AMRL-TR-66-121
AD0638183

USE OF THE EXPERIMENTAL METHOD FOR EVALUATIONS OF PERFORMANCE IN MULTI-MAN SYSTEMS

George E. Passey, PhD
Earl A. Alluisi, PhD

Lockheed-Georgia Company

W. Dean Chiles, PhD

Aerospace Medical Research Laboratories

August 1966

Distribution of this document is unlimited.

AEROSPACE MEDICAL RESEARCH LABORATORIES
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO

NOTICES

FOREWORD

This report was prepared by Dr. George E. Passey and Dr. Earl A. Alluisi of the Lockheed-Georgia Company and Dr. W. Dean Chiles of the Aerospace Medical Research Laboratories. The research was carried out under Contract AF 33(657)-10506 and under previous contracts between the Lockheed-Georgia Company, Marietta, Georgia and the Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio. This work was supported under Project 1710, "Human Factors in the Design of Training Systems," Task 171002, "Performance Effects of Environmental Stress." This paper was read by Dr. Passey at the Seventh Annual Meeting of the Human Factors Society, held at Palo Alto, California on 23-25 October 1963. This report was originally published as AMRL Memorandum P-67, February 1964.

ABSTRACT

This report discusses the use of the experimental method as a technique for arriving at solutions to human factors engineering problems encountered in the design of multi-man systems. Of specific concern are the methodological decisions that must be made in the design of the research. Factors that are likely to influence these decisions are considered as well as the implications of these decisions with respect to the validity and generality of the data thus obtained. These various decision points are illustrated through use of data on group performance during long-term confinement.

No matter what approach he may eventually choose, once he has clearly defined his problem, the human factors engineer must begin his attack on that problem with a search for relevant data. If he is fortunate, he finds the necessary data either among the items stored in his mental inventory or in the library. If he is only typically lucky, he may, at best, come upon "somewhat" relevant data that can be generalized, extrapolated, or "argued by analogy." Not infrequently, even the well-trained and lucky human factors engineer finds that relevant data are not available, or that data which appeared at first to be relevant are not really so because certain parameters were not controlled.

In the absence of relevant data, what should the human factors engineer do? Obviously, he should do whatever is necessary to obtain the needed data; where the library has failed, the laboratory must be made to succeed. If he decides to conduct experimentation to obtain the data he considers necessary, he will be faced with a number of methodological questions--questions, the answers to which will have important effects on his research and subsequent application. For example, should the experimental situation faithfully resemble the specific system in question, or should it represent an abstraction of those factors that the researcher feels are most critical to the question? The more specific situation has the advantage of high face validity, but if the specificity is too great, additional research on the same question may be necessary when the now-new system is significantly modified or becomes obsolete and is replaced.

During today's symposium, I would like to discuss some of these methodological questions. I would like to point out some of the implications (constraints and assumptions) of the decisions made in response to them. We firmly believe it important that the decisions be made overtly, and with some rational weighing of their implications, for they are decisions that determine the usefulness of our results. As H. M. Johnson so cogently pointed out, pre-experimental assumptions are indeed determiners of experimental results. In order to lend substance to our discussion of these methodological questions and decisions, I shall use examples drawn from our own experience in conducting experimental research on crew performance. Fortunately we were in a position in which we were not being hounded by the design engineer for answers to be applied to systems going into production the following week, though we feel we would not have made significant changes in our methodological decisions had that been the case.

At any rate, we would have resisted vigorously any pressures to forget about the experimental niceties.

The problem to which we were directing our attention had the following general characteristics. (1) The class of systems with which we were concerned involved multi-man crews; e.g., there was a requirement for more than 24 man-hours of work per day. (2) The mission durations were to be 5 days or longer. (3) Group-dependent, as well as individual, performance was to be involved. And, (4) the environment in which the crew would operate would be rather restrictive at best. These general characteristics are broadly descriptive of a number of present and potential systems ranging from the airborne command post to the remote ground radar station — the nuclear powered aircraft to the satellite or space vehicle.

In designing research of this sort it is our opinion that every effort should be made to maximize the number and variety of systems to which the results can be generalized. This, of course, must be tempered by the needs of the system or systems that are the sources of justification (or funds) for doing the research. In many instances this approach may require an expert job of selling, both to those in management who have to approve the work and to the operational user of the data gathered.

