Japanese Psychological Research
2002, Volume 44, No. 3, 134–143

Effects of task difficulty and time-on-task on mental workload

SHIGERU HAGA
Department of Psychology, Rikkyo University, Toshima-ku, Tokyo 171-8501, Japan

HARUO SHINODA
Department of Education, Ibaraki University, Miko 310-0056, Japan

MITSUTERU KOKUBUN
Toyota Central R & D Laboratories, Nagakute, Aichi 480-1192, Japan

Abstract: Twelve subjects performed a tracking task and a memory search task simultaneously on a computer screen. The dual task continued for approximately 10 min and was repeated three times, interrupted by a short break for subjective ratings: the NASA Task Load Index (NASA-TLX) and the Check List of Mental Condition. Event related potentials (ERP) evoked by the presentation of memory task stimuli were also recorded. All the subjects participated in three experimental sessions, which varied in difficulty of tracking task. Results demonstrated that the NASA-TLX and ERP were sensitive only to the change in task difficulty and were not affected by time-on-task or interaction between task difficulty and time-on-task.

Key words: mental workload, fatigue, NASA Task Load Index, event-related potentials, railways.

When assessing workload in industry, such as manufacturing and transportation, time-on-task, as well as task demand, is an important loading factor. For example, the workload of railway driving when operating a high-speed train should be assessed in terms of information processing load and time on continuous driving. It must be noted that the consequences of task performance over a prolonged period, such as fatigue increment and vigilance decrement, have a complex relationship with the effects of task demand, when the demand is mainly mental.

There were many studies in 1960s and 1970s dealing with the effects of time-on-task on vigilance or sustained attention (see Warm, 1984). Some of these studies manipulated task demand such as event rate and described the change in performance over time-on-task (e.g. Parasuraman, 1979).

In recent mental workload studies, on the other hand, the effects of time-on-task have been neglected. This indifference probably comes from the North-American view of the concept of mental workload. Jex (1988) defined mental workload as "the operator's evaluation of the attentional load margin" (p. 11); Eggemeier (1988) defined it as "the degree of processing capacity that is expanded during task performance" (p. 41); and Wickens (1992) wrote that "the concept of workload is fundamentally defined by this relationship between resource supply and task demand" (p. 41, l.23–25). This is because in North America, mental workload assessment is required mainly for system design, especially in developing or appraising an aircraft's cockpit and a nuclear power plant's control panels. There are several measures and assessment techniques that are

JPR_016 Pages: 10

said to be sensitive to mental workload. Among them are heart rate variability (HRV), event-related potentials (ERP), dual-task methods, and the two major rating scales known as the SWAT (the Subjective Workload Assessment Technique) and the NASA-TLX (Task Load Index). The authors think that they are all resource-oriented (see Haga, 1993; for further discussion on this point).

Ergonomists in Continental Europe have more concern about the consequences of mental workload after some period of task accomplishment. The ISO 10075, for example, which was composed by a committee chaired by German ergonomist Nachreiner, defined the terms related to mental workload, including mental fatigue, monotony, reduced vigilance, and mental satiation (International Standardization Organization, 1991). Then, ISO/DIS 10075–2 proposed guidelines for designing a "work system", and the guidelines are organized into sections titled "Guidelines concerning fatigue", "Guidelines concerning monotony" etc. (International Standardization Organization, 1996). This means that the Standard mainly aims at preventing impairing effects on workers under a prolonged workload.

As for physical workload, it is conceptually reasonable to assume that the amount of accumulated workload effect on the worker (i.e. typically, fatigue) can be estimated by integrating work intensity by time-on-task. As for mental workload, on the other hand, accumulated size of the effect of workload may not correspond to measured size of momentary workload. However, there are few studies which estimate change in measures of mental workload over time-on-task.

