NASA CONTRACTOR REPORT 177354

Mental Workload Measurement: Event-Related Potentials and Ratings of Workload and Fatigue

M. A. Biferno

CONTRACT NAS2- 11860
June 1985

NASA

NASA CONTRACTOR REPORT 177354

Mental Workload Measurement: Event-Related Potentials and Ratings of Workload and Fatigue

M. A. Biferno
Douglas Aircraft Company
3855 Lakewood Boulevard
Long Beach, California 90846

Prepared for
Ames Research Center
under Contract NAS2-11860

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

Table of Contents

Event-Related Potentials and Ratings of Workload and Fatigue

Michael Biferno
McDonnell Douglas Corporation, Douglas Aircraft Company
Long Beach, California

ABSTRACT

Event-related potentials were elicited when a digitized word representing a pilot's call-sign was presented. This audi ory probe was presented during 27 workload conditions in a 3x3x3 design where the following variables were manipulated: short-term memory load, tracking task difficulty, and time-on-task. Ratings of workload and fatigue were obtained between each trial of a 2.5 hour test. The data of each subject were analysed individually to determine whether significant correlations existed between subjective ratings and ERP component measures. Results indicated that a significant number of subjects had positive correlations between: (1) ratings of workload and P300 amplitude, (2) ratings of workload and N400 amplitude, and (3) ratings of fatigue and P300 amplitude. These data are the first to show correlations between ratings of workload or fatigue and ERP components thereby reinforcing their validity as measures of mental workload and fatigue. Since ratings of fatigue and workload were significantly correlated for 16 of 20 subjects, future studies of workload would benefit from examining the relationship between them.

INTRODUCTION

For reasons of safety, the design of highly automated systems requires that consideration be given to the workload and fatigue of the operator (Weiner and Curry, 1980; Lyman and Orlady, 1981). Incorporation of workload data into the thinking of the design engineer, however, requires that workload be quantified and measured in valid, reliable and standardized ways.

Ratings of workload and fatigue have been available for many years but few of the techniques have established their validity or reliability with standard psychometric techniques (Wierwille, 1979). Two exceptions are techniques developed by agencies of the United States government: Bipolar-adjective rating scales (Hart, Battiste and Lester, 1984), and Subjective Workload Assessment Technique, otherwise known as SWAT (Reid, Eggemeier, Shingledecker, 1981). Laboratory studies have established the construct validity and test-retest reliability of these measures (Childress, Hart and Bortolussi, 1982; Eggemeier, Crabtree, Zingg, Reid and Shingledecker, 1982).

The standard practice in evaluating operator workload is to ask a trained operator about his or her work. Many subjective assessment techniques are available but there are a variety of methodological problems associated with each of them (Williges and Wierwille, 1979). Some of the disadvantages include biased reporting, distortions and forgetting. The use of subjective measures often interferes with the process under investigation by imposing demand characteristics which can enhance or degrade the behavior being studied (Stave, 1977; Walster and Aronson, 1967). This interfering nature of subjective measures makes them most difficult to employ when an impartial workload analysis is required. These problems aside, the design engineer will continue to rely on subjective measures because they are readily available, persuasive,

easily administered, low cost, and they can be easily interpreted. Engineering will continue to use subjective measures until something better is available. Something better would be a measurement technique that is more valid, more reliable or less susceptable to biasing.

Electrocortical measures of workload may offer a better technique under some circumstances (Moray, 1979; O'Donnell, 1979). The electrocortical measure of particular interest is the P300 component of the event-related potential (ERP) which is sometimes called the late positive component (Donchin, 1979). ERPs can not be easily biased since subjects are not aware of variations in their own electrocortical activity and therefore cannot modify their ERP activity in a highly selective fashion. ERP measurement can also be relatively unobtrusive and noninterfering if properly implemented. ERP measurement does not require conscious mediation and they can be recorded in ways which blend into many work environments. For example, ERPs can be recorded from speech stimuli which are a part of the operator's normal communication duties (Biferno and Bigham, 1982).

Despite these advantages, physiological measures are not widely employed to quantify workload for a number of reasons. The most important reason is that no practical measure exists. Although ERPs are known to be related to brain events and to human information processing activites, the complex relationship among the many types of workload and ERP activity is only now beginning to be explored. Although a number of experiments have demonstrated a relationship between levels of workload and the amplitude of the P300 component, this is only a demonstration of its construct validity and indicates that more research is warranted.

Some experiments have found that the P300 amplitude increases with increased workload, while others have found that it decreases. When the P300 is elicited by stimuli which are part of a secondary task, P300 amplitude is reduced when the primary task workload is high (Isreal, Chesney, Wickens and Donchin, 1980; Isreal, Wickens, Chesney and Donchin, 1980; Natani and Gomer, 1981; Biferno, 1985). When elicited by stimuli which are part of a primary task, the P300 amplitude has been found to increase when the primary task workload is high (Horst, Munson & Ruchkin, 1984). At least two other ERP components reflect changes in task demands. The amplitude of the P200 increases as a verbal processing task becomes more difficult (Poon, Thompson and Marsh 1976) and the N200 increases in latency when subjects are required to mentally count the occurrance of a stimulus class (Biferno, 1982).

Despite success in demonstrating the construct validity of ERP and subjective workload measures, the relationship between them has not been well studied. Experiments which manipulate task demands (workload) typically do not measure both and it is unclear whether ratings of workload are correlated with the P300 or any other component. It may be the case that both measures covary with task demands but are not correlated with each other. The amplitude of the P300 component has been found to be correlated with subjective ratings of expectancy or confidence in a judgement (Pritchard, 1981; Horst, Johnson and Donchin, 1980) which suggests that ERP components may be related to other subjective states such as perceived workload or fatigue (Hashimoto, Kogi and Grandjean, 1975; Gauthier and Gottesmann, 1983).

The primary purpose of this experiment was to answer two questions: (1) Are ratings of workload correlated with ERP measures of workload? (2) Are ratings of fatigue correlated with ERP measures of fatigue? There are at least two

general approaches for answering these questions. Correlations can be performed across a group of subjects to determine whether a relationship exists for members of that population or correlations can be performed on the data of individual subjects. The generalization of results to populations, based on individual subject correlations, can be accomplished by determining the frequency of subjects showing the correlation in question and then determining whether the frequency of significant correlations is more than expected by chance.

We selected the second approach for two reasons: (1) Group correlations were unlikely to attain significance because of individual differences unrelated to our experimental procedures. That is, there was not a strong reason to believe that subjects who came to the experiment with large ERP component amplitudes (an individual trait) were likely to report higher workload ratings than would subjects who emitted small ERPs to environmental events. (2) We were interested in observing patterns of individual-subject correlations. Knowing that individuals define and experience workload differently, we might expect more than one pattern of correlations to emerge from a large group of subjects. If the group correlation approach were taken, the different patterns of correlations would be masked or eliminated altogether.

