


2005

Dissociable Aspects of Mental Workload: Examinations of the P300 ERP Component and Performance Assessments

Carryl L. Baldwin
Old Dominion University

Joseph T. Coyne
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/psychology_fac_pubs

 Part of the [Cognition and Perception Commons](#), [Cognitive Psychology Commons](#), and the [Experimental Analysis of Behavior Commons](#)

Repository Citation

Baldwin, Carryl L. and Coyne, Joseph T., "Dissociable Aspects of Mental Workload: Examinations of the P300 ERP Component and Performance Assessments" (2005). *Psychology Faculty Publications*. 78.
https://digitalcommons.odu.edu/psychology_fac_pubs/78

Original Publication Citation

Baldwin, C. L., & Coyne, J. T. (2005). Dissociable aspects of mental workload: Examinations of the P300 ERP component and performance assessments. *Psychologia*, 48(2), 102-119. doi:10.2117/psysoc.2005.102

DISSOCIABLE ASPECTS OF MENTAL WORKLOAD: EXAMINATIONS OF THE P300 ERP COMPONENT AND PERFORMANCE ASSESSMENTS

Carryl L. BALDWIN¹⁾ and Joseph T. COYNE¹⁾

¹⁾*Old Dominion University, U.S.A.*

Advanced technologies have enabled the choice of either visual or auditory formats for avionics and surface transportation displays. Methods of assessing the mental workload imposed by displays of different formats are critical to their successful implementation. Towards this end a series of investigations were conducted with the following aims: 1) developing analogous auditory and visual versions of a secondary task that could be used to compare display modalities; and 2) to compare the sensitivity of neurophysiological, behavioral and subjective indices of workload. Experiments 1 and 2 confirmed that analogous auditory and visual secondary oddball discrimination tasks were of equivalent difficulty as indicated by P300 amplitude, RT, accuracy and subjective ratings of workload. Experiments 1–3 revealed that RT and accuracy for target detections were generally more sensitive to changes in primary task difficulty than P300 responses and subjective ratings. However, Experiment 3 indicated that P300 amplitude was sensitive to increased perceptual demands (resulting from driving in heavy fog versus clear visibility) not revealed by changes in either behavioral or subjective indices. Together the results of the current investigations indicate that a battery of assessment techniques will provide the most sensitive assessment of workload in complex environments.

Key words: mental workload assessment, display modality, transportation, neurophysiological measures, adaptive automation

Advances in information technologies currently allow information to be displayed to operators in either visual or auditory formats. Examples include graphical versus radio based weather information systems, text-based versus radio communications between pilots and air traffic controllers (ATC), and electronic map versus auditory route guidance in navigation systems. Adaptive automation systems have the potential to present the most effective display format for a given operator's state in response to changes in endogenous or exogenous workload conditions. The safe implementation of these new display formats will require the ability to compare mental workload demand across displays of different modalities under a variety of environmental conditions.

Mental workload assessment has played an important role in ergonomics research for nearly a half century (Knowles, 1963; Meshkati, Hancock, & Rahimi, 1990; Williges & Wierwille, 1979). Mental workload is generally agreed to be a multidimensional,

This work was supported in part by NASA Langley Research Center (Grant NAG 1-03020 and NNL04AA06G). The authors wish to acknowledge the technical support of Dr. Fredrick G. Freeman, Old Dominion University and Dr. Lawrence Prinzel, NASA Langley Research Center in conducting these investigations.

Correspondence concerning this article should be addressed to Carryl L. Baldwin, Department of Psychology, Old Dominion University, Norfolk, VA23529-0267, U.S.A. (e-mail: cbaldwin@odu.edu).

multifaceted construct (Baldwin, 2003; Gopher & Donchin, 1986; Hancock & Caird, 1993; Kramer, 1991) often referencing the relationship between task structure and demands and the time available for performing the given task(s). To date a single optimal assessment method has not been identified.

The examination of changes in event-related brain potential (ERP) components as a method of understanding human information processing and mental workload in particular has had a long history. The P300 ERP component, a positive wave deflection occurring approximately 300 ms after a stimulus event, has been of particular interest in mental workload assessment.

Based on the theory that humans have a limited capacity processing system (Kahneman, 1973; Moray, Dessouky, Kijowski, & Adapathya, 1991; Wickens, 2002), P300 amplitude in a binary discrimination paradigm (or oddball task) is thought to reflect the amount of available attentional resources which can be devoted to the task. That is, as fewer attentional resources are available (due to the concurrent performance of a task increasing in difficulty), amplitude of the P300 response to a secondary discrimination task can be expected to decrease.

A primary advantage of assessing workload with an ERP response to an oddball discrimination is that unlike most secondary tasks, the ERP paradigm does not require an overt response (Baldwin, Freeman, & Coyne, 2004; Isreal, Chesney, Wickens, & Donchin, 1980). The operator may be asked to keep a mental tally of the number of oddball targets or to even to ignore the stimuli. An assessment technique that does not require an overt response could be utilized in an adaptive automation system to change task parameters or display modalities during periods of workload transition. Real time assessments of workload could be used to drive an adaptive interface which would present operators with the most effective system configuration for their current state. A driver, for example, might be provided only terse auditory guidance instructions from a navigational system during periods of high workload. Electronic maps or more detailed auditory guidance instructions could be used during periods of relatively low workload.

Despite numerous investigations, the merits of the P300 component in relation to other indices of workload warrant further examination. The P300 component is associated with perceptual and cognitive processes involved with stimulus evaluation and categorization while remaining relatively uninfluenced by response selection and execution (McCarthy & Donchin, 1981). Consequently, depending on the nature of task demands P300 amplitude may decrease with the introduction of an additional task, but may fail to differentiate between workload levels within a given task (Isreal et al., 1980).