In the experiment reported below, task difficulty and time-on-task were introduced as independent variables of workload, and sensitivities of various mental workload measures to each variable are examined. We predict that resource-oriented measures are sensitive to task difficulty but not to time-on-task, while measures of mental fatigue and low arousal level will show an opposite pattern of sensitivity. Among the measures of the first type, secondary task performance, ERP, and NASA-TLX are selected because they are most frequently used mental workload measures and representative of behavioral, physiological, and subjective techniques, respectively. For second type measures, we used the scales of the Check List of Mental Condition (CLMC) constructed by Shinoda (1991). Electrocardiogram (ECG) was also recorded, because heart rate and heart rate variability (HRV) are influenced by autonomic nervous system activities, which react to task difficulty as well as time-on-task.

Method

Subjects

Twelve graduate and undergraduate students with normal vision volunteered to participate in the experiment. Ten were male and two were female; the age range was 20–28. All subjects had good experience in physiological measurement and had been well trained in the experimental tasks.

Tasks and apparatus A dual-task method was used combining pursuit tracking as the primary task with a memory search as the secondary task, both of which were performed on a computer screen (Figure 1).

The pursuit tracking task required the subject to keep a target (a small white disc) moving on a circumference between a pair of cursors (white lines) controlled by turning a small wheel. The target moved at a fixed speed, changing its direction at random intervals. When the target was out of bounds, the color of the target and the cursors turned red as a warning. The difficulty of the primary task was varied by the order of the wheel-cursor control function: zero-order (location) control (Condition E), the first-order (speed) control (Condition M), and the second-order (acceleration) control (Condition D). In Condition E, the cursor moved right when the subject turned the wheel clockwise, and moved left when the subject turned the wheel counterclockwise. The distance of the cursor's movement was proportional to the rotation angle of the wheel. In Condition M, the direction and speed of the cursor's

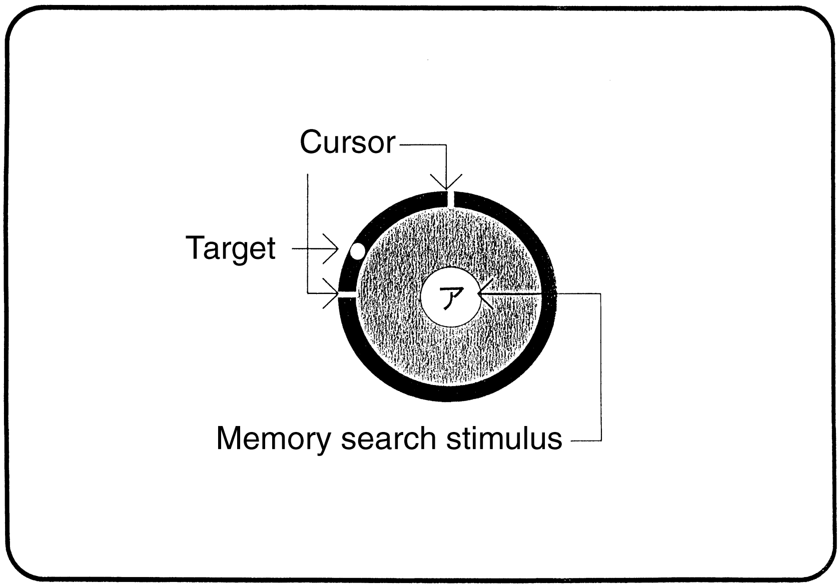


Figure 1. Computer display for the tracking and memory tasks.

movement was proportional to the rotation angle of the wheel. For instance, when the subject turned the wheel clockwise and stopped turning, the cursor kept going right at a constant speed, and when he/she wanted to stop the cursor, he/she had to turn the wheel back to the "home position". In Condition D, the cursor kept increasing its speed in proportion to the angle of the wheel. Therefore, when the subject wanted to stop or slow the cursor, he/she had to turn the wheel to the direction opposite to the cursor's movement in order to brake (accelerate in the opposite direction) the cursor.

The memory search task required the subject to memorize five *katakanas* (Japanese phonetic characters) before the tracking task started and to press the key while performing the tracking task when one of them appeared in the round window at the center of the tracking circle. A katakana was displayed in the window for 500 ms at random interstimulus intervals varying from 3 to 7 s ($M = 5$ s). Katakanas were presented 90 times during a consecutive block of tracking tasks in a random order, and 10 of them belonged to the memory set (target characters which required the subject's response). The memory set was randomly chosen for each block by the computer program.

