THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 5(1), 63–86

Influence of Task Demand Characteristics on Workload and Performance

P. A. Hancock, G. Williams, and C. M. Manning
University of Minnesota

S. Miyake
University of Occupational and Environmental Health, Japan

Two experiments are reported that examined the influence of variation in task demand on performance and workload. The first experiment considered how the manipulation of prior level of task demand affected subsequent workload and performance. The second experiment examined the effects on performance and workload of increments in the level of task demand. Results from the first study indicated that prior level of imposed task difficulty did affect response in a manner consistent with a scaling of workload in relation to previous task conditions. The second study demonstrated the primacy of absolute demand level over increments in that demand as influencing operator response. Overall, our results indicate that workload and performance are sensitive to multiple characteristics of the task and not instantaneous demand level alone. These findings are important in explaining why association and dissociation occur between task demand, operator efficiency, and perceived workload in differing performance contexts. The importance of these findings for the aviation psychologist in assessing pilot and operator workload is articulated.

Substantive changes in contemporary aviation have occurred as a result of the use of computer-based technological innovation. On the human side, a key question for the aviation psychologist remains the assessment and prediction of operator or pilot workload associated with these technically ad-

Requests for reprints should be sent to P. A. Hancock, Human Factors Research Laboratory, 141 Mariucci Arena, 1901 Fourth Street, SE, University of Minnesota, Minneapolis, MN 55455.

vanced systems (Kantowitz & Casper, 1988). The early focus of workload studies was on high-task demand situations. Such tasks characterize much that is done in single-seat, high-performance cockpits and in selected phases of commercial flight, most typically takeoff and landing. The central concern of these investigations was whether the imposed demand exceeded pilot or crew capability and what happened to response efficiency under such circumstances. Mental-workload evaluation also represented the explicit recognition of the energetics aspects of human capability (Freeman, 1948; Kahneman, 1973; Minsky, 1984). Such approaches are in direct contrast to linear information-processing models as avenues for explaining performance variation (cf. Lachman, Lachman, & Butterfield, 1979). In particular, workload and associated energetics concepts represent the acceptance that human operators cannot be regarded as simple linear transducers whose performance efficiency simply fluctuates directly with imposed task demand. It has become the accepted position that mental workload is a multidimensional and mediational construct that can provide important insight into response capability (see Gopher & Donchin, 1986; Hancock & Meshkati, 1988; Moray, 1979; O'Donnell & Eggemeier, 1986). The conceptual and practical use of workload measures became a hotly debated topic among adherents and critics. However, despite more than a decade of intense research, much remains uncertain about workload (see Wickens, 1993). This is not necessarily an indictment of workload per se because similar statements are true of many energetics facets of behavior, including important topics like attention and stress.

The confluence of two practical factors served to stimulate workload research in aviation. The first concerned the design innovations in the cockpit of modern aircraft. The second is the question of performance–workload dissociation. It is perhaps the single most repeated assertion in the study of human factors that automation has changed the operator's role from momentary controller to system manager (cf. Jordan, 1963). What became clear for workload research was that little was known about the costs associated with the apparently low demand or underload of enforced monitoring of these automated and semiautomated systems. The work of Warm and his colleagues demonstrated two critical factors. The first was that enforced monitoring is a stressful experience with a high level of associated workload (see Hancock & Warm, 1989; Warm, Dember, Gluckman, & Hancock, 1991).[1] The second factor was that in sustained attention or monitoring tasks, work-

[1]The recent work of Scerbo and his colleagues (Sawin & Scerbo, 1993) has shown that the context of performance and attitude toward performance are critical factors in the level of experienced workload. Typically, task demands in aviation are critical and the penalty for missed signals can be great, hence the typical sustained attention task in aviation bears the hallmarks of high-stress, high-workload potential (Hancock & Warm, 1989). The attitudinal observation suggests that design manipulations may be enacted that result in enjoyable interaction with concomitant effects on perceived load and potentially performance efficiency.

load was directly influenced by the psychophysical characteristics of task demand (see Becker, Warm, Dember, & Hancock, 1991, this issue; Warm, Dember, Gluckman, & Hancock, 1991). This assertion of a direct association stands in particular contrast to the findings of Wickens and his colleagues (e.g., Derrick, 1988; Yeh & Wickens, 1988) that workload dissociates from task demand; that is, under certain conditions, performance could improve as workload increased and vice versa.

Dissociation is a particularly disturbing phenomenon for workload researchers because it implies that the isomorphism, or linked mapping, between performance level and workload fails in just those situations in which the aviation psychologist relies most heavily on such measures to effect some action or recommend some design modification. As these doubts have percolated through the literature and influenced practitioners to a more cautious interpretation, workload seems to have been overtaken by a new conception: situation awareness. Much debate has followed on the definition of a construct that, like workload, also has its advocates and detractors, in many cases the same researchers involved in the workload debate. In some definitions, workload is included in situation awareness, as are the other aforementioned energetics constructs such as attention, stress, and fatigue. Our position (see Smith & Hancock, in press) is that situation awareness, like other revivals of energetics constructs including workload, is part of the process of the rehabilitation of consciousness in psychology (see Ornstein, 1977), which at one time explicitly sought to excise mental characteristics as explanatory phenomenon (Watson, 1913). Whatever its guise, we still do not have sufficient knowledge about the overall energetic state of the individual to make confident assertions about reactions under differing demand conditions. One primary reason for this is that we have not paid sufficient attention to the differing characteristics of the situated task demand or the *context* in which activity occurs. Typically, we have manipulated only the absolute "level" of task demand and compared performance as that level has been changed, usually between trials, blocks, or conditions. Hence, our need to understand more thoroughly the properties of the workload response with respect to different characteristics of the task and situation at hand.

PROPERTIES OF WORKLOAD RESPONSE

Despite the assertions about its multidimensional nature, one simple way to think of workload is akin to an analog signal that follows on the fluctuations of task demand, at least in some fashion. From this view, it may be that the direct link between workload and performance, an associated relationship, is maintained in certain conditions in which the lag between demand and response is necessarily small. Dissociation, on the other hand, may result from an increase in the lag between change in demand and subsequent response. We cite this one example here, but lag is only one of many

potential characteristics of the demand performance–workload relationship. We have illustrated some further characteristics in Figure 1. Within the region of acceptable performance, we have traditionally been concerned with the absolute value of workload as shown by (1) in Figure 1. We have attempted to establish the boundaries of acceptable workload, the so-called workload redlines, and to understand what task-related factors drive workload into such unacceptable regions (denoted by the hashed regions of Figure 1). Although momentary workload value is clearly one diagnostic of performer state, several other facets of task demand–workload relation are of use. Some of these linkages have already been shown to exert effects, such as the history of demand; see (5) (Matthews, 1986; Miyake, Hancock, & Manning, 1992). These findings suggest that other facets such as future expectation (6) (see Harris, Hancock, & Arthur, 1993) and level and location of recovery (4) may also prove of value in predicting the overall level of operator workload in specific conditions. In our experiments, we examined two facets of the demand–performance–workload relation. The first is concerned directly with the influence of the historical or past level of task demand on subsequent performance and the perception of workload. The second is the level of demand combined with increments in that demand (2). Clearly, the latter is also a manipulation of the trend illustrated as (3) being