There are instances, of course, in which the number of copies of a system will be quite small (say, 5 or fewer), the cost of the individual copy will be high, and the number of personnel who will have to be trained to operate the system will be small. In such cases one might argue that exact simulation, not only of projected tasks, but also of configurations should be the guiding principle. However, one does this at the risk of decreasing the ultimate scientific value and general applicability of data so obtained.

When we finally sit down to design the experiment in detail, perhaps the first question we should ask ourselves is, "What is it that the human element will be expected to provide in this system?" When we have answered this question, at least in broad outline, we should then set about to translate these notions into performance requirements. In this pursuit, we feel that the keynote should be the sampling of performance abilities, as opposed to specific tasks — psychological and psychomotor functions, as opposed to exact operational skills. In other words, we should set as our goal the assessment of those characteristics of the human operator which make him indispensable to the particular system or systems in question.

At a fairly general level, and within the context of our confinement research, the answer to this question was that the operator would make his major contribution to our system through application of his capacities as a receiver and processor of information, a decision maker, and, of course, as a back-up monitor of system operation.

Upon entering the translation phase we encounter a number of very important questions, some of which are primarily related to experimental factors and some of which are primarily related to design or system factors. In each of these cases our answers must not only be defensible but must also be workable. By way of illustration, we believe it can be said with reasonable confidence that there are no important instances in which the validity coefficient can exceed the reliability coefficient. Thus, one of the researcher's first problems is to achieve mechanical and electrical reliability in his simulated tasks and then to demonstrate the statistical reliability of the performance measures obtained with those tasks.

The question of developing a valid measure of the psychological performance function in which we are interested is an old one to the human factors researcher. He is often confronted with the question, "What the hell does that have to do with flying an airplane?" All too frequently the easy way out seems to be to make the task display and the responses so obviously copied after a well known flying task that even the hardest-nosed "operational type" will not object to it. This leads very readily to the notion that the best approach is to build or buy a simulator that duplicates faithfully all of the instruments and controls to be included in the system under consideration. However, it seems to be an unfortunate fact that the requirements for building a simulator that will duplicate the vehicle tasks and also yield reliable performance data are incompatible with the constraints imposed by the typical budget available to the human factors researcher.

In addition, when we exceed a certain degree of complexity in a simulated system, we run the risk of proving conclusively that a college sophomore can not learn to fly an F-102 simulator with only 6 hours of training. In other words, we may have to use the relatively small number of busy operational people who happen to be qualified on a system that is sufficiently similar to the one which we are simulating.

On the other hand, we might approach the question of validity from the point of view of the factor analyst. We deduce the psychological factors that are involved in this particular performance situation and then pull out tried and true tasks with established factor loadings on these functions. The major problems in this approach are as follows: (1) Factor loadings on most of these tasks have been demonstrated to vary as a function of the level of training of the operator and his basic ability to perform the task. (2) On essentially none of these tasks have the factor loadings been obtained when that particular task was performed as a part of a realistic task complex. (3) Few such tasks have exhibited useable validity coefficients against operational criteria. Finally, (4) relatively few of these tasks possess a satisfactory degree of face validity from the point of view of motivating the subject.

Turning again to our specific research, we translated the broadly stated functions to be required of the operator into six performance tasks. First, we selected a mental-arithmetic task as being a measure of information processing in which the manipulation of numbers was required; it also provided a measure of immediate memory in that the use of paper and pencil was not permitted. Second, we selected a target-identification task as representative of the general kind of perceptual problem that would confront the operator; it, too, contained an immediate memory requirement and also required group-based responses. Third, we selected an information seeking and temporal coordination task which is identified as the code-lock task in the appendix. Fourth, we selected a warning-lights monitoring task which involved both the response to the introduction of signals and to the removal of signals. Fifth, we selected a vigilance-type monitoring task that was dependent upon the auditory sense modality. And sixth, we selected a more complex monitoring task that required the integration of a stochastic process over time and the detection of shifts in the mean value of that process.

Once these task ideas were converted into hardware, our next step was to obtain four kinds of information about them; namely, (1) the rate at which subjects become proficient on them when presented individually, (2) the test-retest reliability of the measures, (3) the intertest correlations, and (4) the interactions among tasks when performed simultaneously.

