ELSEVIER

Biological Psychology 52 (2000) 221–240

BIOLOGICAL PSYCHOLOGY

www.elsevier.com/locate/biopsycho

Combined eye activity measures accurately estimate changes in sustained visual task performance

Karl F. Van Orden [a,*], Tzyy-Ping Jung [b,c], Scott Makeig [a]

[a] *Medical Information Sciences and Operations Research Department, Naval Health Research Center, San Diego, CA, USA*

[b] *Institute for Neural Computation, University of California, San Diego, San Diego, CA, USA*

[c] *Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA, USA*

Received 12 March 1999; received in revised form 15 April 1999; accepted 9 November 1999

Abstract

Five concurrent eye activity measures were used to model fatigue-related changes in performance during a visual compensatory tracking task. Nine participants demonstrated considerable variations in performance level during two 53-min testing sessions in which continuous video-based eye activity measures were obtained. Using a trackball, participants were required to maneuver a target disk (destabilized by pseudorandom wind forces) within the center of an annulus on a CRT display. Mean tracking performance as a function of time across 18 sessions demonstrated a monotonic increase in error from 0 to 11 min, and a performance plateau thereafter. Individual performance fluctuated widely around this trend — with an average root mean square (RMS) error of 2.3 disk radii. For each participant, moving estimates of blink duration and frequency, fixation dwell time and frequency, and mean pupil diameter were analyzed using non-linear regression and artificial neural network techniques. Individual models were derived using eye and performance data from one session and cross-validated on data from a second session run on a different day. A general regression model (based only on fixation dwell time and frequency) trained on data from both sessions from all participants produced a correlation of estimated to actual tracking performance of $R = 0.68$ and an RMS error of 1.55 (S.D. = 0.26) disk radii. Individual non-linear regression models containing a general linear model term produced the cross-ses-

* Corresponding author. Present address: Space and Naval Warfare Systems Center, Code D44209, 54325 Patterson Road, San Diego, CA 92152-7150, USA.

E-mail address: vanorden@spawar.navy.mil (K.F. Van Orden)

0301-0511/00/$ - see front matter

PII: S0301-0511(00)00043-5

sion correlations of estimated to actual tracking performance of $R = 0.67$. Individualized neural network models derived from the data of both experimental sessions produced the lowest RMS error (mean = 1.23 disk radii, S.D. = 0.13) and highest correlation ($R = 0.82$) between eye activity-based estimates and actual tracking performance. Results suggest that information from multiple eye measures may be combined to produce accurate individualized real-time estimates of sub-minute scale performance changes during sustained tasks.

Keywords: Eye activity; Fatigue; Visual performance; Fixations; Blink activity; Pupil diameter

1. Introduction

Loss of alertness associated with fatigue and sleep pressure is a concern for any system that requires sustained monitoring by a human operator for efficient and safe operation. A method of objectively monitoring operators for signs of fatigue would be useful in transportation, security, and process control environments where lapses in attention can prove disastrous, and where continuous measurement of task performance is often not feasible. For example, fatigue has been shown to be a serious problem for automobile and truck operators (O'Hanlon, 1978; McDonald, 1984), and until Intelligent Highway Vehicle Systems (IHVS) are in place, there will be no direct measure of operator performance (such as lane drift) in real time. Furthermore, many tasks are of an intermittent nature (e.g. process control monitoring) making direct performance measurements unavailable for monitoring an operator's alertness level.

Recently, methods for objective alertness monitoring have been proposed based on measures of operator actions (Wierwille et al., 1994), electroencephalographic (EEG) activity (Makeig and Inlow, 1993; Makeig and Jung, 1996), and eye activity measures (Stern et al., 1994; Morris and Miller, 1996). Several eye activity parameters have been shown to be sensitive to time on task, which is linked indirectly to the onset of drowsiness in monotonous task environments. For example, using electro-oculographic (EOG) techniques, Stern et al. (1984, 1994) reported that blink duration and blink rate typically increase while blink amplitude decreases as a function of cumulative time on task. Other EOG studies have found that saccade frequencies and velocities decline as time on task increases (Schmidt et al., 1979; McGregor and Stern, 1996). In these studies, subjects performed monotonous tasks (e.g. vigilance tasks with infrequent events requiring responses) for periods of about 2 h, and performance and eye activity data were averaged over consecutive segments of 5–10 min.

Other recent studies have reported on the relation of eye activity to performance in simulated transportation environments. Morris and Miller (1996) demonstrated the sensitivity of EOG measures to fatigue in aircraft pilots during a 4.5-h flight consisting of eight 17.5-min flight maneuver segments and eight 10-min straight and level segments, presented alternately and separated by 1-min course adjustment segments. Ten participants were moderately sleep deprived, having been required to report to the laboratory at 01:00 h and remain awake until the experimental session

at 13:00 h. Mean performance error rate increased steadily from the beginning of the experiment. Integrated over the duration of the flight maneuver segments, blink amplitude, blink rate, long-closure rate (frequency of closures >200 ms), and saccade rate, in descending order, were the best predictors of flight segment performance, accounting for 64% of the variance. For the straight and level flight segments, long-closure rate and blink amplitude accounted for 65% of the error variance. Wierwille et al. (1994) have also reported that a measure of eyelid droop (percent time that the eyelid covers 80% or more of the pupil) may be a useful component of eye activity for the determination of drowsiness during simulated driving tasks.

Using video analysis techniques, other investigators have shown that pupil diameter decreases as a function of subjective drowsiness (Lowenstein and Lowenfeld, 1962; Yoss et al., 1970). Beatty (1982), however, found no change in tonic pupil diameter during a 48-min auditory vigilance task in which target sensitivity showed a small but significant decrease. Given the concurrent sensitivity of pupil diameter to changes in cognitive workload (Peavler, 1974), it remains to be determined whether pupil diameter is a useful measure of fatigue in dynamic visually-oriented tasks.

Previous research (Makeig and Inlow, 1993; Makeig and Jung, 1996) has demonstrated that performance on an auditory detection task shows minute-to-minute fluctuations as well as significant time-on-task decrements. Thus integrating continuous task performance and psychophysiological measures over several minutes or more neglects meaningful moment-to-moment and sub-minute performance variability. Furthermore, most variance in the electroencephalogram spectrum during continuous auditory performance was shown to be highly correlated to performance changes in a one-dimensional manner, suggesting that drowsiness is predominantly a one-dimensional state-change affecting performance on continuous tasks (Makeig and Jung, 1995).