One difficulty in comparing subjective measures with other workload measures is the diverse number of ways in which workload is defined. When individuals are asked to rate their experience of workload along a number of subjective dimensions, they structure their ratings in many different ways (Hart, Childress, and Hauser, 1982). The application of principle component analysis to workload rating data was done by Hart and her collegues (1982) and the result was the identification of several workload factors. The three factors which can be inferred to account for the largest amounts of variance in her rating data were: "Fatigue/stress", "Time-pressure/number-of-tasks", and "Mental-busy/effort". The reduction of dimensionality in the workload metrics provide a convenient method for operationally defining workload along a few general dimensions. This enables the formulation of a simplified experimental design with a few manageable factors and some description of how the independent variables should be manipulated in order to produce workload ratings like those which contributed to the factors.

Our strategy was to manipulate workload along the three dimensions which were outlined above while obtaining workload ratings, fatigue ratings, and ERPs. Individual subject correlations were then performed among the measures and the relative frequency of each correlation was assessed. Workload levels were manipulated to insure that the ERP and rating data contained systematic variability due to task-related variables and to insure that each subject experienced low, medium, and high levels of workload.

METHOD

Design. Workload was manipulated in three ways with each factor having three levels. Selection of the independent variables was based on a study (Hart, Childress and Hauser, 1982) which employed principal component analysis to identify some of the major factors which form the basis of a person's subjective report of workload. The independent variables were: (1) short-term memory load (subjects had to briefly remember consonant strings of two, four or six items), (2) difficulty of a compensatory tracking task (difficulty was varied in a Jex critical tracking task by varing lambda (L), .3L, .6L or .9L), and (3) time-on-task (it was assumed that increasing levels of fatigue could be

observed during early, middle and late trials). The 3x3x3 repeated measures design yielded 27 different workload conditions and each of 20 subjects received 3 trials of each condition for a total of 81 trials.

Subjects. A total of 24 subjects were recruited from Long Beach State University. All were volunteers and paid about $6.20 per hour for the two four hour sessions. Four subjects were discarded for the following reasons: two subjects failed to return for the second day of testing, one subject could not perform the memory task, and one subject would not follow instructions. Of the 20 subjects remaining, seven were run in the tasks more than once because of equipment failures (four subjects) and excessive edge violations with the tracking task (three subjects).

Equal numbers of males and females were assigned to use their left/right hand to perform the tracking task. Their ages ranged from 18 to 28, they were right-handed and English was their native language. The order of performing two baseline tests on the first day was also balanced with half of the people performing the tracking task first and half performing the memory task first.

Stimuli and Procedures. Each subject was seated at a table in a sound attenuated and darkened room. A small (7 x 10 cm) cathode-ray tube (CRT) was placed directly in front of the subject's eyes. A luminous horizontal line whose vertical displacement was controlled by means of an isometric joystick was displayed on this CRT. The controls, display, and forcing functions were modeled after the Jex critical tracking task (Jex and Clement, 1979). The joystick was oriented horizontally so that forces applied up and down would control vertical movements of the horizontal line. The position of the joystick relative to the screen (left/right) was counterbalanced.

A directional microphone protruded from a panel below the CRT and was pointed toward the subject's mouth at a distance of about 5 - 10 cm. A 10 button keyboard, which was used to report ratings of workload and fatigue after each trial, was located to the left (right) of the CRT. The keyboard was oriented horizontally with the "1" key to left and the "10" key to the right. Taped to the surface of the table, and centered in front of the subject, was a typed description of the bipolar-rating scales. The microphone and CRT were immediately above and behind the rating-scale descriptions. Subjects could comfortably reach the joystick and rating keys. They could also see the CRT display and make voice responses into the microphone without moving their head or eyes. There was always sufficient illumination in the room during the test for subjects to read the rating-scale descriptions.

The start of each trial was signaled by a brief tone which was followed by two, four or six consonants spoken by a digitized-speech unit (Digitalker model DT-1000), presented through an earphone placed in the subject's right ear (Sony model MDR E255). The consonants had a .75 s interstimulus interval (ISI) and an approximate intensity of 80 dbA. White noise was presented to the left ear at a level of approximately 60 dbA to mask background noises in the room. Subjects were instructed to immediately repeat back the consonant string to minimize the number of trials lost due to memory errors. If a subject failed to repeat back the consonants in the same order that they were presented, the experimenter would press a key to recycle the same consonant string. The consonants could be recycled at the request of the subject as often as needed.

The tracking task began after a subject accurately repeated the consonant string. For approximately 60 s, the subject performed the critical tracking

task at one of the three fixed levels of difficulty, the actual difficulty levels were determined for each person during their practice session.

During tracking, subjects would listen for the occurrence of one of two words. They would hear either "40" or "14" via an earphone placed in their right ear and their task was to say "Roger" when they heard "14" and say nothing when they heard "40". This activity simulated an aircraft communication where the vocal response was the pilot's acknowledgement of hearing his call-sign (14) while ignoring communications directed to other aircraft. A total of six words (14s or 40s) were presented during each trial. The call-sign (14) occurred 33 1/3 percent of the time, while the number of call-signs varied from one to three on any given trial and the occurrence of each word was presented in an unpredictable order with a ISI which varied between 5.0 and 12.5 s.

A DEC 11/23 laboratory computer controlled the presentation of the experimental stimuli and managed the data collection of the ERP and subjective-rating data. Immediately preceeding the onset of the call-sign, the voice-reaction-time clock was started and recording of the ERP was initiated. Voice reaction time and ERPs were measured in relation to the onset of the 610 ms call-sign.

The tracking task ended at the completion of a trial and the word "check" was presented. This was the signal to report the memory items and then prepare to give workload ratings. After a fixed amount of time (about 5.0 s), the words "Please rate, rate A" were presented via the earphones. This signaled them to reflect on the previous trial and generate a workload rating using the first bipolar-rating scale which was lettered "A" on the scale-description list. When any of the 10 keys were depressed, the Digitalker echoed the numerical value assigned to the key. Erroneous rating entries could be reported to the experimenter for manual correction during data analysis. The same sequence of events was followed until all ten ratings were entered into the computer. After entering the tenth rating, the next trial began automatically. Each trial lasted about two min. this sequence was repeated without a break (Stave, 1977) for 81 trials or about 2.5 hours. Since there were an average of two call-signs per trial and three trials for each of the 27 experimental cells, the averaged ERP waveforms were based on a maximum of six sweeps per cell and the averaged ratings were based on three sets of ratings per cell.

Baseline tests and practice. Each subject was run a total of two sessions with the first day consisting of two baseline tests and a substantial amount of practice on the memory and the tracking tasks. The purpose of the baseline tests was to measure changes in ratings and ERPs when the two types of workload were manipulated separately. Each baseline test employed the same three levels of workload as the multifactorial experiment. Each baseline test had a total of 27 trials and lasted about 50 minutes. Half of the subjects received the memory test first, while the other half received the tracking test first. The results of the baseline tests are not included in this report.

Dependent variables. Three classes of dependent variables were measured: (1) ERP components, (2) subjective ratings, and (3) behavioral performance. The ERP and behavioral performance measures were obtained while subjects experienced one of the 27 different workload conditions. The subjective ratings were obtained immediately after each trial.