As a result of these investigations into the nature of the tasks, two tasks which I have not mentioned were dropped from the battery because they did not possess satisfactory reliability. At that time our judgment was that these two tasks did not merit the expense involved in modifying them so that they would be acceptable. The remaining task battery, as adumbrated a minute ago, possessed a very impressive degree of reliability (as shown in Table 1 of the appendix); it apparently measured essentially orthogonal functions, and, through the selection of various task combinations to be performed simultaneously, it permitted a relatively broad range of work loads to be imposed upon the operator. In addition to this, the performance asymptote could be reached with a reasonable amount of training. For those concerned about a face validity, let me indicate that we also found in our early studies with operational personnel that the consumer acceptance of the tasks and the task-loads used was quite encouraging.

Let us turn now to some of the non-task aspects of our problem area as outlined in the introductory remarks. You will recall that I said we were interested in long-duration missions in an aerospace vehicle. This, of course, immediately implies a requirement for a considerable degree of complexity of the life-support aspects of the system. Thus, very early in the game we were faced with the question, "To what extent should we be concerned about the inclusion of such factors in our experimentation?"

The obvious, but not very informative answer to this question is that we should be concerned "at least to the extent that such factors might influence performance." I won't go into all of the possibilities that we considered and rejected, but will summarize the final philosophy we adopted; it was as follows: If the design engineers concerned with all the life-support factors (ranging from nutritional requirements to waste disposal) are able to achieve the goals they have set for themselves, then, in effect, they will have simulated the environment we would employ in our studies. In other words our subjects would be restricted in only two ways -- socially and physically.

One might ask, of course, how do we know that the results we get in this environment are the same as those that we would obtain with a closed ecological system mounted on a shake-table at the end of the boom on a centrifuge? Our honest answer must be that we don't know that they are the same. But there were two things that led us to conclude that we should be satisfied with this aspect of our methodology — satisfied without things like the centrifuge or shake table. You have already guessed the first of these — the cost (and the money available to us). The second was this: Our main interest was in providing a situation that could be used to subject the operator to what we considered to be some of the important rigors of aerospace vehicle operations and, in so doing, to give the null hypothesis a chance to be rejected. In this respect it is comforting to recall that we can never prove the null hypothesis, no matter how faithful the simulation.

Once we had settled upon the broad outlines of our first studies we ran into additional problem areas. I will take just one example of these and show how it affected the usefulness of our data. We were preparing to conduct our first 4-day confinement study. In this study we were primarily interested in determining the optimum durations of the work period and the rest period. At this point the question arose, "How should we schedule the sequencing of tasks and task loads? Should we pick out some arbitrary potential mission profile, or should we insure frequent repetition of tasks and task loads over a 24-hour period?" Our decision was that we should use a basic 2-hour program which would be repeated 12 times per day which is shown in Table 2 of the appendix. I think the data in Figure 3 rather clearly substantiate the wisdom of our choice.

The diurnal variations shown in Figure 3 could not have been obtained had the tasks not been presented more-or-less equally distributed over the 24 hours of each day. Although I have shown only the diurnal variations obtained in the performance of arithmetic computations, let me point out that these variations occurred in all other measures of performance as well as in the physiological indices of behavior (heart rate, respiration rate, temperature, and skin resistance level). The same diurnal variations both in the underlying physiology (as shown in Figure 4 of the handout) and in performance of arithmetic computations as shown in Figure 5 were evident in the 15-day tests of the 4-2 work-rest schedule that followed the initial 4-day tests. Incidentally, the subjects in the 15-day study were operational personnel from the Strategic Air Command.

The 4-2 work-rest schedule had been selected for detailed 15-day study because it appeared to provide the best trade-off of the requirements for high-level performance and operational efficiency, the latter being measured in terms of man-hours of work per man. Our first 4-day studies had sought to measure the trade-off between the length of the work period and the length of the rest period when the ratio of work-to-rest was held constant. Thus, four conditions were studied; in one, the subjects worked 8 hours and were off-duty for 8 hours. In another condition, the subjects worked 6 hours and were off 6; the third and fourth conditions involved 4-4 and 2-2 work-rest schedules, respectively.