For example, in an early investigation involving an oddball task in combination with a tracking task, P300 amplitude decreased with the introduction of the tracking task however, no decreases in P300 amplitude were found when tracking task difficulty was increased by manipulating the forcing-function bandwidth controlling velocity and acceleration components of the tracking task (Isreal et al., 1980).

The P300's lack of sensitivity to increased tracking difficulty was interpreted as evidence that the processes necessary to carry out the discrimination task did not rely on the same resources required by the increased demands of the tracking task. Specifically,

the tracking task tapped response-related resources while the P300 discrimination and counting task required perceptual resources (Isreal et al., 1980).

Wickens (1990) has emphasized that in certain circumstances electrophysiological measures may provide information either not feasibly obtained or not obtainable at all through behavioral performance measures. The conditions in which a dissociation in sensitivity to task difficulty occurs between ERP measures and performance measures is of primary interest to the current set of investigations. When dissociation is found between performance measures and P300 responses the two measures can be thought to reflect different aspects of workload and thus if used in combination can provide a more complete assessment of task load.

As previously indicated, P300 measures are thought to be sensitive to perceptual and cognitive processing load while being relatively unaffected by response-load manipulations (Coles, Smid, Scheffers, & Otten, 1995; Isreal et al., 1980; McCarthy & Donchin, 1981; Wickens, 1990). Therefore, if a given task becomes more difficult primarily because of increased response-related demands, performance measures are likely to change while P300 measures remain unaffected (Wickens, 1990). The current investigations illustrate task parameters aimed at assessing the concurrence between behavioral performance measures, P300 amplitude, and subject workload ratings.

Stimulus intensity and target probability have previously been demonstrated to increase P300 amplitude in both auditory and visual oddball sensory detection paradigms (Polich, Ellerson, & Cohen, 1996). Conversely, Kramer, Sirevaag, and Braune (1987) found that performance data (RT, accuracy) generated by an auditory detection task were not sensitive to changes in simulated flight task difficulty; however, peak amplitude for the P300 component was sensitive to manipulations in flight task difficulty. If sensitive to external task demands, ERP components have the potential advantage of not requiring an overt response by the participant (Mangun & Hillyard, 1995). However, both ERP components and performance measures may fail to assess the operators' internal perception of workload. Subjective assessment techniques provide a window into internal sources of workload (Hancock & Caird, 1993). Task relevant endogenous or internal aspects may be heavily influenced by the temporal demands of task performance as captured in a definition provided by Hancock and Caird (1993). They conceptualize workload as having three dimensions: 1) time for action; 2) perceived distance from the desired goal; and, 3) the level of effort required to achieve the desired goal.

The current investigations had two primary aims. First, we sought to determine if analogous versions of an auditory and visual discrimination task were of equal difficulty. It was reasoned that if analogous versions of an assessment task could be developed then future research could utilize these tasks in assessing the mental demand stemming from displays of different modalities. The second aim was to examine the relative sensitivity of the P300 ERP component, behavioral performance measures consisting of response time (RT) and accuracy for target detections and subjective ratings of the mental workload imposed by a series of complex tasks. In each investigation, a dual task paradigm involving a sensory (visual or auditory) discrimination task was performed in conjunction with a complex task involving either a simulated ATC to pilot communications flight task

(hereafter referred to simply as the flight task) or a simulated driving task. In each, the complex flight or driving task was designated as the primary task through repeated instructions and the sensory discrimination task was designated the secondary task. The sensory discrimination task consisted of an oddball paradigm in which participants were required to detect and respond to the infrequent “target” stimuli.

A battery of mental workload assessment measures were examined for relative sensitivity and potential dissociation. The physiological measure, P300 amplitude and latency, behavioral measures of RT and accuracy to targets and subjective ratings of workload were included in the battery. Subjective ratings of workload were obtained from responses on the NASA Task Load Index (NASA TLX). It was predicted that P300 amplitude would be greater for target versus distracter stimuli. More importantly, it was predicted that as the mental resource demands increased from single task to dual task trials and with increasing difficulty of the primary task, that amplitude of the P300 component would decrease. Similarly, it was predicted that detection errors and RT would increase and that subjective ratings of workload would also increase when participants were required to perform both tasks, relative to only the sensory discrimination task and most importantly with increased difficulty of the primary task. A general description of the methods and procedure of each investigation is presented next, followed by more specific details and results of each investigation.

GENERAL METHODS

Oddball Discrimination Task:

The secondary oddball task involved discriminating between a frequently and infrequently presented sensory stimuli presented via computer. Two analogous versions of the oddball task (visual and auditory) were modeled after methodology utilized by Kramer et al. (1987). The frequent or distracter stimuli were presented on average approximately 70% of the time and the infrequent or target stimuli were presented the remaining 30% of the time. Stimuli were presented at a rate of 1 per 1700 ms for 50 ms. Participants were instructed to indicate the presence of a target (infrequent stimuli) by pressing a response button. RT and accuracy for target detections were calculated for subsequent analysis.

Auditory detection task. Pure tones of differing frequency were presented in the auditory detection task. The distracter tone consisted of a 1500 Hz tone presented for 50 ms at an amplitude of approximately 65 dB C SPL. The target tone was a 1000 Hz tone presented at the same amplitude for the same duration.

Visual detection task. For the visual detection task, a color flashed on the entire 15-inch viewable area of a laptop computer located to the right of the participant as he/she faced the flight simulation monitor in Experiments 1 & 2 and just to the right of the driver’s seat at dashboard level in Experiment 3. The frequent or distracter color was green and the target color was red. Presentation rates and stimuli durations were equivalent to the auditory detection task conditions with distracter colors (green)

presented approximately 70% of the time and the infrequent target color (red) presented the remaining 30% of the time at a rate of 1 per 1700 ms for a duration of 50 ms. Participants were again instructed to indicate the presence of the targets by pressing a response button.