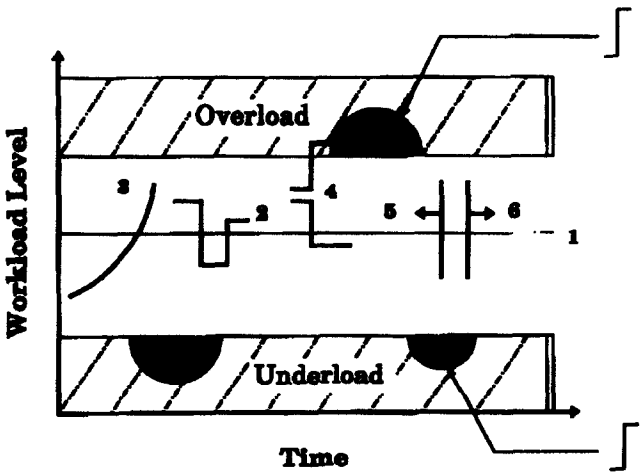

FIGURE 1 The illustration shows facets of the workload response with respect to comparable changes in task demand. Typically, instantaneous level of workload, shown as (1), is of central concern. However, workload may be sensitive to increments in task demand (2) or rate of change of that demand (3). It might be influenced by history (5) or future expectation (6). Also, where the stable demand is established with respect to an individual's capability (4) is of potential concern. Fracturing workload redlines (light hashed areas) for either overload or underload is of critical importance. The effects of prolonged residence in those regions (dark hashed areas) is what gives rise to much concern for the aviation psychologist.

rate of change of task demand. Operators are frequently more sensitive to change and rate of change, rather than the absolute level of a variable; therefore, we have a rationale for the expectation that increment in task demand is an important variable in influencing workload response. Our experiment is also part of a general programmatic investigation that we have pursued on workload transition events (see also Hancock, 1989; Hancock et al., 1989; Miyake, Hancock, & Manning, 1992; Scallen, Duley, & Hancock, 1994). Recognition of the importance of workload transitions is clearly growing (see Howell, 1992; Huey & Wickens, 1993; Warm, 1993). However, as yet relatively few experimental findings have addressed transitions in workload and the task-based characteristics that influence them.

EXPERIMENT 1: EFFECT OF PRIOR DEMAND ON CURRENT PERFORMANCE

Method

Experimental participants. Twelve right-handed men from the University of Minnesota faculty, staff, and student body volunteered to participate in this study. Their ages ranged from 22 to 38 years with a mean age of 31.1 years. All subjects were in professed good health at the time of testing.

Experimental task and procedure. A simulated flight task was shown on a computer monitor. An aircraft icon, which was driven by random forcing functions, appeared inside a sight circle (cf. McRuer & Jex, 1967; Smith, 1967). Subjects controlled the movements of the aircraft with a joystick and attempted to keep it aligned at the center of the sight circle. Thus the task was a two-dimensional compensatory tracking. The tracking area was 157 mm × 110 mm. There were three levels of flight task difficulty: high (H), medium (M), and low (L), which differed in respect of the amplitude and cutoff frequency of the forcing functions as determined by pilot experimentation. Each subject participated in four separate sessions at least a day apart. Each session was comprised of three 5-min tracking trials (Hammerton, 1981). The first session was practice, and all three of the tracking trials were at the medium level of difficulty (M1M2M3). In the three experimental sessions that followed, the first and last of the three tracking trials were always of medium difficulty. However, the middle tracking trial was either of high, medium, or low difficulty. This regimen is illustrated in Figure 2. Over the course of the three sessions, each subject participated in the three experimental conditions: M1H2M3, M1M2M3, and M1L2M3. The order of administration of these conditions was randomized across participants.

Three measurements of subjective response were taken: (a) Critical Flicker Fusion (CFF) values (Hosokawa, Makizuka, Nakai, & Saito, 1989;

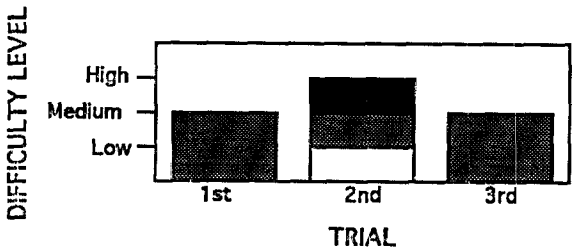

FIGURE 2 Schematic illustration of the experimental procedure and comparisons as accomplished in Experiment 1.

Saito, Hosokawa, Saito, Nakai, & Inzuka, 1988); (b) Subjective Workload Assessment Technique (SWAT) ratings (Reid & Nygren, 1988); and (c) National Aeronautics and Space Administration Task Load Index (NASA–TLX) ratings (Hart & Staveland, 1988). CFF was measured at the beginning of each session to provide a baseline value. CFF was then measured at the end of each trial. Each CFF score was derived from an average of five readings, excluding the highest and lowest recordings. Percentage of change was calculated as the difference between the baseline value and the observed experimental value. The SWAT and the NASA–TLX were installed into the tracking-task program. SWAT and TLX scores were taken after each 5-min trial. SWAT scores were processed with a SWAT program (Reid, 1989). The average weighted workload scores (TLX) were computed from NASA–TLX ratings (Hart & Staveland, 1988). Difference scores for both the SWAT and TLX were obtained by subtracting the value for the first trial from that of the third trial. These differences are also reported. Tracking data were gathered from the measurement of the deviation of the aircraft icon from the center of the sight circle. This value was calculated by a two-dimensional root mean square error (RMSE) of tracking. This value was sampled every 200 ms. Percentage changes of RMSE from the third tracking trial compared to the first tracking trial were used to evaluate performance change.

Experimental hypotheses. We hypothesized that the historical profile of task demand would subsequently influence both performance and workload on that same task. Further, we postulated that previous experience of low demand would improve performance and reduce workload on a subsequent trial at the medium level of demand compared to performance at the same medium demand before the low-demand exposure. Experience of an interpolated high-task demand was hypothesized to have the opposite effect. We expected that when participants experienced no change in demand under the M1M2M3 control condition, then performance and workload would not change between the first and third trials.

Results

Performance and subjective workload measures from only the first and the third trial (i.e., both medium-demand-level conditions) were considered as the purpose of this study was in evaluating the influence of an historical demand level. That is the influence of a previous load on subsequent response (see Figure 2). Analysis was based on within-subject comparisons. The statistical significance of the difference of means was analyzed by a one-way analysis of variance (ANOVA) and multiple comparisons; $p < .05$ was considered to represent a statistically significant difference. Mean results, with one standard error, are shown in each illustration.