ERP components. A total of eight ERP measures were obtained from four ERP components. The latency and peak-to-peak amplitude of the N100, P300, N400, and P500 components were measured with an automatic scoring program which

scanned each ERP waveform for the: (1) N100 - the most negative minima between 70 and 160 ms, (2) P300 - the most postive maxima between 200 and 440 ms following the N100, (3) N400 - the most negative minima between 320 and 565 ms following the P300, and (4) P500 - the most positive maxima between 435 and 705 ms following the N400. These values were obtained from previous experimental work performed in our laboratory. Latencies were measured from stimulus onset until the point when the peak occurred. Amplitudes were measured peak-to-peak where: P300 was the voltage at the P300 peak minus the voltage at the N100 peak, N400 was the voltage at the N400 minus the voltage at the P300 peak, P500 was the voltage at the P500 peak minus the voltage at the N400 peak. N100 amplitude was the voltage at the N100 peak minus the voltage at the first time point in the A/D sampling.

ERP recording. Brain potentials were recorded from monopolar leads referenced to linked earlobes. The midline sites of Fz, Cz, and Pz were sampled (Jasper, 1958) at 200 Hz for a 1000 ms epoch. All correlational analyses were performed on data obtained from the Pz site. The Pz site was selected because the P300 is usually largest, eye movement artifacts are smallest, and pilot data in our lab suggested that correlations with performance are usually high at that location. Grass silver-disk electrodes were employed with a saline bentonite paste covered with gauze and a saline preparation to prevent drying. The sites were prepared with a saline solution and electrode impedances were less than 10K ohms. Grass model 511J AC amplifiers were employed with a 1/2 high and low frequency cut off of 100 Hz and .1 Hz, respectively. Eye blinks were monitored from locations above and below the left eye via Beckman Ag-AgCl mini-cup electrodes. They were sampled in the same manor as the ERP data except that eye blink analysis was terminated at 650 ms because some subjects exhibited movement artifacts associated with saying "Roger" after this time.

ERP data were discarded for the following reasons: (1) an eye blink occurred during the sampling epoch, (2) an edge violation (tracking task) occurred during the sampling epoch or (3) a memory error occurred during the trial. A memory error required that more than one of the recalled consonants was wrong. In the case of a memory error, all ERP sweeps for that trial were discarded.

Subjective ratings of workload and fatigue. The subjective rating scales employed in this experiment were provided by NASA-Ames. They consisted of a set of bipolar adjectives or dimensions which were rated with a ten-point scale (Hart, Battiste and Lester, 1984). Fatigue was one of the rating scales employed in this test and the other was a derived workload rating which took into account the way each subject structured workload in relation to nine component dimensions. Before beginning the test, each subject made paired comparisions between the workload scale descriptions and indicated which member of the pair contributed more to their experience of workload. The more frequently a dimension was chosen, the more weight it would be given when computing the derived workload rating. The weighted-workload rating was computed by multiplying the bipolar ratings by the weights (ranging from: "0" not related to "8" highly related), summing these values, and then divided by the value of the weights (36).

Behavioral Performance. Tracking performance and voice reaction time were recorded during each trial. Average tracking error and the number of edge violations during the ERP epoch were scored and saved. Average tracking error was computed as the area under-the-curve in the compensatory tracking task. Area under-the-curve was computed as the sum of errors (deviations from center line) for each time point during the ERP sampling epoch. Edge violations were

defined as a failure to maintain control of the critical tracking task, allowing the tracking symbol to move off the screen. Whenever an edge violation occurred during ERP sampling, the ERP data was discarded. Voice reaction time was measured from the onset of the call-sign until a voice relay detected the subject's vocalization of "Roger".

RESULTS

There were two criteria for concluding that a significant relationhip between ERP components and subjective ratings had been found: (1) individual-subject correlations for a particular relationship must attain statistical significance, and (2) a significant number of subjects must attain a correlation on that relationship. Individual Pearson correlation coefficients had to be .381 or greater to attain significance with alpha set at .05 (df=25 with correlations performed on the bivariate data from 27 cells); and at least four out of the 20 subjects had to have a significant correlation with any pair of ERP/rating measures before a relationship was claimed to exist.

The binomial distribution was employed to determine the probability of obtaining "x" significant correlations, where each test employs an alpha level of .05. Figure 1 shows the computation of the probability of x, where x equals the number of significant correlations obtained after performing 20 tests with an alpha of .05 (Siegel, 1956). The p(x) is .013 that 4 of the 20 correlations would be significant by chance. The probabilities of obtaining 1, 2, 3 or 4 significant correlations when N=20 are: .377, .188, .059 and .013 respectively.

$$p(x) = (N/x)P^{x}Q^{(N-x)}$$

When $(N/x) = N!/x!(N-x)!$

IF $P = .05,\ Q = .95,\ N = 20,\ x = 4$

THEN $$p(4) = \frac{20!}{4!\ 16!}(.05)^{4}(.95)^{16}$$

$$p(4) = (4845)(.000002)(.4401262)$$

$$p(4) = .013$$

Figure 1. The probability of obtaining four significant correlations when each correlation has an alpha of .05 and 20 correlations are performed.

Correlations with workload ratings. A summary of the Pearson product-moment correlations, performed between workload ratings and ERP component, are shown in Table 1. Two of the ERP measures were correlated with workload ratings more frequently than would be expected by chance. Four subjects were found to have correlations between the P300 amplitude and workload ratings, while significant correlations between the N400 amplitude and workload ratings were found for four additional subjects. Therefore, 8 of 20 subjects showed significant correlations between ratings of workload and brain potential measures of workload.

	Latency				Amplitude			
Subject	N100	P300	N400	P500	N100	P300	N400	P500
1	.077	-.089	.226	.215	.017	.435*	.204	.046
2	-.034	-.130	-.008	.023	-.260	.517*	.319	.203
3	.097	.332	.148	-.184	-.088	.223	.006	.025
4	.318	-.062	-.091	.133	-.133	.135	-.077	-.176
5	-.250	.389*	.485*	.483*	.284	.264	.432*	.525*
6	-.075	-.113	.111	-.101	.140	-.132	-.034	-.014
7	.173	-.001	-.166	-.269	.223	.464*	.139	.318
8	.335	.135	-.124	-.151	.116	.025	-.010	.165
9	.234	-.457*	-.212	.023	-.110	-.105	.448*	.432*
10	.061	.138	-.269	.161	.155	.205	.189	.243
11	-.403*	.273	.103	.398*	.069	.042	-.223	.122
12	-.126	-.453*	-.173	.263	.063	-.050	.420*	.454*
13	.080	.056	.253	.402*	-.011	-.264	.254	-.152
14	-.122	.126	.003	-.252	-.245	.178	.215	.061
15	.057	-.196	.009	-.276	-.136	.126	.187	.080
16	-.022	.139	.006	.244	.248	-.130	-.184	-.002
17	.082	.229	.049	.065	-.227	.386*	.149	.272
18	.228	.018	-.035	.220	-.148	.160	.042	-.087
19	.199	.034	.480*	-.278	-.277	-.009	.525*	.299
20	.018	-.114	-.042	.082	-.083	.182	-.102	.167
f	1	3	2	3	0	4	4	3

Table 1. Correlations between weighted-workload ratings and each ERP component measure for each subject. * = indicates significant correlation, $p<.05$, two-tailed; f = frequency of significant correlations.