A block of dual-tasks continued for approximately 10 min and was repeated three times in an experimental session.

The experimental tasks were programmed using N88-BASIC (NEC; Tokyo, Japan) and ran on a PC-9801RA21 personal computer (NEC). The stimuli were presented on a 14′ color CRT display GVM1415 (SONY, Tokyo, Japan), with the subject seated 100 cm away. Reaction time, duration of stimulus presentation, and interstimulus intervals were measured and/or controlled using a TIR-6 (98) timer board (CONTEC, Osaka, Japan) with a precision of 10 ms.

Performance measures

For performance measures, total duration of time that the target disc was out of the boundaries was recorded for the primary task, and reaction time between the onset of the target character and depression of the key by the subject for the secondary task.

Subjective measures

The Japanese version of NASA-TLX (Haga & Mizukami, 1996) and the CLMC (Shinoda, 1991; Haga, Shinoda, Kokubun & Fujinami, 1994) were used for subjective measures.

The NASA-TLX is one of the most widely known and used mental workload scales and it consists of six subscales: (i) mental demand, (ii) physical demand, (iii) own performance, (iv) temporal demand, (v) effort, and (vi) frustration (Hart & Staveland, 1988; Miyake & Kumashiro, 1993). Before rating workload, any possible pair of subscales is presented and the subject is asked to choose the more important loading factor for him/her when performing the given type of task. Weights of the subscales are calculated from the results of the paired-comparison, and the ratings of workload on the six subscales are combined into a single measure named Weighted Workload (WWL) score using the weights. In addition to the six original subscales, the Japanese version of NASA-TLX used in this experiment had a supplementary scale titled "Overall Workload" on which the subject rated his/her holistic feeling of workload.

The paired-comparison of the subscales was conducted after practice in the first session of each subject. When making the importance judgment, the subject was instructed to think about such computer-based tasks as those just practiced (including the tracking and the memory tasks). The same set of weights for the six TLX scales was used under all three difficulty conditions. After each dual-task trial block, the subject was asked for ratings on six TLX scales for the tracking and the memory tasks separately.

The subject operated a mouse to move a pointer along the individual scales displayed on the computer screen, and clicked the mouse button at a rating point. The computer program converted the position of the mouse-controlled pointer on the scales into digital values ranging from 0 to 100 and calculated the WWL scores using the "weight file" for individual subjects.

The subject was also asked to rate his/her psychophysiological state before the first block and after each block. Six subscales from the CLMC were used for this purpose: (i) awareness (drowsy *vs* clear), (ii) mood (feeling bad *vs* feeling good), (iii) relaxation (tense *vs* relaxed), (iv) comfort level (uncomfortable *vs* comfortable), (v) irritability (not irritated at all *vs* extremely irritated), and (vi) fatigue (not fatigued at all *vs* extremely exhausted). The subscales were presented on the computer screen and the subject rated his psychophysiological state by moving a pointer on the scales with a mouse. The computer converted the position of the pointer to a number ranging from 0 to 100 and recorded it. The difference in the ratings from the initial state (ratings before the first trial) was considered as the measure of effect of workload on a subject's psychophysiological state.

Physiological measures

Electroencephalogram (EEG), electrocardiogram, respiratory curve, and vertical electro-oculogram (EOG) were recorded.

An EEG was recorded for the purpose of obtaining ERP evoked by the memory task stimuli. It was measured with reference to linked ears, with four Ag-AgCl electrodes placed at frontal, central, parietal, occipital regions (Fz, Cz, Pz, Oz in the international 10–20 system). An NEC San-ei 6R12 electroencephalograph (bandwidth: 0.5–30 Hz, time constant: 0.03 s) and an NF-5870-PCM data recorder were used for this purpose.

EOG responses were picked up by sub- and supra-orbital electrodes, and were amplified and recorded with the same device used for the EEG. The EOG was used as a reference data in the analysis of the EEG.

To obtain an ERP, approximately 80 EEG responses to the secondary task stimuli were averaged at an analysis time of 1024 ms (giving a resolution of 2 ms/point) including a 100-ms prestimulus period preceding secondary task presentation, using a personal computer (NEC PC-9801RA21) through A/D board (ANALOG-PRO-DMA; Canops, Kobe, Japan). EEG responses contaminated by blinks or large vertical eye-movements (over 100 μV EOG fluctuations) were not used for the analysis.

