ELSEVIER

Biological Psychology 40 (1995) 33–71

BIOLOGICAL PSYCHOLOGY

Event-related potentials as indices of display-monitoring performance

Leonard J. Trejo*, Arthur F. Kramer, Josh A. Arnold

Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, USA

Accepted 3 November 1994

Abstract

We evaluated event-related potentials (ERP) as indices of performance in three visual display-monitoring tasks: (a) signal detection, (b) running memory and (c) computation. Using factor analysis, we developed a global measure of performance (PF1) for each task. Task-relevant and irrelevant-probe stimuli elicited ERPs, which included components P1, N1, P2, P300, slow waves, and fronto-central negativities. In tasks (a) and (b), P300 amplitude in the task-relevant ERPs increased when the task was engaged, and was greater for accurate- than for inaccurate-response trials. In tasks (a) and (c), the irrelevant-probe ERPs also differed among task and performance conditions. To relate ERP measures to PF1, we developed linear regression models distinguished by three factors: general versus individual-subject, stimulus relevance, and signal-to-noise ratio (SNR). Model accuracy and reliability were highest for individual-subject, relevant-stimulus and high-SNR models, where average R^2 values for the three tasks were 0.44, 0.46, and 0.38, respectively. We discuss implications of the models for performance monitoring and implications of the ERP effects for human information processing.

Keywords: Event-related potentials; Display-monitoring; Performance; Stimulus relevance; Signal-to-noise ratio

1. Introduction

In many important tasks performed by human operators, performance quality varies over time, often falling below acceptable limits. Parasuraman (1986), sum-

* Corresponding author.

SSDI 0301-0511(95)05103-H

marizing the results of 20 human performance studies involving operational inspection and monitoring tasks, found that operators missed a disturbingly large fraction of signals in many tasks. Across tasks, the average percentage of missed signals exceeded 30%. Increases in workload or task difficulty may increase the likelihood of human error. For example, in a CRT-monitoring task, Jerison & Pickett (1964) found that the percentage of missed signals increased from 10% at an event rate of five per minute to about 70% at an event rate of 30 per minute.

Such performance variability may have serious consequences. For example, a missed or misclassified signal on an air traffic controller's display could result in an aircraft collision. Other important tasks affected by human performance variability include defense (radar, sonar, electronic warfare), communications, power plant operations and piloting of vehicles. In many of these tasks, the likelihood of such errors could be reduced if real-time indices of performance quality were available.

Although the performance of a trained operator depends strongly on task variables such as the information rate and signal-to-noise ratio (SNR) of task-relevant stimuli, psychological constructs such as perceptual, cognitive and motor processes are also important. An indication of these processes is provided by brain event-related potentials (ERP). ERPs reflect mental processes and are known to be related to human performance, including signal detection, confidence ratings, target identification and recognition, memory, tracking and mental computation (Hillyard, Squires, Bauer & Lindsay, 1971; Kok & DeJong, 1980; Kramer, Wickens, Vanasse, Heffley & Donchin, 1981; Parasuraman & Beatty, 1980; Parasuraman, Richer & Beatty, 1982; Ruchkin, Johnson, Mahaffey & Sutton, 1988). The aim of this study was to evaluate the utility and reliability of real-time inferences about display-monitoring performance that may be made using ERPs. To assess the generality of ERP-performance relationships, we evaluated tasks that presented three different classes of demands on the subjects: signal detection, running memory and mental computation.

Practical interest in ERPs stems from concepts, models, and experimental data that explain variation in human performance in terms of internal constraints, mechanisms and physiological states. Broadbent (1970) proposed the concept of limited capacity, which constrains the quality of performance in resource-limited tasks (Norman & Bobrow, 1975). For a given task environment, capacity may be thought of as one or more pools of specific resources available to meet task demands (Wickens, 1984). This concept of limited capacity is useful for interpreting relationships between ERPs and performance. In dual-task paradigms, division of resources among the tasks leads to performance trade-offs, which are indexed by ERP components such as the P300 (Blankenship, Trejo & Lewis, 1988a; Hoffman, Simons & Houck, 1983; Israel, Chesney, Wickens & Donchin, 1980). Furthermore, the P300 is selectively sensitive to perceptual/cognitive demands. (Donchin, Kramer, & Wickens, 1986). Many other studies (reviewed by Gopher & Donchin, 1986; Kramer, 1990; Parasuraman, 1990) have also reported relationships between ERP components and performance.

ERPs are often elicited by task relevant or secondary task probes. However, ERPs elicited by task-irrelevant probes have also been successfully used as indices of pro-

cessing demands in humans and animals (Defayolle, Dinand & Gentil, 1971; Garcia-Austt, Bogacacz & Vanzulli, 1964; Kramer, Trejo & Humphrey, in press; Oatman, 1971; Papanicolaou & Johnstone, 1984). In an example relevant to the present study, root-mean-square ERP amplitude over fronto-central areas with a latency of 330 ± 25 ms for a random irrelevant visual-flash stimulus was about 40% lower when subjects performed a complex radar simulation than during a passive baseline period (Trejo, Lewis & Blankenship, 1987). In addition, active-baseline differences in probe ERP amplitude were correlated with performance levels in the task (Trejo, Lewis & Blankenship, 1990). A replication of these effects (Blankenship, Trejo & Lewis, 1988b) with an active baseline control condition indicated that the differences between baseline and simulation irrelevant-probe ERP amplitudes were not response-related.

Attention is also associated with a set of specific ERP effects that bear on the quality of performance. Firstly, efficient task performance requires selective attention to task-relevant events and inattention to extraneous stimuli, such as probes. Attention to relevant stimuli, either spatial or non-spatial, amplifies a range of ERP components, including P1, N1, P2 and N2 as well as slower, broad negativity of latency 150–300 ms (Eason, Harter & White, 1969; Harter & Aine, 1984; Van Voorhis & Hillyard, 1977). With limited capacity, any attention to probe stimuli should detract from primary task performance and enhance components of the probe ERP. Probe stimuli may also reflect allocation of resources in preattentive sensory mechanisms. In hearing, such a mechanism can be demonstrated in experiments where attention is directed to one train of stimuli (say in one ear) and withheld from another train in which a few deviant stimuli occur (Näätänen, 1982). The deviant stimuli produce a small P300 (the P3a) with a shorter latency and more frontal distribution than that observed for attended stimuli (Squires, Squires & Hillyard, 1975). The P3a is preceded by the N2b, a negativity of latency 180–220 ms. The occurrence of the N2b-P3a has been suggested to indicate the activation of preattentive mechanisms sensitive to differences among unattended stimuli.

The approach we took in this study was to record ERPs elicited concurrently by task-relevant stimuli and irrelevant-probe stimuli in three different tasks. The goal of the study was to determine for each stimulus type the quantitative relationship between variations in the amplitude and latency of ERP components and performance on the tasks across subjects as well as for individual subjects. To this end, we manipulated the difficulty of the task relevant stimuli so as to produce variations in performance. In the signal detection task, difficulty was varied by lowering the contrast of targets presented on a CRT. In the running memory task, difficulty was varied by increasing the number of intervening stimuli between a cue and a target in a delayed letter-matching task. In the computation task, difficulty was varied by increasing the empirically measured complexity of mental division problems. In all tasks, a single global measure of performance was derived using factor analysis of the group multivariate performance data (e.g., measures of reaction time, accuracy, confidence) and validated using factor analyses of the individual subjects' performance data.

We examined the relationships between ERP components and performance at two

levels of ERP SNR. At the single-trial level, stepwise linear regression models were fitted to the single-trial performance data using ERP component amplitudes and latencies from single-trial ERPs. At a higher SNR level, stepwise linear regression models were fitted to 10-trial running means of the performance data using 10-trial running means of the ERP component amplitudes and latencies. Across tasks and levels of ERP SNR, ERP components for both relevant and irrelevant stimuli were significantly related to task performance by the linear regression models. However, the models based on running means of the ERP components and the performance data were much more accurate and reliable than the models based on single-trial data. In addition, models based on estimates of relevant-stimulus ERP components were much more accurate and reliable than the models based on irrelevant-probe ERP components. For these reasons, this paper will focus on the analyses of the running-mean ERPs elicited by the task-relevant stimuli and only briefly discuss the single-trial and irrelevant-probe based ERP data. Some interesting differences in model accuracy and reliability were also observed between the regression models for different tasks.

2. Method

2.1. Subjects

The subjects were eight right-handed male volunteers from the U.S. Navy ranging in age between 19 and 44 years. Each subject's vision was tested prior to experimentation. Corrective lenses were worn as required for visual acuity of 20/20 or better. Each subject had training or experience relevant to visual display monitoring tasks (radar, sonar, or electronic warfare). Questionnaire results indicated that none of the subjects had experienced head trauma, dizziness, fainting, or equilibrium problems. With the exception of coffee, cigarettes, and sodas, none of the subjects had taken any medication or drug in the 24 h preceding test sessions.

2.2. General Aspects of Tasks

The subjects performed three tasks: signal detection, running memory with letters, and computation (mental arithmetic). In each task a computer presented transient visual stimuli in discrete trials that required immediate responses. Trials were paced by the computer and the difficulty levels for correct processing of the stimuli were quantized at different levels in each task. These levels were scaled in a pilot study to yield measurably different error rates.

Each subject was seated in front of a 19-in color CRT display at a fixed viewing distance of 50 cm. The subject's head was stabilized to prevent movement by using a combined chin rest and head support. Eye fixation was monitored with an infrared closed-circuit television eye tracker/pupillometer system. Testing was performed in a quiet, electrically shielded room, and to prevent auditory distractions, the subject wore headphones and listened to white noise of 70 dB SPL. Ambient room illumination was 0.5 footcandle.

In all tasks, the display background was a neutral gray matched to the D6500 standard (Wyszecki & Stiles, 1982) with a luminance of 6.25 ft-L. A central crosshair and two white concentric rings (resembling radar range rings) were continuously displayed at a contrast of 66%. The distance between the center of the crosshair and the inner ring subtended 2 degrees of visual angle. The distance between the center of the crosshair and the outer ring subtended 4 degrees. Individual symbols (letters, digits, radar tracking symbols) subtended 42 min of arc. Symbol contrast and location were task-dependent (see below).

Performance was continuous within blocks ranging from 2.55–3.67 min in duration. Each block consisted of 50 or 72 trials of duration 2.1 s separated by a randomly varying intertrial interval with a mean duration of 955 ms and a range of 525–1384 ms. Task-relevant stimulus duration was 50 ms for the signal detection task and 200 ms for the other tasks. The interval between relevant stimuli had a mean of 3055 ms and a range of 2626 to 3484 ms.

Each task-relevant stimulus was followed by one of two irrelevant stimuli (probes) within the same trial. These probes consisted of an abrupt color change of the entire display background for a duration of 50 ms. The range rings and cross hair on the display remained visible. Frequent probes were white flashes occurring on 80% of the trials. Rare probes were green flashes occurring on 20% of the trials. The luminance of both flashes was 12.4 ft-L, or 0.3 log units higher than the background. The sequence of rare and frequent probes was randomized with the constraint that rare flashes could not occur consecutively. The interval between relevant and irrelevant stimuli varied uniformly around a mean of 1000 ms with a range of 526–1576 ms.

Subjects performed each task in three sessions held on separate days. The first session consisted of training, which familiarized the subjects with the tasks and served to stabilize mean reaction times (RT) and error rates. The stability criterion was less than a 10% change in performance across three consecutive blocks of trials. This typically required 8–14 blocks of trials. The second and third sessions were test sessions.

Each test session included one baseline block of trials, in which subjects observed the display without making any responses. In baseline blocks, the correct responses were undefined. The display and the stimuli used for the baseline condition were identical to those used for testing. However, only before testing blocks were subjects instructed how to respond. For detection responses, subjects used their right index and middle fingers to press telegraph keys labeled T and NT for target and nontarget stimuli. Detection responses were allowed any time up to 1800 ms after relevant stimulus onset but typically occurred no earlier than 250 ms. The computer recorded RT for these responses with a precision of 1 ms. Subjects used a three-button computer mouse with the left hand to signal other responses (confidence ratings, mental arithmetic) when required. Mouse-button responses were allowed any time up to the beginning of the next trial but the response times were not recorded.

At least nine blocks of trials were presented during each test session (one baseline block and eight test blocks), yielding at least 400 test trials plus 50 baseline trials per session. Target stimuli were variably mapped to responses in a manner that called

for controlled processing of the detection responses. The target/nontarget ratio was always 50/50 and the sequence was randomized. For all tasks, subjects were instructed to respond as quickly as possible without sacrificing accuracy.

2.3. Signal Detection Task

The display, input device configuration, symbols for task-relevant stimuli, and variable mapping strategy for the signal detection task are shown in Fig. 1. In each block of trials, a pair of triangles, with apexes either both up or both down, were presented at three different contrast levels: easy = 0.17, medium = 0.43, and hard = 0.53. Half the triangles of each contrast contained a small dot in the center, which subtended 4 min of arc. On each trial, the computer presented one of the six triangles for 50 ms at one of eight preset positions just outside the inner range ring. These positions correspond to radar bearing angles of 0, 45, 90, 135, 180, 225, 270 and 315 degrees or, equivalently, to four visual axes: horizontal, vertical, right oblique and left oblique. The sequence of triangles and positions was randomized and balanced within blocks.