ERP Recording Equipment:

Electroencephalographic (EEG) activity was recorded from three sites (Fz, Cz, and Pz according to the International 10/20 system) and linked mastoids were used as references. Electroocular activity recorded from electrodes placed above and below the left eye was evaluated and used to edit ocular artifacts from the EEG data file. Electrode impedances were maintained below 5 k ohms. A NuAmps amplifier system in conjunction with Neuroscan 4.0 software was used for data collection and analysis.

General Procedure:

Participants were first introduced to the primary task simulations. They completed an orientation scenario to acquaint them with the controls and handling characteristics of the simulated vehicle (plane or car). Details of the specific vehicle simulation tasks utilized in each of the investigations are provided in the respective methods sections below. Participants were then given practice executing either the ATC flight commands (Experiments 1 & 2) or maintaining control of the simulated car (Experiment 3). Participants were then given practice on the respective oddball discrimination tasks and then practiced performing the simulated vehicle task concurrently with the oddball discrimination task. Following the completion of all practice trials, participants completed the respective experimental blocks in counterbalanced order. Baseline trials of each of the detection tasks were included at the beginning and end of the dual task trials and baseline flight/driving trials (flying or driving only trials) at each difficulty level were included in the counterbalanced experimental blocks. The entire experimental paradigm for each investigation lasted approximately two hours.

EXPERIMENT 1

Rationale

The first investigation sought primarily to determine the equivalency of the analogous auditory and visual versions of the oddball discrimination task. Toward this end, Experiment 1 compared the sensitivity and discrimination capabilities of P300 responses, RT and accuracy for target detections, and subjective workload ratings of workload to analogous visual and auditory secondary discrimination tasks while participants performed the tasks alone or in combination with a simulated flight task varying in difficulty. Participants flew a simulated flight path scenario while responding to either single or multiple auditory and text-based ATC commands.

Methods

Participants:

Seven undergraduate and graduate psychology students between 18 and 40 years of age voluntarily participated in the experiment. All participants had normal or corrected to normal vision and hearing as determined by self-report.

Flight Simulation Task:

Participants' primary task was to carry out flight instructions received either textually or verbally. Flight tasks were carried out using Microsoft Flight Simulator 2002. Altitude and heading flight instructions were accomplished using the autopilot, and speed instructions were carried out using the throttle control. These commands instructed participants to make altitude (± 1000 ft), airspeed (± 20 knots), or heading (± 45 degree) changes.

Flight task difficulty was manipulated by the number of ATC commands given (either textually or verbally) at one time. In single task conditions, participants were given one command and were allowed to execute that command before the next command was given. In the multiple task condition, participants were given three commands at a time, which they were to execute before being given the second set of commands. Each single command block consisted of 6 single commands, and each multiple command block consisted of 2 sets of three commands. There were a total of 8 blocks of commands divided across Modality (text/verbal), Difficulty (single command/multiple command), and Task (primary task only/both primary and secondary or dual task). The direction of the changes and type of change required by the ATC message were presented in a computer-generated randomized order in the single command conditions.

Textual and verbal commands were preceded by an auditory tone. Textual commands would then appear on the screen for 6 seconds in the single command conditions and 12 seconds in the multiple command conditions. Once participants read the instructions and understood the instructions they were to respond by pressing an acknowledge button. A repeat button, which displayed the message again, was also made available to participants. Once participants completed the command they were to press a button labeled acquire. There was a 10 second period between command completion and the subsequent command.

Verbal commands were presented using a similar method. Duration of the single commands was approximately 4 seconds and multiple command duration was approximately 12 seconds. The same acknowledge, acquire, and repeat controls were available to the participants.

Procedure:

The experiment began with a brief practice session and initial baseline trials for the visual and auditory secondary tasks. This was followed by training on the primary flight task. Participants completed single commands in one modality followed by multiple commands in the other modality. Cross-modal pairings were used to combine the primary flight task with the sensory detection task. Therefore, when participants were presented with the text ATC commands they concurrently performed the auditory oddball task and vice versa when hearing ATC commands they concurrently performed the visual oddball task. Upon completing the training for both primary and secondary tasks the participants proceeded to complete the eight experimental flight blocks, and an additional baseline trial for the secondary tasks.

Results

P300 Component

The primary focus of the ERP analysis was the P300 component. Peak amplitude was defined as the largest positive deflection occurring between 300 to 500 ms after stimulus presentation and peak latency was defined as the time in ms that this peak deflection occurred. Mean peak amplitude was significantly larger for target stimuli than for distracters, $F(1, 6)=9.06$, $p<.05$. There was no significant main effect for electrode

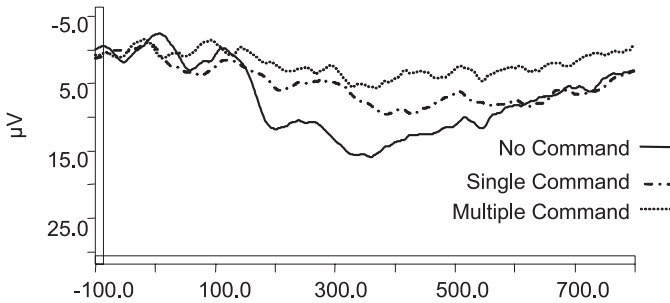


Fig. 1. Auditory evoked potential for each task at Cz in Experiment 1.