An ECG and a respiratory curve were recorded for 3 min between the fourth and the sixth minute in the rest period and during the dual-task performance. The ECG was recorded bipolarly from left and right forearm with the time constant 0.03 s, and fed into a PCM data recorder. The respiratory curve was observed with a thermistor pickup plastered under the nostril, and recorded on the PCM recorder with the time constant 3 s. From the ECG record, heart rate (per min) and coefficient of variance of R-R interval (CV-RR) were obtained, using the respiratory curve as a reference.

Procedure

Each of the 12 subjects participated in three experimental sessions with different tracking-difficulty conditions. The order of conditions applied was counterbalanced between the subjects.

In a session, the subject entered an electrically shielded room with all the electrodes set, and sat relaxed with eyes closed, while physiological data was collected for 3 min. Then he rated his psychophysiological state with CLMC, which was followed by practice trials and the NASA-TLX's importance-judgment stage.

The subject performed three 10-min blocks of the dual-task in a session under one of the three tracking-difficulty conditions. After each block, he/she worked with the CLMC and the NASA-TLX's workload-rating stage. The TLX rating was required both for tracking task and the memory search task separately. The experimental session was concluded by another 3-min physiological data collection.

Since the time-on-task was controlled by repetition of dual-task blocks in this experiment, performance indices, subjective ratings, and physiological measures obtained during or immediately after Blocks 1, 2, and 3 were considered to show the effect of the shorter (10 min), the middle (20 min), and the longer (30 min) time-on-task, respectively.

Results

Performance measures

As shown in Figure 2, the proportion of time when the target was out of the boundaries (error rate) was highest under Condition D and lowest under Condition E. Two-way ANOVA (3 difficulty conditions × 3 blocks) showed a significant interaction effect ($F_{4,44} = 10.18$, $p < 0.01$). Multiple comparison using Fisher's Least Significant Difference (LSD) test proved a significantly higher error rate under Condition D than the other conditions over the three blocks. A significant effect of time-on-task was seen under Condition D, where error rate in Blocks 2 and 3 were higher than in that of Block 1.

As for the memory task performance, mean reaction time increased as time-on-task became longer in Condition D. When comparing the three conditions in Block 3, the reaction time was the longest in Condition D (Figure 3). ANOVA detected a significant interaction effect ($F_{4,44} = 2.86$, $p < 0.05$), and a multiple comparison proved a significantly longer reaction time under Condition D than under the other conditions in Block 3.

Subjective measures

Figure 4 shows WWL scores from NASA-TLX for the tracking task. As the tracking became more difficult, the WWL score for the task became significantly higher. It stayed considerably constant over the 30-min session. A two-way ANOVA (3 × 3) showed a significant main effect of difficulty ($F_{2,22} = 20.66$, $p < 0.01$), and for any pair of difficulty conditions, the more difficult was rated significantly higher than the less difficult (LSD multiple comparison). No effect of time-on-task was seen in the WWL scores for the tracking task.

WWL scores for the memory task were not affected either by difficulty of the primary task simultaneously performed or by time-on-task. Subjective ratings of overall workload showed similar results to the WWL scores.

Figure 5 shows changes in ratings on six scales of CLMC. Various effects were seen here of time-on-task in combination with task difficulty.

The averaged rating of "Awareness" under Condition D went up at first as the block repeated but finally returned to the base-line level. Under Conditions M and E, on the other

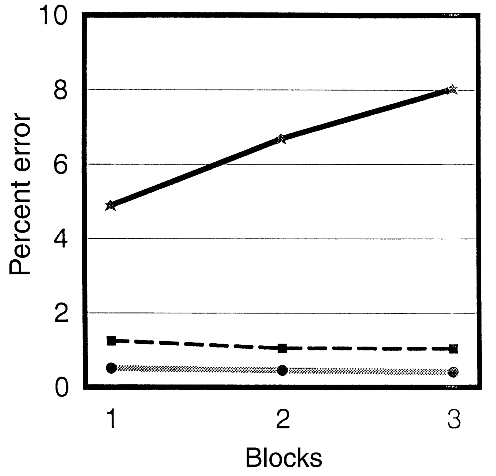


Figure 2. Mean error rate (percentage of time when the target was out of the boundaries) of the tracking task. Cond. = Condition.