Two responses were required. The subject first responded to symbols by using the right index finger to press either the T key or the NT key. RT was measured only

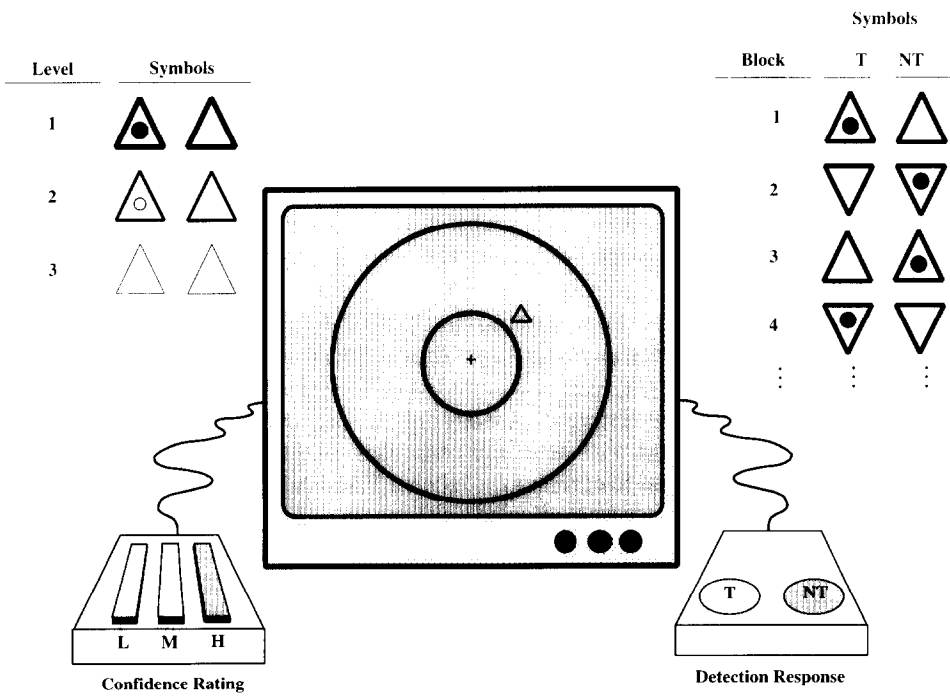

Fig. 1. Display, input device configuration, symbols for task relevant stimuli, and variable mapping strategy for the signal detection task.

for these detection responses. Immediately afterwards, the subject used the mouse with his left hand to provide a subjective three-point rating (low, medium, or high) of his confidence in the accuracy of his detection response. At the beginning of each block, the computer displayed a diagram to instruct the subject as to which set of triangles was the target set. In half the blocks, targets were the triangles with the dot. Triangles without the dot were targets in the other blocks. The association between dots and targets alternated on each successive block. Additionally, the orientation of the triangles (up or down) was alternated on each successive block.

2.4. Running memory task

The display, input device configuration, symbols for task-relevant stimuli, and variable mapping strategy for the running memory task are shown in Fig. 2. The relevant stimuli were six capital letters (B, F, H, J, N and X). As in the detection task, stimuli were presented parafoveally at one of eight bearing angles — however the stimulus duration was 200 ms. The stimuli were presented singly in a pseudo-random order. The objective of the task was to identify the match or mismatch of a letter that had just appeared on the screen with one that had appeared a number of trials back. The number of trials back that the subject had to remember ranged from one

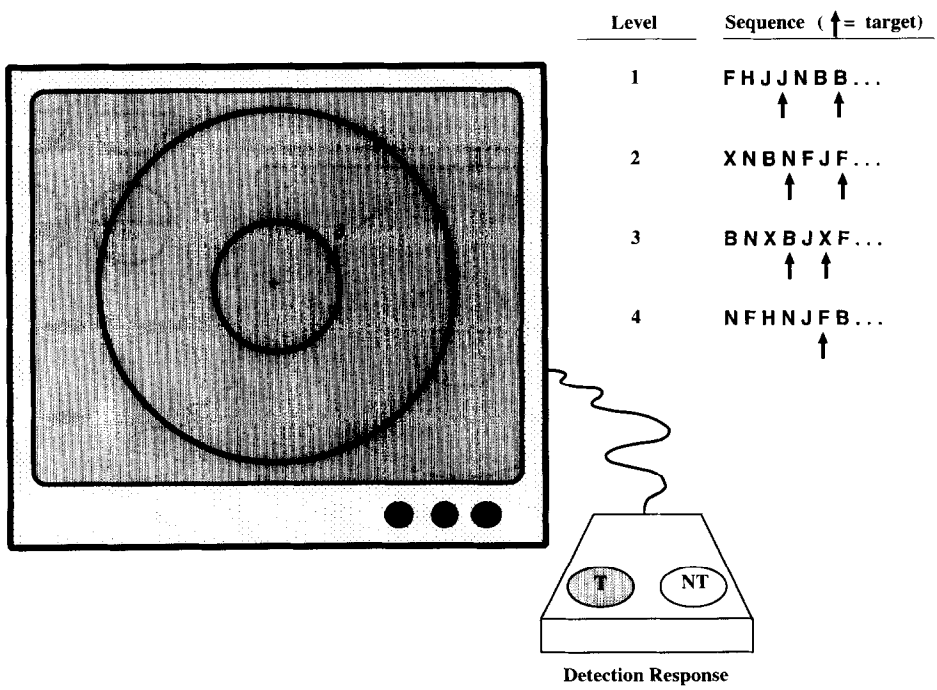

Fig. 2. Display, input device configuration, symbols for task relevant stimuli, and variable mapping strategy for the running memory task.

to four, depending on his ability, which was determined during training. Six subjects' maximum difficulty level was three-back. One subject's maximum level was four-back, and another was two-back. Subjects were tested at two adjacent difficulty levels: four blocks at an easy level (e.g., two-back) and four blocks at a difficult level (e.g., three-back). Only one response was required — the subject pressed T for a match or NT for a mismatch. Accuracy and RT were measured for each trial. The running memory task was the only one in which difficulty varied at the block level instead of the trial level, as in the other two tasks.

2.5. Computation task

The display, input device configuration, symbols for task-relevant stimuli, and variable mapping strategy for the computation task are shown in Fig. 3. The relevant stimuli were 15 pairs of numbers. On each trial a single pair was selected from a pseudo-random sequence and presented foveally for 200 ms in the form of a division problem. One number was above the other number, divided by a straight line. On half the trials the larger number was above the line, and on the other half, the larger number was below the line. Subjects were told how to respond at the beginning of each block. Each trial required two responses. The first was a detection response in

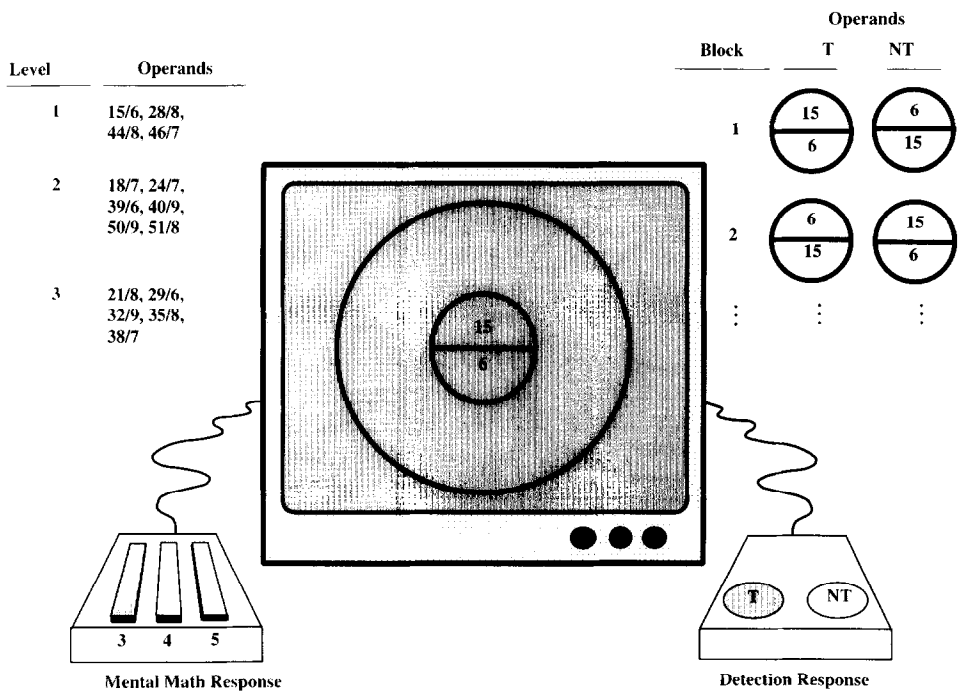

Fig. 3. Display, input device configuration, symbols for task relevant stimuli, and variable mapping strategy for the computation task.

which the subject pressed T for target or NT for nontarget. Targets were defined by the configuration of the number pairs: on odd blocks, the targets were defined by larger numbers on top, and on even blocks targets were defined by larger numbers below. RT was measured only for the detection responses. In addition to the detection response, subjects performed the division problem indicated by each number pair. They always had to divide the larger number by the smaller number (regardless of configuration) and report the remainder (always 3, 4, or 5) by pressing the appropriate button on the mouse. The three difficulty levels — easy, medium and hard, had four, six and five number pairs, respectively. The assignment of number pairs to difficulty levels was determined in pilot studies and from training data, and blocked as shown in Fig. 3.

2.6. Physiological recording

During each trial the EEG was recorded for 2 s, from 200 ms before the task-relevant stimulus (pre-stimulus) to 1800 ms afterward (post-stimulus). The EEG electrode sites were Fz, C3, C4, Cz, Pz, O1, O2 and the right mastoid or A2 (International 10–20 system, Jasper, 1958). All electrodes were referred to the left mastoid, A1. Recording was performed with tin electrodes embedded in a nylon cap (Electro-Cap International, Inc.), Grass amplifiers (Grass Model 12A, Neurodata Acquisition System), and a computer programmed to digitize and record single EEG epochs on a hard disk. During recording the signals were amplified 20 000 times, low-pass filtered with a cut-off of 100 Hz, and sampled at 1000 Hz.

The EOG was also recorded from two bipolar pairs of Ag-AgCl electrodes. One pair measured the vertical component between sites above and below the right eye. Another pair measured the horizontal component between sites that were about 2 cm lateral to the outer canthus of each eye.

2.7. Signal processing

Off line, the EEG epochs were decimated to a sampling rate of 500 Hz and filtered with a zero-phase digital filter with a cutoff frequency of 51 Hz (0 dB at 43 Hz, −89 dB at 60 Hz). The filtered epochs were then arithmetically re-referenced to average mastoids by synchronously subtracting half the amplitude of the A2-A1 recordings from the recordings for each of the scalp electrodes.

The re-referenced epochs were then corrected for artifacts produced by EOG blink and eye-movement potentials using a modification of the method described by Gratton, Coles, & Donchin (1983). The modification consisted of using the cross-correlation function between a half-cycle cosine blink template and the vertical EOG recordings to identify blink periods. Regions of the cross-correlation function that exceeded a threshold value of 0.3 corresponded to blinks, and were corrected with a separate propagation factor from other regions of the recording. Each single epoch was modeled as a linear sum of EEG, propagated blink potentials (when present), propagated non-blink vertical EOG potentials and propagated horizontal EOG potentials. A multiple linear regression procedure estimated the contributions of the

three artifact sources, which were then subtracted. Unlike the Gratton et al. (1983) procedure, no correction for average ERP propagation into the EOG recordings was applied. Effects of EOG correction were assessed by comparing average ERPs created from EOG-corrected epochs with average ERPs created from epochs free of any EOG artifacts greater than 50 μV relative to the median pre-stimulus voltage. No attenuation or distortion of ERP components due to the correction procedure was observed.

Each processed single epoch was displayed on a computer screen in a multi-electrode format and scored for remaining artifacts by trained observers. Such artifacts included amplifier saturation, electrode pops, large muscle or EKG activity, and the occurrence of eye movements or blinks during the presentation of the relevant or irrelevant stimuli. A 50-μV criterion, relative to the median pre-stimulus amplitude, was used to reject spuriously large voltages. Only single epochs that were free of discernible artifacts were included in the analyses.

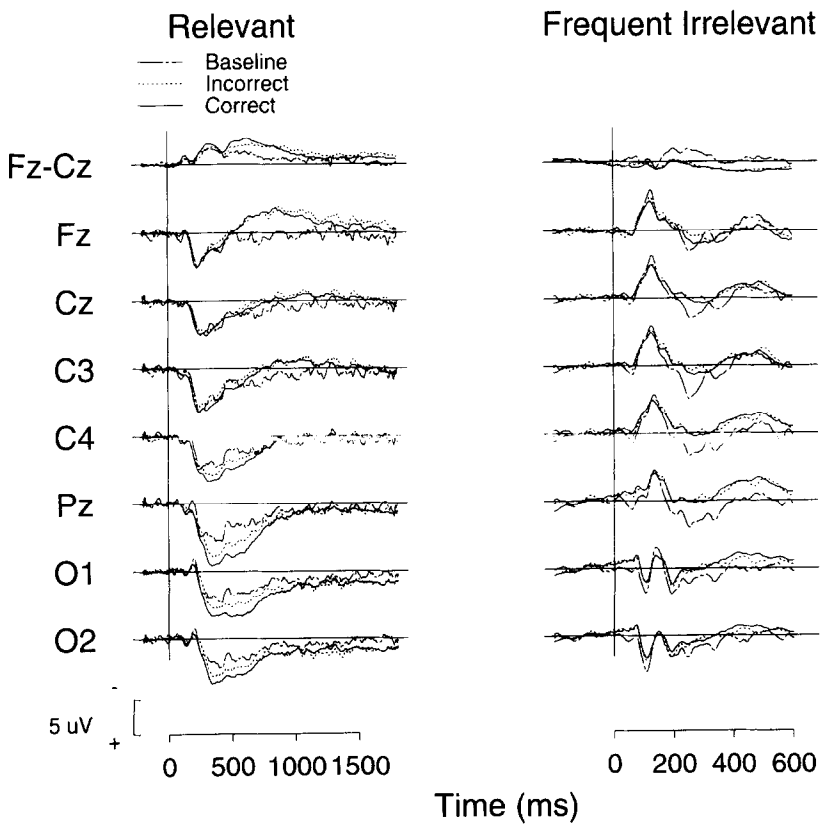

Fig. 4. Grand average ERPs for task relevant and frequent irrelevant probe stimuli in the signal detection task. Electrode Fz-Cz was a bipolar derivation. All others were referred to average mastoids. Separate averages are plotted in different line styles for three task performance conditions: baseline (–·–·–), incorrect response trials (– – –), and correct response trials (——). Values on the ordinate are in μV, as shown by the scale bar (negative is up). Values on the abscissa represent time after stimulus onset in ms.