The data of the 4-day studies indicated that the trade-off in the kind of work demanded in our situation favored the shorter work period, even though this involved a shorter rest period. Subsequently, two additional conditions were studied (6-2 and 4-2 work-rest schedules), and the conclusion reached was that the 4-2 schedule was essentially the most demanding (most efficient) one that could be used with practicality if the length of the mission was to exceed 4 days by any great amount.

The data collected in the 15-day confinement studies of the operational crews following the 4-2 schedule indicated that (with proper controls of selection and motivational factors) men could be obtained who could follow such a schedule without serious decrements in performance for periods at least as long as 15 days, and probably for 30 days. In order to measure something of the degree to which selection and motivation would affect the performances obtained, another study was conducted with a crew of 6 men working the 4-2 schedule for 15 days. In this case, rather than operational personnel, Air Force Academy Cadets were used. These Cadets, representing a more select population than the operational crews, were, in addition, quite highly motivated to show that "The Academy" could "beat all."

The results can be summarized as follows: Of the measureable trends in performance, approximately 2/3 were statistically significant in each of the two studies. Of these statistically significant trends, 90% of those obtained in the earlier study were negative, indicating decrements in performance with continued work under the conditions of experimentation. Only 78% of those obtained with the Cadets were negative, however. This difference, i.e., the difference between the 90% and the 78% negative trends, reflects, we believe, the differences that might be expected to obtain in the performance of the sorts of missions simulated as a function of selection and motivation. It also demonstrates the degree to which the selection and motivation of subjects must reflect the "real" situation if valid generalization is to be made.

Finally, a 30-day study was conducted with two 5-man crews of USAF pilots who worked alternating shifts on a schedule of 4 hours on duty and 4 hours off. It was found again in this case that about 2/3 of the possible trends in performance with time were statistically significant. However, only 8% of these significant trends were negative. Since this group of subjects fell between those of the other two groups in selectivity and motivation, we believe the differences in the findings can be taken as an indication of differences in the effects of the two different work-rest schedules. That is to say, apparently, the 4-4 schedule is essentially no stress at all; on the other hand, the 4-2 schedule is clearly stressful; and so much so that only very unusual subjects are able to prevent the occurrence of decrements in their performance of tasks of the kind used in our studies.

On the basis of the results such as these, and results obtained in further studies conducted this year, and also because the 4-4 schedule permits a safety factor (e.g., crewmembers can make up for an illness or other loss of man-hours by reverting to a 4-2 schedule where necessary), we feel that the 4-4 work-rest schedule should be recommended for use wherever high-alert performances are required on an around-the-clock basis. The conclusions reached, and the recommendations made can be generalized to a large number of different systems, both vehicular and nonvehicular.

This is the major point that we have tried to make. If we take the experimental approach to solving our human factors engineering problem, methodological decisions will have to be made. And in making these decisions, the researcher implicitly determines the extent of generalization and the degree of applicability of his research.

BIBLIOGRAPHY

1. Bevan, W., & Patton, R. M. Selected bibliography: fatigue, stress, body change and behavior. USAF WADC Tech. Rep., No. 57-124 (AD 118091).

2. Adams, O. S. Aircrew fatigue problems during extended endurance flight Phase I: Planning. USAF WADC Tech. Rep., No. 57-510.

3. Adams, O. S., Levine, R. B., & Chiles, W. D. Research to investigate factors affecting multiple-task psychomotor performance. USAF WADC Tech. Rep., No. 59-120 (AD 213592).

4. Adams, O. S., & Chiles, W. D. Human performance as a function of the work-rest cycle. USAF WADD Tech. Rep., No. 60-248.

5. Adams, O. S., & Chiles, W. D. Human performance as a function of the work-rest ratio during prolonged confinement. USAF ASD Tech. Rep., No. 61-720.

6. Alluisi, E. A., Hall, T. J., & Chiles, W. D. Group performance during four-hour periods of confinement. USAF MRL Tech. Doc. Rep., No. 62-70.

7. Alluisi, E. A., Hall, T. J., Hawkes, G. R., & Chiles, W. D. Human group performance during confinement. Lockheed-Georgia Company Engrng. Rep., No. 6024.

8. Chiles, W. D. & Adams, O. S. Human performance and the work-rest schedule. USAF ASD Tech. Rep., No. 61-270.

9. Alluisi, E. A., Hall, T. J., Hawkes, G. R., & Chiles, W. D. Human group performance during confinement. USAF ASD Tech. Rep., No. 63-87.