By observing the relation of minute and sub-minute scale changes in performance to psychophysiological measures it may be possible to develop eye activity-based models that could be used in real-time alertness monitoring systems. In the present study, we first attempted to determine whether sub-minute scale fluctuations in several video-based eye activity measures were correlated with concurrent changes in visuomotor compensatory tracking performance. As previous research on eye movement and pupil measures has shown, sustained attention to a monotonous task may lead to performance fluctuations and eye activity changes that are predominantly the result of increases in drowsiness. Observed patterns of eye-activity measure changes observed during periods of relatively poor performance (e.g. prolonged blink durations) are generally consistent with drowsiness and sleep onset[1]. Other portions of behavioral variance, explained or unexplained, may reflect

[1] Unpublished data from our laboratory show that changes in the EEG spectrum (e.g. increases in θ band activity) during 15-min bouts of the same visual tracking task under sleep deprived conditions are tightly correlated with decrements in visual tracking task performance and are highly consistent with onset of drowsiness (Makeig and Jung, pers. commun.).

changes in subject strategy and/or level-of-effort not directly related to drowsiness. The data analysis approach we used adopts a general strategy first used to successfully characterize real-time electroencephalographic (EEG) power spectral changes associated with drowsiness (Makeig and Inlow, 1993; Makeig and Jung, 1996; Jung et al., 1997). Here, moving estimates of eye activity and compensatory tracking task performance were used to develop regression models and to train neural network models individually for each participant. The accuracy of these models in estimating changes in tracking performance during second sessions on the same participants was then assessed.

2. Method

2.1. Participants

Twenty-nine paid volunteers participated in the study (17 women and 14 men, mean age, 23.1 years).

2.2. Materials

A two-dimensional visual compensatory tracking task (Makeig and Jolley, 1996) was presented on a 12-inch black and white display and controlled by an 80386 computer. The target was a white annulus (7.0 mm in diameter to the inside edge with a thickness of 4.0 mm) positioned in the center of the display located ≈ 1 m from the participant. The tracking stimulus was a white disk 7.0 mm in diameter. The participant's task was to keep the disk in or near the center of the target annulus using a trackball whose movement supplied a restorative force to the disk in the direction of trackball motion. The position of the disk was a function of its previous position and velocity, plus the actions of three forces. The first was a buffeting force which continuously changed in magnitude and direction. The buffeting force was the sum of six sine waves at different amplitudes, incommensurate frequencies and phase angles. The phase angles for each frequency were chosen quasi-randomly by the program at the start of each session. These components had cycle lengths ranging from 1.9 to 19.0 s, with amplitudes proportional to their periods. The second force was simulated gravity acting on a centrally located unseen circular mound. The third force acting on the disk was directed by user input via the trackball. This force was proportional to, and oriented parallel to, the vector representing the trackball cursor movement since the previous time step. All forces, including a small momentum force simulating viscous drag, and a small central repelling force were integrated into a spring-mass-dashpot equation to calculate the disk's new position. The equations specifying the forces described above are available in Makeig and Jolley (1996). The compensatory tracking task software is also available from the authors. The gain applied to trackball movement was set to a comfortable level such that small trackball movements could compensate for disk movements produced by other forces sufficiently easily to produce

relatively good baseline performance by most participants (e.g. RMS error of about one disk diameter).

Visual activity was monitored using an Applied Sciences Laboratory SU4000 eye tracking system. The first 15 participants wore head-mounted optics (an infrared light source colinearly aligned with a camera mounted above reflective glass), which fed an eye image into the image processing system. The remaining 14 participants were positioned in a chin-and-head rest and eye activity was obtained with a remote optics system (near infrared, colinearly aligned) connected to the eye tracking system described above. The system calculated the location and diameter of the pupil reflection and the location of the corneal reflection at a sampling rate of 60 Hz. The eye tracking computer received synchronization signals from the computer running the visual tracking task for the alignment and merging of tracking performance and eye activity data sets. The testing room was quiet and dimly lit.

2.3. Design and procedure

Participants completed visual tracking sessions on two separate days. The first 15 participants completed only the visual tracking task. On the first day, these participants were told of the general features of the visual tracking task, completed necessary informed consent material, and then performed six 2-min visual tracking training trials. Previous data had shown this amount of training to be sufficient to train participants to asymptote on the task (Makeig and Jolley, 1996). Participants were then fitted with the eye-activity measurement hardware, and after a brief calibration procedure, began a 53-min visual tracking session. The entire first session lasted ≈ 100 min. Participants returned on a different day within 1 week of their first session to complete a second 53-min visual tracking session. The second session did not include further task training.

Subsequently, 15 more participants volunteered for an experiment involving two additional sessions to measure performance on an auditory detection task, plus a separate training session for training and orientation for both tasks. EEG and EOG measures were recorded to assess the similarities and differences in brain activity as a function of performance in the two tasks — those data are not reported here. Participants completed consent documentation and training on the visual tracking task on the first day. In subsequent sessions, which occurred on different days, participants were first fitted with electrodes for the EEG/EOG channels, completed calibration procedures, and then performed either the visual tracking or the auditory detection task for 53-min. These sessions, including set-up, calibration and testing, lasted 2 h. All participants were tested at mid-day or in the afternoon.

2.4. Scoring

For every participant, the mean distance of the disk from the center of the target ring, was computed in a 1-min square-weighted (boxcar) window that was moved through the data in 10-s steps. This smoothing was conducted to eliminate fluctuations in disk position produced by the participants responses (via the

trackball) to perturbations caused by the pseudorandom winds (sum of sinusoids forcing function) acting on the target disk. The smoothed data more accurately reflected local estimates of tracking error. The data were transformed using a squashing function:

$$2\Big/\sqrt{\pi}\int_{0\to x} e^{-t^2}\,dt$$

to de-emphasize fluctuations in target disk position at the extreme low end of the performance range, where target distance from the annulus could vary widely due to the influence of the forcing function in the absence of participant input (e.g. when the subject became drowsy and ceased performing the task). The upper asymptote of the squashing function was set to the experimentally-determined mean distance of the disk from the screen center without subject input (9.4 disk radii).

Fig. 1 presents one participant's compensatory tracking and trackball movement data during a 53-min session. Tracking error (the mean distance of the disk from the target) was small in the early stages of the session, but increased dramatically