Since the workload correlations were evenly divided between the P300 and N400 measures, a series of post-hoc analyses was performed to shed light on possible explanations for this dichotomy. Inspection of the subject assignment sheet was done to see if any differences were present. At first glance the striking result was that all four P300 correlators were females and all four N400 correlators were male. Other conditions of interest were also examined (hand used for tracking task, order of practice for the two subtasks, age, and time-of-day tested), but no clear differences were noted.

Post-hoc analyses of P300 and N400 subjects. It was hypothesized that the two groups might have defined or experienced workload differently and therefore be monitoring different physiological processes in rating workload thereby producing the different ERP performance. Since the bipolar-rating technique required each subject to rank-order the importance of the 9 rating scale dimensions, it was possible to statistically compare the groups based on their unique definitions of workload. Table 2 gives the weights for each subject on each of the rating dimensions with the average ratings and rank of each dimension provided in the right-hand columns for each group. A Spearman rank-order correlation was performed to determine whether the two groups structured (rank-ordered) the workload dimensions in a similar way. A significant correlation of .775 indicates that they did.

The performance data of these two groups was then examined (see Table 3). A general trend started to emerge: the male subjects performed better on the tracking task, memory task and responded faster in the voice reaction time task than the female subjects (951 vs. 1297 ms; $F(1,6) = 5.73$, $p = .053$). These results support an informal observation that some males became very involved in the tracking task and were less inhibited in vocally responding "Roger" to their simulated call-sign. Inspection of the scale dimension rankings (Table 2) gives additional support to a performance set difference. The largest difference score between the workload weights of the two groups was with the dimension entitled "performance".

	P300 Subjects						N400 Subjects					
Scale Dimension	1	2	7	17	$\bar{X}$	Rank	5	9	12	19	$\bar{X}$	Rank
Task Difficulty	3	8	3	4	4.5	4	4	2	6	3	3.8	6
Time Pressure	8	3	3	7	5.2	2	2	5	6	4	4.2	4.5
Performance	5	1	2	5	3.2	6	4	8	8	2	5.5	3
Mental/Sensory Effort	4	6	0	2	3.0	7.5	1	1	6	0	2.0	8
Physical Effort	0	5	5	0	2.5	9	4	2	2	5	3.2	7
Frustration	7	0	7	8	5.5	1	8	6	2	8	6.0	1
Stress Level	4	2	8	6	5.0	3	7	7	3	6	5.8	2
Fatigue	3	4	6	3	4.0	5	6	4	0	7	4.2	4.5
Activity	2	7	2	1	3.0	7.5	0	1	3	1	1.2	9

Table 2. Weights assigned to the nine workload scale dimensions by subjects who attained a significant correlation between workload ratings and a brain potential measure.

	P300 Subjects					N400 Subjects				
Measure	1	2	7	17	$\bar{X}$	5	9	12	19	$\bar{X}$
Reaction Time	997	1277	1388	1526	1297	863	924	1214	803	951
Tracking Error	1.3	1.5	1.9	1.7	1.6	.9	.3	2.4	1.4	1.2
Edge Violation	6	15	11	16	12.0	4	2	27	12	11.2
Memory Error	0	1	0	6	1.8	0	0	0	0	0
Workload Rating	2.1	2.3	4.4	5.4	3.6	3.5	1.4	3.3	5.4	3.4
Fatigue Rating	.8	.9	1.2	2.0	1.2	1.3	.4	1.8	1.5	1.2

Table 3. Performance comparison of subjects who attained a significant correlation between workload ratings and a brain potential measure. Ratings are in arbitrary units; Reaction time in ms; Tracking error in arbitrary units; Edge violation and Memory error are frequencies.

Since the two groups differed in ERP performance, a between groups ANOVA was performed on their ERP data. The N400 subjects had longer P300 latencies than the P300 subjects ($F(1,6) = 6.15$, $p<.05$, 301.2 versus 276.0 ms). This result suggests that the N400 correlations may be due to a shift in the P300 peak latency and the N400 correlations are more properly considered a P300 phenomenon. The subgroups did not respond differently with any of the other ERP measures.

These results could have far-reaching implications for the measurement of workload with ERP measures. The question is raised: Do the two groups exhibit different types of workload correlations because of a subject trait (which might be relatively stable) or is it due to some subject state (which might be unstable and situationally dependent)?

Correlations with fatigue ratings. A summary of Pearson product-moment correlations, performed between fatigue ratings and each ERP component, are shown in Table 4. Only the P300 amplitude component was correlated frequently enough to be considered a statistically reliable finding. The correlations were positive and similar in magnitude to the workload rating correlations. In fact, 3 of the 4 subjects having significant fatigue correlations also had significant workload correlations with the P300 component.

	Latency				Amplitude			
Subject	N100	P300	N400	P500	N100	P300	N400	P500
1	.094	-.245	-.013	.171	.153	.498*	.391*	.236
2	-.029	.132	.138	.317	.076	.159	-.072	-.147
3	.189	.281	.184	-.241	.102	.076	-.041	-.260
4	.497*	-.178	-.264	-.095	-.079	.213	-.008	-.075
5	-.105	.341	.546*	.185	.172	-.146	.062	.077
6	-.044	.063	.132	-.085	.125	-.051	.012	-.195
7	.181	-.077	-.129	-.240	.228	.416*	.163	.300
8	.487*	.274	-.026	.027	-.011	.101	-.044	.219
9	.197	-.298	-.253	-.106	-.054	-.117	.346	.513*
10	.088	.188	-.212	.077	.174	.104	.177	.256
11	-.134	.395*	.480*	.068	-.102	-.025	-.240	.055
12	-.185	-.373	-.211	.136	-.137	.057	.371	.504*
13	.130	.247	.241	.579*	-.233	-.158	.065	.129
14	.083	-.133	.349	.084	-.249	-.250	-.023	.024
15	.059	-.105	.019	-.290	-.231	.038	-.001	-.028
16	-.097	.158	.064	.380	-.012	.085	.054	.142
17	-.058	.242	-.021	.046	-.054	.467*	.232	.401*
18	.058	-.135	-.054	.099	-.075	.091	.284	.224
19	.057	.152	.447*	-.313	-.081	.024	.327	.085
20	.246	.137	.178	.474*	-.089	.450*	.420*	.156
f	2	1	3	2	0	4	2	3

Table 4. Correlations between fatigue ratings and each ERP component measure for each subject. * = indicates significant correlation, f = frequency of significant correlations.