Subjective workload measures. The absolute scores for SWAT and TLX are shown in Figure 3. It is clear that for both SWAT and TLX, the scores are highly correlated with task difficulty. As shown in Figure 4, SWAT and TLX scores on the last trial of medium difficulty (M3) increased after

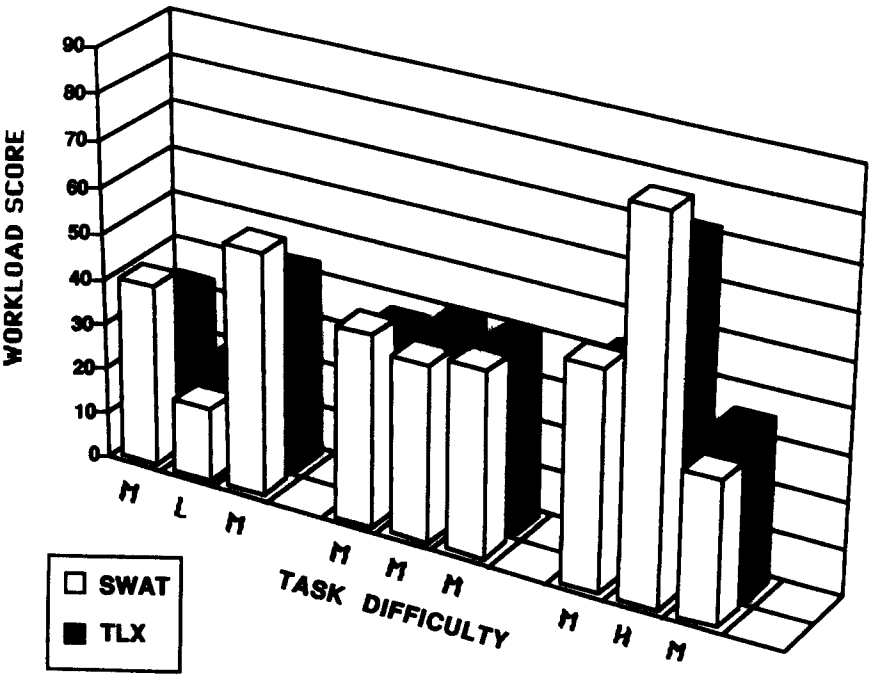

FIGURE 3 Changes in weighted SWAT workload (SWAT) and overall NASA–TLX workload (TLX). Mean responses for all participants ($n = 12$) are shown. Only difference scores (% change) between M1 and M3 trial are illustrated here.

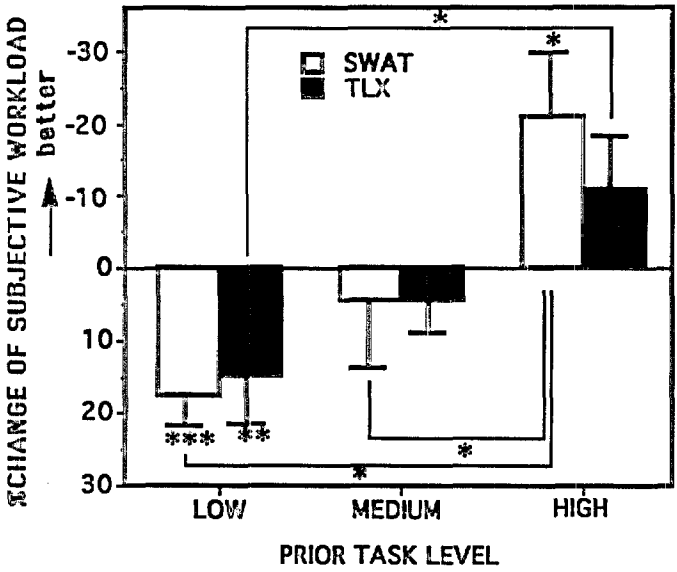

FIGURE 4 Percentage change in subjective workload scores. T-bars are 1 standard error. $*p < .05$, $**p < .01$, $***p < .001$.

exposure to the easier level of the task (L2). However, the workload scores for each subjective assessment technique decreased after exposure to the more difficult level (H2). Specifically, there were significant differences between SWAT scores from the first medium-difficulty trial (M1) to the third trial (M3). In the case of previous exposure to low-demand conditions on the second trial (L2), SWAT scores significantly increased from the baseline trial to the comparison trial. This significant effect also held for the SWAT score when the middle trial was at the high-demand level (H2), except that the subsequent score was depressed. For the TLX score, only the experience of an interpolated low-demand (L2) condition had a significant effect. Ryan's multiple comparisons revealed that percentage changes of SWAT after H2 level are significantly different from those after L2 and M2 level tasks. Furthermore, significant differences were found between TLX scores after L2 and M2 levels. These results mean that after the easier task is interpolated in the sequence, subjects experienced more workload, and after the more difficult task, they experienced less workload on a task that is objectively of the same difficulty. That is, all comparisons are made between responses on a task of common demand (M1–M3). In short, prior events color experienced workload. They also influence performance, as is reported subsequently.

Performance measures. Percentage changes in the respective performance measures are shown in Figure 5. Combined time lead (CTL) de-

creased significantly after the high-difficulty task, $p < .05$. Decrement of time lead means that subjects' responses became slower or their predictions on the task were worse. In contrast, after the L2, time lead increased but not significantly. RMSE decreased after L2 and increased after H2. However, unlike workload, these changes did not reach the predetermined level of significance. These results suggest that after the easier task, performance becomes better, and after the more difficult task, it becomes worse, but the statistically reliable effect was only for CTL after the H2. The learning effects on the performance are shown in Figure 6.

Percentage changes of RMSE and CTL of the first trial (M1) in the second and the third sessions from the first session are shown. RMSE in the second and third session is significantly better than that in the first session, $p < .05$. No significant change was found in CTL measures. It should be noted that the RMSE scale is shown so that better performance increases on the y axis. These results indicate that learning was still occurring in the task investigated.

Psychophysiological measure. The percentage change in the CFF value is shown in Figure 7. After every trial in every session, CFF value

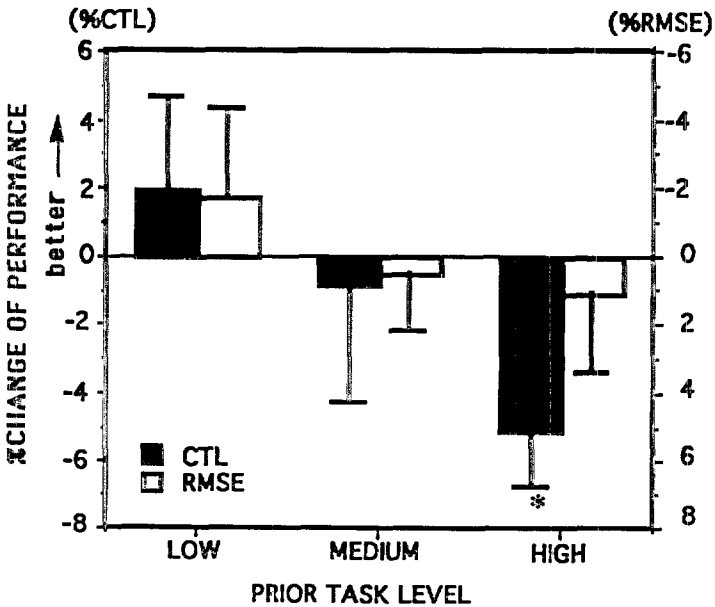

FIGURE 5 Percentage changes of combined time delay (CTL) and root mean square error (RMSE).