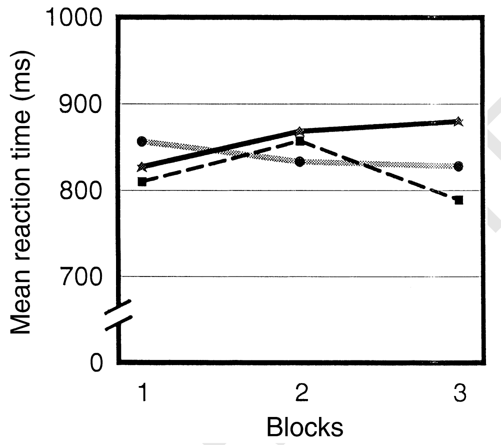


Figure 3. Mean reaction time (ms) for the memory search task. Cond. = Condition.

hand, the ratings tended to go down as time-on-task was prolonged. Main effect of difficulty was significant in ANOVA ($F_{2,22} = 10.15$, $p < 0.01$) and ratings under Conditions E and M were significantly lower than those under Condition D (LSD multiple comparison).

As for the "Mood" and the "Comfort Level" scales, ratings dropped after Block 1 under all the conditions, but showed no significant effect of task difficulty nor of time-on-task.

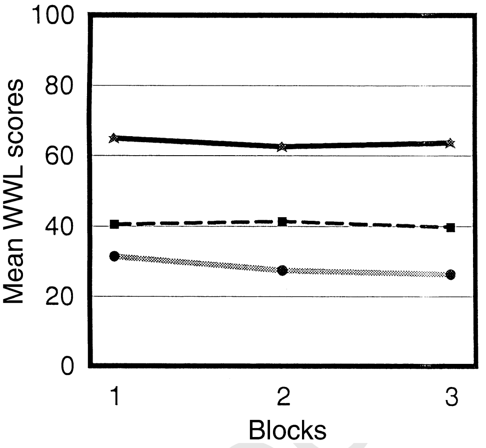
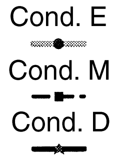


Figure 4. Mean Weighted Workload Scores (WWL) of NASA task load index TLX for the tracking task. Cond. = condition.



Ratings of "Irritability" under Condition D were significantly higher than those under Conditions M and E (LSD Multiple Comparison). Two-way ANOVA detected a significant main effect of task difficulty ($F_{2,22} = 4.21$, $p < 0.05$).

As for the "Relaxation" scale, ratings returned to the "relaxed" direction as the block repeated after they had moved toward "tense" under Conditions E and M, while subjects under Condition D did not regain relaxation. ANOVA showed a significant main effect of task difficulty ($F_{2,22} = 4.18$, $p < 0.05$), and a multiple comparison proved significant difference in mean rating scores between Condition D and Conditions E and M.

The "Fatigue" score increased as a function of blocks under any of the difficulty conditions. The ANOVA showed a significant main effect of time-on-task ($F_{2,22} = 10.95$, $p < 0.01$), and multiple comparison proved that scores (change from the baseline) obtained after