Since both subjective ratings were correlated positively with the same ERP components, the intercorrelation between these variables was examined. The degree of correlation between ratings of fatigue and weighted-workload for each subject is shown in Table 5. A high degree of correlation between the scales can be observed for some subjects, while no correlations exist for others. The general trend is for a moderate correlation (average $\underline{r}$ = .596) with 16 of 20 individuals showing a significant relationship between ratings of workload and fatigue.

Subject	r	Significant
1	.643	*
2	.183	
3	.799	*
4	.767	*
5	.351	
6	.763	*
7	.961	*
8	.685	*
9	.545	*
10	.907	*
11	-.313	
12	.915	*
13	.623	*
14	.395	*
15	.880	*
16	.471	*
17	.926	*
18	.212	
19	.791	*
20	.417	*

Table 5. Pearson product-moment correlations between ratings of fatigue and weighted-workload for each subject. * = $p<.05$

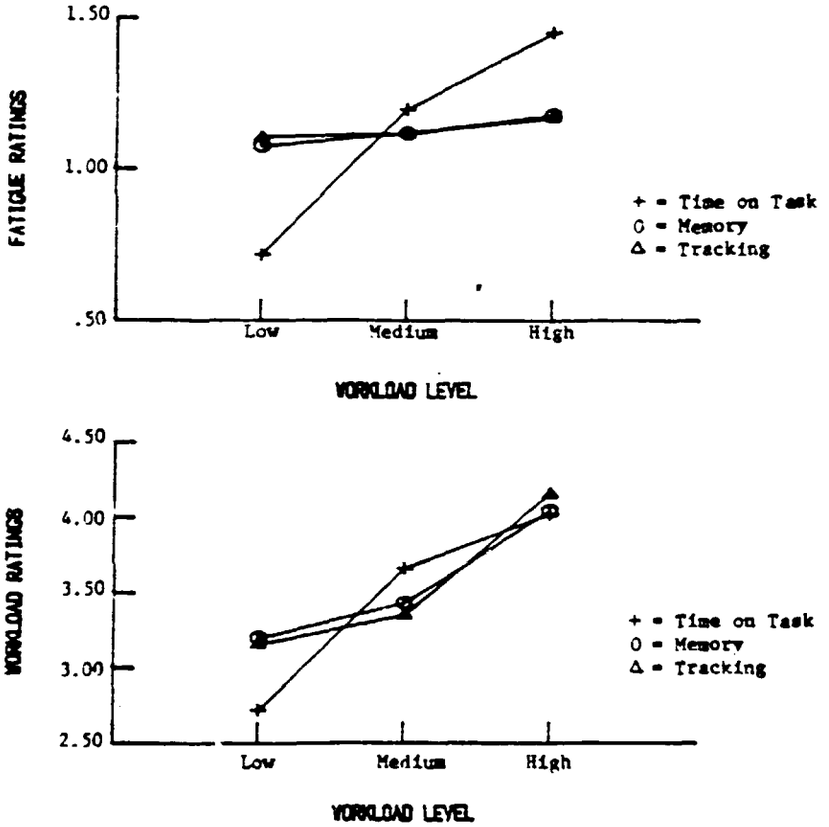

Figure 2. Fatigue ratings increase from early to late trials (top) and workload ratings increase as task demands increase (bottom).

Although fatigue ratings were highly correlated with workload ratings, they were influenced differently by the three types of workload. Fatigue ratings strongly increased with time-on-task (see Figure 2, $F(2,38) = 47.68$, $p<.0001$), moderately increased with greater memory load ($F(2,38) = 5.73$, $p<.01$), but did not change with increased tracking difficulty ($F(2,38) = 1.85$, $p<.20$). Workload ratings, which include the fatigue subscale, increased strongly with all three types of workload: memory load ($F(2,38) = 26.12$, $p<.0001$), tracking difficulty ($F(2,38) = 37.51$, $p<.0001$), and time-on-task ($F(2,38) = 14.75$, $p<.0001$).

In summary, the P300 and N400 components of the ERP were positively correlated with workload ratings, while only the P300 was positively correlated with fatigue ratings. Of the 8 subjects showing a correlation between workload ratings and ERP measures, half exhibited a correlation with the P300 amplitude and the other half with the N400 amplitude. Post-hoc analyses suggest that the the mechanism of the difference was increased P300 latency for the N400 group and that the different pattern of correlation for the two groups may be due to a subject trait (e.g., sex) or due to differences in subject state (e.g., performance set).

Performance and workload. Average tracking error, frequency of edge violations in the tracking task, and voice reaction time were the behavioral measures of performance employed in this experiment (Figure 3). A 3x3x3 ANOVA performed on each of these measures revealed significant increases due to increased time-on-task: (1) voice reaction time ($F(2,38) = 3.25$, $p<.05$), (2) tracking error ($F(2,38) = 10.87$, $p<.005$), and (3) edge violations ($F(2,38) = 7.01$, $p<.005$). Increased difficulty of the tracking task also resulted in significantly increased: (1) tracking error ($F(2,38) = 101.36$, $p<.0001$), and edge violations ($F(2,38) = 48.80$, $p<.0001$).

Significant interactions were obtained with both tracking measures. There was a tendency for tracking error and edge violations to increase with longer time-on-task with the exception of the highest tracking load which showed a drop in error near the end of the test (tracking error x time-on-task interaction, $F(2,38) = 6.05$, $p<.00$ and an edge violation x time-on-task interaction, $F(2,38) = 9.92$, $p<.0001$) Behavioral observation of the subjects, during the test, suggested that they were having great difficulty performing the most demanding tracking task near the end of the experiment, due to fatigue. The diminished error at this point is interpreted as being due to higher levels of effort being applied to maintain adequate performance.

ANOVA of ERP components. A 3x3x3 ANOVA was performed on each of the ERP measures to determine whether they reflected workload changes. Grand average waveforms based on the data from all 20 subjects can be seen in Figure 4. The solid line represents the low-memory-load/low-tracking-load condition, while the dotted line represents the high-memory-load/high-tracking-load condition.

A 3x3x3 ANOVA performed on the data of all 20 individuals revealed significant ERP changes due to time-on-task and tracking difficulty, but not for memory load. Increased time-on-task (fatigue) was associated with increased N100 latency (100.9 ms, 104.2 ms, 107.7 ms; $F(2,38) = 7.38$, $p<.005$), increased N400 latency (431.6 ms, 442.1 ms, 441.9 ms; $F(2,38) = 3.95$, $p<.05$), and increased P500 latency (599.3 ms, 609.1 ms, 617.7 ms; $F(2,38) = 3.32$, $p<.05$). P300 amplitude tended to increase with increased time-on-task (10.5 μv, 9.3 μv, 11.9 μv; $F(2,38) = 2.63$, $p=.08$). Increased tracking difficulty was associated with increased P300 latency (284.6 ms, 296.0 ms, 299.1 ms; $F(2,38) = 3.63$, $p<.05$) and N400 amplitude (16.1 μv, 16.9 μv, 18.7 μv; $F(2,38) = 3.34$, $p<.05$).