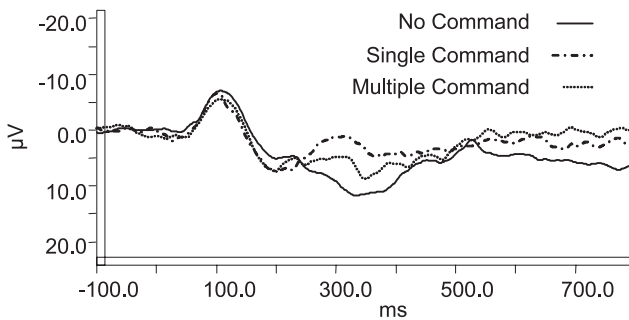


Fig. 2. Visual evoked potential for each task at Cz in Experiment 1.

(Fz, Cz, and Pz) $F(2, 12)=0.03, p>.05$. However there was a significant interaction of electrode and trial type $F(2, 12)=7.02, p<.05$. The Fz electrode was not sensitive to peak amplitude differences between target trials and distracter trials, however both the Cz and Pz electrodes were sensitive to trial type. All subsequent analyses are for responses to target trials only.

There was a significant main effect of task load, $F(2, 12)=5.80, p<.05$. Mean amplitude in μV was 14.07 for the secondary-task alone, 8.24 for the single command conditions, and 7.44 for the multiple command conditions. Though the trend was in the expected direction across all conditions, only the difference between performing the ERP task alone (single task) and both the ERP task and the flight tasks concurrently (dual task) was significant. The peak amplitude was not significantly different between single and multiple command conditions in the flight task. The ERP waves for site Cz for both the auditory and visual secondary-tasks are presented in Figs. 1 and 2, respectively.

Behavioral Measures

Accuracy and RT measures were also taken from the secondary discrimination task. There was no significant main effect of modality on accuracy ($F(1, 6)=.2, p>.05$); however, there was a main effect of modality on response time ($F(1, 6)=27.88, p<.05$). Response times were longer for the auditory discrimination task (515 ms auditory and 436

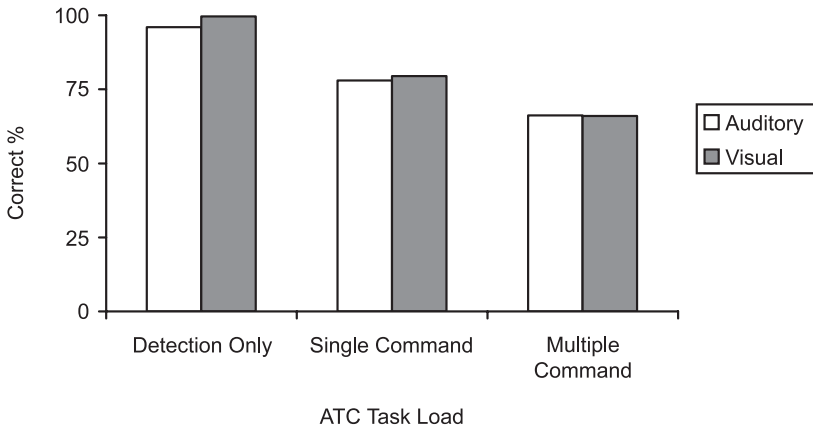


Fig. 3. Mean accuracy for each task as a function of secondary-task modality in Experiment 1.

ms visual).

There was a significant main effect of task load for both the accuracy and response time data, ($F(2, 12)=21.33, p<.05$ and $F(2, 12)=25.73, p<.05$, respectively). Both measures could discriminate between single task and dual task conditions. However, only the accuracy data revealed significant differences between the primary task difficulty manipulation of single versus multiple ATC commands (79% and 66% respectively). Fig. 3 presents accuracy data in each task load condition.

Subjective Workload

Participants completed a computerized version of the NASA TLX at the end of each condition. The unweighted averages of the six TLX dimensions were analyzed using a three way repeated measures Analysis of Variance (ANOVA) (2 command modality \times 2 presence/absence of secondary-task \times 2 number of commands). There was no significant main effect for either command modality, or number of commands, ($F(1, 6)=.02, p>.05$, and, $F(1, 6)=.41, p>.05$, respectively).

A main effect was found for the presence of the secondary-task, $F(1, 6)=50.40, p<.05$. Subjective ratings were significantly larger when the secondary task was presented in combination with the flight task (dual task conditions) as compared to when the flight task was performed alone.

Primary-Task Performance

Several performance measures were taken from the primary flight task. These included the time taken to acknowledge the command, the time taken to accomplish the command, the number of times the message was repeated, heading errors (± 20 degrees), speed errors (± 15 knots), and altitude errors (± 100 feet). Due to the lack of flight experience and use of the autopilot only large deviations were called errors. All of the performance measures were analyzed using three way repeated measures ANOVAs (2 command modality \times 2 presence/absence of secondary-task \times 2 number of commands).

Primary-task analysis was principally conducted to determine if there were performance decrements caused by the addition of the secondary-task. The presence and absence of either secondary task (auditory or visual) did not result in a significant difference between the number of repeats or heading, speed, and altitude errors. There were also no significant differences for the time to acknowledge the message or the time to accomplish the commands.

Analysis of the performance data revealed differences between the single and multiple command conditions. The time required to acknowledge and accomplish multiple commands was significantly longer than the time required for single commands ($F(1, 6)=28.83, p<.05$, and $F(1, 6)=20.21, p<.05$). No significant differences were revealed by any of the dependent measures for the command modality (verbal/textual).

Discussion

Results indicate that the auditory and visual versions of the oddball discrimination task were of equal difficulty. These findings are encouraging as this was a primary aim of Experiment 1 and the results provide preliminary support for the use of this cross modality paradigm as a technique for comparing the mental workload required by visual versus auditory displays.