APPENDIX

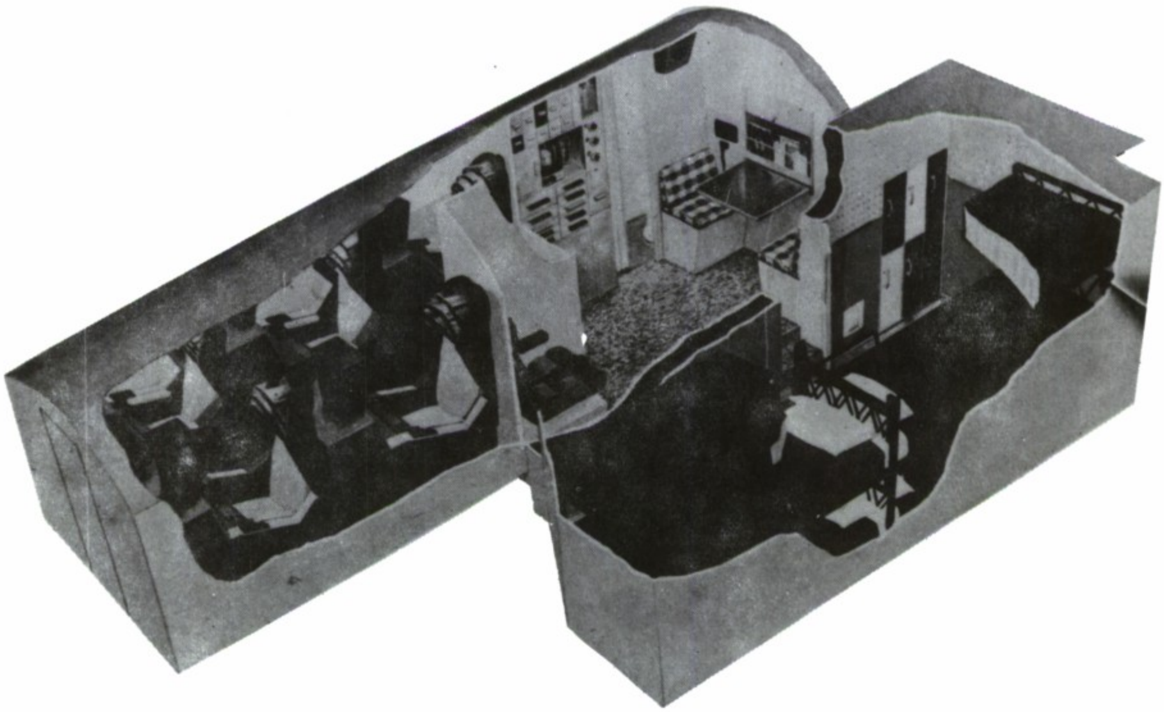

Figure 1. Confinement facility showing work stations, leisure area, and sleeping quarters.