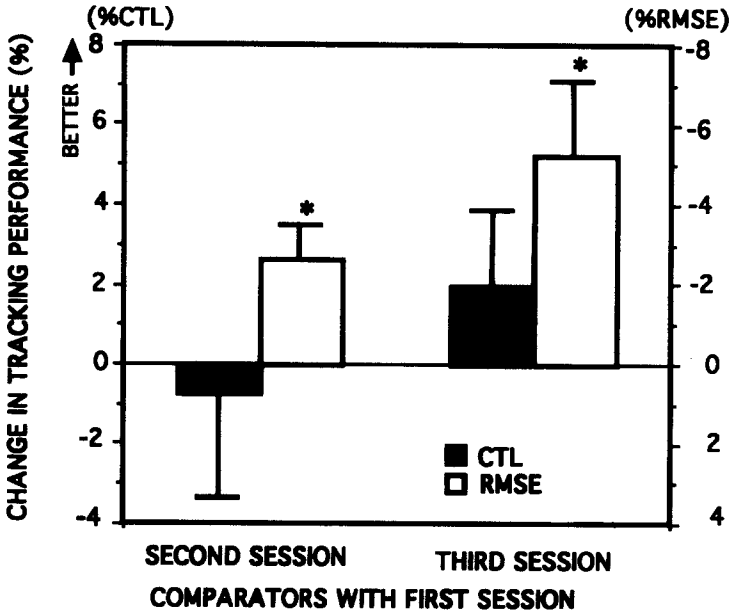

FIGURE 6 Learning effect on performance. RMSE of M1 in the second and third session are significantly better than that in the 1st session. $*p < .05$.

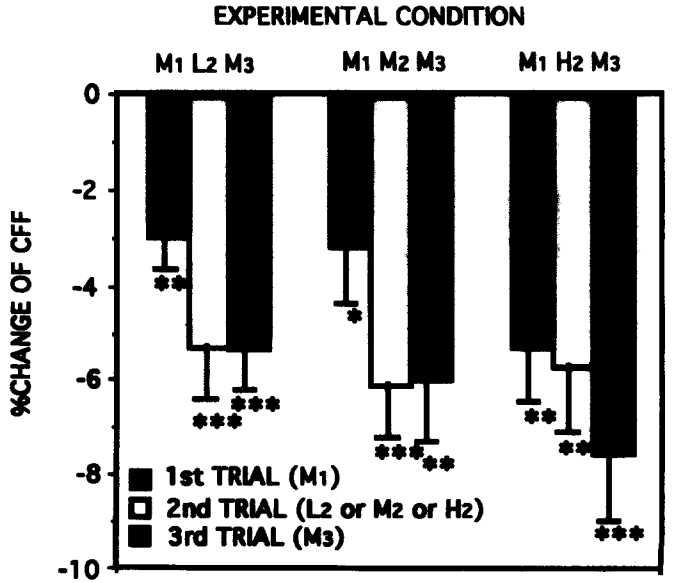

FIGURE 7 Percentage changes of Critical Flicker Fusion Frequency (CFF). All CFF values show significant decrease from the baseline (before task) value. $*p < .05$, $**p < .01$, $***p < .001$.

significantly decreased from the baseline value (measured before the task commenced). No significant relation was found between task difficulty and CFF decrement. Percentage CFF after each trial was pooled for all sessions and is shown in Figure 8. All CFF changes represent significant reductions, $p < .001$. Significant differences were found between the first trial and both the second and third trial. This result suggests that amount of decrease in CFF value depends on the time on task rather than any particular characteristic of the task itself. CFF reflected an overall level of task-related fatigue that was independent of the specific historical task profile. We now turn to the influence of task-load increment effects.

EXPERIMENT 2: EFFECT OF INCREMENTED DEMAND ON PERFORMANCE

Method

Experimental participants. The participants in this experiment were 15 students from the University of Minnesota. There were 9 men and 6 women. The mean age of the sample was 24 years with a standard deviation of 5. Subjects were volunteers, and all were in professed good health at the time of testing.

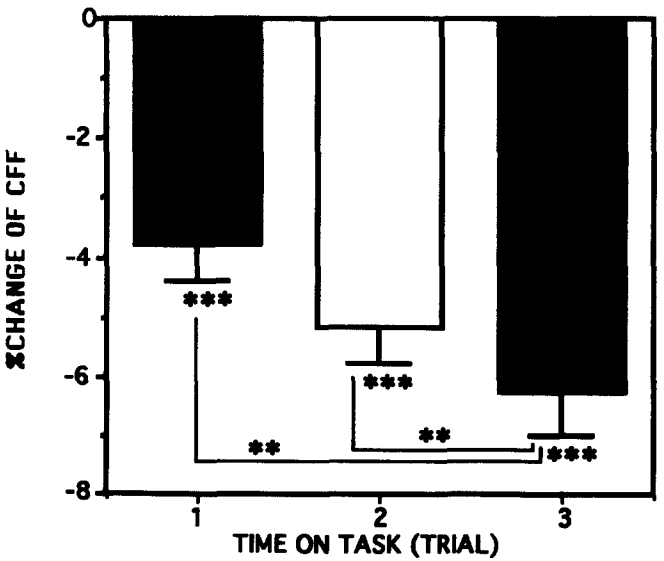

FIGURE 8 Percentage changes of CFF versus time on task. ***$p < .001$, **$p < .001$.

Experimental procedure. The experimental platform that permitted evaluation of our hypotheses was MINUTES (MINnesota Universal Task Evaluation System; see Hancock et al., 1992). This environment consists of three major subtasks: monitoring, resource management, and tracking, each of which can be controlled through script commands. These subtasks are illustrated on the MINUTES display shown in Figure 9. Subjective assessment of workload was collected by having subjects complete SWAT (Subjective Work Assessment Technique) tests, which appeared in a window in the MINUTES display. Subjects were trained on the individual subtasks that would be completed during the experimental sessions. The tasks were completed using joystick and keyboard controls. The monitoring task required keyboard responses to indicator lights and gauges. Resource management required monitoring and control of fuel tanks and pumps to maintain a constant target level of fuel in the outer two of the five tanks. Tracking required joystick manipulation to maintain a crosshair at the center of a display. Task-load baseline and increment levels were determined by varying the frequency of the indicator light and gauge-state changes, frequency and duration of pump failures, and changes in the gain of tracking and sensitivity of the joystick, as detailed subsequently. The SWAT tests provided subjective assessment of time load, stress, and mental effort as described earlier (Reid & Nygren, 1988).

The experimental protocol employed a within-subject design. All subjects completed three sessions that lasted 12 min each. A 2-min break was permitted between each session. Each session presented three routines, each consisting of two segments: a baseline level of task load (100 sec) followed by the same baseline plus an incremental load (100 sec). After each segment, subjects were presented with three SWAT scales (20 sec). The baseline level and incremental levels are explained subsequently. The presentation of the routines was controlled by each subject receiving all the conditions in a pseudorandom order (i.e., each subject received a unique order of presentation).