2.8. ERP measurement

Grand average ERPs were computed for each electrode site as a function of stimuli, task conditions and detection response accuracy (Figs. 4, 5, & 6). Separate averages were computed for the baseline condition, and for correct-response trials and incorrect-response trials in the task performance conditions, irrespective of task factors such as difficulty, target or position. These averages are not the object of the regression analyses we will describe below. Instead they serve to define the ERP structure for measurement and interpretation of effects.

Since our goal was to build empirical regression models that account for performance in terms of ERP data, we took a liberal approach to defining the features of the ERP to be measured. From grand averages and single-subject averages, we chose a set of intervals (windows) from multiple electrode sites for ERP measurements (Table 1). Window selection was guided by two criteria: (a) selected windows cor-

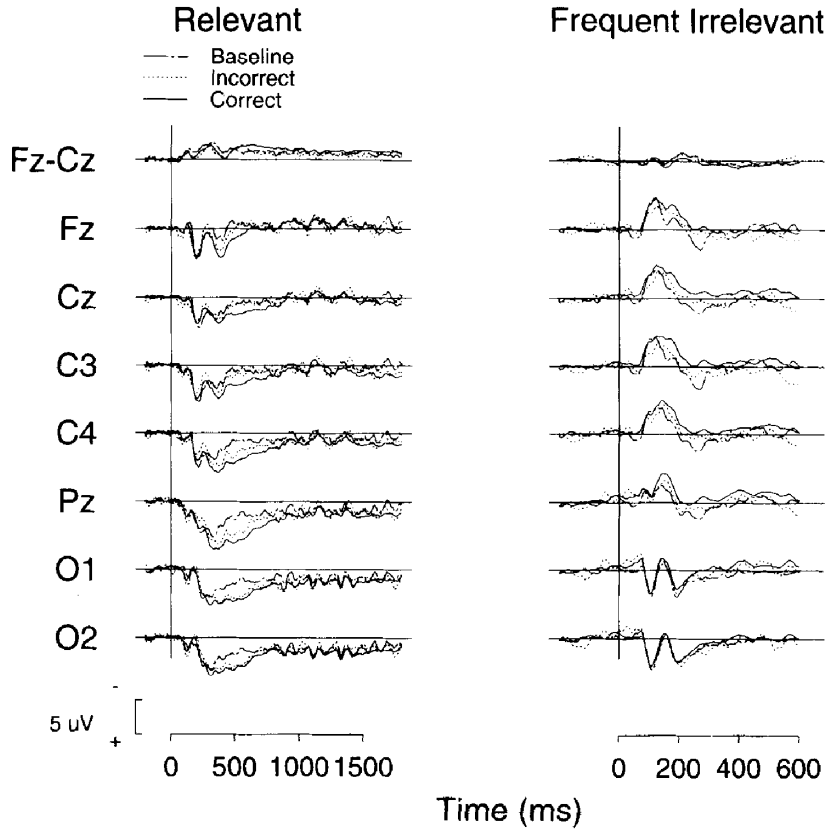

Fig. 5. Grand average ERPs for task-relevant and frequent irrelevant-probe stimuli in the running memory task. See Fig. 1 for details.

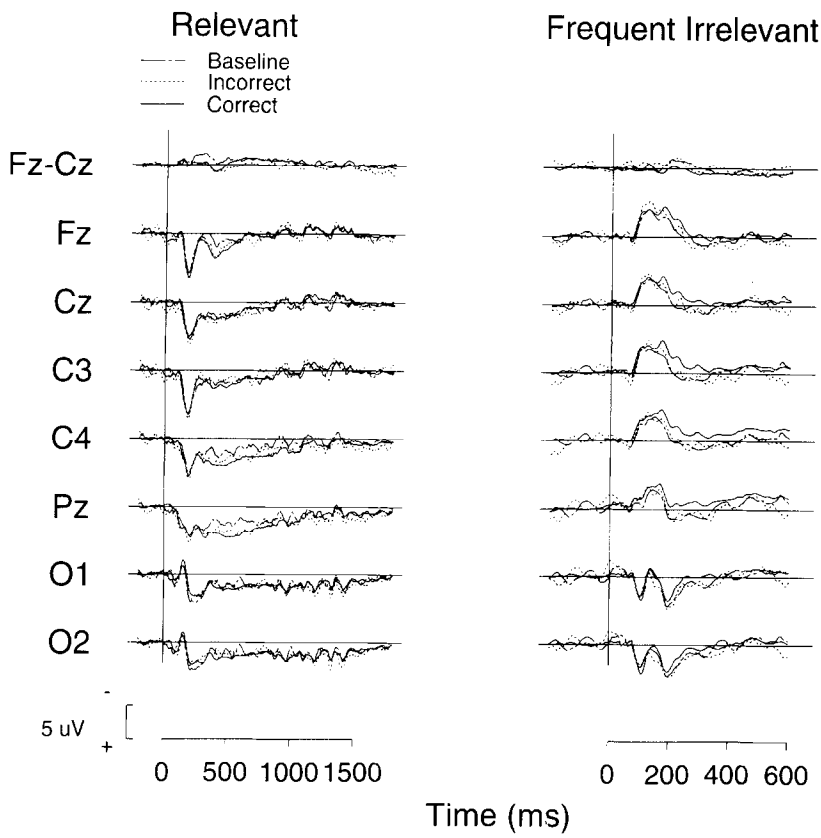

Fig. 6. Grand average ERPs for task-relevant and frequent irrelevant-probe stimuli in the computation task. See Fig. 1 for details.

responded to known ERP component latency and scalp topography or (b) selected windows exhibited a clear peak or slow wave across several subjects and tasks. The rationale for matching the measures in Table 1 to ERP components and peaks is detailed below.

2.9. Relevant stimulus ERP features

Across tasks, the most prominent feature of the relevant-stimulus ERP averages was a centro-parietal positive wave beginning about 200 ms post-stimulus and continuing for up to 860 ms. This component tended to be smallest during the baseline conditions, and largest for trials in which correct responses were made in the active conditions. For the signal detection and running memory tasks, the scalp topography of the middle phase of this component matched that of the P300. For the computation task, this component was less sharply defined and mainly appeared to reflect slow wave activity. We measured this component five ways.

Table 1
ERP windows measured[a] in all tasks

Window name	Electrode site	Polarity	Measurement interval (ms)		
			Start	End	Center
Relevant stimuli					
PRE-STIM	Pz	+	−150	0	−75
FW1	Fz-Cz	−	80	210	145
FW2	Fz-Cz	−	210	360	285
FW3	Fz-Cz	−	460	710	585
FW4	Fz-Cz	−	710	1710	1210
N1	Fz	−	80	210	145
P2A	Fz	+	190	290	240
N2	Fz	−	250	400	325
SW1	Fz	−	550	1550	1050
P3A	Fz	+	350	600	475
P3B	Pz	+	280	480	380
P3C	Pz	+	460	860	660
SW2	Pz	+	860	1760	1310
P1	O2	+	80	180	130
P2B	O2	−	150	250	200
Irrelevant stimuli					
FW1	Fz-Cz	+	100	200	150
FW2	Fz-Cz	+	200	600	400
N1	Cz	−	75	175	125
N2	C3	−	200	300	250
P3	Pz	+	200	350	275
SW1	Pz	−	200	600	400
SW2	Pz	−	450	550	500
P1	O2	+	80	180	130
P2	O2	+	170	370	270

[a]ERP measures: 1, baseline to peak amplitude; 2, peak latency; 3, average amplitude; 4, root mean square amplitude.

1. P3A corresponds to the positive deflection at electrode Fz between 350–600 ms which is seen most clearly in the ERPs for the running memory task.

2. P3B designates the positive deflection seen at electrode Pz between 280–480 ms which is best seen in the ERPs for the signal detection and running memory tasks.

3. P3C corresponds to the second hump of the positive deflection at Pz which follows P3B and is apparent (but not sharply defined) in the ERPs for all tasks.

4. SW1 measures the slow negative wave that is maximal at Fz between 550–1550 ms. It is clearly defined only in the ERPs for the signal detection task.

5. SW2 measures the positive DC shift seen over posterior electrodes between 860–1760 ms.

The occipital electrodes showed an early positive deflection about 130 ms post-stimulus which we designated as P1. Since there was no indication of hemispheric asymmetry for P1, we measured it at O2. The P1 was more pronounced in the run-

ning memory and computation tasks than in the signal detection task, presumably due to the longer stimulus duration (200 ms) employed for these tasks than in the signal detection task (50 ms). The P1 was followed by a positive deflection at about 200 ms, which we designated as P2B. This peak was most sharply defined at electrode O2 in the ERPs for the signal detection task, but for all tasks it merged with the ensuing P300-slow-wave complex. At more frontal locations (e.g., at Cz and Fz) a sharp positive deflection with a latency of about 240 ms was observed in the ERP averages for all tasks. This peak, which we designated as P2A, was clearly present at Fz for all three tasks.

The ERP at Fz also contained negative-going peaks centered at about 145 ms and 325 ms immediately preceding and following the P2A, respectively. These peaks, which we designated as N1 and N2, are best seen at Fz in the ERPs for the running memory task (Fig. 5).

For compatibility with an earlier study (Trejo et al., 1990), ERPs for a bipolar derivation, Fz-Cz, were computed off-line and also averaged (Figs. 4–6). The Fz-Cz derivation showed three negative peaks for relevant stimuli in the range between 0–600 ms, and a slow long-lasting negativity between 600–1800 ms in some conditions. We refer to these areas of the ERP underlying these peaks as 'frontal windows,' which we designated as FW1 (145 ms), FW2 (285 ms) and FW3 (585 ms). We refer to the slow DC shift at Fz-Cz between 710–1210 ms as FW4.

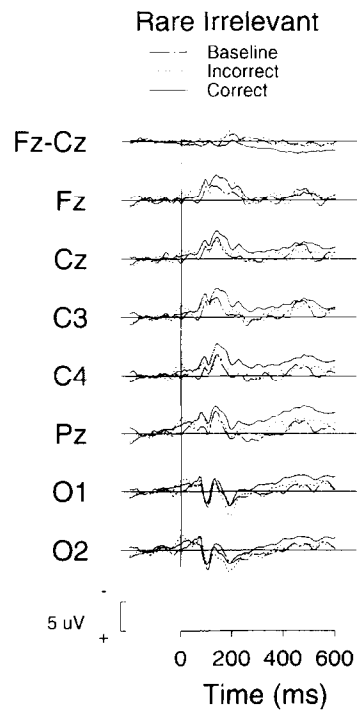

Fig. 7. Grand average ERPs for rare irrelevant-probe stimuli in all tasks. See Fig. 1 for details.

2.10. Irrelevant probe ERP features

In all tasks, the average frequent and rare irrelevant-probe ERPs contained an occipital P1-N1-P2 complex and a frontal N1 (Figs. 4–6, Fig. 7). The rare irrelevant-probe ERP averages were derived from relatively few ERPs and have a correspondingly low SNR. So we averaged the rare irrelevant-probe ERP averages across tasks using weights proportional to the number of trials averaged in each task. The rare irrelevant-probe ERP averages for the baseline period suggested a small P300 with a maximum at Pz. There were no obvious and consistent differences among baseline, correct or incorrect averages in the early components of the irrelevant-probe ERP averages. However, a centro-parietal slow wave between about 200–600 ms (designated SW1) provided some differentiation among the averages. Correctly processed trials resulted in average ERPs that were more negative in this latency range than incorrect or baseline trials. In addition, we defined two measurement windows at the Fz-Cz derivation, FW1 (150 ms) and FW2 (400 ms). These windows were also included for compatibility with the Trejo et al. (1990) study.

2.11. ERP measures

The set of windows listed in Table 1 covers traditional ERP components such as the N1 and P300, plus some novel deflections such as the negative peaks present in the relevant-stimulus ERPs at the Fz-Cz derivation (FW1, FW2 and FW3). Four measurements were made in each window: (a) maximum baseline-to-peak amplitude (AMP), (b) latency at maximum baseline-to-peak amplitude (LAT), (c) average amplitude (AVG) and (d) root-mean-square amplitude (RMS; Trejo, 1988). For the AMP and LAT measures, a polarity was pre-defined for each window based on the polarity expected from the average ERP waveforms.