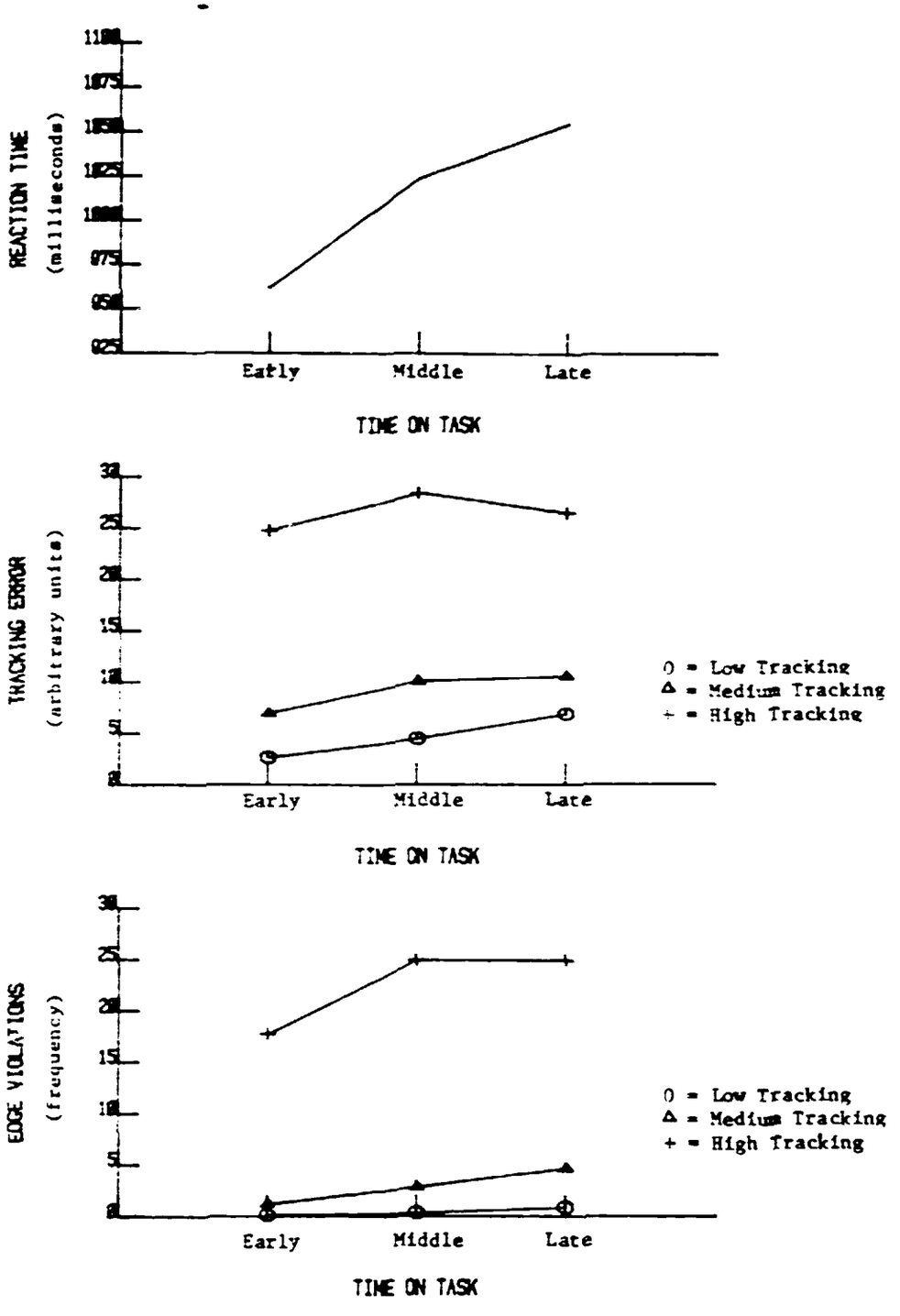

Figure 3. Voice reaction time increases with time-on-task (top), tracking error increases with tracking difficulty and time-on-task (middle), and frequency of edge violations increase with tracking difficulty and time-on-task (bottom).

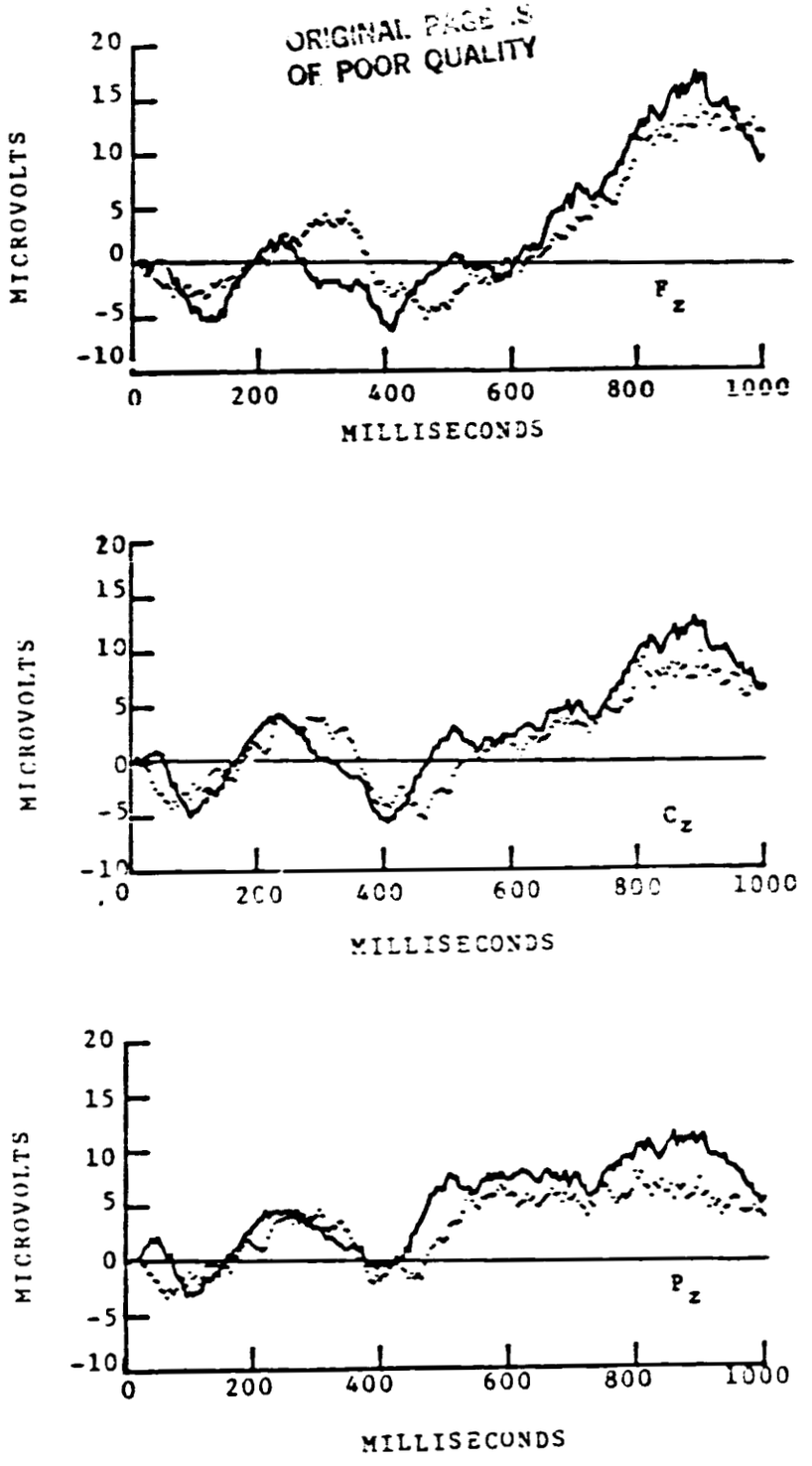

Figure 4. Average ERP at Fz, Cz and Pz sites in the lowest and highest workload conditions: Low memory load and low tracking difficulty (solid line), High memory load and high tracking difficulty (dashed line).