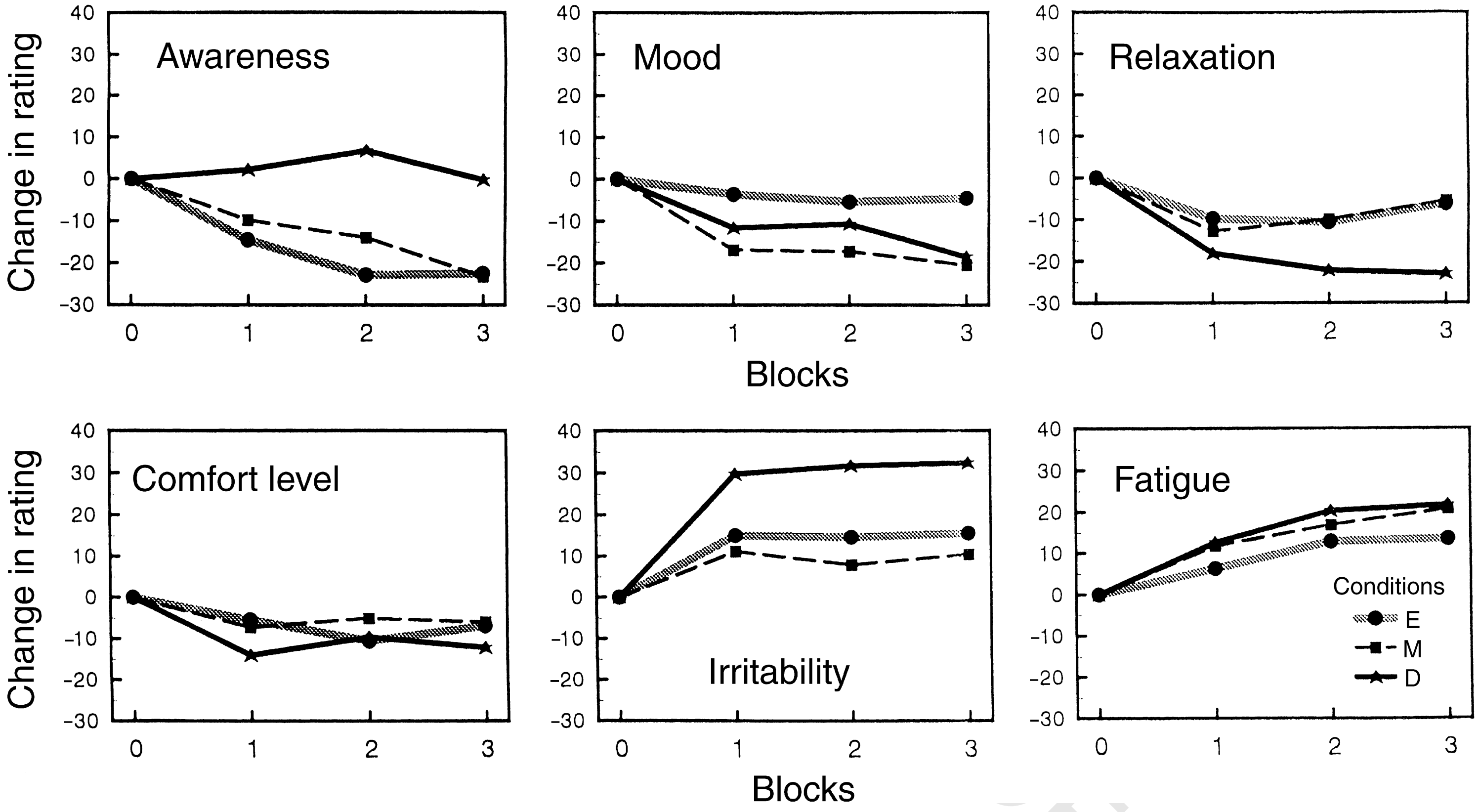


Figure 5. Changes in ratings on the Check List of Mental Condition. Scores (0–100) rated after each block was compared with one rated before the first block for each scale and the difference was averaged over the 12 subjects. E = Location, M = speed, D = acceleration.

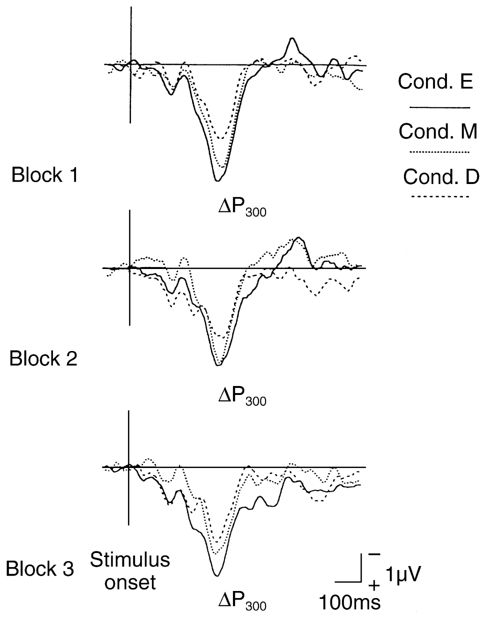


Figure 6. The grand-averaged waveforms of event related potentials for the presentation of the memory task stimulus, recorded at Pzs of the 12 subjects. Cond. = condition, E = location, M = speed, D = acceleration.