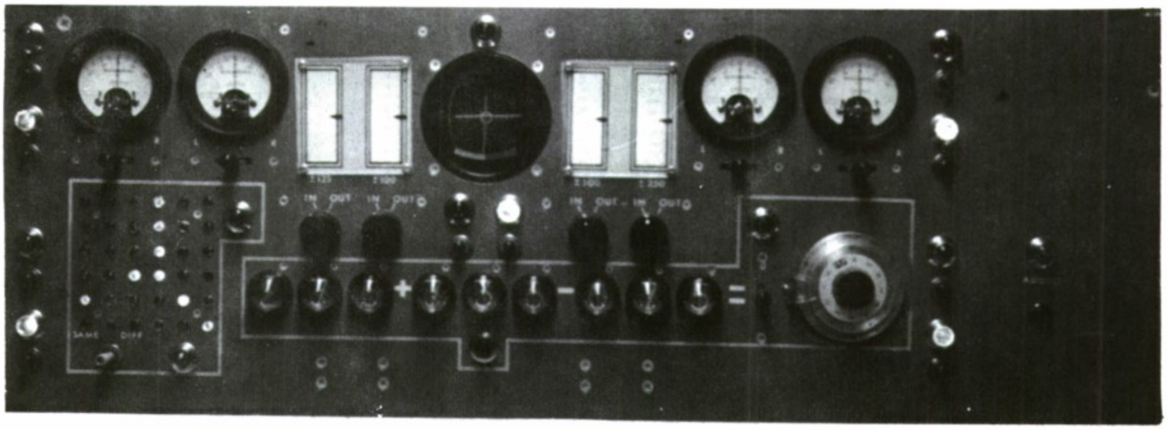

Figure 2A. Performance Panel Containing Original Task Battery

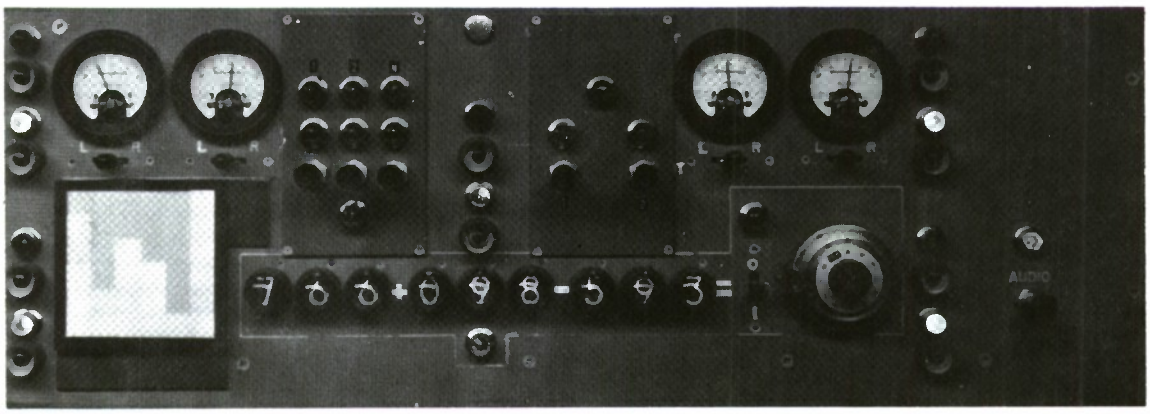

Figure 2B. Performance Panel Containing Modified Task Battery

PERFORMANCE BATTERY AS PRESENTLY CONSTITUTED

PROBABILITY MONITORING

In many flight situations (e.g., rough air), the position of an instrument scale or pointer may fluctuate. In these situations the task of the operator involves not only the simple detection of out-of-tolerance signals at a signle point in time, but also (a) the combination of the instantaneously acquired information with previously observed reading, and (b) a decision, based upon this combination of information, about the average value of the indication and its relation to some average reference value. The probability monitoring task was developed to reflect this emphasis.

A display consisting of four moving pointer indicators with a circular scale was selected. The indicators are located in pairs in the upper right and upper left portions of the instrument panel. Each pointer fluctuates in a random fashion about a mean value generating a Gaussian distribution with a known variance. The shifts in mean value of the pointer settings were of such magnitude that the probability of their being randomly drawn from a population of indications having a mean of zero bias was less than .05.

Indices of performance available are (a) number of false responses, (b) number of correct responses, and (c) time to detect a bias. In our studies biases have been introduced on only a single meter at a time at the rate of approximately nine per hour. A three position toggle switch allows for response to each display and activation of the switch provides for arresting pointer movement at the mean of the distribution then being generated. Such action provides for knowledge of results. Upon release of the switch any bias present is removed. Undetected biases are removed when the next programmed bias is introduced to any meter.

WARNING-LIGHTS MONITORING

Reaction time and probability of detection of discrete signals have been used in a number of studies as measures of attention or vigilance.

The task chosen extends the difficulty of the detection of, and response to, this type of signal by increasing the number of possible locations of the signals. Broadbent indicated that one result of fatigue is a decrease in the extent of the perceptual field from which information may be extracted. Similarly, one might argue that a decrease in the extent of the perceptual field to which a subject attends may also become restricted. The present task assesses, in part, the process of attention and vigilance using warning light indicators similar to those employed as oxygen warning lights, pump pressure warning lights, fuel system warning lights, radio altimeter lights, and the like.

Ten warning and indicator lights, five red and five green, were located in pairs at central and perhpheral points on the instrument panel. The subject is required to respond to any normally lighted green lights which are extinguished, and to any normally extinguished red lights which are illuminated.