The significant interactions obtained with the ERP data involved two components: the N100 amplitude and the N400 latency. The N400 latency showed a significant time-on-task main effect and a three-way interaction (time-on-task x tracking load x electrode site, $F(8,152) = 2.71$, $p<.01$). The interaction suggested that N400 latency increased with time-on-task but only when the tracking load was high. N400 latency did not consistently increase with time-on-task for the low and moderate tracking load when all three electrode sites were compared. N100 amplitude was implicated in the three remaining interactions. Since all of them were small ($p<.05$) and concern a measure which did not correlate with workload or fatigue, they will not be mentioned further.

Eyeblink analysis. The total number of blinks was small due to the stringent criteria employed in the eye blink rejection procedure. The cumulative frequency of eye blinks for each type of workload is listed in order of increasing workload (low, medium, high): memory load (30, 28, 43), tracking difficulty (36, 35, 30), and time-on-task (23, 39, 39). The general trend was toward increased blinking as subjects became more fatigued or as they experienced higher memory load, but they appeared to blink less under conditions of higher tracking task difficulty.

DISCUSSION

P300 correlation with workload. The data provide an affirmative answer to the two primary experimental questions. Ratings of workload and ratings of fatigue were correlated positivity with ERP component measures. These results provide another line of evidence which indicate that the P300 is related to the concept workload. The strength of the relationship between ratings of workload and P300 amplitude (14 to 26 percent of the variance) is not sufficient to make them interchangeable measures of workload, but the correlations appear to reflect a common underlying process in at least four subjects.

N400 correlation with workload. Of special interest is the finding that just as many subjects had a correlation between the N400 component and workload ratins. This suggests that other measures of mental workload exist besides P300 amplitude. Even if the N400 correlations were due to P300 latency increases, P300 amplitude was not related to workload ratings for these subjects. This indicates that a composite ERP measure may improve the degree of association between ratings of workload and brain potential measures.

One reason the N400 has not previously been associated with workload changes may be due to the type of stimulus employed to elicit the ERP. Simple tones, flashes, and the visual presentation of text have been typical in workload research. Since these stimuli are processed relatively quickly, the N400 component may not have been observed. When speech stimuli are employed, on the other hand, the N400 is often one of the largest components and can be identified frequently in a single ERP sweep.

Since two different relationships between workload ratings and measures of brain activity were observed for different groups of subjects, it is tempting to suggest that they represent two distinctly different types of subjects or two different cognitive states. A method of testing each of these hypotheses would be to retest the eight subjects who exhibited significant correlations and determine whether they still maintained a correlation with the same component. This could be followed by an experiment in which the subject's performance set (e. g., speed/accuracy instructions) was manipulated. If subjects exhibited a stability of correlation with the same ERP component, then

a subject trait variable would be indicated as responsible for the P300/N400 dichotomy in workload correlations. However, if subjects exhibited a flexibility in the ERP component which correlated with workload ratings, then a state variable would be indicated as responsible for the P300/N400 dichotomy in workload correlations.

P300 correlation with fatigue. A correlation between ratings of fatigue and P300 amplitude are not a great surprise, but a positive correlation was unexpected. Positive correlations may reflect a requirement for greater processing resources, and correspondingly larger P300 amplitudes, in order to maintain performance on the task. Another possibility is that the larger P300 amplitudes may be due to increased drowsiness with its associated increase in low-frequency/high-amplitude background EEG. No support was found for the hypothesis that increased fatigue would be associated with smaller P300 amplitudes because of greater interference in processing the call-sign.

It seems that the study of fatigue was virtually synonomous with the study of workload for three of the subjects. Subjects 1, 7, and 17 had moderate to high correlations between their ratings of workload and fatigue (.643, .961, and .926). These results suggest that progress in subjective workload assessment would be made by a more detailed study of fatigue.

Individual subject analysis. This method evaluates the correlations for individual subjects and then determines whether a sufficient number had attained significant correlations. There are several advantages to this approach. First, individual-subject correlations describe relationships present in single subjects. Second, since the correlations are performed within-subject, highly variable subjects do not disrupt the inferential mechanism for deciding the presence of relationships between ratings and ERP components. Individuals who respond in nontypical ways may not exhibit any significant correlations, but this outcome has a small impact on the final interpretation of the results. On the other hand, when between-subject correlations are employed, two or three unusual subjects (out of 20) could turn an otherwise highly significant Pearson correlation into one which is near zero. A third advantage of this approach is that single subject correlations performed on large groups of subjects enable sub-groups to emerge and reveal differences in how people respond to the same experimental task. There is evidence that people structure workload differently and this procedure might enable subgroups to be identified, such at the P300/N400 subgroups of this experiment. One improvement in the analyses would be to increase the number of sweeps per experimental condition. Perhaps a greater number of significant correlations would have been obtained if the averaged ERPs were based on more than six sweeps per cell. Alternatively, some subjects exhibit very clear single-trial ERPs in response to speech stimuli (about 1/3) and correlations could be performed on the raw data of these subjects.

ANOVA of ERP components. The ERP components reflect a moderate sensitivity to the types of workload manipulated in this test. The clearest effects were due to time-on-task (fatigue) where three measures showed increases: N100 latency, N400 latency, and P500 latency. These ERP changes, in conjunction with the relatively good correlation between ratings of fatigue and workload suggests why significant correlations were obtained between workload ratings and the P300 amplitude. Other processes besides fatigue were operating, however, since tracking difficulty also influenced the P300 latency and N400 amplitude. Although the influence of memory load was observed in a number of statistical interactions, its influence on the ERP was not strong.

REFERENCES

Biferno, M. A. (1982). N200 latency in event-related potentials elicited by tone and synthetic speech stimuli. Psychophysiology, 19, 551.

Biferno, M. A. (1985). Short-term memory influences on event-related potential components and lateralization. Unpublished manuscript, McDonnell Douglas Corporation, Douglas Aircraft Company, Long Beach, California.

Biferno, M. A. and Bigham, T. R. (1982). Speech-related potentials elicited by synthetic speech stimuli. Psychophysiology, 19, 306-307.