Results indicate the P300 component as well as accuracy, RT, and subjective ratings were sensitive to the presence or absence of the primary flight task. Specifically, mean peak amplitude of the P300 in the auditory oddball paradigm was decreased with the introduction of a text based ATC command task while P300 amplitude elicited by a visual oddball task was sensitive to the introduction of a speech based ATC command task. For both auditory and visual paradigms, amplitude was highest and latency was shortest in baseline conditions, where no ATC commands were given. However, only accuracy distinguished between single and multiple task command conditions.

It was expected that P300 amplitude would distinguish between difficulty levels of the flight task resulting from responding to single versus multiple ATC commands. The P300 results in the current investigation did not support this prediction. A nonsignificant trend towards decreased P300 amplitude between single and multiple command conditions however, was in the expected direction. One explanation for these results may be that the current difficulty manipulation relied primarily on increased response execution resources rather than increased perceptual cognitive resources. Previous research has indicated that P300 responses are relatively immune to increased response-load demands (Coles et al., 1995; McCarthy & Donchin, 1981).

A second explanation for the failure of the P300 component to distinguish between single and multiple command conditions could be low statistical power due to the small sample size and large individual differences. However, previous research (Kramer et al., 1987) found significant differences in P300 responses with a comparable sample size. Further, if P300 responses are to be used as an on-line index of mental workload for application in adaptive systems, they will need to provide robust measurements

distinguishable without large sample sizes.

Yet another interpretation is that perhaps the cross modal paradigm utilized in Experiment 1 resulted in a secondary discrimination task that utilized resources which were relatively independent from primary task resources. In order to investigate this possibility, Experiment 2 was conducted. Experiment 2 utilized the same basic task paradigm with the exception that discrimination task and the primary ATC command flight task were presented in the same modality.

EXPERIMENT 2

Rationale

Experiment 1 provided initial evidence that the two secondary discrimination tasks (visual and auditory) were of relatively equivalent difficulty and were suitable for comparing the sensitivity of different workload assessment techniques. Experiment 2 was based on assumptions stemming from Wickens' multiple resource theory (Wickens, 1984, 1991) that two tasks sharing the same modality (auditory/visual) would draw upon resources from the same limited capacity reservoir. It was hypothesized that implementation of a secondary task requiring the same perceptual processing modality as the primary task would result in a more sensitive index of changes in primary task difficulty (O'Donnell & Eggemeier, 1986). Thus in Experiment 2, the visual discrimination task was presented in combination with the visual version of the primary flight task and conversely the auditory discrimination task was presented in combination with the auditory version of the primary flight task. In all other respects the primary task and secondary tasks were equivalent to Experiment 1. It was predicted that P300 amplitude would decrease and that RT, errors and subjective ratings of workload would increase with increased primary task difficulty stemming from the presentation of multiple versus single ATC commands during the primary flight task.

Methods

Participants:

Seven undergraduate and graduate psychology students ranging in age from 18 to 40 years voluntarily participated in the experiment. All participants had normal or corrected to normal vision and hearing as determined through self-report.

Flight-Task Simulation:

The flight task utilized in Experiment 2 was the same as that used in Experiment 1 and the procedures were equivalent with the noted exception that intra-modal pairings of the oddball discrimination task and ATC command task were used rather than the cross-modal pairings utilized in Experiment 1.

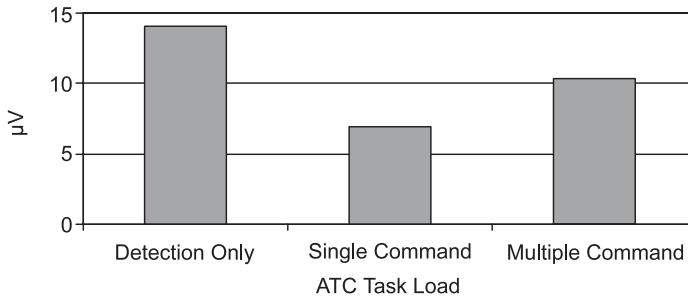


Fig. 4. P300 peak amplitude as a function of task load in Experiment 2.

Results

P300 Component

As in Experiment 1, the primary focus of the ERP analysis was the P300 component. P300 mean peak amplitude was significantly larger for target stimuli than for distracters, $F(1, 6)=60.07, p<.05$. Responses to target stimuli were analyzed in a two way repeated measures ANOVA (2 secondary task modality \times 3 task difficulty).

As in Experiment 1, there were no significant differences in either peak amplitude or peak latency for modality. There was a significant main effect of task difficulty on peak amplitude, $F(2, 12)=4.76, p<.05$, but not on peak latency. There was a significant difference in peak amplitude between the single command and baseline conditions. However, as in Experiment 1, P300 response did not differentiate between the single and multiple command difficulty manipulation. The P300 amplitude as a function of task load is presented in Fig. 4. The trend in peak latency data, although not significant, was in the predicted direction with a mean latency in milliseconds and standard deviation in parenthesis of 375 (54), 394 (53), and 412 (48) in the discrimination task only, the single command, and multiple command conditions, respectively.

Behavioral Performance

Accuracy and RT measures for the secondary discrimination tasks revealed no significant main effects of modality on accuracy, $F(1, 6)=1.41, p>.05$, nor RT, $F(1, 6)=3.88, p>.05$. There was a significant main effect of task load for both accuracy and RT data, $F(2, 12)=14.40, p<.05$, and $F(2, 12)=43.53, p<.05$, respectively. Both measures discriminated between all three levels of difficulty. The data revealed significant differences between single task (detection task only) performance and dual task performance as well as between the single and multiple command conditions of the flight task. There was a significant performance decrement (i.e. an increase in RT and decrease in accuracy) when the flight task was added and when the difficulty of the flight task increased. The accuracy and response time data are presented in Figs. 5 and 6, respectively.