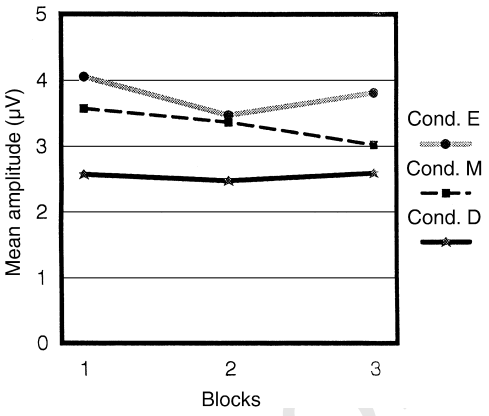


Figure 7. P300 amplitude values of event related potentials for the memory task stimuli, averaged over the 12 subjects. Cond. = condition, E = location, M = speed, D = acceleration.

Blocks 2 and 3 were significantly higher than those after Block 1. Here no effect of task difficulty was observed.

Physiological measures

The grand-averaged waveforms of ERP to the presentation of subsidiary task stimulus over 12 subjects are shown in Figure 6. The amplitude of P300 tends to decrease corresponding to increasing of difficulty on the primary task. This can be seen more clearly in Figure 7, where P300 amplitude values of ERP are graphed out. The results of ANOVA showed that the main effect of difficulty was significant ($F_{2,22} = 4,71$, $p < 0.05$), and the amplitude significantly decreased in condition D in comparison with that in condition E by multiple comparison test. The effect of time-on-task was not significant.

ECG data were analyzed for 10 subjects because the respiration rate of the remaining two was not stable during the task. Mean heart rate of the 10 reduced as time-on-task was prolonged. Two-way ANOVA detected a significant main effect of time-on-task ($F_{2,18} = 5.29$, $p < 0.05$) and a multiple comparison proved that mean heart rate in Block 3 was significantly lower than preceding blocks. CV-RR, on the other hand, did not change systematically.

Discussion

As predicted, WWL scores of NASA-TLX and ERP evoked by secondary task stimuli were sensitive only to the factor of task difficulty and were not affected by time-on-task. In contrast, mean heart rate dropped and "Fatigue" rating on the CLMC scales rose as a function of time-on-task.

In Blocks 1 and 2, secondary (memory) task performance was not negatively related to primary (tracking) task difficulty, but in Block 3, at last, the effect of primary task load appeared on the secondary task performance; mean reaction time under Condition D became significantly longer than under the less demanding conditions. As seen in Figures 2 and 3, both tracking performance and memory search deteriorated as a function of blocks under Condition D, suggesting the possibility of decrement of available resource due to prolonged hard work. This speculation is in line with early dual-task experiments by Brown and

Poulton (1961), who assessed the decrement of "spare capacity" due to fatigue by means of decline of secondary task performance after some duration of automobile driving. It is noteworthy that the subsidiary task performance is not only sensitive to primary task difficulty, but can also be affected by the factor of time-on-task.

Among the CLMC scales, "Awareness", "Irritability", and "Relaxation" were mainly affected by the factor of task difficulty, while "Fatigue" was only related to the factor of time-on-task. From the former three scales, however, it was observed that a subject's psychophysiological state changed as the block of dual-task repeated, and that the direction and the degree of this change varied with the level of task difficulty. "Awareness", for example, rated after Block 3 under Condition D was almost the same as at the beginning of the session, while it dropped over time-on-task under the less difficult conditions. This is because a certain level of task difficulty is preferable for maintaining the arousal level, at least for tasks to be performed for approximately 30 min.

As predicted, the sense of fatigue clearly increased as a function of time-on-task. It can be observed in the graph (Figure 5) that the sense of fatigue grows more rapidly under Conditions D and M than under Condition E but this difference did not reach statistical significance.