Detection is signalled by depressing a switch below the affected light; this leads to a restoration of the normal condition. The principal index of performance is the time to detection (response latency). In our studies, only a single light has been placed in non-normal condition at a given time, about 10 critical signals have been presented during each hour of operation. Non-normal conditions remaining undetected for a period of two minutes are restored.

AUDITORY VIGILANCE

Failures of attention may manifest themselves in tasks which require vigilance for signals of an auditory, as well as of a visual nature. Evidence of the breakdown of auditory perception in a task of selective listening to two simultaneously presented messages, is given by Broadbent. He found that listeners were able to distinguish which of two voices, simultaneously presented, contained a relevant message only if they were warned in advance to listen only for one voice. Presumably, lapses of such selective attention for the relevant message will result in breakdown of performance on this type of task.

In this task the subject is required to note unequal gaps occurring in transmission of a continuously presented intermittent auditory signal, and to indicate detection by depression of a switch on the operator panel. The indices of performance are number of correct responses, number of deviant signals given, and total number of responses; the principal criterion used has been the percentage of correct detections. In order to be considered as a correct response the deviant signal must be reacted to within a period of 30 seconds. We have employed "beeps" of a 1175 cps tone presented intermittently for 0.25 second with a silent period of 0.90 second between "beeps". The critical signal used was a single off-period of about 50% greater, or 1.30 seconds. Critical signals in our studies have been presented eight times per hour.

ARITHMETIC COMPUTATION

Bartlett contended that "mental fatigue" would appear as deterioration in the regularity with which successive steps of performance in mental tasks are carried out. Computations involving addition and subtraction are necessary in many phases of vehicle operation, as in navigation, ranging and the like. To assess this function as well as immediate memory, involved in sequential operations, a mental arithmetic task was designed.

Three 3-digit numbers are displayed in the lower central portion of the instrument panel, on circular, single-digit numerical indicator tubes, arranged horizontally in triads. The subject is required to add two 3-digit numbers and to subtract from this sum a third 3-digit number; he records his answer by setting the last 3-digits of the answer on concentric dials. The first digit, either one or zero, is recorded by activating a three position toggle switch.

The index of performance is the total number of correct solutions. In the experimental programs in our laboratory a constant rate of presentation of three problems per minute has been employed.

TARGET IDENTIFICATION

Perceptual functions of a relatively complex variety are likely to deteriorate under conditions involving fatigue. Bartlett regards the disorganization of perceptual performance as an important criterion of fatigue. A task to examine this function was developed and as a part of the task the requirement of individual decision as well as command decision was incorporated involving supply of individual decision to the vehicle commander who then rendered a decision based on these inputs.

An array of 36 close-butted, square lights, forming a 6 by 6 matrix, was used in the presentation of contoured figures consisting of lit and unlit elements giving the appearance of solid bargraphs. Noise-free stored images of "targets," were used with other figures drawn from the same basic set but perturbed by visual "noise" to represent noisy sensed images of "targets." A noise-free target was followed by two noisy targets and the subject was required to determine whether the original target was contained in noisy target A, noisy target B, or neither.

The index of performance would be the number of correct identifications. In our studies a 5-second display was given for the noise-free target and a 2-second display was allowed each of the noisy targets. The three displays were separated by dark intervals of five seconds and two seconds, respectively. After the second noisy target was displayed, a 14-second response period was allowed and then the cycle was repeated.

Another feature of performance on this task was its incorporation into a group performance effort. Depression of the subject's control also activated an indicator at the commander's station giving him knowledge on which to base a command decision. The decision made by the commander was reflected in the display of each crewmember. The commander alone received information as to the correctness of his decision. Noise permutations were essentially random. They could be varied in level.

CODE-LOCK SOLVING

A group performance task without individual performance features was developed. It required the crew to discover the proper sequential order for depressing five push-buttons -- one for each crewmember.