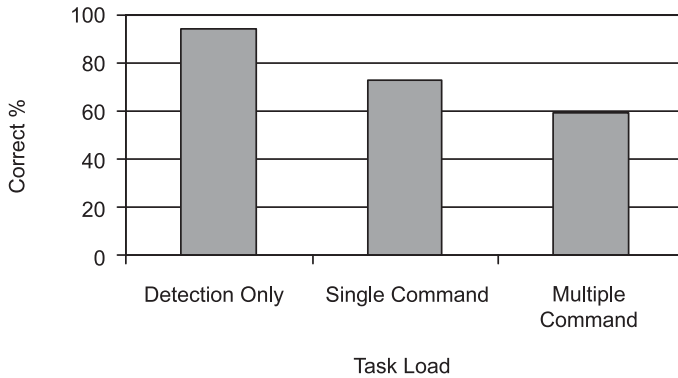


Fig. 5. Mean accuracy of target detections in Experiment 2.

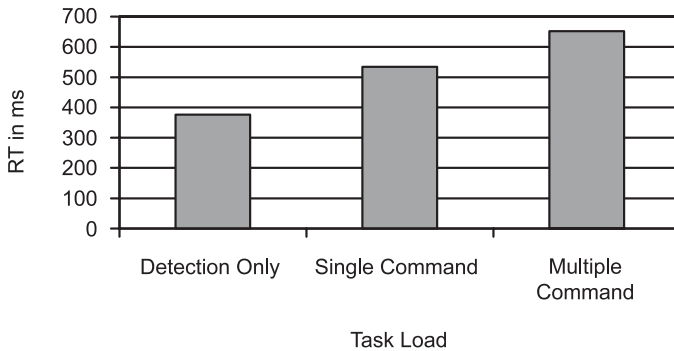


Fig. 6. Mean RT to target detections in Experiment 2.

Subjective Ratings

The unweighted average of the six NASA TLX dimensions were analyzed using a 2 command modality (text, speech) \times 3 secondary discrimination (visual, auditory, none) \times 2 task load (single, multiple) repeated measures ANOVA. There was no significant main effect for either discrimination task modality or command modality ($F(1, 6)=1.11, p>.05$, and $F(1, 6)=5.78, p>.05$ respectively).

A main effect was found for task load, $F(2, 12)=7.52, p<.05$. Subjective ratings were significantly larger when either secondary task (visual or auditory) was presented in combination with the flight task as compared to when the flight task was performed alone. However, no differences were observed between single and multiple command conditions.

Primary-Task Performance

Primary task measures included the time taken to acknowledge the command, the time taken to accomplish the command, the number of times the message was repeated, heading errors (± 20 degrees), speed errors (± 15 knots), and altitude errors (± 100

feet). Performance measures were analyzed using three way repeated measures ANOVAs (2 command modality \times 2 presence/absence of secondary-task \times 2 number of commands). As in Experiment 1, the presence or absence of the secondary task in either modality did not result in a significant difference between the numbers of heading, speed, or altitude errors. There were also no significant differences for the time to acknowledge the message or the time to accomplish the commands.

Analysis of the performance data revealed differences between the single and multiple command conditions. The time required to acknowledge and accomplish multiple commands was significantly longer than the time required for single commands ($F(1, 6)=9.28, p<.05$, and $F(1, 6)=26.99, p<.05$). No significant differences were revealed by any of the dependent measures for the command modality (verbal/textual).

Discussion

Consistent with the findings of Experiment 1, the auditory and visual detection tasks appear to be of relatively equal difficulty. Participants performed both tasks with equal accuracy, RT, and demonstrated equivalent P300 amplitudes and latencies for each version of the detection task. No differences were found in subjective ratings of difficulty or in primary flight task performance measures as a function of the discrimination task modality. These results provide further support for the suitability of the sensory detection tasks for comparing the mental workload of many of the emerging display technologies that shift information previously presented through voice/radio into a visual format.

Secondary task behavioral performance measures of accuracy and RT were sensitive to changes in flight task difficulty. However, contrary to our predictions the P300 response was not sensitive to changes in flight task difficulty. These results are also comparable to the findings of Experiment 1. Relative to P300 responses and subjective ratings, behavioral measures of RT and accuracy were more sensitive to increased primary task difficulty in Experiments 1 & 2. This finding indicates that the cross modal pairings utilized in Experiment 1 did not account for the lack of P300 sensitivity. Though previous investigations have demonstrated sensitivity with a comparable sample size, the lack of sensitivity in the P300 response in the first two experiments may be due to a lack of statistical power. An alternative explanation is that the difficulty manipulation in the first two experiments failed to require additional perceptual and cognitive resources, relying instead primarily on increased response execution resources. Experiment 3 examined these possibilities.

EXPERIMENT 3

Rationale

Experiment 3 was conducted to further examine the potential sensitivity of the P300, behavioral measure and subjective rating assessment battery to changes in workload. The

primary task was changed in order to examine additional task demand components while continuing to maintain a complex, realistic paradigm. A simulated driving task which allowed manipulation of perceptual demands (presence or absence of fog) and cognitive-response related demands (presence or absence of traffic) independently was utilized. Additionally, the participant size was increased in Experiment 3 in order to determine if the lack of sensitivity to increased primary task demands found in the first two experiments was possibly a result of low statistical power.

Methods

Participants:

Data from fourteen volunteer participants ranging in age from 18-35 years were analyzed. All were licensed drivers and reported normal hearing and visual abilities.