As for physical workload, accumulation rate of workload effects (fatigue) should be positively related to intensity of work. Whether the analogy to mental workload of this relationship is true or not would depend upon the duration of work and the amount of resource required for the task. When a task requires a high level of concentration, as the one in this experiment did, mental fatigue will build up more rapidly under more demanding conditions. On the other hand, when a task requires less concentration, the pace of fatigue accumulation might be faster under less demanding conditions.

In summary, some workload measures were sensitive only to task difficulty, while some to time-on-task, and others to both. It was also demonstrated that the relationship between the task demand and its effects on the worker after some duration of performance is fairly complex. Therefore, further study on the time factor in workload is necessary before bringing mental workload techniques into wider practice. First of all, effects of longer time-on-task should be examined because duration of work in the real world is much longer than 30 min The authors have already conducted laboratory experiments in which subjects' time-on-task was as long as 60 min and 90 min, and have found similar results to those reported above. In addition, measurement of workload and its effects in a real-life task of over 2 h is planned. Furthermore, the authors are trying to develop new subjective workload scales for train drivers on the basis of the above concept and findings.

References

Brown, I. D., & Poulton, E. C. (1961). Measuring the spare "mental capacity" of car drivers by a subsidiary task. *Ergonomics*, **4**, 35–40.

Eggemeier, F. T. (1988). *Properties of workload assessment techniques*. In P. A. Hancock & N. Meshkati (Eds), *Human mental workload* (pp. 41–62). Amsterdam: North Holland.

Haga, S. (1993). Measurement of mental workload and attentional resource. *Japanese Journal of Ergonomics*, **29**, 349–352. (In Japanese).

Haga, S., & Mizukami, N. (1996). Japanese version of NASA Task Load Index: sensitivity of its workload score to difficulty of three different laboratory tasks. *Japanese Journal of Ergonomics*, **32**, 71–79. (In Japanese with English abstract).

Haga, S., Shinoda, H., Kokubun, M., & Fujinami, K. (1994). Mental workload measurement by evaluating effect of task on psychophysiological state. *Proceedings of the International. Ergonomics association*, **6**, 242–244.

Hart, S. G., & Staveland, L. E. (1988). *Development of NASA-TLX* (Task Load Index): results of empirical and theoretical research. In P. A. Hancock & N Meshkati (Eds), *Human mental workload* (pp. 139–183). Amsterdam: North Holland.

International Standardization Organization (1991). *Ergonomic principles related to mental work-load – general terms and definitions, ISO 10075: 1991*. Geneva: International Standardization Organization.

International Standardization Organization (1996). *Ergonomic principles related to mental work-load*

– design principles, ISO 10075-2: 1996. Geneva: International Standardization Organization.

Jex, H. R. (1988). Measuring mental workload. *Problems, progress, and promises*. In P. A. Hancock & N. Meshkati (Eds), *Human mental workload* (pp. 5–39). Amsterdam: North Holland.

Miyake, S., & Kumashiro, M. (1993). Subjective assessment of mental workload. *Japanese Journal of Ergonomics*, **29**, 399–408. (In Japanese).

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science*, **205**, 924–927.

Shinoda, H. (1991). *The psychophysiological measurement of mental workload.* Doctoratal Dissertation, Tsukuba University, Ibaraki, Japan. (In Japanese).

Warm, J. S. (Ed.) (1984). *Sustained attention in human performance*. New York: John Wiley.

Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd ed.). New York: Harper Collins.

(Received Feb. 26, 2001; accepted Mar. 4, 2002)

Author Query Form

Journal: Japanese Psychological Research

Article: 016

Dear Author,

During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers on the query sheet if there is insufficient space on the page proofs. Please write clearly and follow the conventions shown on the attached corrections sheet. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

Many thanks for your assistance.

Query Refs.	Query	Remarks
1	Is the text OK: **E = Location, M = speed, D = acceleration.**	
2	Is the text OK: **E = Location, M = speed, D = acceleration.**	
3	Is the text OK: **E = Location, M = speed, D = acceleration.**	