3. Results

3.1. Multivariate behavioral analyses

We performed multivariate analyses of variance (MANOVAs) on the performance data for each of the tasks (Table 2). In all tasks, factors included test session, stimulus difficulty and target. In the signal detection task and the running memory task, stimulus position was also entered as a factor. In order to conserve degrees of freedom, position was entered as a four-level factor, coded by the axis along which a symbol appeared (vertical, horizontal, right oblique or left oblique). Dependent behavioral measures included accuracy and RT in all tasks. Effects on confidence ratings in the signal detection task and mental math accuracy in the computation task were also tested. The degrees of freedom were corrected for violations of the sphericity assumption where appropriate (Geisser & Greenhouse, 1958).

There was no main effect of session in any task, indicating that the training had led to stable performance levels. As expected, stimulus difficulty significantly affected performance in all tasks. The target/nontarget distinction affected perfor-

Table 2
MANOVA effects of task variables on performance measures

Source	Signal detection		Running memory		Computation	
	df[a]	*F*	*df*	*F*	*df*	*F*
Session (S)	3, 5	4.91	2, 6	1.75	3, 4	0.50
Difficulty (D)	6, 24	7.79***	2, 6	5.48*	6, 24	14.54***
Target (T)	3, 5	11.87*	2, 6	5.12	3, 5	3.50
Position (P)	9, 46.39	2.43*	6, 40	3.00*	—	—
S × D	6, 24	1.44	2, 6	0.97	6, 20	0.50
S × T	3, 5	0.08	1, 7	0.00	3, 4	1.53
S × P	9, 46.39	0.58	6, 40	2.06	—	—
D × T	2, 28	13.65***	2, 6	0.00	6, 24	1.83
D × P	18, 113.62	2.00*	6, 40	1.49	—	—
T × P	9, 46.39	2.91**	6, 40	0.86	—	—
S × D × T	6, 24	3.45*	1, 7	0.49	6, 20	8.53***
S × T × P	9, 46.39	0.62	6, 40	1.78	—	—
S × D × P	18, 113.62	1.51	6, 40	0.91	—	—
D × T × P	18, 113.62	0.71	6, 40	0.90	—	—
S × D × T × P	18, 113.62	0.78	6, 40	0.79	—	—

[a]Fractional *df* indicate that Geisser-Greenhouse corrections were applied.
*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

mance only in the signal detection task. Where applicable, as in signal detection and running memory, stimulus position also affected performance. Several two-way interactions were significant in the signal detection task, including difficulty × target, difficulty × position, and target × position. A three-way interaction of session × difficulty × target was significant in the signal detection and computation tasks. Due to the small number of subjects, we did not pursue further analyses of the significant interactions.

3.2. Univariate behavioral analyses

3.2.1. Signal detection task

Univariate analyses of variance were performed on each dependent performance measure for the main effects and two-way interactions indicated by the MANOVAs. In the signal detection task (Table 3, Fig. 8), the main effect of stimulus difficulty resulted in significant decreases in accuracy and confidence ratings and increases in RT. The main effect of target was significant for RT and confidence but not for accuracy. Targets produced significantly lower mean RT and higher mean confidence ratings than nontargets. There was also a significant main effect of position on accuracy and RT (Fig. 9). For both measures, performance was best (high accuracy, low RT) for stimuli presented on the horizontal axis, worst for the vertical axis, and intermediate for the oblique axes. The means were in the same direction for mean confidence ratings (Fig. 9), but the effect failed to reach significance.

Table 3
ANOVA effects of task variables on signal detection performance measures

Source		Accuracy	Reaction time	Confidence
	df	*F*	*F*	*F*
Difficulty (D)	2, 14	30.5***	16.92***	6.70**
Target (T)	1, 7	0.48	41.70***	24.10**
Position (P)	3, 21	7.56**	3.96*	1.54
D × T	2, 14	20.3***	0.00	0.00
D × P	6, 42	3.42**	4.67***	1.22
T × P	3, 21	3.42*	5.14**	3.52*

*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

The difficulty × target interaction indicated by the MANOVA was significant only for accuracy (Fig. 8). At low difficulty, target accuracy was higher than nontarget accuracy, but the reverse occurred at high difficulty. This reversal did not appear in the confidence or RT data. The position × difficulty interaction indicated

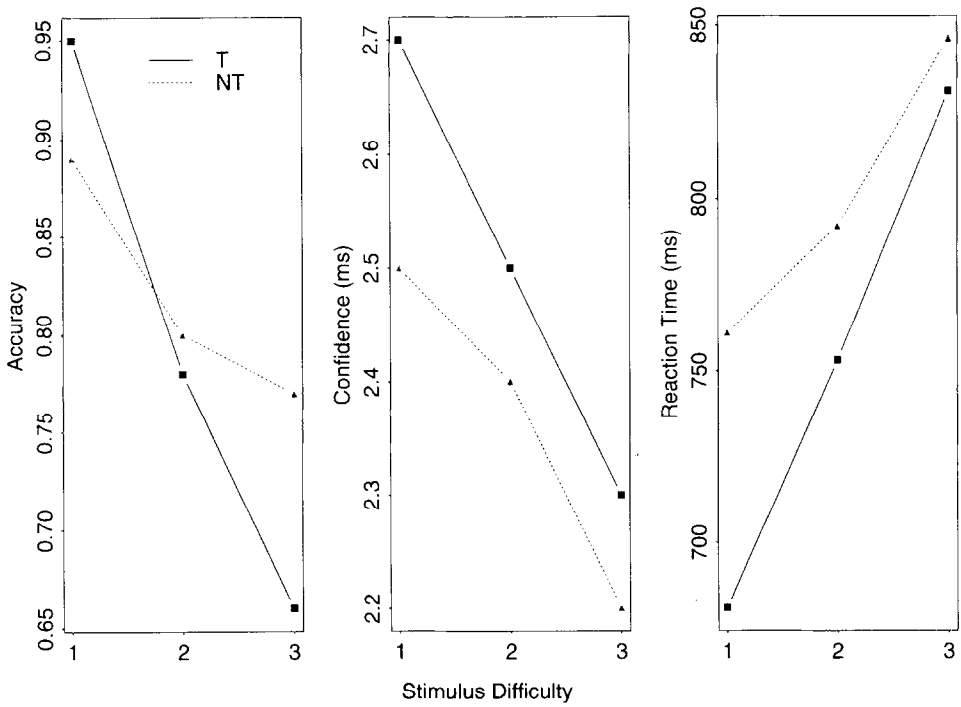

Fig. 8. Effects of stimulus difficulty on univariate performance measures in the signal detection task for target and nontarget stimuli. Left, accuracy (proportion correct); middle, confidence (three-point scale); right, RT for detection response (ms).

by the MANOVA was significant for accuracy and RT (Fig. 9). For both measures, the difference between the horizontal and vertical axes increased with difficulty. In addition, the advantage of the left oblique axis versus the right oblique axis at the low difficulty level was reversed at high difficulty. Again, the mean confidence ratings were in the same direction as these effects, but failed to reach significance.

The target × position interaction indicated by the MANOVA was significant for all three measures (Fig. 10), leading to a complex set of effects. For both the accuracy measure and RT, the effect of stimulus position was greater for targets than for nontargets. RT for targets on the vertical axis was about 50 ms higher than RT for targets on the horizontal axis. However, for nontargets there was only about 25 ms difference between mean RT values on the vertical and horizontal axes. The confidence rating means for targets and nontargets interacted differently with the position factor than did the means for accuracy and RT. Firstly, the position effects for targets and nontargets are about the same size. Secondly, the differences between horizontal and vertical axes are clearly in opposite directions for targets and nontargets,

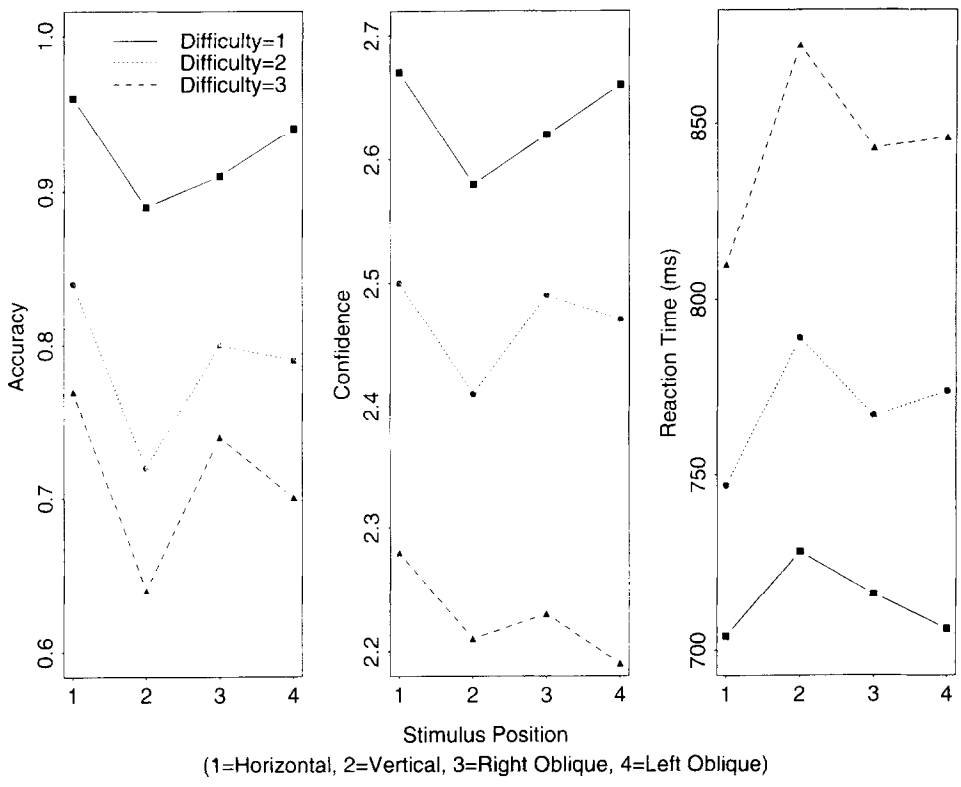

Fig. 9. Effects of stimulus position on univariate performance measures in the signal detection task as a function of stimulus difficulty. Left, accuracy (proportion correct); middle, confidence (three-point scale); right, RT for detection response (ms).

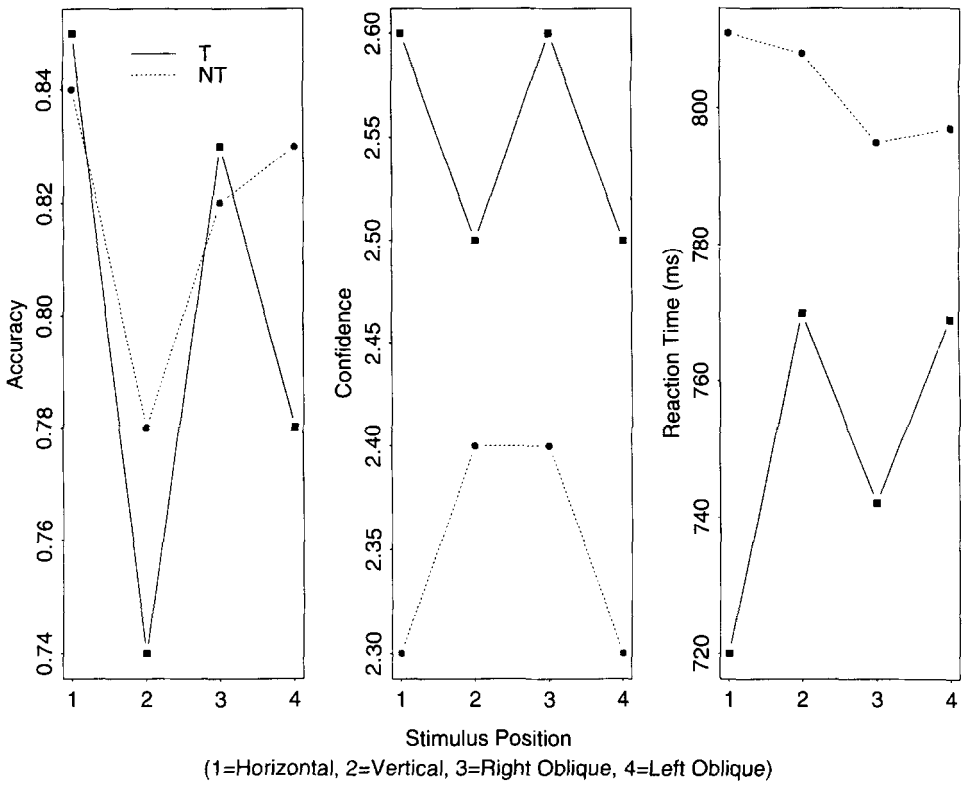

Fig. 10. Effects of stimulus position on univariate performance measures in the signal detection task for target and nontarget stimuli. Left, accuracy (proportion correct); middle, confidence (three-point scale); right, RT for detection response (ms).

i.e., confidence at horizontal was greater than confidence at vertical for targets, but just the reverse was true for nontargets. Interestingly, there appears to be a performance advantage for targets on the horizontal meridian as compared to the vertical meridian for all three measures. The same was not true for nontargets, where the sign of the horizontal-vertical differences varied among the measures.

3.2.2. Running memory task

In the running memory task the main effect of stimulus difficulty indicated by the MANOVA was significant for both accuracy, $F[1, 7] = 10.15, p < 0.0154$, and RT, $F[1, 7] = 8.23, p < 0.0240$. Accuracy was higher for difficulty level one than for level two and, correspondingly, RT was lower for level one than for level two (Fig. 11).

The effect of position indicated by the MANOVA was significant for RT (Fig. 11), $F[3, 21] = 3.08, p < 0.0496$. As in the signal detection task (Fig. 9), mean RT was highest for stimuli presented on the vertical meridian, lowest on the horizontal meridian, with the oblique meridians falling in between. Although the data suggest a po-

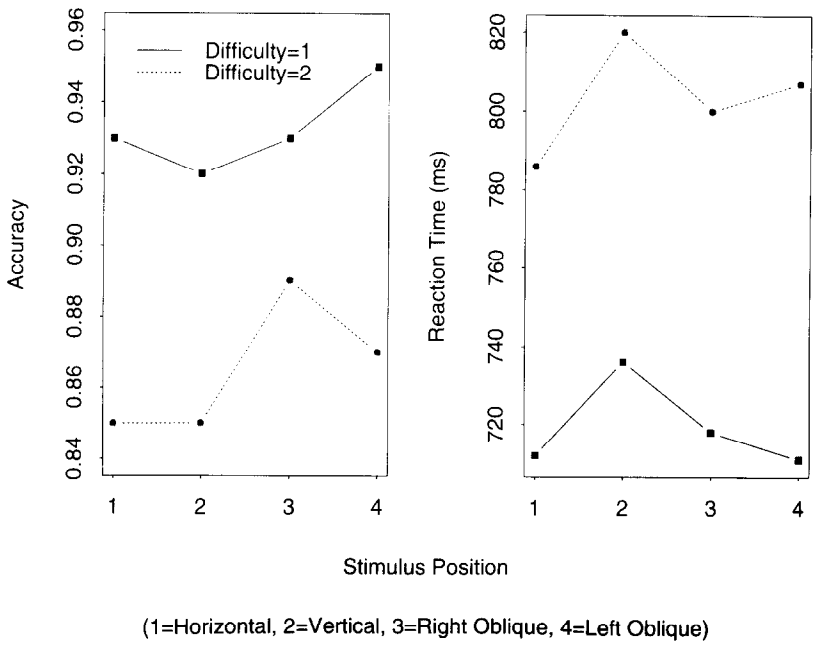

Fig. 11. Effects of stimulus position on univariate performance measures in the running memory task as a function of stimulus difficulty. Left, accuracy (proportion correct); right, RT for detection response (ms).

sition × difficulty interaction (Fig. 11), no significant two-way interactions were indicated by the MANOVA results for the running memory task.

3.2.3. Computation task

In the computation task, the main effect of difficulty indicated by the MANOVA was significant for mental arithmetic accuracy $F[2, 14] = 55.69$, $p < 0.0001$. Mean accuracies for the easy, medium and hard levels were 0.47, 0.36 and 0.22, respectively. Difficulty did not significantly affect detection response accuracy (for target/non-target configurations) or the RT for that response. No significant two-way interactions were indicated by the MANOVA.

3.3. Factor analyses

In order to summarize performance and provide a general measure that could be correlated with ERP indices, a factor analysis of the dependent measures was performed. Using the principal components method, factors were extracted from the correlation matrix of the single-trial performance measures for all subjects within each task (Table 4). In the signal detection and running memory tasks, a single unrotated factor accounted for a majority of the variance in the single-trial perfor-

Table 4
Factor analyses of task performance measures

Factor	Statistics		Factor pattern		
	Eigenvalue	Proportion of variance	Accuracy	Confidence	Reaction time
Signal detection					
1	1.55	0.52	0.52	0.81	−0.79
2	0.89	0.29	0.85	−0.21	0.34
3	0.56	0.19	−0.07	0.55	0.51

Factor	Statistics		Factor pattern	
	Eigenvalue	Proportion of variance	Accuracy	Reaction time
Running memory				
1	1.19	0.59	0.77	−0.77
2	0.38	0.41	−0.64	−0.64

Factor	Statistics		Factor pattern		
	Eigenvalue	Proprotion of variance	Math accuracy	Detection accuracy	Reaction time
Computation					
1	1.19	0.40	0.77	−0.02	−0.77
2	1.00	0.33	−0.13	−0.99	−0.10
3	0.81	0.27	−0.63	0.14	−0.63

mance data. In the computation task, two factors were required to explain more than 50% of the variance. However, the second factor loaded almost exclusively on the detection accuracy variable which the ANOVAs had shown to be insensitive to manipulations of task variables. For this reason, only the first factor from the computation task was used as a global performance measure.

Primary factors in all tasks exhibited some similarity in structure. Each factor weighted accuracy measures positively and RT negatively, indicating that speed and accuracy were positively correlated. The confidence rating in the signal detection task was also weighted positively, indicating a correlation with accuracy. Thus, high scores on the primary factor in that task indicated fast, accurate and confident responses. In the other two tasks, high scores on the primary factor indicated fast and accurate responses. For the computation task, however, only math accuracy (not detection accuracy) affected primary factor scores significantly. The negative loading for detection response accuracy in the first factor suggests that math accuracy was compromised when subjects took too long to make the detection response, perhaps due to the forced pace of the task. Primary factors for each task were validated by performing the factor analyses within subjects and comparing the within-subjects factor loadings to the across-subjects loadings. In the signal detec-

tion and running memory tasks, primary factor patterns for all subjects matched the across-subjects pattern. In the computation task, the primary factor pattern for six of the eight subjects matched the across-subjects pattern. Of the two subjects who differed from the group, one (#7) showed a higher loading for detection accuracy than for math accuracy. The other (#4) showed nearly equal loadings for both accuracy measures.

Standardized scoring coefficients for the primary performance factor were computed and applied to the single-trial data for all three tasks. (Due to the good agreement among six of the eight subjects in the computation task, the across-subjects primary factor was applied to all subjects' data for that task also.) The result was a single measure, Performance Factor 1 (*PF*1), which reflected overall performance quality for any single trial. The equations defining *PF*1 were:

$$PF1_S = 0.33 \text{ Accuracy} + 0.53 \text{ Confidence} - 0.51 \text{ RT} \quad (1)$$

$$PF1_M = 0.65 \text{ Accuracy} - 0.65 \text{ RT} \quad (2)$$

$$PF1_C = 0.65 \text{ Math Accuracy} - 0.02 \text{ Detection Accuracy} - 0.51 \text{ RT} \quad (3)$$

where the subscripts S, M, and C, denote signal detection, running memory, and computation, respectively.

3.4. Sensitivity of factor scores to task variables

As for the raw behavioral data (accuracy, RT, etc.), repeated-measures ANOVAs were performed for each task to determine the sensitivity of PF1 factor scores to task variables. In the signal detection task (Table 5), significant main effects included difficulty, target and position. These were similar to the effects on the raw performance data. Performance, as measured by PF1, decreased almost linearly with task difficulty (easy = 0.31, medium = 0.00 and hard = −0.35) and was higher for targets (0.07) than nontargets (−0.09). PF1 was best for stimuli on the horizontal meridian (0.09),

Table 5
ANOVA effects of task variables on signal detection performance Factor 1

Source	*df*	*F*
Difficulty (D)	2, 14	16.61***
Target (T)	1, 7	27.65**
Position (P)	3, 21	5.49**
D × T	2, 14	2.77
D × P	6, 42	4.60**
T × P	3, 21	5.81**

*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

worst on the vertical meridian (−0.12) and intermediate for the oblique meridians (left = 0.00, right = −0.01).

3.4.1. Signal detection task

Of the three two-way interactions that were significant in the analyses of the raw signal detection performance measures, only two were significant for PF1 (Table 5, Fig. 12). The difficulty × position interaction indicated an increase in the effect of the position factor with increasing stimulus difficulty. PF1 was nearly level across positions for easy targets. For more difficult targets, PF1 was higher on the horizontal meridian than on vertical meridian, with the oblique axes having intermediate PF1 scores. A similar increase of position effects with increasing difficulty occurred in the raw performance data (Fig. 9). The target × position interaction (Fig. 12) indicated weak or absent position effects for nontargets, and larger position effects for targets. The position effect for targets resembled that of difficult stimuli in general with best performance on horizontal axes, poorest performance on vertical axes and intermediate performance on oblique axes.

3.4.2. Running memory task

In the running memory task, there was a significant main effect of stimulus difficulty on PF1, $F[1, 7] = 12.21$, $p < 0.0101$. Mean values of PF1 for easy and difficult memory problems were −0.18 and 0.18, respectively. No other main effects

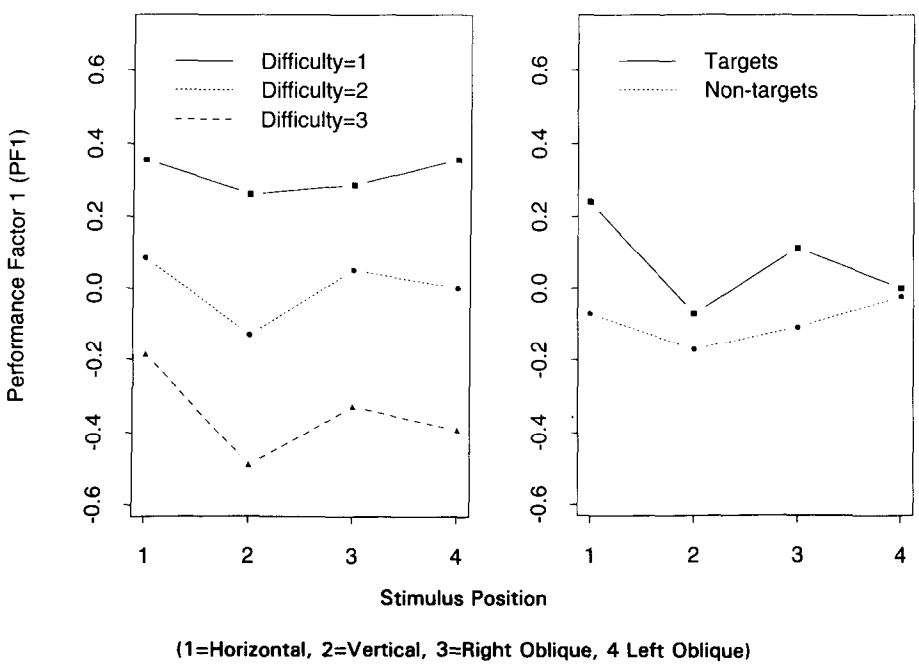

Fig. 12. Effects of stimulus position on PF1 performance measure in the signal detection task as a function of stimulus difficulty (left) and for target and nontarget stimuli (right).

or two-way interactions were significant. However, a three-way interaction of session × target × position was significant $F[3, 21] = 3.16$, $p < 0.0462$.

3.4.3. Computation task

In the computation task, the main effect of difficulty on PF1 was significant, $F[2, 14] = 28.32$, $p < 0.0001$, in agreement with the MANOVA results for the raw performance data. However, unlike the MANOVA results, the main effect of target was also significant for PF1, $F[1, 7] = 11.87$, $p < 0.0108$. Mean PF1 values for targets at the easy, medium, and hard levels were 0.27, 0.02, and −0.14, respectively. The corresponding PF1 values for nontargets were 0.09, −0.05, and −0.18, respectively. Although the nonsignificant MANOVA results precluded a significance test for the target main effect in the ANOVAs on the raw performance measures, we did observe significant F ratios for both mental math accuracy and reaction time and a nonsignificant low F ratio for detection accuracy. Thus, the large loadings for mental math accuracy and reaction time on PF1 (Table 4) led to the significant main effect of the target factor on PF1. In addition, there was a significant two-way interaction of difficulty × target, $F[2, 14] = 6.76$, $p < 0.0332$. This was expressed as a limiting effect of difficulty on the target-related differences in PF1: at the easy difficulty level, mean PF1 values for targets were 0.18 higher than nontargets, but this difference diminished to 0.07 at medium difficulty and to 0.04 at the hard difficulty level.

3.5. Effects of task performance conditions on ERP amplitudes

To identify major sources of ERP variance related to task performance conditions, we performed a limited analysis of the ERP measures, which focused on the average amplitude measure (AVG) for each feature of the relevant- and irrelevant-probe ERPs (Table 1). For each feature, a repeated-measures one-way ANOVA was performed with condition as the factor. Condition had three levels. The first level corresponded to the baseline blocks of trials. The second and third levels corresponded to the incorrect-response trials and correct-response trials in the active task performance blocks. Detection response accuracy was used to assign trials to conditions for the signal detection and running memory tasks. Math accuracy was used to assign trials to conditions in the computation task. We also performed two planned comparisons of the means as follows: baseline-active (B-A), where 'active' is the mean of the correct and incorrect levels, and incorrect-correct (I-C).

3.5.1. Signal detection task

For the relevant-stimulus ERPs, three features depended on the condition factor in the signal detection task: P3B-AVG, P3C-AVG, and SW1-AVG (Table 6). For the P3B-AVG measure, the main effect of condition was significant. Mean P3B-AVG was 4.04 μV in the baseline condition, 6.86 μV in the incorrect condition and 8.25 μV in the correct condition. The B-A and I-C differences were both significant.

For the P3C-AVG measure, the main effect of condition was significant. Mean P3C-AVG was 2.29 μV in the baseline condition, 4.33 μV in the incorrect condition

Table 6
Effects of condition on ERP average-amplitude measures

Condition[a]		Signal detection		Running memory		Computation	
	df	*F*	ϵ^b	*F*	ϵ	*F*	ϵ
Task-relevant stimulus ERPs							
P3B							
Overall	2, 14	8.27*	0.63	13.02**	0.85	—	—
B-A	1, 7	8.41*	—	27.29**	—	—	—
I-C	1, 7	7.26*	—	—	—	—	—
P3C							
Overall	2, 14	8.47*	0.62	14.73***	0.91	—	—
B-A	1, 7	8.57*	—	34.52***	—	—	—
I-C	1, 7	7.92*	—	—	—	—	—
SW1							
Overall	2, 14	7.03*	0.55	—	—	—	—
B-A	1, 7	7.51*	—	—	—	—	—
Irrelevant-probe ERPs							
FW2							
Overall	2, 14	5.48*	0.78	—	—	—	—
B-A	1, 7	7.23*	—	—	—	—	—
N2-AVG							
Overall	2, 14	4.24*	0.97	7.59*	0.71	—	—
B-A	1, 7	7.07*	—	16.68**	—	—	—
P3-AVG							
Overall	2, 14	10.49**	0.83	—	—	4.42*	0.99
B-A	1, 7	14.63**	—	—	—	—	—
I-C	1, 7	—	—	—	—	6.80*	—
SW1-AVG							
Overall	2, 14	11.30**	0.79	—	—	—	—
B-A	1, 7	14.85**	—	—	—	—	—
SW2-AVG							
Overall	2, 14	7.22*	7.22	—	—	—	—
B-A	1, 7	9.27*	—	—	—	—	—

[a]B — A: comparison of means in the baseline and active conditions; I — C: comparison of means for correct- and incorrect-response trials in the active conditions.
[b]Geisser-Greenhouse epsilon.
*$p < 0.05$
**$p < 0.01$
***$p < 0.001$

and 5.56 μV in the correct condition. As for P3B-AVG, the B-A and I-C differences for P3C-AVG were both significant.

For the SW1 measure (a late frontal negative slow wave), the main effect of condition was significant. Mean SW1-AVG was 0.35 μV in the baseline condition, −2.39 μV in the incorrect condition and −1.95 μV in the correct condition. Unlike the two P3 measures, the I-C difference was not significant for SW1-AVG. Only the B-A difference was significant.

For the irrelevant-probe ERPs, five average amplitude measures were sensitive in the condition factor in these analyses. There was a significant main effect of condition on FW2-AVG. Mean FW2-AVG was −0.49 μV in the baseline condition, 0.93 μV in the incorrect condition and 1.06 μV in the correct condition. Of the two comparisons, only the B-A difference was significant.

There was a significant main effect of condition on N2-AVG. Mean N2-AVG was 1.51 μV in the baseline condition, −0.25 μV in the incorrect condition and 0.05 μV in the correct condition. Only the B-A comparison was significant.

The main effect of condition on P3-AVG was significant. Mean P3-AVG was 2.40 μV in the baseline condition, −0.31 μV in the incorrect condition and −0.03 μV in the correct condition. Only the B-A comparison was significant.

The main effect of condition on SW1-AVG was also significant. Mean SW1-AVG was 0.93 μV in the baseline condition, −1.46 μV in the incorrect condition and −1.52 μV in the correct condition. Only the B-A comparison was significant.

Finally, the main effect of condition was significant for SW2-AVG. Mean SW2-AVG was −0.39 μV in the baseline condition, −2.69 μV in the incorrect condition and −2.85 μV in the correct condition. Again, only the B-A comparison was significant.

3.5.2. Running memory task

The main effects of condition on two relevant-stimulus ERP features, P3B-AVG and P3C-AVG, were significant in the running memory task. Mean P3B-AVG was 3.78 μV in the baseline condition, 5.32 μV in the incorrect condition and 6.37 μV in the correct condition. Mean P3C-AVG was 1.90 μV in the baseline condition, 3.25 μV in the incorrect condition and 4.33 μV in the correct condition. The pattern of results was the same as that observed in the signal detection task. However, unlike the signal detection task, only the B-A differences were significant.

For the irrelevant-probe ERPs, only one feature, N2-AVG, was significantly affected by the condition factor in the running memory task. Mean N2-AVG was 1.33 μV in the baseline condition, 0.17 μV in the incorrect condition, and −1.06 μV in the correct condition. Only the B-A difference was significant.

3.5.3. Computation task

No effects were observed for AVG measures of the relevant-stimulus ERP features. Only the P3-AVG measure for the irrelevant-probe ERPs was significantly affected by the condition factor. Mean P3-AVG was 1.00 μV in the baseline condition, 0.87 μV in the incorrect condition, and −0.84 μV in the correct condition. Surprisingly, in this case, only the I-C difference was significant.

3.6. Regression analyses

We performed a complete analysis of single-trial regression models, but we concluded that the proportions of variance explained by the models and the SNR of the

ERP measures were too low to be of practical value.[1] To increase the SNR of the ERPs, we applied a running-mean process to the series of single-trial ERPs in each block of trials for all subjects. We constrained the process to average ERPs over a window that contained a maximum of 10 trials (mean interval of 3.1 s, range 2.6–3.5 s). Thus, the process replaced each ERP with an average ERP based on itself and the ERPs from the preceding nine trials. The first nine of each block were omitted. A minimum of seven artifact-free ERPs in each window were included in each ERP average. When the 10-trial window contained more than three ERPs with artifacts, no average ERP was computed, a gap was left in the sequence of running-mean ERPs at the position of the current trial, and the window was advanced. After artifact rejection, there were too few ERPs for the rare irrelevant-probe ERP measures to be included in the models. So only the frequent irrelevant-probe ERP measures were included in the running means.

We applied the same running-mean process to the PF1 measure. The result was a series of running-mean ERPs and PF1 which could be used to develop regression models. Because our purpose was to examine the effect of increasing ERP signal-to-noise ratio on the reliability of the ERP-based models, we did not apply the running-mean process to the task factors, nor did we force them into the regression models[2].

[1] To provide a reference for estimating how much the regression models are improved by using 10-trial running means versus single-trial ERPs, we provide the range of R^2 values for the corresponding single-trial regression analyses.

In the signal detection task, the general model of PF1 based on ERP measures alone was significant with an R^2 of 0.09. When irrelevant-probe ERP measures were excluded from the model, R^2 was 0.07. A model based only on measures of ERPs for the irrelevant probes was significant but the R^2 was only 0.03. Individual models based on relevant-stimulus ERP measures had an average R^2 of 0.21. For the irrelevant-probe ERP measures the corresponding average R^2 was 0.13.

For the running memory task, the general model of PF1 based on ERP measures alone was significant with an R^2 of 0.16. When irrelevant-probe ERP measures were excluded from the model, the R^2 decreased to 0.13. A model based only on measures of ERPs for the irrelevant probes was significant with an R^2 of 0.07. Significant individual regression models based on relevant-stimulus ERP measures alone were obtained in all eight subjects, with an average R^2 of 0.17. Individual models based on irrelevant-probe ERP measures alone had an average R^2 of 0.07.

For the computation task, the general model of PF1 based on ERP measures alone was significant with an R^2 of 0.10. For the model based on relevant-stimulus ERP measures alone, the R^2 was 0.09. A model based only on irrelevant-probe ERP measures was significant, but the R^2 was only 0.03. Significant individual regression models based on relevant-stimulis ERP measures alone were obtained in all eight subjects, with an average R^2 of 0.12. Individual models based on irrelevant-probe ERP measures alone had an average R^2 of 0.06.

[2] In the initial single-trial regression analyses, task factors and interactions that had been significant in the ANOVAs were used to model PF1 alone and also forced into regression models based on ERP measures. In the general model for the signal detection task, task factors alone led to a significant model with an R^2 of 0.21. When we combined task factors with ERP measures, the R^2 increased to 0.42. The corresponding models for the running memory task were also significant, with R^2 values of 0.11 and 0.27, respectively. For the computation task, again the models were significant, with corresponding R^2 values of 0.04 and 0.19. In all tasks, ERP measures clearly explained as much or more unique variance in PF1 than the task factors and their interactions.

In these analyses, a forward-selection stepwise approach (SAS PROC STEPWISE) was used to develop linear regression models that explain trial-to-trial variations in overall task performance (PF1) in terms of relevant-stimulus ERP measures or irrelevant-probe ERP measures. For each task, a general model (for all subjects) and individual models for each subject were developed. For all models, the significance was assessed using an F ratio test with α set at 0.01. In addition, because each running-mean ERP combined data from 10 single-trial ERPs, we divided the corresponding degrees of freedom for significance tests by 10.

In addition to improving SNR, the running-mean process may lead to improved regression models as a result of adding observations that share large proportions of variance with their neighbors. However, the procedure is valid for on-line prediction, particularly when the models are validated with independent observations, as shown below.

Our regressions were performed as part of a conservative cross-validation analysis. This procedure used the data from the odd-numbered blocks of trials as a screening sample to build the models and the data from even-numbered blocks of trials to calibrate the models and assess shrinkage of the model R^2. In general, the individual running-mean ERP models of PF1 developed on odd-numbered blocks explained PF1 nearly as well in even-numbered blocks for most subjects (Tables 6, 7, & 8). In addition, general models of PF1 based on relevant-stimulus ERP measures cross-validated with little or no shrinkage of the R^2 in the running memory task and the computation task, but not in the signal detection task.

3.6.1. Signal detection task

No general model of PF1 cross-validated. For the relevant-stimulus ERP measures, the stepwise algorithm produced individual regression models that cross-

Table 7
Cross validated* regression models of signal detection performance

Subject	Screening sample			Calibration sample			Shrinkage
	F	df	R^2	F	df	R^2	
Relevant stimuli							
1	9.96	17, 545	0.24	7.57	17, 559	0.19	0.05
2	16.59	16, 364	0.42	14.34	16, 379	0.38	0.04
4	14.85	16, 383	0.38	18.55	16, 398	0.34	0.04
5	32.83	19, 405	0.61	11.64	19, 423	0.34	0.27
6	23.84	16, 381	0.50	14.74	16, 394	0.37	0.13
7	27.52	17, 470	0.50	13.33	17, 486	0.32	0.18
Irrelevant stimuli							
1	9.61	7, 298	0.18	3.07	7, 364	0.06	0.13
2	13.25	12, 327	0.33	5.75	12, 396	0.15	0.18
6	10.69	9, 309	0.24	7.60	9, 372	0.16	0.08

*$p < 0.0001$

Table 8
Cross validated* regression models of running memory performance

Subject	Screening sample			Calibration sample			Shrinkage
	F	*df*	R^2	*F*	*df*	R^2	
Relevant stimuli							
All	66.53	36, 4617	0.34	66.51	36, 4636	0.34	0
1	15.40	15, 558	0.47	26.50	15, 571	0.41	0.06
2	31.23	19, 556	0.52	29.35	19, 574	0.49	0.02
3	28.04	16, 558	0.45	20.17	16, 571	0.36	0.08
4	26.04	19, 682	0.42	27.93	19, 693	0.44	0
5	21.52	24, 551	0.48	16.65	24, 573	0.41	0.07
6	25.57	17, 558	0.44	23.60	17, 573	0.41	0.03
7	26.23	19, 479	0.51	24.46	19, 496	0.48	0.03
8	17.57	17, 558	0.35	18.78	17, 571	0.36	0
Irrelevant stimuli							
5	12.62	10, 437	0.22	6.57	10, 537	0.11	0.11
7	13.52	8, 366	0.23	9.07	8, 450	0.14	0.09

*$p < 0.0001$

validated in six subjects (Table 7). R^2 ranged from 0.24–0.61 for the screening sample (mean = 0.44) and from 0.19–0.37 for the calibration sample (mean = 0.32). The shrinkage in R^2 ranged from 0.05–0.27 (mean = 0.12) and tended to increase in proportion to the value of R^2.

For the irrelevant-probe ERP measures, the stepwise algorithm produced individual regression models that cross-validated in three subjects. R^2 ranged from 0.18–0.33 for the screening sample (mean = 0.25) and from 0.06–0.16 for the cali-

Table 9
Cross validated* regression models of computation performance

Subject	Screening sample			Calibration sample			Shrinkage
	F	*df*	R^2	*F*	*df*	R^2	
Relevant stimuli							
All	91.79	28, 4272	0.38	86.42	28, 4286	0.36	0.02
1	13.57	16, 559	0.28	10.69	16, 574	0.23	0.05
2	13.61	14, 560	0.55	43.68	14, 572	0.52	0.03
3	27.56	23, 552	0.53	21.63	23, 571	0.47	0.06
4	19.09	21, 266	0.60	11.64	21, 286	0.46	0.04
5	12.62	14, 549	0.24	10.10	14, 562	0.20	0.04
6	14.11	23, 552	0.37	9.93	23, 571	0.29	0.08
7	10.82	15, 555	0.23	11.50	15, 564	0.23	0
8	9.94	17, 557	0.23	9.28	17, 572	0.22	0.02
Irrelevant stimuli							
4	12.36	6, 230	0.24	6.88	6, 254	0.14	0.08

*$p < 0.0001$

bration sample (mean = 0.12). The shrinkage in R^2 ranged from 0.13–0.18 (mean = 0.13).

3.6.2. Running memory task

A general model based on relevant-stimulus ERP measures cross-validated. The R^2 for this model was 0.34 with no shrinkage. The stepwise algorithm included 36 ERP measures. Of these, only three variables had relatively large influence, as indicated by the partial R^2: FW3-RMS (Fz-Cz, 460–710 ms), P3A-AVG (Fz, 350–600 ms), and N1-AVG (Fz, 80–210 ms). The corresponding partial R^2 values for these three variables were 0.10, 0.07 and 0.04, respectively. The 33 remaining measures had partial R^2 values of 0.02 or less. No general model based on irrelevant-probe ERP measures cross-validated.

For the relevant-stimulus ERP measures, the stepwise algorithm produced individual regression models that cross-validated in all subjects (Table 8). R^2 ranged from 0.35–0.52 for the screening sample (mean = 0.46) and from 0.36–0.49 for the calibration sample (mean = 0.42). The amount of shrinkage in the models ranged from 0.0–0.08 (mean = 0.04) and bore no clear relationship to the value of R^2.

For the irrelevant-probe ERP measures, the stepwise algorithm produced individual regression models that cross-validated in two subjects (Table 8). In both of these models, R^2 values were about 0.2 for the screening sample and 0.1 for the calibration sample.

3.6.3. Computation task

A general model based on relevant-stimulus ERP measures cross-validated, with R^2 values of 0.38 for the screening-sample and 0.36 for the calibration-sample (Table 9). The stepwise algorithm included 28 ERP measures. Of these, three variables had relatively large influence, as indicated by the partial R^2: FW4-AMP (Fz-Cz, 710–1710 ms), FW2-AVG (Fz, 210–360 ms), and FW3-AVG (Fz, 460–710 ms). The corresponding partial R^2 values for these three variables were 0.11, 0.07 and 0.06, respectively. The remaining measures had partial R^2 values of 0.03 or less. No general model based on irrelevant-probe ERP measures cross-validated.

For the irrelevant-probe ERP measures, the stepwise algorithm produced only one individual regression model that cross-validated. R^2 values were comparable to those of the running memory task: 0.24 for the screening sample and 0.14 for the calibration sample.

For the irrelevant-probe ERP measures, the stepwise algorithm produced only one individual regression model that cross-validated. R^2 values were comparable to those of the running memory task: 0.24 for the screening sample and 0.14 for the calibration sample.

4. Discussion

4.1. Relationship of ERP measures to task performance

Our results show that, given sufficient ERP SNR, as with our running-mean ERP measures, we can reliably estimate a global measure of display monitoring perfor-

mance using linear regression models. The success of this approach across subjects and tasks was clearly greater for models based on ERPs elicited by task-relevant stimuli than for models based on ERPs elicited by irrelevant-probe stimuli. For the relevant-stimulus ERPs, general models significantly estimated performance in the running memory and computation tasks. Individual models reliably estimated performance in six of eight subjects in the signal detection task and in all subjects in the running memory and computation tasks. The R^2 values for these models ranged from 0.19–0.52, with many values in the 0.4–0.5 range.

In contrast, with the low SNR that is characteristic of single-trial ERPs, models of task performance were not generally reliable and explained small proportions of variance. For the individual screening-sample models based on relevant-stimulus ERP measures in the signal detection task, the single-trial data yielded an average R^2 of 0.21 whereas the running-mean data yielded an average R^2 of 0.44. The corresponding comparison of R^2 values was 0.17 as against 0.46 for the running memory task, and 0.12 as against 0.38 for the computation task. If we take the ratios of the R^2 values for running-mean models to single-trial models in the three tasks, we obtain values of 2.1, 2.7 and 3.2, respectively. These ratios suggest that the improvement in the R^2 due to 10-trial ERP averaging is a factor of about 2–3, depending on the task. This is close to the theoretical value of the increase in SNR between one and 10 trials, which should be $\sqrt{N_1\text{-}N_2}$, or $\sqrt{9} = 3$ (Regan, 1989, p. 56).

While this is an interesting parallelism, it is an oversimplification because extrapolation to larger values of N would lead to R^2 values greater than one. More likely, there is a nonlinear relationship between SNR and R^2 in such models, with an upper limit on R^2 set by the covariance of the ERP component-generating processes and task performance. Nevertheless, it appears that techniques that increase the SNR of ERP measures, such as time-varying filters (Hohenberger, 1988), principal component measures or wavelet measures will lead to improvements in the accuracy and speed (i.e., fewer trials required) of ERP-based performance estimation. For example, a reanalysis of the signal processing data from this experiment using principal components or wavelet-based feature extraction of the running-mean ERP data led to a significant, cross-validated, general model of our PF1 measure (Trejo & Shensa, 1993). This result was not obtained using the straightforward amplitude and latency measures we applied in this paper.

Some insight into the covariance between ERP components and performance may be provided by the standard deviation of the R^2 values among the individual models. If we assume that the covariance between ERP and performance measures depends on subject factors, such as the strategy employed by the subject, then the standard deviation of R^2 values among the subjects and tasks sets a limit on how close to 1.0 the expected value of R^2 can be. Using the R^2 values from the relevant-stimulus ERP models and the running-mean data, we observed that the standard deviation of R^2 values across the signal detection, running memory, and computation tasks was 0.10. With this standard deviation, if the R^2 estimate were 1.0, a reasonable one-sided confidence interval for the true value of R^2 would be $1.0 \pm 1.65\sigma$ or between 0.84 and 1.0. Conservatively, then, the lower limit of this interval, or 0.84, may serve as an upper limit of the expected value of R^2 for regression models

of the type we have developed, using asymptotically high estimates of the ERP SNR. This number should be considered only in the context of our experiments, where factors such as eye movements, blinks, head and body motion, and distracting stimuli were carefully controlled. Under real-world conditions, the R^2 limit is probably lower.

Our results show that ERP-based regression models of task performance are generally improved by tailoring the models to the data of individual subjects, as compared to general models. However, the degree of improvement in R^2 was task-dependent. The biggest difference was seen for the signal detection task where no general model cross-validated and the mean cross-validated R^2 for the individual models was 0.44. For the running memory task, the R^2 value for the general model was 0.34, or 0.12 lower than the average R^2 values for the individual models (0.46). In the computation task, however, the R^2 of the general model nearly matched the average R^2 of the individual models (0.36 as against 0.38). These comparisons may fail to reflect the true degree of improvement possible with individual versus general models because, while the stepwise regression procedure was individualized, the set of ERP measures was not. More improvement might be found if the ERP measures were selected independently for each subject.

4.2. Implications for performance assessment

Our results show that linear combinations of single-trial ERP amplitude and latency measures are not likely to provide a useful real-time index of task performance. The R^2 values obtained with single-trial data were low, but more importantly, single-trial based models were not reliable. However, our single-trial ERP measures were crude by present signal processing standards. Application of more sophisticated measures, such as wavelets or principal components could lead to improved performance assessment using single-trial ERP data.

On the other hand, the R^2 values and reliability of the models based on running-mean ERPs we observed suggest that such models are useful for a quasi real-time index of performance. For example, in the signal detection task mean values of PF1 under different task conditions ranged between −0.55 and 0.61. With an R^2 value of 0.38, as seen in subject two, the standard error of prediction for PF1 was 0.53. Thus the regression estimate of PF1 could perform a binary discrimination between the highest and lowest average performance conditions for this subject. We also observed that over time, many 10-trial epochs had PF1 values below or above these mean levels. Thus, over time, additional levels of performance discrimination could be provided by the regression estimate of PF1. Still better discrimination of performance levels could be provided for tasks or subjects in which the regression models had higher R^2 and lower standard errors of prediction. As for single-trial based models, some improvement in performance estimates may also be obtained in the running-mean based models by using improved ERP signal processing methods.

Finally, the addition of significant task information, such as the position of the stimulus or designation as target/nontarget, should also lead to more accurate models of performance. In many real-world tasks, such information is readily avail-

able, particularly in computer-controlled tasks. Although task factors were not forced into the running-mean ERP models, it was clear from the single-trial models that task factors alone explained considerable variance in PF1. Furthermore, this variance was not entirely shared with that explained by ERP measures, as shown by the increases in R^2 we observed when adding ERP measures to models based on task factors.

4.3. Theoretical considerations

4.3.1. Relevant stimulus ERPs

As discussed in the introduction, notions of limited capacity or resource allocation have been proposed to explain relationships between task demands or task difficulty and the amplitude of ERP components such as the P300. Our data provide new insight into some of these relationships. In the signal detection and running memory tasks, two of our parietal average amplitude measures (P3B-AMP and P3C-AMP) for the relevant-stimulus ERPs were greater when the task was actively engaged than in the baseline condition. Also in both tasks, these amplitude measures were greater for correct trials than for incorrect trials in the active condition, although this difference narrowly missed significance in the running memory task ($p = 0.07$). P3B-AMP and P3C-AMP are the measures that most-closely match the latency and scalp distribution criteria of P300 (Trejo et al., 1991). Although we did not analyze P300 × difficulty directly, these results suggest that P300 for the task-relevant stimuli decreased as a function of task difficulty because difficulty had the largest effect on accuracy of all task factors (Figs. 8, 9 & 11).

Such a decrease in P300 amplitude with an increase in primary task difficulty disagrees with the idea that P300s elicited by task-relevant stimuli simply index the processing resources required for the task. For example, Kramer, Wickens, Vanasse, Heffley, & Donchin (1981) found that increases in the difficulty of a step-tracking task led to increases in the amplitude of the P300 elicited by the primary task stimuli (movements of the cursor being tracked) but to decreases in the amplitude of the P300 elicited by secondary task probes. This result was explained in terms of a trade-off of processing resources between primary and secondary tasks. Since we used no secondary task, the differences in P300 amplitude between accurate- and inaccurate-response trials that we found cannot simply be due to a trade-off of resources among tasks. In addition, the P300 for the baseline conditions, where the difficulty level was effectively zero, was smaller than in the active conditions, regardless of accuracy.

An alternative to a simple resource-allocation model is that, in a decision-making task, the P300 generator also produces an output that increases with the information available to make decisions about the task-relevant stimuli. Support for this view has been reported in the form of greater P300 amplitudes for more confident signal detection decisions than for less confident decisions (Parasuraman et al., 1982; Sutton, Ruchkin, Munson, Kietzman & Hammer, 1982). Additional support for this view was provided by an experiment in which the amount of information in the task-relevant stimuli was directly manipulated (Ruchkin, Johnson, Canoune, Ritter & Hammer, 1990). Our confidence-response data in the signal detection task also sup-

port this view because confidence was positively correlated with response accuracy, as shown by the positive loadings on PF1 of these variables in the factor analysis.

This 'information' model could be combined with the resource-allocation model as follows. The increases in P300 amplitude between baseline and active conditions in our tasks indexed the commitment of processing resources to task demands. The decrease in P300 amplitude between accurate and inaccurate response trials indexed a decrease in the amount or quality of information available to the decision-making process. In the signal detection task, this decrease in information is clearly related to the degradation of the stimuli caused by reducing the contrast on the more difficult trials. In the running memory task, the comparison letters had to be held in memory longer during the difficult conditions than in the easy conditions. During this longer retention interval, the comparison letters in memory could be decreased by decay of iconic storage (Sperling, 1960) or displaced by following letters (Waugh & Norman, 1965), either of which would decrease the information. Still another explanation for difficulty-related decreases in P300 amplitude in complex tasks can be derived from trade-offs of internal processes such as memory scanning and rehearsal (for an example, see Mecklinger, Kramer & Strayer, 1992).

The preceding argument is not necessarily refuted by the absence of P300 effects in the computation task. For one thing, a P300 component was not clearly present in the ERP averages for this task. More importantly, this task differed from the other tasks in complexity and in the type of processing. In this task, difficulty was a function of abstract properties of the stimuli, and not necessarily of information which could be degraded or which decayed over time. Possibly, the performance of this task relied more heavily on long-term memory, performance being better for familiar number pairs than for less familiar pairs.

In addition to the P300-related effects, in our signal detection task there was an effect of task performance conditions on the amplitude of SW1, which corresponds to a late frontal negative slow wave. SW1 amplitude, however, showed only the task-engagement related increase and no differences as a function of response accuracy. Several such negative slow waves have been reported. Our SW1 results agree with Rohrbaugh, Syndulko & Lindsley (1978), who reported a frontally negative 'after-wave' for visual and auditory stimuli. This slow wave, which was similar to our SW1 in scalp distribution and latency, was larger when subjects were required to silently count stimuli or perform a discrimination than in passive conditions. We also saw no evidence for the late negative slow wave reported by Ruchkin et al., (1988), which increased in amplitude with conceptual difficulty. Unlike our SW1, their negative slow wave had a centro-posterior distribution and reversed in polarity at more frontal locations. However, identification of our SW1 measure with other slow waves is complicated by possible cancellation among multiple slow waves, which occurs in ERP averages. Special methods may be required to deal with such cancellation (Loveless, Simpson & Näätänen, 1987).

There were no significant effects on smaller ERP components, such as P1, N1 and P2. However, our data do not provide for strong tests for such effects due to the low number of subjects we used.

4.3.2. Irrelevant probe ERPs

For the most part, the amplitudes of irrelevant-probe ERP features were sensitive to the difference between passive and active conditions, i.e. task engagement. In the signal detection task, the FW2 measure, which extends from 200–600 ms at Fz-Cz, was negative in the baseline condition and positive in the active conditions. The task-engagement-related change in this measure is consistent with workload-related irrelevant-probe ERP effects at Fz-Cz with a latency of 330 ms (Blankenship et al., 1988b; Trejo et al., 1987 and 1990). A change in amplitude in the FW2 window appears to index the engagement of resources required to perform a visual task, but does not differentiate well between task performance levels. Trejo et al. (1990) observed, however, that the degree of change between passive and active conditions was correlated with the mean level of task performance across subjects.

In two tasks, signal detection and running memory, the amplitude of the N2 component, as measured by N2-AVG, became more negative in the active conditions than in the baseline conditions, but did not vary as a function of response accuracy. This effect is unclear for two reasons. Firstly, the N2 peak was poorly defined, appearing as a shoulder on the descending limb of the N1. Secondly, the following P300 canceled the N2 negativity in varying degrees depending on task conditions. In two tasks, signal detection and computation, this P300 effect was significant. The P300 tended to be greater in the baseline conditions than during task engagement. In the computation task, the P300 also differentiated the accurate- and inaccurate-response averages. So we cannot rule out the possibility that the N2 effects were confounded with workload effects on the probe P300 itself, which are less ambiguous and have been reported elsewhere (Blankenship et al., 1988a; Kramer, Trejo & Humphrey, in press).