Driving Simulation Task:

The primary task for Experiment 3 was a driving simulation task presented on a General Electric Capital I-Sim® Patrol II simulator. The simulator is equipped with a full set of operational controls including steering wheel, brake and accelerator pedals as well as lights, turn signal indicators, etc. The side screens allow presentation of a 180 degree horizontal field of view. Side rearview mirrors allow the driver to monitor traffic from all directions.

Driving Scenarios:

Four roadway scenarios were presented. A low traffic density freeway scenario and a moderate traffic density urban scenario were presented in either clear visibility or heavy fog. Each scenario lasted approximately 5 minutes.

Oddball Discrimination Task:

Due to the consistent results obtained between the analogous visual and auditory version of the oddball paradigm used in Experiments 1 & 2 a determination was made to examine only the visual version of the discrimination task. This enabled greater experimental manipulations of primary task demands while maintaining the duration of the total paradigm to within 2 hours.

Procedure:

The procedure was essentially equivalent to the first two experiments with the exceptions already noted. Participants were introduced to the driving simulator and allowed practice on both the driving task by itself, followed by the oddball task and then performing both tasks concurrently. Experimental conditions were presented in a randomized semi-counterbalanced order.

Results

A series of 2 (roadway type: freeway versus urban) by 2 (visibility: clear versus heavy fog) repeated measures ANOVAs were analyzed for each of the dependent measures of discrimination task performance. An alpha level of .05 was again used to determine statistical significance for all analyses.

P300 Component

P300 amplitude for targets was again the main focus of the ERP analysis. A main

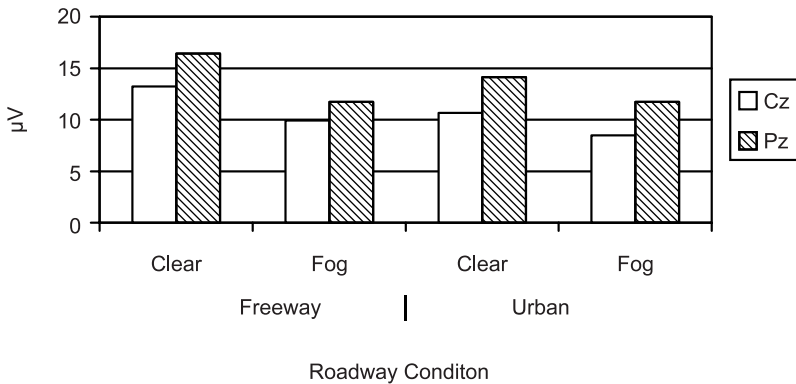


Fig. 7. P300 amplitude at Cz and Pz as a function of road type and visibility in Experiment 3.

effect for P300 amplitude for the visibility manipulation was observed at sites Cz, $F(1, 8)=26.15$, $p<.05$, and Pz, $F(1, 8)=12.99$, $p<.05$. P300 amplitude was not sensitive to the road type manipulation at either Cz or Pz, nor was a significant interaction between road type and visibility observed. Fig. 7 illustrates the mean P300 amplitude for sites Cz and Pz as a function of road type and visibility.

Behavioral Measures

Accuracy and RT to target discriminations were again analyzed as behavioral measures of the secondary task. There was a significant main effect for road type (freeway versus urban) for both accuracy and RT to the detection task, $F(1, 8)=81.68$, $p<.05$ and $F(1, 8)=7.99$, $p<.05$, respectively. Participants correctly detected more targets in the freeway scenarios (90% and 89% in clear visibility and fog, respectively) relative to the urban scenarios (69% and 75% in clear visibility and fog, respectively). Similarly, participants responded faster to targets during the freeway scenarios (493 ms and 475 ms in clear visibility and fog, respectively) relative to detections during urban scenarios (536 ms and 583 ms, respectively). The manipulation of visibility did not result in a main effect for either accuracy or RT. However, a significant interaction between road type and visibility was observed for accuracy, $F(1, 8)=10.89$, $p<.05$. No interaction was observed for RT.

Subjective Ratings

The unweighted average of the six NASA TLX dimensions revealed a significant main effect for road type, $F(1, 12)=6.91$, $p<.05$. The urban roadway scenarios were rated as significantly more difficult than the freeway scenarios. Unweighted mean TLX ratings across the six workload dimension for the urban scenarios were 4.3 and 4.5 for the clear visibility and fog conditions, respectively, compared to ratings of 3.1 and 2.6 for clear visibility and fog in the freeway roadway conditions. No other main effects or interactions were significant for the NASA-TLX composite measure.

Discussion

Results of Experiment 3 revealed dissociation between workload sensitivity of the P300 responses and behavioral task measures. P300 amplitude was sensitive to increased driving task demand due to reduced visibility (presence of fog) but was not sensitive to changes in road type (urban versus freeway). Conversely, RT and accuracy to the secondary task discriminations were sensitive to changes in road type (the traffic density manipulation) but were not sensitive to the visibility manipulation. Similar to behavioral task measures, subjective workload ratings were sensitive only to the road type manipulation.

GENERAL DISCUSSION

The results of three experiments combining an oddball sensory discrimination task with a complex simulated transportation task were reported. Experiments 1 and 2 compared analogous auditory and visual versions of the discrimination task. Results of these experiments indicate that the auditory and visual versions of the task are of relatively equivalent difficulty. P300 amplitude, RT, accuracy and subjective ratings for the two tasks were roughly equivalent in the auditory and visual versions of the discrimination task. These findings indicate that analogous versions of the task appear suitable for comparing the mental workload stemming from displays of different modalities (i.e., data link text displays versus radio voice ATC communications).

