). Seven three-person teams of officers participated in the study. Each team member was assigned a position (SCC, MEU, or ISR) which he or she maintained throughout the entire study. Each team participated in four two-hour sessions. The first two sessions were training sessions designed to familiarize the officers with the simulation software, the general outline of the mission, and the specifics of their own roles as decision makers. The third session, while primarily designed as a training session, was also a valuable source of data. During this session, the teams worked through two scenarios in the DDD simulation. The first scenario was a training scenario. The second was a new scenario designed to test the team's understanding of the situation. This same second scenario was then repeated during the fourth and final session under different test conditions. Thus, we had direct comparisons between the third and fourth sessions.

The fourth session was designed to be the major source of experimental data. When team members arrived for this session, they were told that many of their assets (e.g., helos, UAVs, etc.) used in the previous sessions were no longer available to them. Consequently, they faced the necessity to decide among themselves how to cover the tasks required in the mission under these reduced conditions. The most obvious way to do this was to combine their individual resources and to share tasks. Teams reached consensus about how to work together by discussing the previous scenarios they had seen and describing how they had utilized their individual assets. They then created a written plan to detail how they expected to work together and to share responsibilities. Finally, they repeated the same scenario that was used at the end of the third session and implemented their new plan of cooperation.

Thus, the research design allowed direct comparison of team behavior under two conditions: autonomous task performance and coordinated task performance. Under the first condition, each team member was free to select a task objective and to pursue it without undue deliberation or constraint by the actions of other team members. Under the second condition, team members were forced to communicate their plans to the other team members so that they could allocate the necessary resources in a timely fashion. Many mission objectives required actions to be taken by two or sometimes three team members simultaneously. If one team member did not deploy a specific asset in a timely fashion, the mission objective would not be achieved.

This design proved to be extremely valuable in examining how cognitive workload changed from one condition to the other. The underlying simulations were identical, thus the same events occurred at the same time and we could monitor how the teams responded to them.

For each run during the third and fourth experimental sessions, all team members were monitored using the unobtrusive networked eyetracking system. Data consisted of all eye movements of each team member, pupil size for both left and right eye measured at 120 Hz, and a video overlay of eye movements on the simulation screen. Each simulation run lasted 20-30 minutes.

Several unexpected problems were encountered during data collection. First, some team members assumed extreme positions to the left or right of the computer display, leaning heavily on one elbow as they looked at the screen. They were not viewable by the eyetracking cameras while they were doing so, and data were lost temporarily while they maintained this position. Second, a few of the officers were unable to read the very small print on the display and had to lean forward to within a few inches of the screen to read messages. The eyetracking cameras could not keep them in focus during these times and these data were also lost.

Examples of workload results are shown in the following figures. Workload was measured by the Index of Cognitive Activity (ICA), a metric based on changes in pupil dilation [2, 5]. The ICA is computed every 30 seconds to show how workload changed over time during the scenarios.

Figure 1 illustrates the difference in cumulative workload for two positions, SCC and ISR, on two different simulations, session 3 and session 4. Each graph shows the two scenarios for the SCC member of a team as well as the same two scenarios for the ISR member of the same team. Teams are identified by letter. For Teams B and D, ISR experienced higher workload than SCC throughout most of the scenario. The cumulative plots shown here begin to rise more steeply for ISR than SCC by the end of 5 minutes (10 observation points). It is interesting to note that the ISR Coordinators in Teams B and D experienced higher workload than the ISR Commander in Team G.

A key objective of the study was to understand how the workload of the various team members changed when they had reduced assets and were forced to coordinate their activities. It was expected that workload would rise as assets were reduced. Figure 2 shows the results for three teams under the two conditions. Surprisingly, some SCCs had *lower* workload under the reduced-asset condition. This unexpected result was explained during the team's follow-up discussion in which these officers volunteered that they had had difficulty keeping all the assets moving around efficiently under the full-asset condition. Thus, by reducing the number of assets they had to manage, they experienced lower workload even though they had to interact more with their team members.

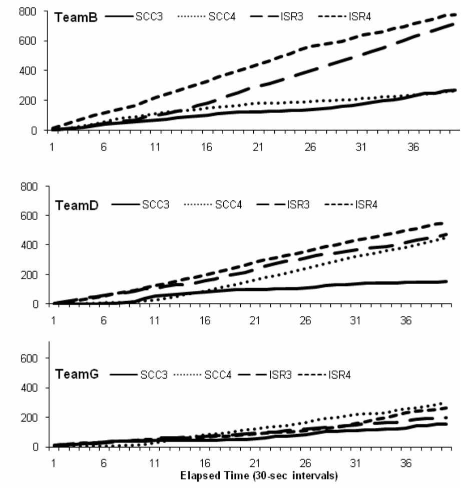


Fig. 1. Cumulative workload for three teams

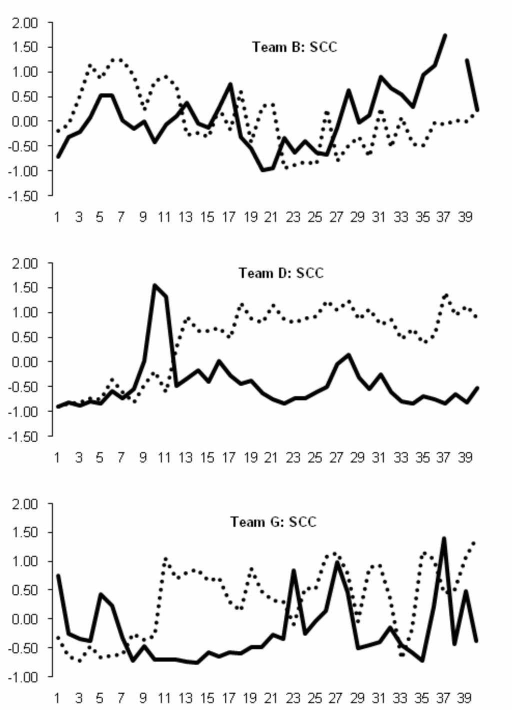


Fig. 2. Original Scenario versus Reduced Assets Scenario: SCC (original is solid line and reduced is dotted line)

4 Assessing Cognitive State

This study examined several different psychophysiological measures of task difficulty and subject fatigue. Only the eyetracking data are described here. The study was led by Dr. Judith Orasanu, NASA Ames Research Center and involved collaboration between several research groups at NASA Ames and EyeTracking, Inc. The task required 5 team members to work together to solve a series of lunar search and recovery problems. It also allowed each individual to score points, so that the individual was working not only for the good of the team but was also trying to maximize his or her own points. Multiple versions of the task were employed and were presented in the following order: Run1 (Moderate), Run2 (Difficult), Run3 (Difficult), Run4 (Moderate), Run5 (Difficult), Run6 (Moderate).

Eye data were recorded for three participants during six experimental runs, with each run lasting 75 minutes. The three participants were part of a larger 5-person team who were jointly tasked with manning 4 lunar vehicles plus the base station. We eyetracked two operators of lunar vehicles (code named RED and PURPLE) as well as the base operator (BLACK). They were tested six times over the course of a 24-hour period during which time they were sleep deprived.

Each participant worked in a separate small room and communicated with other team members through a common view of the lunar landscape on the computer display and through headsets. The head-mounted Networked Evaluation System (NES) was used in this study, with each participant undergoing a brief calibration prior to each run. The eye data and workload were then sent in real-time to the central processing CWAD server where all data were time stamped and synchronized for subsequent analysis.

A large quantity of data was collected. For each participant on the experimental task of interest, we have a total of 450 minutes of data (6 runs x 75 minutes), which is 27,000 seconds or 13,500,000 individual time points (taken every 4 msec). The data were subsequently reduced to 1-minute intervals by averaging the variables across successive 60 seconds. Seven eye-data metrics were created: Index of Cognitive Activity (ICA) for both eyes, blink rates for both eyes, fixation rates for both eyes, and vergence. All variables were transformed by the hyperbolic tangent function to produce values ranging from -1 to +1. These seven metrics have been employed successfully in the past to examine the cognitive states of individuals in diverse situations including solving math problems, driving a car (simulator), and performing laparoscopic surgery.

The six runs were performed by the subjects as three sets of two runs, with each set containing a moderate run and a difficult run. The first set occurred in the first few hours of the study when the subjects were not fatigued; the second set occurred under moderate levels of fatigue; and the third set occurred during the last few hours of the study when the subjects experienced severe levels of fatigue. A patented process based on linear discriminant function analysis was carried out for each subject in each of the three sets to determine whether the eye data were sufficient for predicting task difficulty.

The first analysis compared Run 1 with Run 2 to determine if the eye metrics are sufficient for distinguishing between the two levels of task difficulty. The linear discriminant function analysis (LDFA) determined the linear functions that best separated the 1-minute time intervals (75 per run) into two distinct categories for each participant. Classification rates were very high, with 85%, 96%, and 100% success rates for BLACK, RED, and PURPLE respectively. The eye metrics clearly distinguish between the initial moderate and difficult scenario. It is possible to estimate from a single minute of performance whether the individual was carrying out the easier task or the more difficult one.

The analysis of the middle set of runs (runs 3 and 4, made under moderate fatigue) also shows successful discrimination between the two levels of task difficulty, with success rates of 99%, 92%, and 90%. And, the analysis of the third set of runs (runs 5 and 6, made under extreme fatigue) shows similar but slightly lower success rates of 85%, 95%, and 86%. Looking across all three sets, it is evident that the eye metrics distinguish between the two levels of the scenario whether participants are alert (first set), moderately fatigued (second set), or very fatigued (third set). The lowest classification rate was 85%, meaning that the eye metrics correctly identified at least 85% of all minutes according to the scenario in which it occurred. It should be noted that *all* minutes of each scenario were included in these analyses, including initial minutes during which the scenarios presumably looked very similar to participants.

The first set of analyses described above looked at task difficulty while holding fatigue constant. Similar analyses look at whether we can distinguish between little fatigue and extreme fatigue while holding task difficulty constant. Two analyses parallel those described above.

The first fatigue analysis looked at levels of fatigue during two moderate runs. It compares the initial moderate run (Run1) with the final moderate run (Run6). The former was the run with least fatigue because it occurred first in the experimental study. The latter was presumably the run with the most fatigue because it occurred after participants had been sleep deprived for approximately 24 hours. LDFA classification rates for this analysis were 85%, 95%, and 86% for BLACK, RED, and PURPLE respectively.

The second fatigue analysis looked at levels of fatigue during two difficult runs. Once again, the first difficult run (Run2) was contrasted with the final difficult run (Run5). Classifications rates here were 100%, 95%, and 100%. The eye metrics were extremely effective in detecting the difference between low and high fatigue states with near perfect classification across the all 1-minute intervals for all three participants on the challenging difficult runs.

A final view of the data illustrates the importance of the Networked Evaluation System. The objective was to determine whether the participants experienced similar levels of workload during the tasks. For this analysis, it is critical that the data be synchronized so that we are comparing precisely the same time interval for every participant.

Figure 3 shows the left and right ICA for the three participants across all six runs. These figures show that the ICA varies considerably within each run, peaking at various times and dropping at other times. These figures also show a dramatic impact

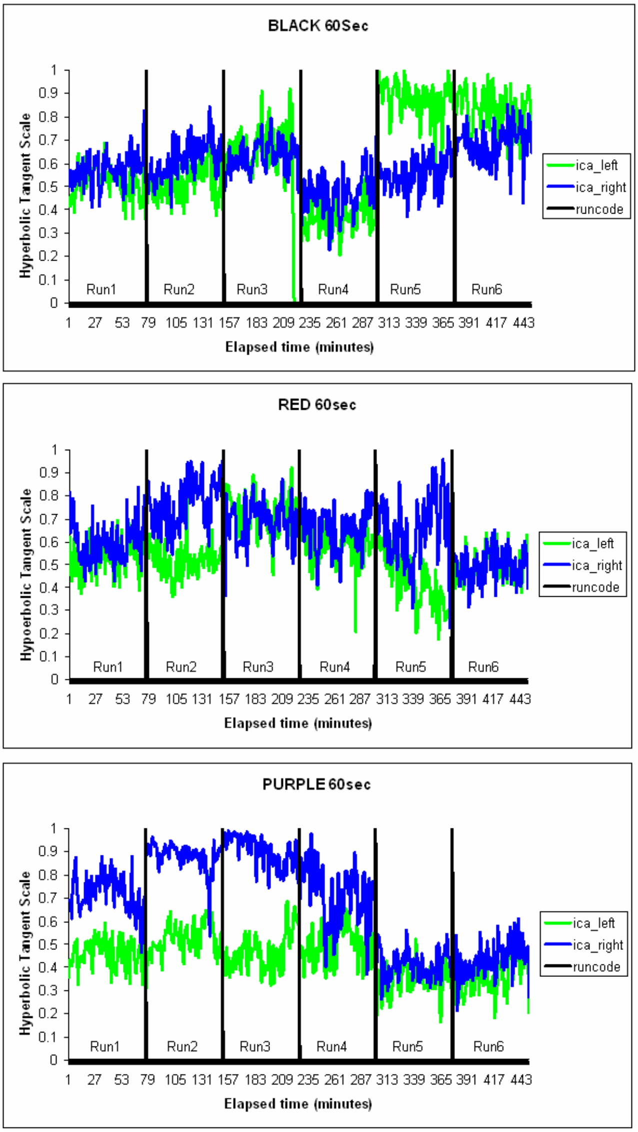


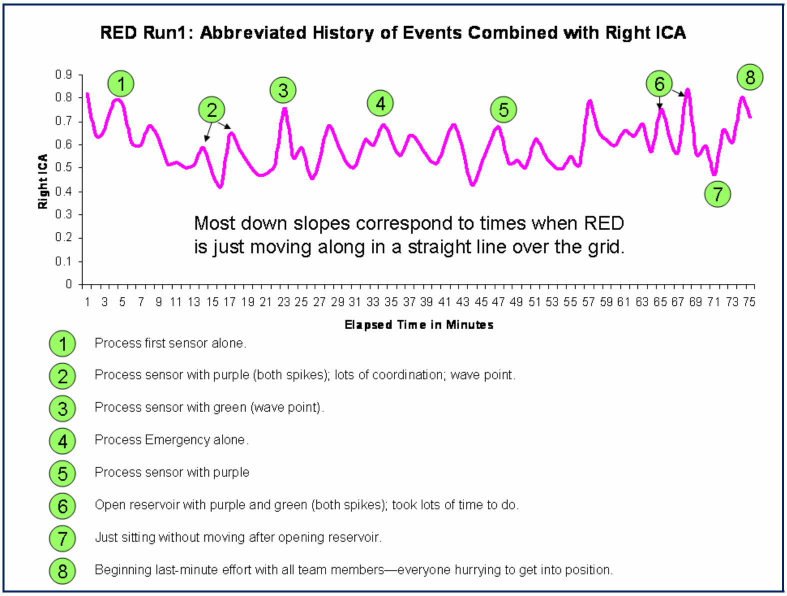
Fig. 3. Left and right eye ICA across the entire six runs

of fatigue on the ICA (see for examples the fourth panel for BLACK, the last panel for RED and the last two panels for PURPLE). And, there are sizable differences between left and right eyes for all three participants.

Each of the panels of Figure 3 could be expanded and mapped against the task details to determine what the participant was doing during periods of high and low workload. Figure 4 contains an annotated graph of the Right ICA for RED during Run1 (first panel of middle graph in Figure 3). This graph has a number of peaks and valleys. Eight peaks were selected for annotation using the screen video from the eyetracking session (audio was not available). For the most part, it is possible to determine from the video what the participant was doing, i.e., working with other team members to process a seismic monitor sensor, working alone to process other sensors, or navigating across the terrain.

We assumed that the many steps required to process a seismic monitor required considerable cognitive processing and that moving in a straight line across the grid required very little cognitive processing. And that is what we observed here. As Figure 4 shows, most of the spikes correspond to times that RED was processing sensors, either alone or in tandem with other team members. Most of the time when she was simply moving from one location to the other the Right ICA was descending. (Some spikes are not labeled because the video did not provide sufficient evidence alone to be sure of the task she was attempting.)

Thus, we are confident that the ICA can locate time periods that are more cognitively effortful for any participant. It should be kept in mind, however, that participants could have been processing information that is neither on the screen nor spoken by the team. In such instances, we might see active processing but not be able to trace its source.


RED Run1: Abbreviated History of Events Combined with Right ICA

Most down slopes correspond to times when RED is just moving along in a straight line over the grid.

1 Process first sensor alone.

2 Process sensor with purple (both spikes); lots of coordination; wave point.

3 Process sensor with green (wave point).

4 Process Emergency alone.

5 Process sensor with purple

6 Open reservoir with purple and green (both spikes); took lots of time to do.

7 Just sitting without moving after opening reservoir.

8 Beginning last-minute effort with all team members—everyone hurrying to get into position.

Fig. 4. Annotated History of Run1 for RED

5 Summary

The Networked Evaluation System worked very well in both environments described here. During both studies, it was possible to monitor the workload of the team members in real time as they performed their tasks. An obvious extension to NES would be to create some sort of alert that can inform either the team member directly or a supervisor when levels of workload are unacceptably high or low. Another option would be to have a direct link between NES and the operating system for the task. If the team member's workload exceeded a defined threshold, the system could reduce the demands on the team member directly without supervisor intervention.

Additional studies are now planned or underway in both environments and will provide more data about how NES can be implemented in real settings. Future studies will focus on how to time stamp automatically critical events for post hoc analyses and how to better capture and display task elements that correspond to high and low workload.

References

1. Marshall, S.: Identifying cognitive state from eye metrics. Aviation, Space, & Environmental Medicine 78(5), 165–175 (2007)
2. Marshall, S.: Measures of Attention and Cognitive Effort in Tactical Decision Making. In: Cook, M., Noyes, J., Masakowski, V. (eds.) Decision Making in Complex Environments, pp. 321–332. Ashgate Publishing, Aldershot (2007)
3. Marshall, S.P.: U.S. Patent No. 6,090,051. U.S. Patent & Trademark Office, Washington, DC (2000)
4. Marshall, S.P.: U.S. Patent No. 7,344,251. U.S. Patent & Trademark Office, Washington, DC (2008)
5. Weatherhead, J., Marshall, S.: From Disparate Sensors to a Unified Gauge: Bringing Them All Together. In: Proceedings of the 1st International Conference on Augmented Cognition, Las Vegas, NV, CD-ROM (2005)