Childress, M. E., Hart, S. G., and Bortolussi, M. R. (1982). The reliability and validity of flight task workload ratings. Proceedings of the Human Factors Society-26th Annual Meeting (pp 319-324).

Donchin, E. (1979). Event-related brain potentials: A tool in the study of human information processing. In H. Begleiter (Ed.), Evoked brain potentials and behavior (pp 13-88). New York: Plenum.

Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., and Singledecker, C. A. (1982). Subject workload assessment in a memory update task. Proceedings of the Human Factors Society-26th Annual Meeting (pp 643-647).

Gauthier, P. and Gottesmann, C. (1983). Influence of total sleep deprivation on event-related potentials in man. Psychophysiology, 20, 351-355.

Hart, G. S., Battiste, V. and Lester, P. T. (1984). POPCORN: A supervisory control simulation for workload and performance research. Proceedings of the 20th Annual Conference on Manual Control.

Hart, S. G., Childress, M. E., and Hauser, J. R. (1982). Individual definitions of the term "workload". Paper presented at the symposium: Psychology in the Department of Defense. U. S. Airforce Academy.

Hashimoto, K., Kogi, K. and Grandjean, E. (1975). Methodology in human fatigue assessment. London: Taylor and Francis.

Horst, R. L., Johnson, R. and Donchin E. (1980). Event-related brain potentials and subjective probability in a learning task. Memory and Cognition, 8, 476-488.

Horst, R. L., Munson, R. C., and Ruchkin, D. S. (1984). Event-related potential indices of workload in a single task paradigm. Proceedings of the Human Factors Society-28th Annual Meeting (pp 727-731).

Isreal, J. B., Chesney, G. L., Wickens, C. D., and Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources and dual-task performance. Psychophysiology, 17, 259-273.

Isreal, J. B., Wickens, C. D., Chesney, G. L., and Donchin, E. (1980). The event-related brain potential as an index of display-monitoring workload. Human Factors, 22, 212-224.

Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. Electroencephalography and Clinical Neurophysiology, 10, 371-375.

Jex, H. R. and Clement, W. F. (1979). Defining and measuring perceptual-motor workload in manual control tasks. In N. Moray (Ed.), Mental workload its theory and measurement. New York: Plenum.

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. Science, 207, 203-205.

Lyman, E. G. and Orlady, H. W. (1981). Fatigue and associated performance decrements in air transport operations. National Aeronautics and Space Administration, Ames Research Center, Moffett Field, California. NASA CR-166167.

Moray, N. (Ed.) (1979). Mental workload: Its theory and measurement. New York: Plenum.

Natani, K. and Gomer, F. E. (1981). Electrocortical activity and operator workload: A comparison of changes in the electroencephalogram and in event-related potentials. McDonnell Douglas Corporation, St. Louis, Missouri. McDonnell Douglas report MDC E2427.

O'Donnell, R. D. (1979). Contributions of psychophysiological techniques to aircraft design and other operational problems. AGARD report number 244, Available from NTIS, 5285 Port Royal Road, Springfield, VA, 22161.

Poon, L. W., Thompson, L. W., and Marsh, G. R. (1976). Average evoked potential changes as a function of processing complexity. Psychophysiology, 13, 43-49.

Pritchard, W. S. (1981). Psychophysiology of P300. Psychological Bulletin, 89, 506-540.

Reid, G. B., Eggemeier, F. T., and Shingledecker, C. A. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society-25th Annual Meeting, 522-526.

Siegel, S. (1956). Nonparametric Statistics: For the Behavioral Sciences. New York: McGraw-Hill.

Stave, A. M. (1977). The effects of cockpit environment on long-term pilot performance. Human Factors, 19, 503-514.

Walster, B. and Aronson, E. (1967). The effects of expectancy of task duration on the experience of fatigue. Journal of Experimental Social Psychology, 3, 41-46.

Wiener, E. L. and Curry, R. W. (1980). Flight-deck automation: Promises and problems. Moffett Field, California: National Aeronautics and Space Administration, Ames Research Center, NASA Technical Memorandum 81206.

Wierwille, W. W. (1979). Physiological measures of aircrew mental workload. Human Factors, 21, 575-593.

Williges, R. C. and Wierwille, W. W. (1979). Behavioral measures of aircrew mental workload. Human Factors, 21, 549-574.

Symbols and Abbreviations

Ag-AgCl	Silver-Silver chloride
A/D	Analog to digital
ANOVA	Analysis of variance
cm	centimeter
Cz	Midline, vertex
dbA	decibel - A weighted
df	degrees of freedom
ERP	event-related potential
f	frequency
$\underline{F}$	$\underline{F}$ ratio
Fz	Midline, frontal
Hz	Hertz
L	lambda
min	minute
ms	millisecond
N100	negative peak at 100 ms
N400	negative peak at 400 ms
P200	positive peak at 200 ms
P300	positive peak at 300 ms
P500	positive peak at 500 ms
$\underline{p}$	probability
Pz	Midline, posterior
$\underline{r}$	Pearson correlation coeficient
s	second
μv	microvolt
$\overline{X}$	mean
<	greater than
*	probability = .05

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CR-177354		

4. Title and Subtitle	5. Report Date
Mental Workload Measurement: Event-Related Potentials and Ratings of Workload and Fatigue	June , 1985
	6. Performing Organization Code: LH
7. Author(s): Michael A. Biferno	8. Performing Organization Report No.
9. Performing Organization Name and Address: Douglas Aircraft Company, 3855 Lakewood Blvd., Code 35-36, Long Beach, California 90846	10. Work Unit No. (TRAIS)
	11. Contract or Grant No.: NAS2-11860
12. Sponsoring Agency Name and Address: NASA, Ames Research Center, Moffett Field, California 94035	13. Type of Report and Period Covered: Contractor Report
	14. Sponsoring Agency Code: 505-35-31

15. Supplementary Notes

Point of Contact: Technical Monitor, Sandra G. Hart, MS 239-3
Ames Research Center, Moffett Field, CA 94035
(415) 694-6072 or FTS 464-6072

16. Abstract

Event-related potentials were elicited when a digitized word representing a pilot's call-sign was presented. This auditory probe was presented during 27 workload conditions in a 3x3x3 design where the following variables were manipulated: short-term memory load, tracking task difficulty, and time-on-task. Ratings of workload and fatigue were obtained between each trial of a 2.5-hour test. The data of each subject were analyzed individually to determine whether significant correlations existed between subjective ratings and ERP component measures. Results indicated that a significant number of subjects had positive correlations between: (1) ratings of workload and P300 amplitude, (2) ratings of workload and N400 amplitude, and (3) ratings of fatigue and P300 amplitude. These data are the first to show correlations between ratings of workload or fatigue and ERP components thereby reinforcing their validity as measures of mental workload and fatigue. Since ratings of fatigue and workload were significantly correlated for 16 of 20 subjects, future studies of workload would benefit from examining the relationship between them.

17. Key Words	18. Distribution Statement
Mental workload Subjective ratings Fatigue Event-related potentials ERP	Unclassified - Unlimited Star Category - 53

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	22	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized