Human Factors Research Reports—AASHO Road Test

I. Field Study of Vigilance Under Highway Driving Conditions

D.A. DOBBINS, J.G. TIEDEMANN, and D.M. SKORDAHL, Human Factors Research Branch, The Adjutant General's Research and Development Command, U.S. Army, Washington, D.C.

• VIGILANCE can be defined as the prolonged ability to detect certain environmental signals. Many individuals who are required to observe infrequent or irregularly spaced visual signals detect a progressively smaller proportion of signals with the passage of time. As found in laboratory studies, the classical decrement in vigilance occurs rapidly during the first thirty minutes of monitoring and then stabilizes at a low detection plateau (14).

In the present study, research opportunity was afforded by the American Association of State Highway Officials (AASHO) Road Test to study signal detection performance under realistic field conditions rather than by the classical laboratory approach. For the Road Test, Army drivers were required to drive trucks on experimental highways under monotonous and fatiguing conditions, thus providing a "natural laboratory" for a study of vigilance. The factors already present in the driving situation which were expected to challenge driver vigilance and consequently allow observation of the extent of performance decline under actual operating conditions included truck noise and vibration, boredom induced by repeated circling of the driving loops, and the sheer physical fatigue caused by driving large and heavily loaded trucks. To these factors were added the energy expenditure demanded by the experimental vigilance task which required each driver to discriminate among nearly 850 signals in order to respond to approximately 210 critical signals over 7 hr of driving.

The specific objectives of the human factors portion of the AASHO investigation were

1. To develop an apparatus test to measure vigilance during the actual driving process.
2. To analyze the vigilance performance of a group of drivers during one full 7-hr driving shift to determine the level of and trends in performance in detecting signals as a function of time spent in monitoring.
3. To determine the extent of individual driver differences in signal detection performance and to estimate the reliability of these differences.

METHOD

Vigilance Tester

A portable apparatus, the U.S. Army Transportation Corps Vigilance Tester, was designed for use on the trucks (for a detailed description of the tester, see 5). The apparatus consisted of a visual signal display unit mounted on the truck dashboard, a foot-operated response pedal, and a combined programer and response recorder unit mounted on the truck bed. The signal display unit was a circular area 5 in. in diameter divided into six red and nine white 24° radial panels with a 150-milliamp light bulb positioned behind each panel. The red panels were arbitrarily designated as the critical areas; that is, requiring a response by the driver whenever a signal appeared in one of the areas. The white panels were designated noncritical signal areas and required no response at the occurrence of a signal. A predetermined program for activating the visual signals (light bulbs) was maintained through the use of a punched tape.

Signals occurred in the display area at a rate of 30 critical and 91 noncritical signals per hour. Intersignal intervals of 5, 10, 15, 20, 30, 50, and 75 sec were used. Both the position of the signals appearing on the display panel and the intersignal intervals that followed were randomized. The duration of both critical and noncritical signals was 1 sec and the driver had 5 sec following each signal in which to respond. Automatic counters were used to record signal and response data.

Subjects

The subjects were 42 enlisted drivers from the AASHO Road Test Support Activity of the U. S. Army Transportation Corps. The only standard used for the selection of drivers was that they had to have been assigned to the particular driving loop for at least one month before testing.

The records of 42 drivers were considered acceptable for the purposes of this report even though many more were tested. The records of many drivers who participated in the experiment were discarded for the purposes of the present study because of interruptions in the driving shifts resulting from malfunctioning of the vigilance tester, truck breakdowns, and need for highway maintenance. The only records used in the statistical analyses were those of drivers who had uninterrupted testing for six consecutive driving periods.

Experimental Procedure

Before beginning a driving shift, drivers were told that lights of 1-sec duration would appear, one at a time, in different red and white panels of the display. Their task was to monitor the display unit while driving the truck and to respond only to light signals appearing in red panels by depressing the foot pedal. Each driver responded to a few practice signals before beginning his driving shift to make sure he fully understood his task.

The monitoring session was a 9-hr driving shift divided into seven periods of 90, 90, 90, 60, 45, 45, and 30 min, respectively—a total of 7½ hr testing time. Five 15-min rest periods and one 40-min meal break separated the driving periods. To prevent confounding day and night effects within a single driving shift, the last 30-min driving period of both the day and the night shift was discarded from the data analysis because it contained the twilight time. The performance measures were taken cumulatively at the end of each period, but no within-period measures could be taken because of recording limitations of the apparatus. Testing for the record began in August 1960 and continued through November 1960.

Performance Measures

The vigilance performance of a driver was evaluated by means of two scores. The first score was "Percent Detections," simply the percentage of critical signals detected, computed as

$$\text{Percent Detections} = \frac{\text{(Number of responses to critical signals)}}{\text{(Number of critical signals presented)}} (100)$$

The second score was "Percent False Detections"—an index of errors of commission. The total number of false detections made to noncritical and imaginary signals was divided by the number of noncritical signals presented as

$$\text{Percent False Detections} = \frac{\text{(Number of responses to noncritical signals)} + \text{(Number of responses to imaginary signals)}}{\text{(Number of noncritical signals presented)}} (100)$$

Responses to imaginary signals were defined as those instances in which the subject depressed the response pedal even though no signal was on the display unit at the time. Although the inclusion of these errors in the numerator of the performance fraction was questionable mathematically, these two types of false detections seemed to represent the same kind of error and hence were pooled.

Statistical Control Procedures and Analysis

Inspection of the frequency distributions of Percent Detection and Percent False Detection revealed marked skew as well as correlation between means and standard deviations. Thus, to make the data amenable to parametric analysis, each driver's vigilance scores were transformed to arcsins ($X = \arcsin\sqrt{P}$) which satisfactorily reduced these irregularities.

Also, t-tests were performed for possible differences between the mean detection levels for two variables imposed by the nature of AASHO Road Test conditions: (a) driving schedule and (b) the particular experimental loop on which driving was accomplished. Because no significant differences were found for either schedule or loop effects, scores for drivers on different schedules and on different loops were merged.

F-ratios revealed a significant difference between mean detection levels of day shifts vs night shifts. The difference in percent detections between day and night drivers is believed to have occurred because of a difference in the amount of contrast between the signal display and the surround during day and night conditions, and is not considered to reflect true day-night differences in alertness. (The surround of the display unit was the driver's view through the windshield of the truck. The surround during the day driving consisted of colors, movements, and glare not present to nearly the same degree during night driving.) Nevertheless, as a result of this analysis, day and night shift results were kept separate in all subsequent analysis to aid in interpretation.

Analyses of variance were performed on the transformed data to test for differences in vigilance between day and night drivers, vigilance score differences among the means of the six driving periods, and interactions (6, pp. 220-232). The analyses of variance were supplemented by tests for significant trends in the driving period means, and for differences between the slopes of the day-period means and the night-period means (6, pp. 247-250).

Changes in intersubject variability from the first to the sixth driving periods were tested by means of t-tests (18, p. 244). The reliability of individual score differences was estimated by internal consistency analysis.

RESULTS

When the six driving periods for all 42 drivers were combined, it was found that approximately 83 percent of all critical signals were detected—unusually high considering the conditions under which the monitoring task was accomplished. The over-all percent of false detection was also considered low, averaging only 4 percent for all drivers. (As previously reported, the night drivers detected a significantly higher percentage of critical signals than did day drivers; however, no significant difference in percentage of false detections was found between the two groups.) Tables 1 and 2 give the mean percent detections and false detections in arcsin form by driving period for day drivers, night drivers, and for the two groups combined.

Detection levels over the six driving periods for all 42 drivers showed a significant increasing trend, which was linear, instead of the hypothesized decrement. In terms of percent false detection, the total group of 42 drivers demonstrated a significant decreasing trend, also linear, over the six-period driving shift. (Only the night drivers contributed to these significant trends, whereas the day drivers showed no significant increase or decrease for either vigilance measure. Once again, the visibility artifact was suspected to have caused the increasing average level of correct detections. Toward the early morning hours, the surround of the display unit became more homogeneous, making detections of critical signals easier for drivers tested at night.) Figures 1 and 2 show mean percent detection score trends and mean percent false detection score trends, respectively, by driving periods. Table 3 is a summary of F-ratios from trend analyses of period means. No trend of a higher order than linear was found to be significant.

Although average percent detection levels were consistently high, wide individual differences in vigilance performance were apparent. These individual differences in

TABLE 1

PERCENTAGE OF CRITICAL SIGNALS DETECTED BY ROAD TEST DRIVERS IN SIX DRIVING PERIODS (Arcsin Scores)

Period	Day Driving (N = 19)			Night Driving (N = 23)			Total (N = 42)		
	Mean	Std. Dev.	Range	Mean	Std Dev.	Range	Mean	Std. Dev.	Range
1	61.2	9 7	39.2-75.9	72 2	11.0	49.4-90.0	67.2	11.7	39.2-90.0
2	60.3	12.3	40.5-90.0	73 9	10.2	54.3-90.0	67.7	13.1	40.5-90.0
3	60.1	11.5	36.1-80.2	75.0	13.3	36.2-90.0	68.3	14.5	36.1-90.0
4	61.7	10.5	45.6-77.3	73.1	10.9	51.5-90 0	67.9	12.1	45.6-90.0
5	61.8	13.4	43 7-90.0	77.6	11.4	52 9-90.0	70.4	14.6	43.7-90.0
6	59.8	17.6	27.1-90.0	78.3	12.3	53.7-90.0	69.9	17 5	27.1-90.0
All	60.8	10.4	27.1-90.0	75.0	9.4	36.2-90.0	68.6	12.2	27.1-90.0

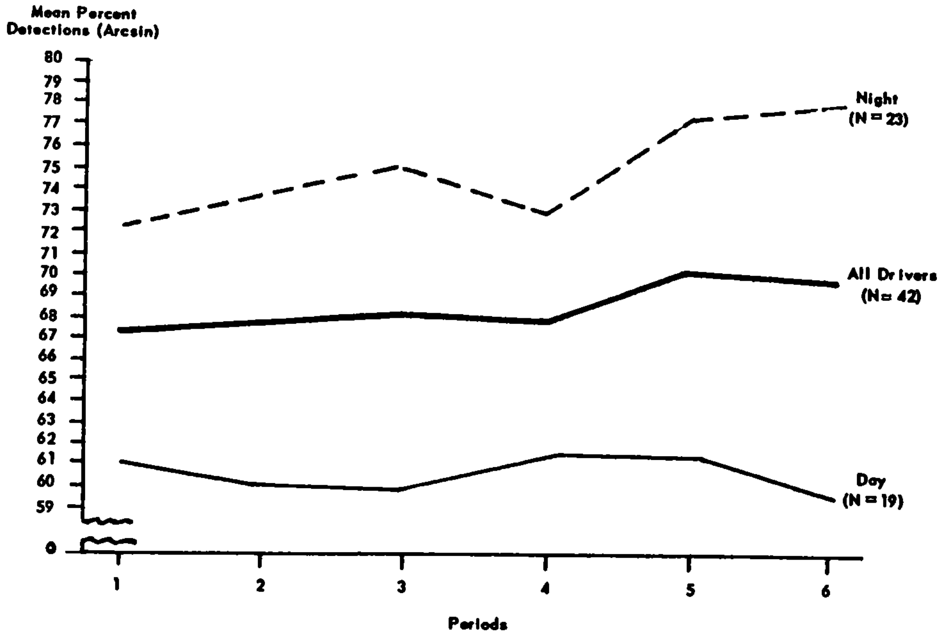

Figure 1. Mean percent detection scores by driving period.

percentage of detections tended to increase as the driving time increased. Tables 1 and 2 give the standard deviations and ranges of transformed scores for each driving period. Intersubject variability in percent detections increased significantly between driving periods 1 and 6. Intersubject variability in percent false detections decreased significantly during the same interval for the total group of 42 drivers. (A significant increase in percent detection variability was obtained for day drivers only, and a significant decrease in percent false detection variability was obtained for night drivers only. A statistical artifact is presumed to have prevented a significant increase in night drivers' percent detection variability. As night drivers' average scores increased during the early morning hours due to better visibility, individual score variation was limited by the 100 percent upper score limit. Day drivers whose visibility remained fairly constant did not show the steady increase in average percent detection scores, and their

TABLE 2

SCORES ON FALSE RESPONSE MEASURE FOR ROAD TEST DRIVERS IN SIX DRIVING PERIODS (Arcsin Scores)

Period	Day Driving (N = 19)			Night Driving (N = 23)			Total (N = 14)		
	Mean	Std. Dev.	Range	Mean	Std. Dev.	Range	Mean	Std. Dev.	Range
1	8.4	5.8	0-20.4	10.4	9.3	0-42.4	9.5	7.9	0-42.4
2	10.8	9.3	0-36.7	9.6	11.6	0-40.6	10.1	10.7	0-40.6
3	7.3	5.6	0-19.2	8.9	8.0	0-31.2	8.2	7.1	0-31.2
4	9.8	14.5	0-66.3	7.0	6.2	0-15.9	8.2	10.8	0-66.3
5	9.1	11.7	0-43.9	6.0	6.5	0-18.3	7.4	9.3	0-43.9
6	4.9	5.0	0-14.5	6.5	5.9	0-17.0	5.8	5.6	0-17.0
All	8.4	5.7	0-66.3	8.1	6.0	0-42.4	8.2	5.9	0-66.3

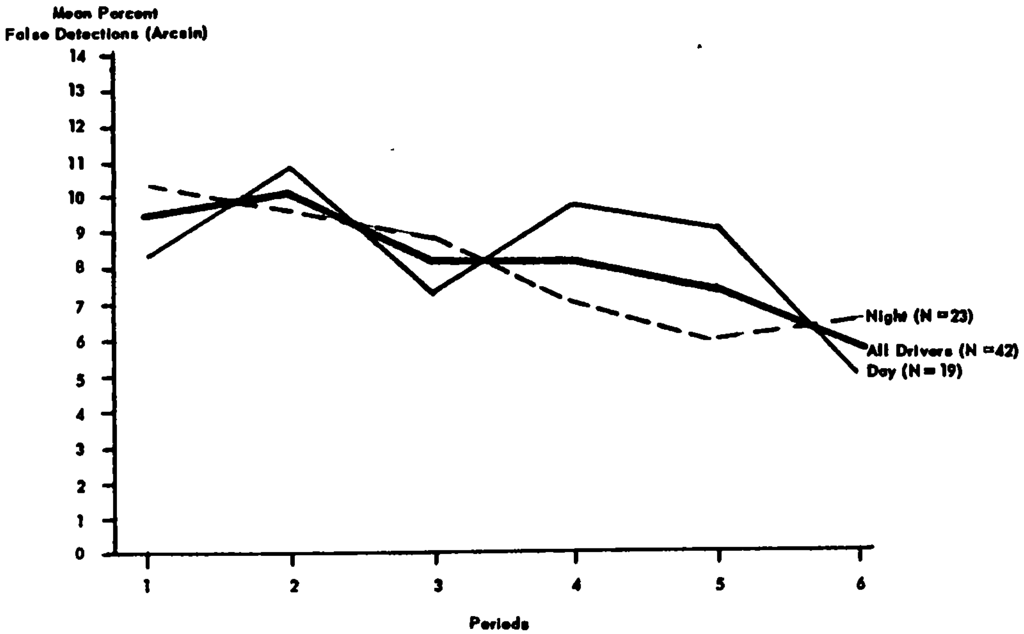

Figure 2. Mean false detection scores by driving period.

performance was consequently more free to vary toward the end of their driving shifts. The significant decrease in variability of false detections may be similarly accounted for.) Table 4 shows the significance of changes in variance between driving periods 1 and 6 for both vigilance measures.

The reliability of individual differences for both vigilance measures was high. Reliability coefficients, computed between scores on odd-numbered and even-numbered driving periods, and augmented by the Spearman-Brown formula, were 0.93 and 0.88 for percent detections and percent false detections, respectively. Table 5 gives the reliability coefficients for both vigilance measures using odd vs even periods and first vs sixth driving periods. The latter analysis showed moderately high stability of percent detection scores (r =0.67) and lower stability of false detection scores (r = 0.25) even between scores separated by 5½ driving hours.

In summary the four major findings of this investigation were (a) the high percentage detection and low percentage false detection levels; (b) the lack of decrement in vigilance as a function of time spent monitoring; (c) the high reliability of individual differences; and (d) increased variability in percentage detection performance as a function of time spent monitoring.

TABLE 3

SUMMARY OF F-RATIOS FROM TREND ANALYSES OF PERIOD MEANS OF VIGILANCE MEASURES (Arcsin Scores)

Detection	Periods	Trend (%)		
		Linear	Quadratic	Residual
Signal	Day only	0 0	0 0	0 4
	Night only	8.4[a]	0 3	0 8
	Total	4 4[b]	0 1	0.5
	Day/night × period	4 8[b]	0 3	0.4
False	Day only	2 0	2 0	1 1
	Night only	6 1[b]	0 1	0.2
	Total	7 7[a]	0 5	0.3
	Day/night × period	0.8	1 5	0 9
df		1/200	1/200	3/200

[a]Trend significant beyond 1 percent point.
[b]Trend significant beyond 5 percent points.

TABLE 4

SIGNIFICANCE TESTS OF DIFFERENCES IN VIGILANCE TEST SCORE VARIANCE BETWEEN DRIVING PERIODS 1 AND 6 (Arcsin Scores)

Detection	Periods	Variance Period 1 (%)	Variance Period 6 (%)	$S^2_6 - S^2_1$ Diff	p[a]
Signal	Day	94 09	309 76	+215 67	0 01
	Night	121 00	151 29	+ 30 29	N S
	Over-all	139 24	306 25	+167 01	0 01
False	Day	33 64	25 00	- 8 64	N S
	Night	86 49	34 81	- 51 68	0 05
	Over-all	62 41	31 36	- 31 05	0 05
df day = 18, night = 22, over-all = 41					

[a]Based on two-tailed t-tests

INTRODUCTION

Prolonged High Detection Levels

Not only did the expected decline in performance as a function of time spent in monitoring fail to occur but a significant linear increase in percent detections was found during the 7 hr of driving. Also, contrary to expectations, a significant linear decrease in percent false detections was found over driving shifts. (A procedural difficulty of the present experiment was the inability to record responses at intervals within driving periods. There is a possibility that some performance decrement occurred within driving periods but did not appear in the response measures because they were recorded at the end of driving periods. However, it is certain that within-period decrement did not occur to a great extent because the over-all high detection levels left little freedom for within-period variance.)

At least two previous experiments have shown that noise and vibration such as were present in the driving situation have a demonstrable effect of monitoring behavior (11, 13). The fact that 100 percent critical signal detection accuracy was not obtained in the present study indicates these factors may have had some influence on the vigilance performance of the drivers. The high detection levels indicate, however, the presence of compensatory factors that tended to overcome these decrement-inducing conditions. New information, in the form of recent research by other investigators, allows more enlightened speculation as to the nature of these compensatory factors.

Signal Characteristics.—Four major signal characteristics seem to be the most probable contributors to the prolonged high detection levels. These are the high rate of presentation of noncritical signals, the highly stimulating signal environment caused by the divided-attention nature of the task, the relatively restricted range of intersignal intervals, and the relatively long signal duration.

The "filter theory" of Broadbent (3) suggests that irrelevant signals during a monitoring task would tend to make the task more difficult because the monitor would be more likely to respond to false signals, and consequently miss more critical signals, with the passage of time. Many noncritical signals were programed into the present experiment. In apparent contradiction to Broadbent's theory, however, two recent studies (7, 16) have shown that the introduction of irrelevant signals into the monitoring task actually enhanced monitoring performance under certain conditions. The effect is marked when the artificial signals are perceptually similar to the real signals. In the present study, critical and noncritical signals were identical in all aspects except for the fact that critical signals appeared in red panels and noncritical signals in white panels. It seems probable that the high rate of perceptually similar noncritical signals used in the present study actually enhanced alertness.

TABLE 5

RELIABILITY COEFFICIENTS FOR VIGILANCE PERFORMANCE MEASURES
(Arcsin Scores)

Driving Period	Measure	Day Driving	Night Driving	Total
Odd vs even	Percent signal detection	0.94[a]	0.88[a]	0.93[a]
	Percent false detections	0.87[a]	0.91[a]	0.88[a]
First vs sixth	Percent signal detection	0.52	0.64	0.67
	Percent false detections	0.50	0.09	0.23

[a] Augmented by Spearman-Brown formula.

Another possible explanatory factor was the high complexity of the stimulus surround caused by the "divided-attention" aspect of the study. Studies of "complex" vigilance (that is, the monitoring of multiple signal sources) have reported no decline in vigilance (10, 11); however, the over-all detection levels reported were much lower than those of the present study. Moreover, the amount of environmental stimulation received from sources other than experimental signals was high and may have enhanced alertness. Hebb's "arousal" theory (8) proposes that high detection levels are a function of an optimal level of cortical arousal. The maintenance of cortical arousal is a function of the total stimulus input, according to Hebb, regardless of whether that stimulation is relevant to the monitoring task at hand. The present results seem to support Hebb's views.

Two other factors of the present experiment that may have helped prevent the appearance of a performance decrement were the relatively long 1-sec signal duration, and the relatively restricted range of intersignal intervals.

Task Characteristics.—Task factors that may have contributed to the lack of performance declines were the high degree of perceptual-motor activity level, progressively diminishing lengths of driving periods during a shift, and the presence of interpolated rest pauses.

Drivers, in performing their normal driving duties simultaneously with the experimental vigilance task, were engaged in a great deal more perceptual-motor activity than that required of laboratory subjects passively monitoring experimental displays. Ray, Martin, and Alluisi (20) have pointed to the conflicting results between active and passive vigilance tasks and have suggested that the degree of active participation required by the subject may make a critical difference in the appearance of performance decrement.

Two other aspects of the road test engineering research design probably helped maintain high detection levels. These were the diminishing length of driving periods as shifts progressed, and the interpolated rest pauses between driving periods. These two factors, planned by road test personnel for the lessening of driver fatigue, probably helped in maintaining alertness.

Subject Characteristics.—In the present study it is suspected that driver motivation may also have played a role in the maintenance of high detection levels. Most drivers seemed interested in the task and many who were not tested requested it. It is also suspected that the sources of motivation were extrinsic to the experimental task itself and could be attributed to relief from driving boredom. Many drivers, because they were also soldiers, had lingering suspicions that their performance might somehow be entered into their official service records, in spite of instructions to the contrary. This suspicion may have been an incentive to do well on the experimental task.

The role of motivation in vigilance has not been systematically studied. However, Adams and Chiles (1) found that highly motivated subjects were able to perform complex monitoring and cognitive tasks under fatiguing work schedules for as long as 15 days without serious performance decrement. Adams and Chiles used a realistic space mock-up apparatus in their experiment. Their task appeared to have much more face validity for their subjects than many others typically used in laboratory studies. This realism may help explain the high performance levels found in their study.

Extent and Stability of Individual Performance Differences

The finding of wide initial individual differences in vigilance levels and the relatively high stability of these individual differences during a complete driving shift suggests classifying the vigilance phenomenon among other relatively enduring aspects of the individual. High score reliability, if found for acceptable retest intervals, will qualify the vigilance phenomenon as a profitable subject area for psychologists interested in the prediction of individual differences. Only recently have some attempts been made to conceptualize the vigilance phenomenon as an attribute of the individual observer. These attempts are discussed in Part II of this paper.

The increased variability in percentage detections during a monitoring period is a common finding in vigilance research. Buckner, Harabedian, and McGrath (4) present evidence that detection variability increases not only in percentage measures but also in measures of threshold sensitivity and response latency. The explanation suggested by these data is that initial individual differences in vigilance levels became more exaggerated due to increasing motivational differences among subjects as boredom and monotony come more into play toward the end of the monitoring period.

CONCLUSIONS

In spite of inhibitory factors present in this study which would lead to a prediction of performance decrement (noise, truck vibration, long hours, boredom, and fatigue), other compensatory factors also present may have combined to cause prolonged high detection levels. The influence of inhibitory factors was apparent in increases in variability of performance, rather than in levels of performance.

Possible compensatory factors discussed were (a) signal characteristics, including high rate of noncritical signals, complexity of over-all stimulus conditions resulting from the divided-attention task, low range of inter-signal intervals, and high signal duration, (b) task characteristics, including high perceptual-motor activity level, interpolated rest pauses, and diminishing length of driving periods; and (c) subject characteristics, including motivation. The lack of within-period performance measures of the present study precluded precise comparisons with laboratory experiments. The present study, however, showed that detection performance began at a high level and stayed at a high level in spite of noxious monitoring conditions. Along with other research on active and complex monitoring tasks, the present study suggests the rapid, severe decrement found in the passive monitoring of laboratory displays may be of limited generality. The results of the latter type of study do not seem to represent human monitoring proficiency adequately when the monitoring task is meaningful and when monitors are fairly active physically. The classical decrement curve may represent a basic and significant perceptual phenomenon under conditions of reduced organismic stimulation. However, it is felt that investigators who advise military and industrial management on such factors as personnel monitoring schedules should attempt more realistic simulation of representative signal environments and conditions of work. Otherwise, there is a danger of seriously underestimating human monitoring capacity on a great many operational tasks.