Baseline Level of Task Demand Conditions

Three baseline rates were used for each of the components of the MINUTES task. The event rates are given in Table 1. In each box, the first entry refers to the number of monitoring events, the second to the number of resource-management events, and the third to the level of tracking difficulty. Each event refers to an entry in the script that creates a change in state to either the monitoring task or the resource-management task. Eight tracking levels were used for controlling the demand of the tracking task.

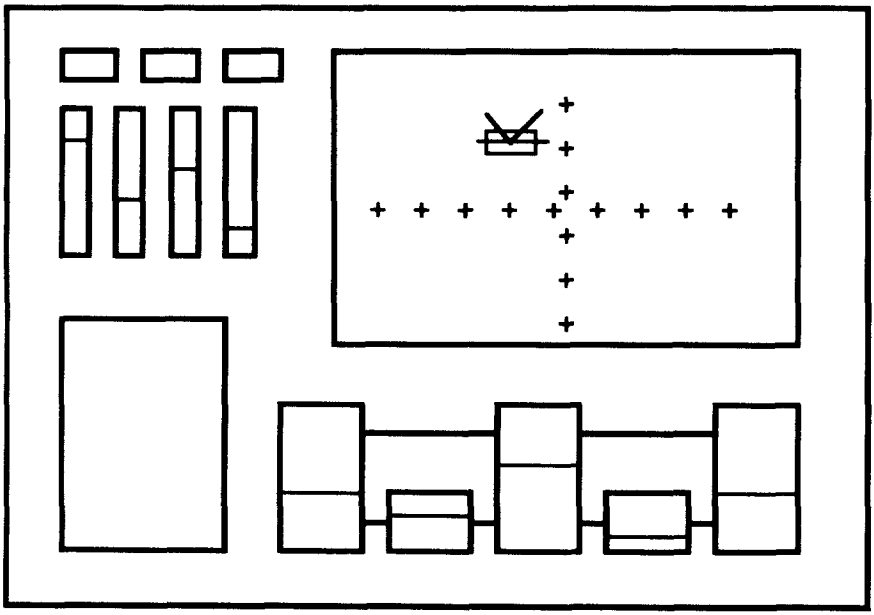

FIGURE 9 Schematic illustration of MINUTES environment showing the three subtask differentiations. The tracking task is at the upper right and the monitoring task is shown upper left. The resource-management task is shown lower right; the message window, lower left.

TABLE 1
Task Combinations

		Baseline Plus Increment		
	Baseline	*Low*	*Medium*	*High*
Low	2 - 1 - 1	6 - 3 - 3	9 - 4 - 5	12 - 5 - 6
Medium	5 - 2 - 2	9 - 4 - 5	12 - 5 - 6	15 - 6 - 7
High	8 - 3 - 4	12 - 5 - 6	15 - 6 - 7	18 - 7 - 8

Incremental Demand Conditions

Three incremental levels were combined with the baseline task levels to form nine experimental conditions, which are also shown in Table 1. The rationale behind the size of the increment was to provide task levels that could compare relative increase in load with absolute task level (e.g., low-baseline level plus medium increment being equivalent to medium-baseline level plus low increment). Equivalent task loads are shown on the diagonals of Table 1.

Results

Data were analyzed for each component task individually (i.e., monitoring, resource management, and tracking). Each set was analyzed using a three-baseline level (low, medium, and high) by three increment level (low, medium, and high) by two before/after increment (i.e., a baseline level) and after (i.e., a baseline plus an increment) ANOVA with repeated measures.

Tracking. ANOVAs for each baseline plus increment condition indicated no significant effects for incremental demand on tracking performance. This observation did not jibe with the known relationship between difficulty and performance or with observation of subjects who had clear problems adjusting to demand after the step change. It was hypothesized that tracking performance effects were transitory and therefore did not appear in the data for the whole period. As a result, an additional analysis was performed on the tracking response for the last 20 sec of the baseline condition versus the first 20 sec of the incremented condition. In this analysis, the interaction between baseline level and before/after condition was significant, $F(2, 16) = 7.395$, $p < .01$. This effect, illustrated in Figure 10, implies that before/after effects were found for the low- and high-baseline conditions but not for the medium-baseline condition. Such effects were confirmed with post hoc t tests. Post hoc t tests were carried out on the before/after data for the baseline level conditions. These tests proved significant for the high-baseline condition, $t(8) = 2.714$, $p < .05$, and for the low-baseline condition, $t(8) = 2.898$, $p < .05$ only.

In considering this interaction, it is important to point out that subjects' performance was best in the easiest condition of the low baseline. However, they showed their next best baseline performance in the high condition, which counters the assertion that performance is always directly linked to demand.

Frequently, subjects adjusted their effort according to demand, and this may be the case in producing the better tracking performance at the high compared to the medium baseline. The interaction then results from this nonlinear baseline influence because the after conditions do follow the pattern linking performance efficiency to level of demand. The significant interaction between before/after conditions and increment level, $F(2, 16) = 5.248$, $p < .05$, illustrated in Figure 11, also reflects this nonproportional baseline effect. In short, these interactions, although interesting, are not ones on which too much theoretical significance should be placed.

The main effect for this subsequent analysis confirmed that the increment does indeed have a significant effect, $F(1, 8) = 13.047$, $p < .01$. The mean change was some 30% decrement in performance in the time span chosen. This leads to two conclusions. First, there is a time order effect in operation here in which performance disturbance is large at the transition event and is

progressively reduced in impact as time progresses at the new demand level. Second, the absence of a main effect over the 100-sec epoch implies that subjects recover and then compensate to the new and higher level of demand. This is a further example of the adaptive strategy in which participants expend effort according to perceived and actual demands. Two immediate conclusions can be drawn from these tracking data. First, increments in demand disturb performance. Second, adaptive responses to these changes cope with such increments progressively. We suggest that the length of the time course of recovery of performance in response to increment in demand is proportional to the degree of the increment itself, a proposition that requires further investigation.

Monitoring. Three sets of data were collected from the monitoring task: response time, response omissions, and false alarms. For response time, there was a significant main effect in the before versus after conditions, $F(1, 14) = 6.510$, $p < .05$. The mean response time for the before condition (i.e. for baseline) was 1.30 sec and for after condition (i.e. baseline plus an increment), the mean time was 1.42 sec. The same significant main effect

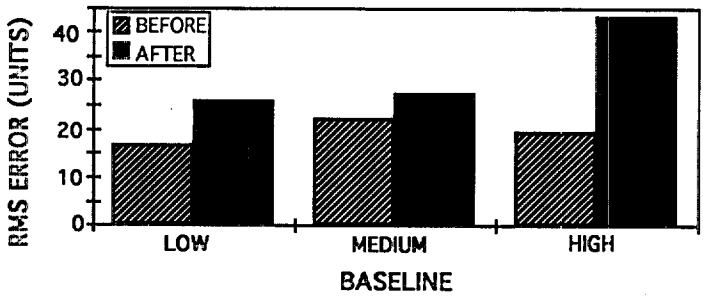

FIGURE 10 Significant interaction for the baseline level and before/after condition on root mean square error of tracking (RMSE).