Comparison of overall sensitivity of the workload metrics used in the current investigations revealed an interesting pattern of results. In general, behavioral measures of RT and accuracy to sensory discriminations were more sensitive to fluctuations in workload than either the P300 component or the subjective ratings. However, the dissociation in sensitivity between P300 amplitude and the other indices observed in Experiment 3 indicate that P300 measures may be used to differentiate between task demands not readily observable with other indices (Wickens, 1990). The visibility manipulation can reasonably be assumed to have primarily increased the perceptual processing demand of the simulated driving task. Only the P300 metric was sensitive to this manipulation. This dissociation further underscores the multifaceted nature of workload present in many complex tasks.

Results of the three current investigations support the findings of previous research that indicate that P300 amplitude and latency are not affected by changes in the response characteristics of an operational task (Coles et al., 1995; McCarthy & Donchin, 1981). Further, the dissociation between P300 response sensitivity and behavioral measures in Experiment 3 provides further support for the observations in previous studies that the P300 component is particularly sensitive to manipulations of perceptual task demands (Isreal et al., 1980).

The sensitivity of P300 response to the oddball discrimination task as implemented in the current paradigm was not adequate as a stand alone method of detecting changes in

operator workload state. An adaptive vehicle interface system would require that both endogenous and exogenous factors contributing to operator state be assessed in real time, preferably without requiring the operator to perform another auxiliary task. The assessment indices investigated in the current series of experiments do not appear suitable for this aim. Further research examining the sensitivity of different ERP components and alternative physiological measures of brain activity is required before such an adaptive interface could be developed.

Ongoing research is currently underway to further examine the task characteristics which may place demands on different aspects of processing resources and to identify workload metrics sensitive to these differential aspects. Results of the current investigations indicate that a battery of workload assessments are needed to fully appreciate the workload involved in complex environments.

REFERENCES

- Baldwin, C. L. 2003. Neuroergonomics of mental workload: New insights from the convergence of brain and behavior in ergonomics research. (Commentary). *Theoretical Issues in Ergonomics Science*, **4**, 132–141.
- Baldwin, C. L., Freeman, F., & Coyne, J. T. 2004, September. *Mental workload as a function of road type and visibility: Comparison of neurophysiological, behavioral, and subjective indices*. Paper presented at the Human Factors and Ergonomics Society, New Orleans, LA.
- Coles, M. G., Smid, H. G. O. M., Scheffers, M. K., & Otten, L. J. 1995. Mental Chronometry and the study of human information processing. In M. D. Rugg & M. G. Coles (Eds.), *Electrophysiology of the mind: Event-related brain potentials and cognition*. (pp. 86–131). Oxford: Oxford University Press.
- Gopher, D., & Donchin, E. 1986. Workload-An examination of the concept. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance*. Vol. II: Cognitive Processes and Performance (pp. 41-41–41-49). New York: John Wiley & Sons.
- Hancock, P. A., & Caird, J. K. 1993. Experimental evaluation of a model of mental workload. *Human Factors*, **35**, 413–429.
- Isreal, 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.
- Kahneman, D. 1973. *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.
- Knowles, W. B. 1963. Operator loading tasks. *Human Factors*, **9**, 155–161.
- Kramer, A. F. 1991. Physiological metrics of mental workload: A review of recent progress. In D. L. Damos (Ed.), *Multiple-task performance* (pp. 279–328). London: Taylor & Francis.
- Kramer, A. F., Sirevaag, E. J., & Braune, R. 1987. A psychophysiological assesment of operator workload during simulated flight missions. *Human Factors*, **29**, 145–160.
- Mangun, G. R., & Hillyard, S. A. 1995. Mechanisms and models of selective attention. In M. D. Rugg & M. G. Coles (Eds.), *Electrophysiology of the mind: Eventrelated brain potentials and cognition* (25th ed., pp. 40–85). Oxford: Oxford University Press.
- McCarthy, G., & Donchin, E. 1981. A metric for thought: A comparison of P300 latency and reaction time. *Science*, **211**, 77–80.
- Meshkati, N., Hancock, P. A., & Rahimi, M. 1990. Techniques in mental workload. In J. R. Wilson & E. N. Corlett (Eds.), *Evaluation of human work: A practical ergonomics methodology* (pp. 605–627). Philadelphia, PA: Taylor & Francis.
- Moray, N., Dessouky, M. I., Kijowski, B. A., & Adapathya, R. 1991. Strategic behavior, workload, and performance in task scheduling. *Human Factors*, **33**, 607–629.
- O'Donnell, R. D., & Eggemeier, F. T. 1986. Workload assessment methodology. In K. Boff, L. Kauffman & J. P. Thomas (Eds.), *Handbook of perception and human performance* (pp. 42-41–42-49). New York:

- Wiley.
- Polich, J., Ellerson, P. C., & Cohen, J. 1996. P300, stimulus intensity, modality, and probability. *International Journal of Psychophysiology*, **23**, 55–62.
- Wickens, C. D. 1984. Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63–101). Orlando, FL: Academic Press.
- Wickens, C. D. 1990. Applications of event-related potential research to problems in human factors. In J. W. Rohrbaugh, R. Parasuraman, & R. Johnson, Jr. (Eds.), *Event-related brain potentials: Basic issues and applications* (pp. 301–309). London: Oxford University Press.
- Wickens, C. D. 1991. Processing resources and attention. In D. L. Damos (Ed.), *Multiple-task performance* (pp. 3–34). London: Taylor & Francis.
- Wickens, C. D. 2002. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, **3**, 159–177.
- Williges, R. C., & Wierwille, W. W. 1979. Behavioral measures of aircrew mental workload. *Human Factors*, **21**, 549–574.

(Manuscript received January 6, 2005; Revision accepted April 29, 2005)