The P300 effect for the irrelevant-probe stimuli suggests that subjects' attention narrowed when the tasks were engaged. That is, they selectively attended the task-relevant stimuli and ignored the probes during the active conditions, but attended both types of stimuli during the baseline condition. This is consistent with a dual-task interpretation, such that subjects performed some monitoring of the probes during the baseline condition, but shed this task during active conditions.

The remaining task-engagement-related changes in amplitude measures of the irrelevant-probe ERPs appear consistent with an increase in negative slow wave amplitude between passive and active conditions. Such an increase would explain the more negative SW1 and SW2 amplitudes in active versus baseline signal detection task conditions. We note that these differences are not likely to be caused by the frontal negative slow wave (SW1) that increased in the relevant-stimulus ERPs as a function of task engagement. The SW1 and SW2 for the irrelevant-probe ERPs were much more posteriorly distributed, with a maximum at Pz, as compared to the Fz maximum for the relevant-stimulus SW1. Thus, the irrelevant-probe SW1 and SW2 measures appear to index a different component than the relevant-stimulus SW1. We know of no other reports showing parietal negative slow wave elicited by irrelevant probes that increases in amplitude when the probes are ignored to a greater degree.

Due to the low numbers of trials and SNR of the rare irrelevant-probe ERPs, we attempted no statistical inferences about them. However, we note that these ERPs show evidence of several of the effects found in the frequent irrelevant-probe ERP averages. In particular, the reversal of polarity in the FW2 window at Fz-Cz, the presence of a positive voltage in the P300 window at Pz and the greater negativity in the SW1 and SW2 windows at Pz are all apparent in the cross-task grand average ERP (Fig. 7). Future experiments may find additional diagnostic value by comparing the rare and frequent irrelevant-probe ERPs. For example, such an approach has proven to be successful in workload estimation with rare-frequent difference components of the auditory ERP, such as the mismatch negativity (Kramer, Trejo & Humphrey, 1995).

4.4. Summary

In this study we examined some of the practical issues and limitations involved in an ERP approach to predicting performance on display-monitoring tasks. In three visual-display monitoring tasks, we found that a linear regression method of predicting performance from estimates of ERP components produced models that explained substantial variance in task performance and generalized to data collected under separate conditions. The level of sophistication of these models was modest, not incorporating higher-order or nonlinear terms. Yet, even this modest approach indicated that a real-time index of relatively high or low performance is currently possible.

Three factors strongly influenced the accuracy and validity of the regression models. Firstly, the number of trials used to estimate ERP components was critical. Single-trial estimates were not generally reliable. Estimates based on running means of about 10 trials were reliable. We attributed the better reliability of the running-mean estimates to improved signal-to-noise ratio (SNR). However, we acknowledge that some of this improvement may arise from increasing the number of serially correlated observations. Future research should independently address these issues. Nevertheless, the validity of the running-mean approach to on-line estimation was shown by the generalization of the models to new data. Secondly, the models were improved when they were tailored to the data of individual subjects. This suggests that implementations of ERP-based monitoring systems ought to be calibrated to the user for best results. Thirdly, we found that the utility of task-relevant stimuli for performance prediction was high whereas the utility of irrelevant probes was low. However, this issue is far from settled, as we did not explore a range of probe types nor did we obtain good estimates of the utility of rare or deviant probes.

Finally, we observed a pattern of ERP effects that is suggestive of two different influences on ERP-workload relationships: allocation and information quality. For task-relevant stimuli, we observed that P300 and a frontally negative slow wave indexed the engagement of processing resources as shown by increases in amplitude between baseline and active conditions. P300 additionally reflected the quality of performance in active conditions, being lower for inaccurate than for accurate-response trials. Since accuracy and difficulty were strongly related, we inferred that

increasing difficulty led to decreases in P300 amplitude, and proposed that such decreases might arise from reduced information quality under difficult conditions. For irrelevant probes, a number of components also indexed the engagement of resources that distinguished baseline and active task conditions. These included a frontal negativity (FW2), N2, P300 and slow waves. Together, these observations suggest that the engagement or application of processing resources produces switch-like changes in some relevant-stimulus and irrelevant-probe ERP components, whereas information quality both influences performance and modulates the amplitude of the P300.

Acknowledgments

This work was supported in part by grants to Leonard Trejo from the Office of Naval Research, monitored by Dr. Stanley C. Collyer and Dr. Terry Allard, and in part by a grant to Arthur Kramer from the Office of Naval Research, monitored by Dr. Harold Hawkins. Data collection and analyses were supported by the Navy Personnel Research and Development Center, San Diego, CA. The opinions expressed here are those of the authors, are unofficial, and do not necessarily reflect the views of the Navy Department.

We gratefully acknowledge Kimberly Fowler for help in the data collection and Marcel Compelube, Carrie Lesh, Richard Ogle, Michelle Mullane, and Renee Sherry for help in the data screening and analyses.

References

Blankenship, M.H., Trejo, L.J., & Lewis, G.W. (1988a). *Brain activity during tactical decision-making: IV. Event-related potentials as indices of selective attention and cognitive workload.* (NPRDC Tech. Note TN-89-6). San Diego: Navy Personnel Research and Development Center.

Blankenship, M.H., Trejo, L.J., & Lewis, G.W. (1988b). *Brain activity during tactical decision-making: V. A cross-study validation of evoked potentials as indices of cognitive workload* (NPRDC Tech. Note TN-89-7). San Diego: Navy Personnel Research and Development Center.

Broadbent, D.E. (1970). Stimulus set and response set: two kinds of selective attention. In D.I. Mostofsky (Ed.), *Attention: Contemporary Theory and Analysis* (pp. 51–60). New York: Appleton-Century-Crofts.

Defayolle, M., Dinand, J.P., & Gentil, M.T. (1971). Averaged evoked potentials in relation to attitude, mental load and intelligence. In W.T. Singleton, J.G. Fox, & D.Whitfield (Eds.), *Measurement of man at work*. New York: Van Nostrand Reinhold Company.

Donchin, E., Kramer, A.F., & Wickens, C.D. (1986). Applications of brain event-related potentials to problems in engineering psychology. In M.G.H. Coles, E. Donchin, & S. Porges (Eds.), *Psychophysiology: systems, processes, and applications*. Middleton, NJ: Till & Till.

Eason, R.G., Harter, M.R., & White, C.T. (1969). Effects of attention and arousal on visually evoked cortical potentials and reaction time in man. *Physiology and Behavior*, 4, 283–289.

Garcia-Austt, E., Bogacacz, J., & Vanzulli, A. (1964). Effects of attention and inattentions upon visual evoked response. *Electroencephalography and Clinical Neurophysiology*, 17, 136–143.

Geisser, S., & Greenhouse, S.W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics*, 29, 885–891.

Gopher, D., & Donchin, E. (1986). Workload — an examination of the concept. In K. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance*. Vol. II. New York: John Wiley.

Gratton, G., Coles, M.G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, 55, 468–484.

Harter, M.R. & Aine, C.J. (1984). Brain mechanisms of visual selective attention. In R. Parasuraman and D.R. Davies (Eds.), *Varieties of Attention*. Orlando: Academic Press.

Hillyard, S.A., Squires, K.C., Bauer, J.W., & Lindsay, P.H. (1971). Evoked potential correlates of auditory signal detection. *Science*, 172, 1357–1360.

Hoffman, J.E., Simons, R.F., & Houck, M.R. (1983). Event-related potentials during controlled and automatic target detection. *Psychophysiology*, 20, 625–632.

Hohenberger, M. (1988). *A time-varying digital filter for evoked potentials and evoked magnetic fields.* (NPRDC Technical Note 88-47). San Diego: Navy Personnel Research and Development Center.

Israel, J.B., Chesney, G.L., Wickens, C.D., & Donchin, E. (1980). P300 and tracking difficulty: evidence for multiple resources in dual-task performance. *Psychophysiology*, 17, 259–273.

Jasper, H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 43, 397–403.

Jerison, H.J., & Pickett, R.M. (1964). Vigilance: The importance of the elicited observing rate. *Science*, 143, 970–971.

Kok, A., & DeJong, H.L. (1980). The effect of repetition of infrequent familiar and unfamiliar visual patterns on components of the event-related brain potential. *Biological Psychology*, 10, 167–188.

Kramer, A.F. (1990). *Physiological metrics of mental workload: a review of recent progress.* (NPRDC Technical Note 90-23). San Diego: Navy Personnel Research and Development Center.

Kramer, A.F., Trejo, L.J., & Humphrey, D. (1995). Assessment of mental workload with task-irrelevant auditory probes. *Biological Psychology, 40,* 83–100.

Kramer, A.F., Wickens, C.D., Vanasse, L., Heffley, E.F., & Donchin, E. (1981). Primary and secondary task analysis of step tracking: an event related potentials approach. In R.C. Sugarman (Ed.), *Proceedings of the Twenty-Fifth Annual Meeting of the Human Factors Society*, Rochester, NY.

Loveless, N.E., Simpson, M., & Näätänen, R. (1987). Frontal negative and parietal positive components of the slow wave dissociated. *Psychophysiology*, 24, 340–345.

Mecklinger, A., Kramer, A.F., & Strayer, D.L. (1992). Event-related potentials and EEG components in a semantic memory search task. *Psychophysiology*, 29, 104–119.

Näätänen, R. (1982). Processing negativity: an evoked-potential reflection of selective attention. *Psychological Bulletin*, 92, 605–640.

Norman, D.A., & Bobrow, D.G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7, 44–64.

Oatman, L.C. (1971). Role of visual attention on auditory evoked potentials in unanesthetized cats. *Experimental Neurology*, 32, 341–356.

Papanicolaou, A.C., & Johnstone, J. (1984). Probe-evoked potentials: theory, method and applications. *International Journal of Neuroscience*, 24, 107–131.

Parasuraman, R. (1986). Vigilance, monitoring, and search. In K. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance*. Vol. II. New York: John Wiley.

Parasuraman, R. (1990). Event-related brain potentials and human factors research. In J.W. Rohrbaugh, R. Parasuraman, & R. Johnson, Jr. (Eds.), *Event-related brain potentials. Basic issues and applications.* New York: Oxford.

Parasuraman, R., & Beatty, J. (1980). Brain events underlying detection and recognition of weak sensory signals. *Science*, 210, 80–83.

Parasuraman, R., Richer, F., & Beatty, J. (1982). Detection and recognition: concurrent processes in perception. *Perception and Psychophysics*, 31, 1–12.

Regan, D. (1989). *Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine.* New York: Elsevier.

Rohrbaugh, J.W., Syndulko. K., & Lindsley, D.B. (1978). Cortical slow negative waves following nonpaired stimuli: effects of task factors. *Electroencephalography and Clinical Neurophysiology*, 45, 551–567.

Ruchkin, D.S., Johnson, R. Jr., Canoune, H.L., Ritter, W., & Hammer, M. (1990). Multiple sources of P3b associated with different types of information. *Psychophysiology*, 27, 157–176.

Ruchkin, D.S., Johnson, R. Jr., Mahaffey, D., & Sutton, S. (1988). Toward a functional categorization of slow waves. *Psychophysiology*, 25, 339–353.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, 74, (Whole No. 11).

Squires, K.C., Squires, N.K., & Hillyard, S.A. (1975). Decision-related cortical potentials during an auditory signal detection task with cued observation intervals. *Journal of Experimental Psychology: Human Perception and Performance*, 103, 268–279.

Sutton, S., Ruchkin, D.S., Munson, R., Kietzman, M.L., & Hammer, M. (1982). Event-related potentials in a two-interval forced-choice detection task. *Perception and Psychophysics*, 32, 360–374.

Trejo, L.J., Lewis, G.W., & Blankenship, M.H. (1987). *Brain activity during tactical decision-making: II. Probe-evoked potentials and workload* (NPRDC Tech. Note 88-12). San Diego: Navy Personnel Research and Development Center.

Trejo, L.J. (1988). *RMS Amplitude Computations in ERP Time Windows* (NPRDC Letter 3900 Ser. 41/85). San Diego: Navy Personnel Research and Development Center.

Trejo, L.J., Lewis, G.W., & Blankenship, M.H. (1990). *Brain activity during decision-making: III. Relationships between probe-evoked potentials, simulation performance, and task performance* (NPRDC Tech. Note TN-90-9). San Diego: Navy Personnel Research and Development Center.

Trejo, L.J., Inlow, M., Stanny, R.R., Morey, W.A., Makeig, S., Kobus, D.A., & Hillyard, S.A. (1991). *The P300 component of the auditory event-related potential: interlaboratory consistency and test-retest reliability* (NPRDC Technical Report TR-91-6). San Diego: Navy Personnel Research and Development Center.

Trejo, L.J., & Shensa, M.J. (1993). Linear and neural network models for predicting human signal detection performance from event-related potentials: a comparison of the wavelet transform with other feature extraction methods. *Proceedings of the Fifth Workshop on Neural Networks: Academic/Industrial/NASA/Defense, SPIE Volume 2204* (pp. 153–161). San Diego: Society for Computer Simulation.

Van Voorhis, S., & Hillyard S.A. (1977). Visual evoked potentials and selective attention to points in space. *Perception and Psychophysics*, 22, 54–62.

Waugh, N.C. & Norman, D.A. (1965). Primary memory. *Psychological Review*, 72, 89–104.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman & D.R. Davies (Eds.), *Varieties of attention*. New York: Academic Press.

Wyszecki, G., & Stiles, W.S. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*. New York: John Wiley & Sons.