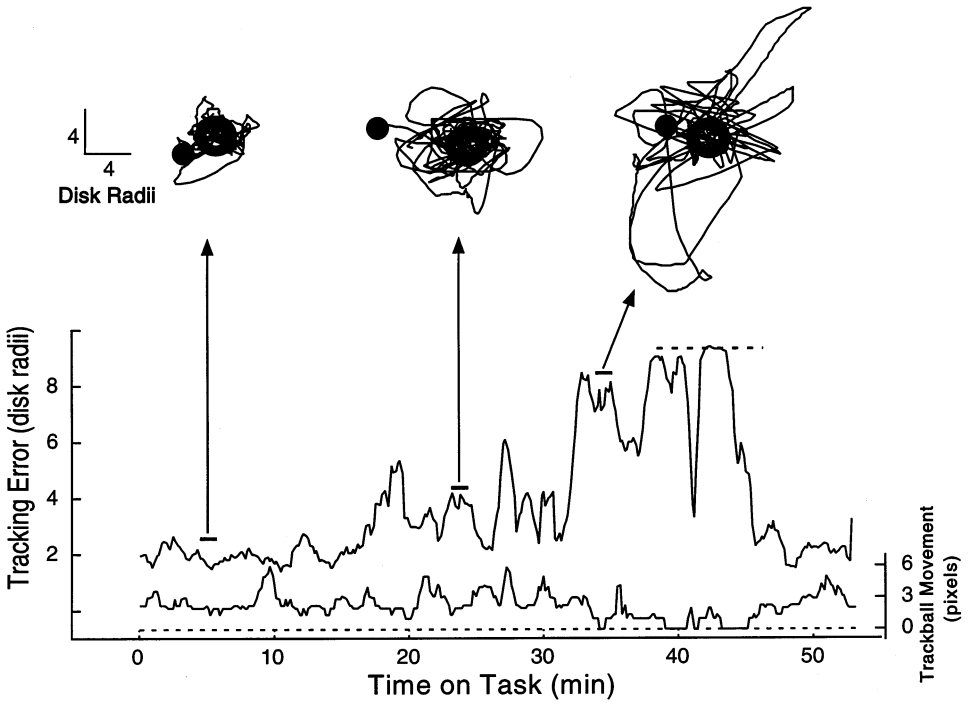

Fig. 1. Lower plot, upper trace: Single session compensatory tracking data from one participant. Tracking error in disk radii is plotted as a function of time in minutes. Dotted line at 9.4 disk radii near minute 44 represents the experimentally-derived mean upper-bound tracking error associated with no input from the participant. Lower trace: Relative effort (in trackball movement) applied to trackball by the participant during the session. Upper plots: Disk position relative to the target annulus for the 1-min period indicated on the center plot.

between 30 and 45 min. Inserted within Fig. 1 are plots of the disk trajectory relative to the center target annulus for 1-min periods during early (good performance), middle (fair performance), and late (poor performance) phases of the session. The upper dashed line shows the maximum error rate in the task (9.4 disk radii), which was reached during two 1- to 2-min periods when the participant occasionally stopped responding altogether (minutes 40 and 43). Notice that the performance record included frequent fluctuations on a circa minute scale, and did not change monotonically with cumulative time on task. Tracking error generally rose when response input declined.

Generally, good performance was characterized by participants keeping the disk at a mean distance of between 1.5 and 2.0 disk radii from the target annulus. During periods of poor or absent performance, tracking error increased to as many as 9.4 radii from the target. Because our ultimate goal was to test performance estimates produced by regression models and artificial neural networks trained on data obtained in a different session from the testing session, only participants who produced periods of poor performance in *both* sessions could provide the variability in performance needed to develop and test suitable tracking performance estimation models. Most participants maintained relatively good tracking performance in both sessions. Participants selected for further analysis were those who demonstrated mean tracking errors between 2.0 and 5.4 disk radii with a range of at least 2.0 disk radii (i.e. from 10th to 70th percentiles) in both sessions. Based on these criteria, nine participants (four females and five males, mean age of 23.1 years) were selected for further modeling and cross-session testing.

Blinks were extracted from the raw eye activity data by identifying partial eye closures as moments when pupil diameter was 35% or less of its mean value during a 240-s baseline period at the outset of the trial. Blinks were defined to be partial or complete eye closures lasting a minimum of 83.3 ms. Setting a threshold of pupil diameter for determination of blinks was required to detect 'near closure' produced by severe eyelid droop. The 83-ms minimum-closure duration criterion prevented brief signal losses from being counted as blinks. Blink duration was defined as the time interval between blink onset and the return of pupil diameter to greater than 35% of baseline. Having identified blink occurrences and durations, we next calculated moving estimates of total blink duration and blink frequency using a 1-min square-window moved through the data in 10-s steps.

Point-of-regard (POR) data were used to calculate spatial locations and dwell times of eye fixations using a standard space-by-time boundary algorithm provided by the eye tracking system manufacturer. This algorithm derived fixations by first finding six successive x and y POR data points with a S.D. of $< 0.5°$ of visual angle. Once the beginning of a fixation point had been established, subsequent POR points were considered as part of the fixation (and contributed to the calculated fixation dwell time and x/y location) if they fell within 1° of the current fixation point. PORs could deviate from (and contribute to the calculation of) the current fixation point by as much as 1.5° provided that at least one of two subsequent points fell within 1° of the fixation point, and that the mean of the most recent three PORs fell within 1° of the current fixation point. PORs falling beyond the 1.5°

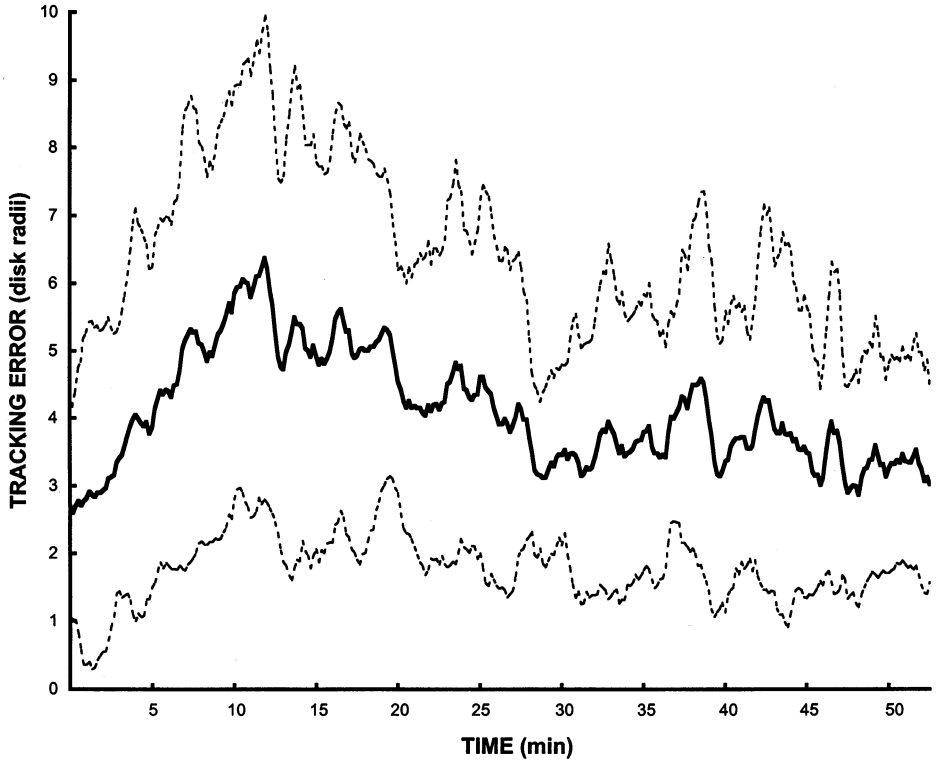

Fig. 2. Mean tracking performance (in disk radii from center of display) as a function of time on task across 18 sessions. Dashed lines represent 1 S.D. from the mean.

boundary did not contribute to the calculated x/y position of the fixation point. The current fixation was terminated when the mean position of the most recent three PORs fell $>1°$ from the current fixation point, or when a blink duration longer than 200.0 ms was observed. From the resulting array of fixation activity, moving estimates of fixation frequency (fixations/min) and total fixation dwell time (or dwell) were derived using 1-min square windows moved through the data in 10-s steps. Moving estimates of mean pupil diameter (excluding closures and signal losses) were calculated similarly. The 1-min window width was found to be adequate for deriving stable estimates for all the measures. For each subject, the moving estimates of blink frequency and duration, fixation frequency and duration, and pupil size, were merged with tracking error data prior to subsequent analyses.

3. Results

Fig. 2 presents the grand mean trend of tracking performance for all 18 sessions. These data represent the best general performance trend that can be derived for the

population of interest, i.e. those participants who demonstrate performance lapses during the tracking sessions. The monotonic increase in tracking error from 0 to 11 min, followed by a stable (and even improving) performance trend is remarkably similar to the performance trend observed for an auditory detection task by Jung et al. (1997). However, individual sessions showed considerable individual variance around the mean trend, as evidenced by the root mean square (RMS) error of 2.3 disk radii we obtained by attempting to estimate the performance data in each of the 18 sessions using the generalized mean trend.

Mean blink, fixation, and pupil data are plotted as a function of relative tracking error in Figs. 3–5, respectively. The data shown were collapsed across both tracking sessions of the nine participants. Tracking error is plotted as deviation from individual mean tracking performance to normalize the between-participant variability in overall visual tracking ability. All five measures exhibited clear linear or nonlinear trends as a function of performance level. An inspection of individual plots indicated that the highly nonlinear function for fixation frequency (Fig. 4) was the result of averaging over highly variable curves obtained from the nine participants.

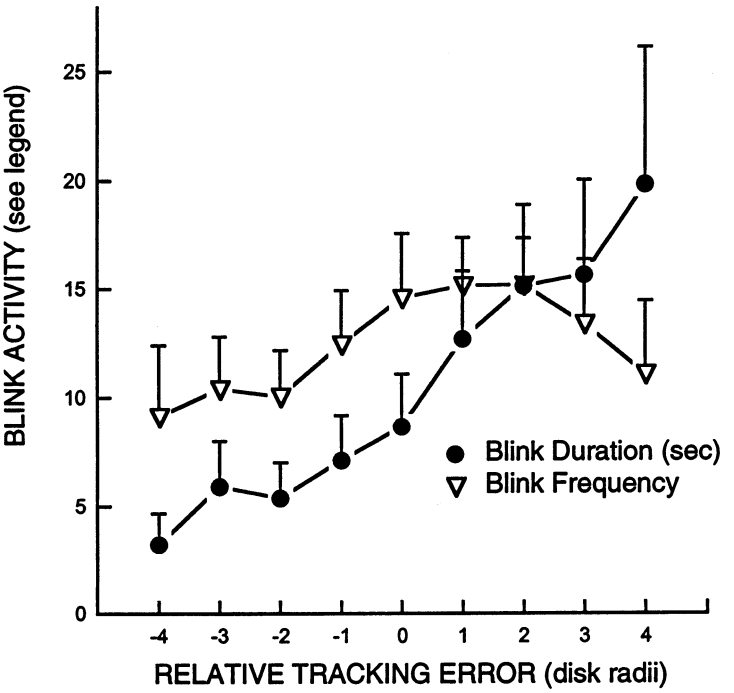

Fig. 3. Mean blink duration and blink frequency as a function of relative tracking error in disk radii for all participants demonstrating performance lapses in both testing sessions. Closure times and frequencies are averaged over the 1-min moving window. Tracking error plotted relative to each participant's mean tracking error over both sessions. Error bars represent 1 S.E. from the mean.

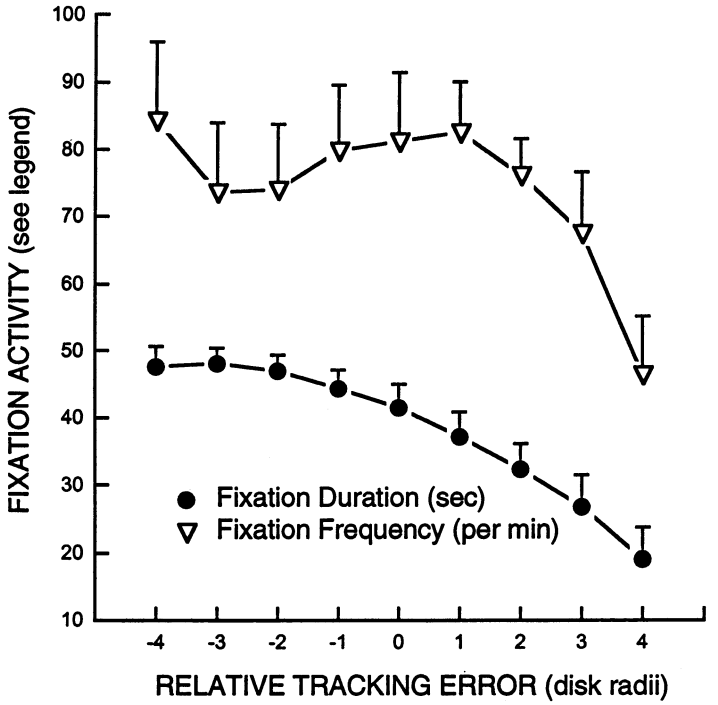

Fig. 4. Mean fixation dwell time and fixation frequency as a function of relative tracking error. Fixation dwell times (solid circles) represent total dwell time in s per min, as calculated by the fixation algorithm used in the study. Fixation frequency data (inverted open triangles) are the mean number of fixations that occur within a 1-min period.

First, individually tailored non-linear regression models were constructed for the data. Non-linear regression analysis examined blink duration and frequency, fixation dwell time and frequency, and pupil diameter, as well as squared terms and cross-products of all the linear terms. For these and subsequent regression analyses, a correlation matrix was constructed to test for multicollinearity among predictor variables. Correlation coefficients of 0.8 or higher between pairs of predictor variables forced the elimination of one predictor in order to stabilize regression weights (Berry and Feldman, 1985). Because multiplicative terms tended to correlate highly with their component parts, deviation scores were calculated for each of the linear terms before calculating the squared and cross-products terms (Jaccard et al., 1990).

A stepwise regression approach was used to build each model. The criteria for inclusion of candidate variables for the stepwise procedure was a correlation with performance of at least $R = 0.20$. The final form of the model was determined by selecting predictor terms that contributed at least 2% improvement in R^2. (A 1% improvement in R^2 was required in the subsequent development of sample-wise general models). Results of the stepwise analysis are presented in Table 1, which

Table 1
Individual participant regression models[a]

P #	Session	BD	BF	FD	FF	PD	Additional terms			R_W	R_{CR}	R_{WA}	R_{WB}	$R_{WA\&B}$
1	A	0.93			0.22		−0.24 (BD × BF)			0.88	0.70			
	B			−0.74						0.74	0.84			
	AB			−0.77								0.84	0.74	0.77
2	A	0.69	0.28							0.84	0.83			
	B			−0.46	0.33	−0.27				0.90	0.68			
	AB		0.21	−0.44		−0.36						0.86	0.87	0.86
3	A	0.64	0.12	0.51	0.66					0.67	0.80			
	B		0.74				0.35 (BD × FF)			0.77	0.47			
	AB		0.45	0.30	0.52		0.54 (BD × FD)					0.63	0.80	0.72
4	A	0.34			0.29	−0.32				0.66	0.29			
	B			−0.87						0.87	0.29			
	AB	0.28		−0.44		−0.28						0.64	0.85	0.74
5	A			−0.89	0.61		0.11 (FF^2)	−0.39 (FD^2)	−0.28 (BF × FD)	0.82	0.57			
	B		0.42	−0.34		−0.40				0.68	0.55			
	AB			−0.67	0.47	−0.31	−0.25 (BF × FD)					0.80	0.69	0.76
6	A			−0.86						0.86	0.77			
	B			−0.70			−0.22 (FD^2)	−0.18 (BD × FF)		0.82	0.74			
	AB			−0.57	−0.25							0.83	0.79	0.79
7	A			−1.25	0.35		0.21 (FD × FF)			0.89	0.61			
	B		0.36	−0.63						0.69	0.79			
	AB	0.24		−0.69								0.85	0.64	0.75
8	A			−0.46			−0.22 (FF × PD)			0.52	0.51			
	B		0.36	−0.51						0.77	0.46			
	AB		0.40	−0.61								0.45	0.77	0.68
9	A			−0.34			0.25 (BD × BF)			0.51	0.60			
	B			−0.97	0.22					0.83	0.47			
	AB			−0.94	0.36							0.48	0.82	0.69
$\bar{X}$										0.76	0.61	0.71	0.77	0.75

[a] BD, blink duration; BF, blink frequency; FD, fixation duration; FF, fixation frequency; PD, pupil diameter; R_w, within-session correlation; R_{CR}, cross-session correlation; R_{WA}, session-A correlation for model based on both sessions; R_{WB}, session-B correlation for model based on both sessions.

presents the standardized (β) model regression weights and correlations for each session, and both sessions combined, for every participant. The individual regression models derived using data from one session and cross-validated on a second session produced within- and cross-session mean correlations (R) of 0.76 and 0.61, and RMS errors averaging 1.3 and 1.6 disk radii, respectively.

The relatively large number of data points (318) in the estimated and actual tracking performance series results in very high statistical power for the assessment of correlations for statistical significance, leading to significant correlations that may not be of use from an applied perspective. Thus, we derived a cut-off statistic to determine the probability that correlations reported in Table 1 departed meaningfully from zero. For every participant, estimated tracking performance in each session was correlated with single session tracking data from every other participant. This process resulted in 16 surrogate correlations for every participant, and 144 correlations for the group of nine participants. The sample of surrogate correlations had a mean $R = 0.035$ and a S.D. of 0.309. Setting the cut-off at 1.96 S.D.s from the mean established a minimum correlation of 0.64 as necessary to reject the null hypothesis that correlation coefficients of estimated to actual tracking

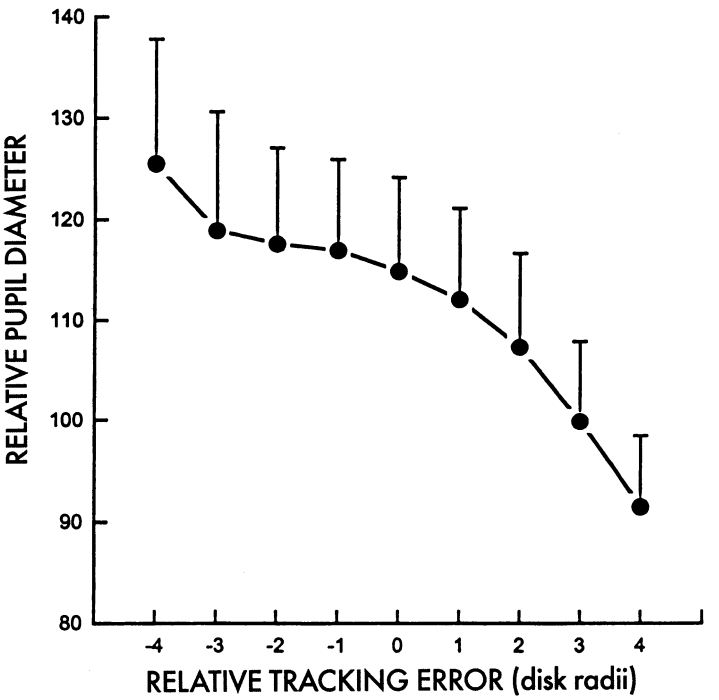

Fig. 5. Relative pupil diameter as a function of relative tracking error. Actual diameter is dependent upon distance of eye tracking optics from the eye. Generally, 110 U is ≈ 4 mm, and the slope of the function relating general units to actual diameter is nearly linear (10 U represents about 1 mm change in diameter).

performance were not significantly different from zero. This surrogate derived statistic was conservative since it took into account trends in the data, including the tendency for tracking error to increase within the first few minutes of each session (Fig. 2).

Because of the relatively large number of predictor variables common across participants within Table 1, a general regression model was developed next using data from both sessions of each participant. Predictor variables were required to account for greater than 1% of the variance for inclusion in the model. Fixation dwell time and fixation frequency accounted for 45 and 3% of the total variance, respectively. The general model, using standardized coefficients, was

$$\Delta\text{Tracking error} = -0.77 \times \Delta\text{Fixation dwell time} + 0.21 \times \Delta\text{Fixation frequency}$$

The general regression model produced estimates of tracking performance that correlated reasonably well with actual tracking data ($R = 0.68$). Mean RMS error for the general model was 1.55 disk radii (S.D. = 0.26).

To determine the extent to which individual differences existed in the data, the general regression model was entered as an independent predictor term in a stepwise procedure to derive individually-modified general models. Using the general model as an independent predictor term enabled some assessment of the extent to which additional predictor variables improved estimation performance. To avoid multicolinearity with the general model term, the predictor variables of fixation dwell time and fixation frequency were excluded from the stepwise process. The results of the analysis, as well as the application of the general model alone to each participant's data, are presented in Table 2. The standardized coefficients are presented for the general model term and for each additional term accounting for at least 2% of total variance. As shown, additional predictor variables improved the accuracy of the combined session-A/session-B models for seven of nine participants. The individually modified general regression models derived using data from one session and cross-validated on a second session produced within- and cross-session mean correlations (R) of 0.75 and 0.67, and mean RMS estimation errors of estimated to actual tracking performance of 1.3 and 1.4 disk radii, respectively.

Finally, to develop models optimized for estimating performance, artificial neural network models were constructed for each participant and for the group. For each participant, two-thirds of shuffled data from one session were used to train a feedforward three-layer (one hidden layer, three hidden units) back propagation network to estimate the performance measure. Data from the remaining third of the session were used to validate the need for further network training. Training was halted when the mean estimation error for the validation data stopped decreasing. Five different nets were trained on these data using different initial node weights. The median (third) best-performing neural net from the within-session validation data was then used to produce tracking performance estimates from eye activity data for the other session. Multiple training is a common practice to avoid selecting a network that terminated recursive training at a local minimum in the data, or of selecting a network that had over fit the training data. Within and cross-session

Table 2
Modified-general and general regression weights and coefficients[a]

P #	Session	GEN	BD	BF	PD	Additional terms		R_W	R_{CR}	R_{WA}	R_{WB}	$R_{WA\&B}$	R_{GENA}	R_{GENB}	$R_{GENA\&B}$
1	A	0.44	0.39			−0.27 (BD × BF)		0.88	0.71						
	B	0.79						0.79	0.80						
	AB	0.80								0.80	0.79	0.80	0.80	0.79	0.80
2	A	0.19	0.53	0.26				0.85	0.84						
	B	0.65			−0.27			0.88	0.82						
	AB	0.45		0.16	−0.37					0.85	0.88	0.86	0.77	0.85	0.82
3	A	0.23		0.39				0.52	0.75						
	B	0.24		0.58		0.23 (BD × FF)		0.79	0.50						
	AB	0.33		0.42						0.51	0.76	0.65	0.37	0.67	0.49
4	A	0.28	0.35		−0.33			0.66	0.83						
	B	0.86						0.85	0.34						
	AB	0.43	0.25		−0.31					0.65	0.84	0.74	0.34	0.86	0.60
5	A	0.63				−0.24 (BF × FD)		0.77	0.54						
	B	0.39		0.40	−0.38			0.70	0.58						
	AB	0.42			−0.24	−0.22 (BF × FD)				0.78	0.60	0.72	0.74	0.50	0.65
6	A	0.86						0.86	0.73						
	B	0.68				−0.20 (BD × FF)	−0.30 (FD^2)	0.81	0.72						
	AB	0.73								0.86	0.73	0.73	0.86	0.73	0.73
7	A	1.03				0.21 (FD × FF)		0.89	0.61						
	B	0.59		0.31				0.65	0.83						
	AB	0.70	0.18							0.88	0.61	0.75	0.88	0.57	0.73
8	A	0.52				−0.21 (FF × PD)		0.57	0.61						
	B	0.57		0.25				0.77	0.53						
	AB	0.63		0.29						0.53	0.77	0.70	0.53	0.74	0.64
9	A	0.37				0.25 (BD × BF)		0.52	0.82						
	B	0.83						0.83	0.47						
	AB	0.68								0.47	0.83	0.68	0.47	0.83	0.68
$\bar{X}$								0.75	0.67	0.70	0.76	0.74	0.64	0.73	0.68

[a] Terms as in Table 1. GEN, general model as described in text; R_{GENA}, correlation coefficient of omnibus general model to session-A data; R_{GENB}, correlation of general model to session-B data; $R_{GENA\&B}$, correlation of general model to data from both sessions.

Table 3
Individual participant artificial neural network correlations[a]

P #	Session	R_W	R_{CR}	R_{WA}	R_{WB}	$R_{WA\&B}$
1	A	0.93	0.74			
	B	0.92	0.77			
	AB			0.86	0.90	0.87
2	A	0.90	0.88			
	B	0.94	0.62			
	AB			0.86	0.90	0.87
3	A	0.73	0.82			
	B	0.91	0.55			
	AB			0.69	0.89	0.81
4	A	0.81	0.54			
	B	0.89	0.32			
	AB			0.79	0.85	0.82
5	A	0.86	0.61			
	B	0.78	0.63			
	AB			0.87	0.83	0.85
6	A	0.89	0.75			
	B	0.89	0.70			
	AB			0.85	0.85	0.81
7	A	0.90	0.61			
	B	0.77	0.68			
	AB			0.88	0.74	0.81
8	A	0.81	0.53			
	B	0.89	0.33			
	AB			0.65	0.81	0.80
9	A	0.67	0.60			
	B	0.88	0.45			
	AB			0.55	0.85	0.79
$\bar{X}$		0.85	0.62	0.78	0.85	0.82

[a] R_w, within-session correlation coefficient; R_{CR}, cross-session correlation; R_{WA}, session-A correlation for model based on both sessions; R_{WB}, session-B correlation for model based on both sessions.

correlations of estimated to actual tracking performance time series for individual-session and across-session neural network models are shown in Table 3. Individual neural net models produced within- and cross-session mean correlations (R) of 0.85 and 0.62, and mean RMS errors of 1.04 and 1.96 disk radii, respectively. Also shown in Table 3 are correlations of estimated to actual tracking performance based on individualized neural networks trained on data from both sessions (trained using the process described previously) for each participant. These models yielded a mean correlation (R) of 0.82 and a mean RMS error of 1.23 disk radii (S.D. = 0.13). Fig. 6 presents neural network-derived estimated and actual tracking performance data from session-A of participant 5.

The relative success of each model type in estimating actual tracking performance during cross-session validation was assessed using a repeated measures analysis of

variance (ANOVA) procedure on the correlations from each participant. No significant differences between the mean cross-session correlations derived from individual, modified general and neural network models were found; $F(2, 34) = 2.79$, $P > 0.05$. Only the mean correlation produced by the individually modified-general model ($R = 0.67$) exceeded the derived cut-off for meaningful correlations of $R = 0.64$. A similar analysis was performed on correlations derived from models based upon both sessions from every participant, as well as the performance of the general model on each participants data, and yielded a significant effect of model type; $F(3, 24) = 15.72$, $P < 0.001$. A Tukey-HSD test for differences among the means indicated that the within-session mean correlation derived using the neural networks ($R = 0.85$) was significantly greater than correlations derived using the individual, modified general, and general models (0.75, 0.74, and 0.68, respectively), $P < 0.005$. Furthermore, the mean correlation of estimated to actual tracking performance data produced by the general model ($R = 0.68$) was found to be significantly lower than the mean correlation produced by the individualizd models ($R = 0.75$), $P < 0.05$. The correlation data, from individual models based upon both experimental sessions, are presented in Fig. 7.

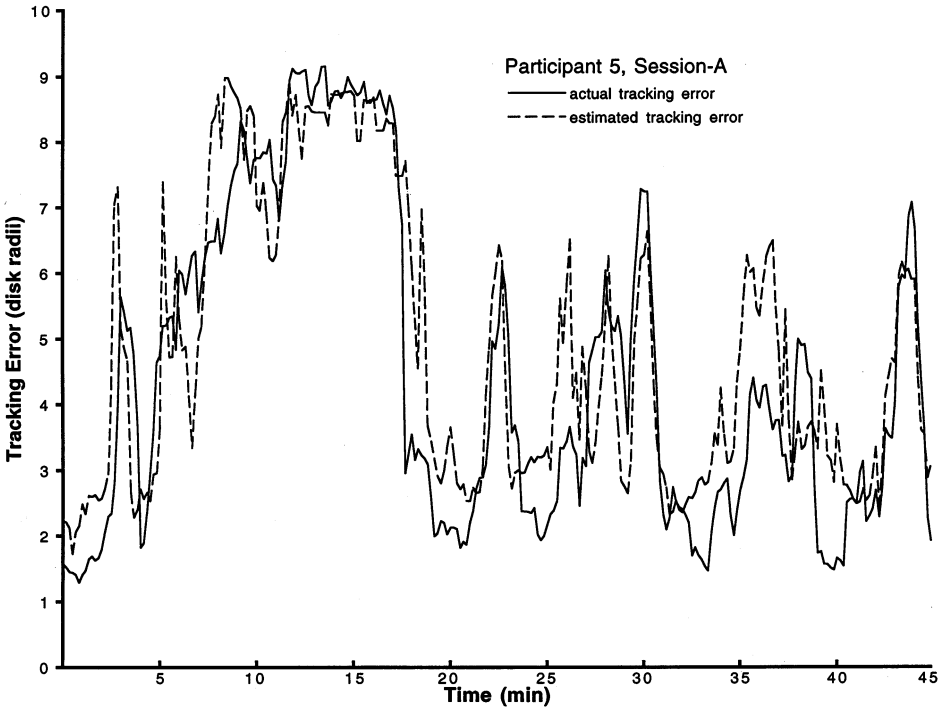

Fig. 6. Actual and estimated compensatory tracking performance for session-A of Participant 5. Estimated data were produced by an artificial neural network trained on data from both of the participant's testing sessions. RMS estimation error for this session was 1.24 disk radii.

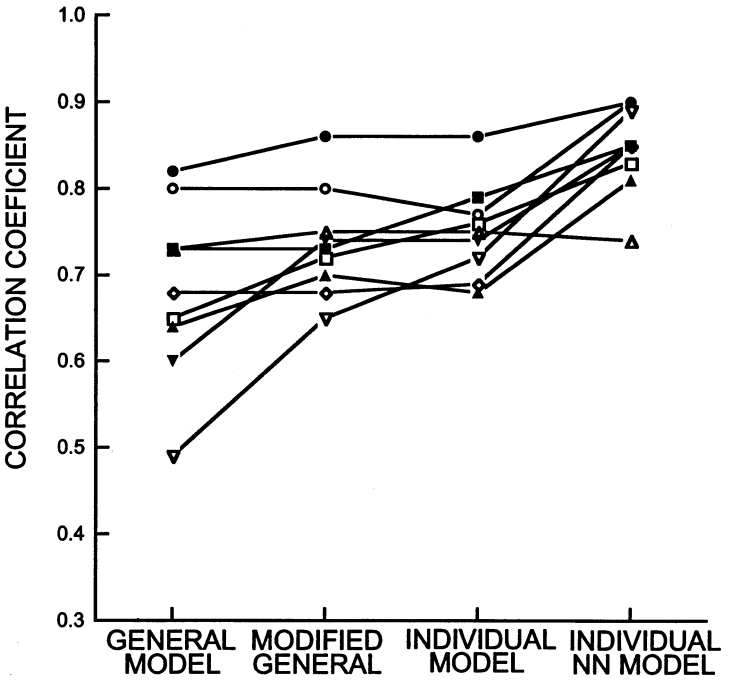

Fig. 7. Individual participant correlation coefficients (*R*s) as a function of model type. The general regression model was based upon data from all participants. All other models were derived for each individual participant, using data from both sessions. The neural network (NN) model produced the highest correlations for eight of the nine participants.

4. Discussion

The present study demonstrated that fluctuations in performance during a sustained visual tracking task occur on a circa minute-scale and can be accurately monitored using multiple eye activity measures. Performance and eye activity changes during the visual tracking task appeared to be consistent with fluctuations in alertness of the participants. Various linear and nonlinear combinations of blink measures (frequency and duration), fixation measures (frequency and dwell time), and pupil diameter were used to develop linear and non-linear regression and neural network models which produced moving estimates of tracking error within sessions and in separate sessions that were highly correlated with actual changes in visual tracking performance. Mean correlations between estimated and actual tracking performance in the present study were similar to those reported for a flight simulation experiment by Morris and Miller (1996) based on two EOG derived parameters: long closure rate and blink amplitude. However, unlike their study, we examined both general and individualized models, and further used models based on data from one session to predict near continuous changes in tracking performance in a different session. Cross-session validation is useful for estimating the

potential utility of individualized and general models for online performance monitoring.

The major component of the general regression model developed from the present data was fixation dwell time, which was negatively correlated with blink duration, and proved to be a better predictor variable than blink duration within most of the regression analyses herein. The best within-session performance was obtained with individually derived neural network models. Performance estimation was most accurate for models derived from sessions including a wide range of performance levels. Such sessions allowed regression modeling and neural net training to more accurately characterize differences in eye activity measures during both good and poor performance periods. As expected, the results of performance estimation using neural network models were comparable to those using regression models. Neural networks, however, extracted non-linear features from the data automatically, making them convenient and adaptive to individual differences in eye activity changes as a function of performance fluctuations. Fig. 6 shows actual and estimated performance time series for one participant. Estimates of tracking error were produced by the neural network model trained on both of the participant's testing sessions. The RMS estimation error for the data shown in Fig. 6 is 1.22 disk radii.

The present data generally support observations from previous studies that changes in individual eye blink (Stern et al., 1984, 1994), eye fixation (McGregor and Stern, 1996), and pupil diameter (Yoss et al., 1970) measures covary with changes in performance due to drowsiness, loss of vigilance and/or increasing time-on-task. Our measure of blink duration was highly correlated with tracking error for many participants, in agreement with research conducted by Stern et al. (1984, 1994), and Morris and Miller (1996) indicating that blink duration increases as a function of time on task. Perhaps related to blink duration, Wierwille et al. (1994) have reported that a measure of eyelid droop, 'perclose', correlates with lane wobble deviations in sustained driving simulator studies. Perclose was defined as the percentage of time (within a several minute integration period) that the eyelid is at least 80% closed. However, in our regression analyses fixation dwell time was a significantly more robust and reliable predictor of tracking performance. In our task, the fixation dwell time measure may have captured smooth pursuit eye movements (e.g. participants visually tracking the moving disk) and/or slow rolling eye movements that were observed in some participants during periods of heightened drowsiness. Neither of these types of eye movements were registered as fixations, further reducing the recorded dwell times. By contrast, the blink duration measure was insensitive to smooth pursuit or rolling eye movements.

Consistent with reports by Lowenstein and Lowenfeld (1962), Yoss et al. (1970) and McLaren et al. (1992), pupil diameter was typically smaller during periods of increased tracking error, although this finding was not consistent across all participants. In general agreement with previous research, blink frequency also increased as a function of tracking error. Finally, fixation frequency, reported by McGregor and Stern (1996) and Morris and Miller (1996) to decline with time on task, was variably correlated to tracking performance across participants. Morris and Miller

found fixation frequency was correlated with performance only in the flight maneuvers portion of their experiment, during which participants needed to focus attention on a greater number of cockpit displays than in the straight-and-level flight segments. Saccade velocities, which have also been found to decrease with increasing time on task (Schmidt et al., 1979; McGregor and Stern, 1996), could not be measured in the present study because of the limited (60 Hz) temporal resolution of our eye tracking apparatus. Analysis of saccade velocity data might possibly be useful, but will require a faster recording system.

To our knowledge, ours is the first study to demonstrate continuous and objective moving-mean estimation of performance changes based on multiple eye activity measures and using signal processing methods that could be applied in online systems. In particular, our results demonstrate that continuous information related to performance under monotonous conditions is available in eye activity measures. Performance monitoring using multiple eye measures can easily outperform generalized (grand mean) performance models for such tasks, as was the case in the present study. Based upon the findings from our highly constrained experimental sessions, general performance trends (e.g. time-on-task) within real world monotonous task environments and over extended periods are even less likely to predict minute-scale performance fluctuations. While the moving-mean estimation approach may be useful in many task domains, it is important to note that the present results and models generalize only to the visual tracking task used in our study. Further development and improvement of real-time eye activity alertness monitoring methods will require analysis of data collected from a greater number of participants within other task-specific environments. Cross-session validation results from the present study, while encouraging, could be improved by using multiple training sessions. Parameter variability across repeated measurements would allow regression or neural network models to detect and then ignore unreliable measures during extended use. Additionally, the development of real-time monitoring systems should also involve an assessment of the extent to which combined eye activity measures are indicative of brief periods of poorer performance by individuals like those who rarely deviated from near-ideal performance in our experiments.

It is possible that still more robust alertness monitoring methods might combine measures of eye activity with other psychophysiological measures such as EEG or electromyographic signals. In most applications, however, routine use of these measures awaits the availability of convenient dry electrode technology, while non-invasive measurement of eye activity may prove a nearer-term possibility.

Acknowledgements

This research was supported by a grant from the Office of Naval Research (ONR.WR.30030-6429). The views expressed in this article are those of the authors and do not reflect official policy or position of the Department of the Navy, Department of Defense, or the US Government. The authors wish to thank Shawn

Wing, Silvina Moncho, William Pugh and Wendy Limbert for their assistance with the study. Technology disclosed herein may be the subject of an invention disclosure owned by the US Government. Inquiries may be directed to Office of Patent Counsel-Code D0012, Space and Naval Warfare Systems San Diego, 53510 Silvergate Avenue, San Diego, CA 92152, USA.

References

Beatty, J., 1982. Phasic not tonic pupillary responses vary with auditory vigilance performance. Psychophysiology 19, 167–172.

Berry, W.D., Feldman, S., 1985. Multiple regression in practice. Sage University Series of Quantitative Applications in the Social Sciences, series no. 07-050. Sage, Newbury Park, CA.

Jaccard, J., Turrisi, R., Wan, C.K., 1990. Interaction effects in multiple regression. Sage University Series of Quantitative Applications in the Social Sciences, series no. 07-072. Sage, Newbury Park, CA.

Jung, T.-P., Makeig, S., Stensmo, M., Sejnowski, T.J., 1997. Estimating alertness from the EEG power spectrum. IEEE Trans. Biomed. Eng. 44, 60–69.

Lowenstein, O., Lowenfeld, I.E., 1962. The pupil. In: Dawson, H. (Ed.), The Eye, vol. 3. Academic Press, New York.

Makeig, S., Inlow, M., 1993. Lapses in alertness: coherence of fluctuations in performance and the EEG spectrum. Electroencephalogr. Clin. Neurophysiol. 86, 23–35.

Makeig, S., Jolley, K., 1996. COMPTRACK: A compensatory tracking task for monitoring alertness, NHRC Tech. Doc. No. 96-3. Naval Health Research Center, San Diego, CA.

Makeig, S., Jung, T.-P., 1995. Changes in alertness are a principal component of variance in the EEG spectrum. NeuroReport 7, 213–216.

Makeig, S., Jung, T.-P., 1996. Tonic, phasic, and transient EEG correlates of auditory awareness in drowsiness. Cogn. Brain Res. 4, 5–25.

McDonald, N., 1984. Fatigue, Safety and the Truck Driver. Taylor and Francis, London.

McGregor, D.K., Stern, J.A., 1996. Time on task and blink effects on saccade duration. Ergonomics 39, 649–660.

McLaren, J.W., Erie, J.C., Brubaker, R.F., 1992. Computerized analysis of pupillograms in studies of alertness. Invest. Ophthalmol. Visual Sci. 33, 671–676.

Morris, T.L., Miller, J.C., 1996. Electrooculographic and performance indices of fatigue during simulated flight. Biol. Psychol. 42, 343–360.

O'Hanlon, J.F., 1978. What is the extent of the driving fatigue problem? In: Driving Fatigue in Road Traffic Accidents. Rep. EUR6065EN. Commission of the European Communities, Brussels, pp. 19–25.

Peavler, W.S., 1974. Individual differences in pupil size and performance. In: Janisse, M.P. (Ed.), Pupillary Dynamics and Behavior. Plenum Press, New York, pp. 159–175.

Schmidt, D., Abel, L.A., Dell'Osso, L.F., Daroff, R.B., 1979. Saccade velocity characteristics: intrinsic variability and fatigue. Aviation Space Environ. Med. 50, 393–395.

Stern, J.A., Walrath, L.C., Goldstein, R., 1984. The endogenous eyeblink. Psychophysiology 21, 22–33.

Stern, J.A., Boyer, D., Schroeder, D., 1994. Blink rate: a possible measure of fatigue. Hum. Factors 36, 285–297.

Wierwille, W.W., Wreggit, S.S., Knipling, R.R., 1994. Development of improved algorithms for on-line detection of driver drowsiness. Convergence 94 Conference. Society of Automotive Engineers, Detroit, October 1994.

Yoss, R.E., Moyer, N.J., Hollenhorst, R.W., 1970. Pupil size and spontaneous pupillary waves associated with alertness, drowsiness, and sleep. Neurology 20, 545–554.