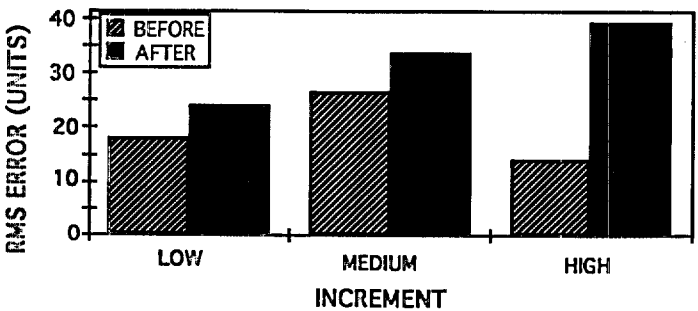

FIGURE 11 Significant interaction between increment level before/after conditioned on root mean square error of tracking (RMSE).

was evident in the false alarm data, $F(1, 14) = 8.199$, $p < .05$. The mean number of false alarms for the baseline conditions was 0.53 and for the baseline plus incremental condition was 1.03. This suggests that an increase in the task load produces an increase in both the time to react correctly to a monitoring cue and the number of false responses, both of which reflect deterioration of capability.

The data that produced perhaps the most interesting result were for response omissions. The misses were converted to proportion scores representing the total number of misses with respect to the total number of possible correct responses. Two main effects, illustrated in Figure 12, proved significant. The before and after conditions $F(1, 14) = 8.617$, $p < .05$, with means of 0.31 and 0.37, respectively, and the level of baseline task load conditions, $F(2, 28) = 32.801$, $p < .01$, with means of 0.21, 0.40, 0.41 for low, medium, and high baseline, respectively. Two interactions also proved significant. The first was for the level of baseline and before/after increment, $F(2, 28) = 4.106$, $p < .05$, and the second was for the increment level and before/after condition, $F(2, 28) = 4.036$, $p < .05$. Post hoc t tests on the first interaction (i.e., the differences between the baseline values and baseline values plus increment for each baseline level) produced significant results for the differences between the low-baseline condition and both the medium baseline condition, $t(14) = 1.995$, $p < .05$, and the high-baseline condition, $t(14) = 2.926$, $p < .05$.

Resource management. No significant results were found in the resource-management data under any of the conditions.

Subjective workload. The three components of SWAT procedure were analyzed: Time Load, Stress, and Mental effort. The SWAT data produced two significant main effects for the before/after condition and for the baseline level conditions. The means for each of SWAT response showed a trend of increasing with respect to workload for each of the scales. The results of each ANOVA are presented in Table 2. Post hoc t tests for each of the SWAT tests were applied to distinguish differences between baseline conditions, and those results are shown in Table 3.

Discussion

Influence of historical profile of task demand. The main findings of the first experiment are important ones. Concerning perceived workload, the results demonstrated that how an individual perceives the present level of workload is influenced directly by what has gone before (see also Matthews, 1986). If, for example, pilots transition from a period of low demand to

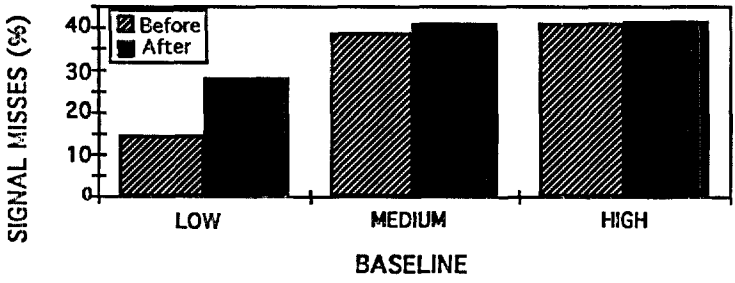

FIGURE 12 Significant interaction between baseline conditions and the before versus after manipulation for the number of signal-response omissions expressed as a percentage value of possible responses.

TABLE 2
Significant Main Effects for Subjective Workload Responses (SWAT)

	Time Load			*Mental Effort*			*Stress*		
Main Effect	*F*	*df*	*p*	*F*	*df*	*p*	*F*	*df*	*p*
Before-after	8.576	1, 13	<.05	47.326	1, 13	<.001	19.519	1, 13	<.001
Baseline level	11.308	2, 26	<.001	11.308	2, 26	<.001	5.260	2, 26	<.05

TABLE 3
Results of Post Hoc *t* Tests for Baseline Effect on Subjective Workload (SWAT) Scales

	Time Load			*Mental Effort*			*Stress*		
Level	*F*	*df*	*p*	*F*	*df*	*p*	*F*	*df*	*p*
Low versus medium	3.606	13	<.005	1.710	13	ns	1.422	13	ns
Low versus high	4.315	13	<.001	4.678	13	<.001	2.797	13	<.05
Medium versus high	1.935	13	ns	2.156	13	<.05	2.120	13	ns

Note. ns = nonsignificant.

medium demand, they would scale their current perceived load as higher than if they had stayed at the medium-demand level all along. The opposite is true for high levels of demand in which a transition from high to medium demand suppresses perceived load. These workload results are evidence of lag in the system or, with respect to human operators, their memory of previous events colors the perception of present events. Given the importance of workload transitions in aviation (Huey & Wickens, 1993), these results are central to the task of the aviation psychologist who seeks to predict acceptable levels of momentary mental workload or to use workload response in aiding the process of design. *The task history or mission profile has to be factored into any approach to assessing current workload level.*

However, the results for perceived workload cannot be considered separately from actual performance. In our case, with one exception, differences in tracking performance largely failed to reach traditional levels of statistical significance. In respect to performance, however, the overall pattern is potentially instructive. It follows that for perceived workload. That is, previous experience of a high level of demand decreased performance efficiency on a subsequent medium-demand condition. We have to be careful here, because the pattern for performance, which looks seductively like the inverted image for workload, cannot be confirmed as stable, especially in the condition in which prior task demand level was low. What can be confirmed, however, is that when prior level of demand was high, individuals subsequently performing a medium level of difficulty rated their workload as relatively low, but their performance was poorer. This is a case of dissociation (Yeh & Wickens, 1988). Typically, we would consider poor performance to be accompanied by high workload and low workload to be reported when efficiency was high. However, this is not the case here. There are a number of reasons that this might be so.

The reasons include both methodological and theoretical influences. With respect to method, note that tracking performance is measured every fifth of a second throughout the 5-min trial. Therefore, the outcome for any one trial is a summation of momentary conditions. All of the subjective measures, including both workload scales, are point measures taken after performance has been completed. Therefore, the very nature of the measurement procedures themselves would appear to encourage perceptual assessment for performance and more memory-based reference for workload. This statement allows that other methods could be used to distinguish whether measurement procedures themselves are solely responsible for the scaling and dissociation effects noted. However, because performance comparisons reported here are actually across repeated trials, some form of performance memory (or lack of it) must be involved in the pattern of findings reported. With respect to theory, it appears that perception of workload in general involves reference to memory of previous conditions. This proposition can be further evaluated by examining workload associated with other performance tasks that specifically employ memory as a key component. At present, we can propose that one reason for the dissociation between workload and performance is the factor of lag. Whereas performance tracks demand with little lag, perceived workload tracks demand with greater lag, so the previous profile of demand has much more influence on the perception of current workload than on current performance. The generality of this statement clearly depends on wider empirical evaluation using differing tasks and alternative strategies for workload assessment.

There were significant changes in CFF; however, these were not correlated with task difficulty but were related to overall time on task. Thus CFF appeared not responsive to the task-load manipulation as represented in our experiment. CFF can be considered to indicate the cortical activity level

(Kogi & Saito, 1971); however, the eyestrain alone of performing a visual task using a VDT can affect the CFF value (Iwasaki & Akiya, 1991). In our study, significant decreases in CFF value after each task trial compared to the baseline were found. These results may be due to subjects' eyestrain because (a) the experimental room was dark and luminance contrast of the VDT and the background was high, (b) luminance of the LED (light-emitting diodes) indicator of the CFF device was also relatively high, (c) some of subjects reported that they felt eyestrain because of the previously mentioned high-contrast screen, and (d) the tracking task was only three trials of 5-min each and it probably did not induce significant fatigue effects. Consequently, we favor a mechanistic rather than energetics explanation for the CFF results at present. From our evidence, we do not advocate the use of CFF to distinguish workload and performance changes.

Influence of increments in demand level. With respect to the second experiment, the overall tenor of our results indicates that the primary driver of performance is the absolute level of task demand over the increment in that demand. However, we must temper this observation because of the number of significant interactions observed. For example, in the tracking data, there was a significant modification of the before versus after increment effect because of the baseline level. It is critical to note, however, that the original baseline levels (i.e., the before conditions) do not exhibit a simple increase in RMSE with baseline demand. The interaction effect consequently seems to represent a threshold characteristic where it is the combination of an increment over a high baseline that triggers a nonproportional increase in RMS error. This is further clarified by the before/after increment interaction. Here we see a differential increment effect that initially might lead us to support a case for the influence of such a manipulation. However, examination of the preincrement baselines indicates that under the high-increment condition, the baseline was depressed such that the interaction appears. This suppression of baseline militates against a strong support for an increment effect in tracking.

Further support for the task-demand-level primacy is seen in the monitoring data. The only significant effects in response time and false alarms reflect this demand characteristic. The pattern for signal omission is somewhat more complex. Although the before/after pattern is maintained, an interaction occurs because of the effects in the low-baseline condition. The difference between the before and after comparison is exacerbated in the low-baseline condition because of the low frequency of misses in the before condition. Again, as with tracking, we favor an explanation that revolves around a suppression of baseline effect rather than emphasizing the increment effect, because the latter influence did not percolate through all baseline levels. Also, the tracking suppression occurred at high-baseline levels compared with the suppression in monitoring signal omissions at the low-

baseline level. This inconsistency argues against strong support for incremental influences. Our conclusion is further buttressed by analysis of the workload data. Each of the SWAT subscales ubiquitously showed the before versus after differences and main effects for baseline load were evident in all scales. Also whereas all pairwise comparisons of workload response under baseline manipulations did not reach significance, the low- versus high-baseline conditions were always reliably distinguished. Overall, our results confirm the primacy of absolute task load over incremental effects.

SUMMARY AND CONCLUSIONS

Our introduction posed the question of which characteristics of task demand the workload might be most sensitive to. In the first experiment, we demonstrated that prior load profile provided a strong moderating influence on workload and that this influence may be one factor in the dissociation story. In the second experiment, we proposed to drive workload via manipulation of task demand level and increment of that level. In these data, demand level seemed to be more important than increment. Of course, some caution is necessary. First, it is possible that present levels of difficulty and increments on that difficulty were not sufficiently differentiated to elicit effects. In essence, the sensitivity of the measure argument will always be with us (Poulton, 1965). Additional work is already, therefore, in progress in a multitask-simulation facility in which the levels of subtask difficulties have been magnified.

Whatever the energetic facet of performance under consideration, the question for the aviation psychologist remains how to assess and predict pilot and crew capability. Workload has been helpful in that it offers a window on efficiency and one that potentially offers information before rather than after the fact. Anyone who wishes to use workload measures, particularly subjective response, must be concerned that there are occasions in which an opinion is expressed that the task is becoming harder while performances are actually improving and vice versa. We suspect from our findings that such events are part and parcel of the nonlinearity of human response. That is, humans in general, and pilots in particular, use their previous experience and their future expectations to scale current events. That these effects are context-specific is most frustrating and is due, we believe, to the nature of the task being performed. For aviation, the hopeful aspect of these findings is that direct association between performance and workload appears mainly in monitoring tasks, which are becoming more predominant in contemporary cockpits. Our evidence implies that memory and prediction scale current events rather than instantaneous incremented change to such conditions. This would be a parsimonious suggestion because it accords with the general idea of being "ahead of the aircraft" and reflects theoretical formulations on biological survival (Holland, 1975/1992). The precise elucidation of these contextually based effects needs further evalua-

tion. Whether this knowledge is produced under the umbrella of workload or under the canopy of situation awareness is essentially irrelevant because both assist the aviation psychologist to serve, support, and protect those who fly.

ACKNOWLEDGMENTS

This research was supported by the Naval Air Warfare Center through N/N62269-92-C-0211 and NASA Ames Research Center through Grant NAG 2-749 to P. A. Hancock. Jeff Morrison and John Gluckman were the NAWC contact monitors, and Sandra Hart was the NASA Grant Technical Monitor. The views expressed are those of the authors and do not necessarily represent those of the named agencies. We thank Erik Arthur for programming help with the MINUTES program for the second experiment and Sue Chrysler for her assistance in all general areas. The first experiment was conducted while Dr. Miyake was a visiting researcher at the Human Factors Research Laboratory on leave from The University of Occupational and Environmental Health, Japan.

REFERENCES

Becker, A. B., Warm, J. S., Dember, W. N., & Hancock, P. A. (1991). Effects of feedback on perceived workload in vigilance performance. *Proceedings of the Human Factors Society, 35,* 1491–1494.

Bendat, J. S., & Piersol, A. G. (1971). *RANDOM DATA: Analysis and measurement procedures.* New York: Wiley.

Derrick, W. L. (1988). Dimensions of operator workload. *Human Factors, 30,* 95–110.

Freeman, G. L. (1948). *The energetics of human behavior.* Ithaca, NY: Cornell University Press.

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* (pp. 41:1–49). New York: Wiley.

Hammerton, H. (1981). Tracking. In D. Holding, (Ed.), *Human skills* (pp. 177–201). New York: Wiley.

Hancock, P. A. (1989). The effect of performance failure and task demand on the perception of mental workload. *Applied Ergonomics, 20,* 197–205.

Hancock, P. A., Harris, W. C., Chrysler, S. T., Arthur, E., Manning, C., & Williams, G. (1992). *Minnesota universal task evaluation system [MINUTES].* HFRL 93–01, Minneapolis: University of Minnesota.

Hancock, P. A., & Meshkati, N. (Eds). (1988). *Human mental workload.* Amsterdam: North-Holland.

Hancock, P. A., Robinson, M. A., Chu, A. L., Hansen, D. R., Vercruyssen, M., Grose, E., & Fisk, A. D. (1989). The effects of practice on tracking and subjective workload. *Proceedings of the Human Factors Society, 33,* 1310–1313.

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors, 31,* 519–537.

Harris, W. C., Hancock, P. A., & Arthur, E. (1993, April). The effect of taskload projection on automation use, performance, and workload. In R. S. Jensen & D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 890A–890F). Columbus: Ohio State University.

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: Elsevier.

Holland, J. H. (1992). *Adaptation in natural and artificial systems.* Cambridge, MA: MIT Press (Original work published 1975).

Hosokawa, T., Makizuka, T., Nakai, K., & Saito, K. (1989). A new pocket-type flicker apparatus. *Japanese Journal of Industrial Health, 31,* 324–329.

Howell, W. C. (1992). Engineering psychology in a changing world. *Annual Review of Psychology, 44,* 231–263.

Huey, B. H., & Wickens, C. D. (Eds.). (1993). *Workload transition: Implications for individual and team performance.* Washington, DC: National Academy Press.

Iwasaki, T., & Akiya, S. (1991). The significance change in CFF values during performance on a VDT-based visual task. In M. Kumashiro & E. D. Megaw, (Eds.), *Towards human work* (pp. 352–357). London: Taylor & Francis.

Jordan, N. (1963). Allocation of functions between man and machines in automated systems. *Journal of Applied Psychology, 47,* 161–165.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kantowitz, B. H., & Casper, P. A. (1988). Human workload in aviation. In E. L. Weiner & D. C. Nagel, (Eds.), *Human factors in aviation* (pp. 157–187). San Diego: Academic.

Kogi, K., & Saito, Y. (1971). A factor-analytic study of phase discrimination in mental fatigue. *Ergonomics, 14,* 119–127.

Lachman, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive psychology and information processing: An introduction.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Matthews, M. (1986). The influence of visual workload history on visual performance. *Human Factors, 28,* 623–632.

McRuer, D. T., & Jex, H. R. (1967). A review of quasi-linear pilot models. *IEEE Transactions on Human Factors in Electronics, 8,* 231–249.

Minsky, M. (1984). *The society of mind.* New York: Simon & Schuster.

Miyake, S., Hancock, P. A., & Manning, C. M. (1992, October). *Effects of prior task load level on subsequent workload and performance.* Paper presented at the 36th Annual Meeting of the Human Factors Society, Atlanta, GA.

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

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. Boff, L. Kaufman, & J. P. Thomas, (Eds.), *Handbook of perception and human performance* (pp. 42:1–49). New York: Wiley.

Ornstein, R. E. (1977). *The psychology of consciousness.* New York: Harcourt Brace.

Poulton, E. C. (1965). On increasing the sensitivity of measures of performance. *Ergonomics, 8,* 69–76.

Reid, G. B. (1989). *Subjective Workload Assessment Technique (SWAT): A user's guide*(U), AAMRL-TR-89-023. Wright-Patterson Air Force Base, OH: Systems Research Laboratory.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati, (Eds.), *Human mental workload* (pp. 185-214). Amsterdam: Elsevier.

Saito, K., Hosokawa, T., Saito, T., Nakai, K., & Inzuka, Y. (1988). *VRT and pocket flicker as a new apparatus for fatigue measurement.* Grant-in-aid for developmental scientific research (2) 62870024 Report.

Sawin, A. D., & Scerbo, M. W. (1993). Vigilance: Where has all the workload gone? *Proceedings of the Human Factors and Ergonomics Society, 37,* 1383–1387.

Scallen, S., Duley, J., & Hancock, P. A. (1994, April). Pilot performance and preference for cycles of automation in adaptive function allocation. *Proceedings of the First Automation Technology and Human Performance Conference.* Washington, DC: Catholic University of America.

Smith, H. J. (1967). Human describing functions measured in flight and on simulators. *IEEE Transactions on Human Factors in Electronics, 8,* 264–268.

Smith, K., & Hancock, P. A. (in press). Situation awareness is adaptive, externally directed consciousness. *Human Factors.*

Warm, J. S. (1993). Vigilance and target detection. In B. H. Huey & C. D. Wickens (Eds.), *Workload transition: Implications for individual and team performance* (pp. 139–170). Washington, DC: National Academy Press.

Warm, J. S., Dember, W. N., Gluckman, J. P., & Hancock, P. A. (1991). Vigilance and workload. *Proceedings of the Human Factors Society, 35,* 980–981.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20,* 158–167.

Wickens, C. D. (1993). Workload and situation awareness: An analogy of history and implications. *Insight: The visual performance technical group newsletter, 14,* 1–3.

Yeh, Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measurement of workload. *Human Factors, 30,* 111–120.

APPENDIX

Let φxi be phase shift between forcing function and control output in X direction and let ΔTx be average time lead (or time delay) between these signals in frequency domain from 0.1 to 0.3 Hz:

$$\Delta T_x = \frac{1}{n} k_{i=j} \left(\frac{\varphi xi + 90}{360} \cdot \frac{1}{f_i} \right) \quad (f_j = 0.1 \text{ Hz}, f_k = 0.3 \text{ Hz})$$

$$= \frac{1}{n} k_{i=j} \frac{\varphi xi}{360fi} + \frac{1}{n} k_{i=j} \cdot \frac{1}{fi} \quad \text{(sec)}$$

Let right side be $\Delta T'x + c$ and $fi = \Delta f \cdot i$, where $\Delta T'x$ is the time delay between forcing function and system output and and Δf is the frequency resolution:

$$c = \frac{1}{4n\Delta f} k_{i=j} \frac{1}{i} \quad \text{(sec)}$$

and

$$\Delta f = \frac{1}{N \times \Delta t} = \frac{1}{512 \times 0.2} = \frac{1}{102.4} \quad \text{(Hz)}$$

because N is the data number (512) and Δt is the sampling interval (0.2 sec).

$$j = \frac{f_j}{\Delta f} = 10, \; k = \frac{f_k}{\Delta f} = 30$$

(Continued)

Therefore,

$$c = \frac{1}{4} \cdot \frac{1}{(k - j + 1)} \times 102.4\ \ 30_{i=10} \frac{1}{1} \times 1{,}000$$

$$= 1{,}421.432 \text{ (msec)}$$

Finally, combined time lead (CTL) is,

$$\text{CTL} = \sqrt{\Delta T_x{}^2 + \Delta T_y{}^2}$$

$$= \sqrt{(\Delta T'_x + c)^2 + (\Delta T'_y + c)^2}$$

Hancock, P.A., Williams, G., Miyake, S., & Manning, C.M. (1995). The influence of task demand characteristics on workload and performance. International Journal of Aviation Psychology, 5 (1), 63-85.