Three jewel indicator lights (red, amber, and green) and a push-button were placed on each subject's panel for the performance of this task. Illumination of the red light was the signal that an unsolved problem was present. The amber light, which was illuminated during the time in which any subject kept his push-button depressed, served as an indication to all subjects that some member of the crew was responding. The problem would be solved only when each of the five subjects had depressed his push-button at the point in the sequential order that was correct for a given problem. The red light was extinguished until an incorrect response was made, at which time it was reilluminated (and the programming apparatus automatically reset the problem to the beginning of the sequence). Once the correct first subject had been identified by the group, that subject had to remember to push his button any time the red light came on again (i.e., whenever an error had been made) in order to permit the search for the correct second subject to proceed. When the correct second subject was found, the first and second subjects had to remember their positions while the search continued for the third correct subject, etc.

When all five crewmembers had responded in the correct order, the green light was illuminated to signal that the problem had been solved. After a 30-second pause the same sequence was required a second time as a test of short term memory. Following a between-problem pause of 30 seconds, the programmer moved to a new problem, the green light went off, the red light came on, and the crew was presented with a new sequence or "code" to solve.

The code-lock task is the only task in the battery that is paced by the crew (self-paced); all other tasks are forced-paced at specified rates. Each code-lock problem is scored in terms of the time required for solution, the total number of responses made, and the number of errors (or programmer-resetting responses) made. In addition, five information measures have been derived and analyzed for this task.

RELIABILITY DATA AND BATTERY MODIFICATION

These tasks all have been found to show rather high 24-hour test-retest reliabilities as exhibited in Table 1. The first four tasks, probability monitoring, warning-lights monitoring, auditory vigilance, and arithmetic computation have remained substantially the same since initial development.

TABLE 1

Reliabilities of the Various Measures of Performance

Task	Measure of Performance	Reliability Coefficients + Split-Half	Predicted
Arithmetic Computation	Percentage Correct Solutions	.957**	.978**
Auditory Vigilance	Percentage Correct Signal Detections	.777**	.874**
Warning-Lights Monitoring	Response Latency to Red Lights	.802**	.890**
	Response Latency to Green Lights	.702**	.825**
Probability Monitoring	Mean Detection Time	.746**	.854**
	Percentage Correct Signal Detect.	.451*	.621**
Target Identification	Percentage Correct Individual Responses (Pattern Perception)	.751**	.858**
Code-Lock Solving	Mean time per Individual Response	.959**	.979**
	Proportion Erroneous Responses	.458	.629*
	Rate of Information Transmission		
	Per Performance Period	.963**	.981**
	While Responding	.961**	.960**
	Relative Information-Transmission Rate		
	Per Performance Period	.964**	.982**
	While Responding	.960**	.980**
	Response Equivocation Rate		
	While Responding	.969**	.984**

+ Each reliability coefficient is based on an N of 25, except those for Code-Lock Solving where N = 15. The predicted reliability coefficients were obtained by application of the Spearman-Brown Prophecy Formula to the appropriate split-half coefficients.

*P less than .05;**P less than .001; all levels of significance are based on use of Fisher's z' transformation.

In the initial tryouts all of these tasks save auditory vigilance demonstrated high reliability. The auditory vigilance task as originally presented involved an on-period of 0.12 seconds and an off-period of 0.44 seconds. The signal to be detected was an off-period of 0.55 seconds. Only four signals were presented for detection per hour and a low reliability in the initial tryout led to the suspicion that too few signals were given and perhaps the interval to be detected under work load stress was too short to allow differential detection. Modification of the task to its present form gives a sufficiently high reliability.

TASKS IN ORIGINAL PERFORMANCE BATTERY NO LONGER RETAINED

The original task battery contained three tasks which were dropped and replaced by only two tasks - target identification and code-lock solving. The compensatory tracking task was dropped because it showed low test-retest reliability over days and because the instrumentation of the task also proved high unreliable. The scale position monitoring task, which showed satisfactorily high reliability, was dropped on two counts; first, it was felt that monitoring activity was perhaps oversampled in the battery and secondly, the maintenance problems and equipment malfunction possibilities were judged to be rather great. The third task called "pattern perception", persists in greatly modified form in the present target identification task.

PATTERN DISCRIMINATION

Perceptual functions of a relatively complex variety are likely to deteriorate under conditions involving fatigue. This task was developed to assess the function of discrimination of relatively complex dot patterns and from an individual performance point of view, the same function as served by the present target identification task. The general level of organization of perceptual processes, as well as the level of awareness of an observer, may be reflected in this type of task. Pattern perception is necessary in operational situations which involve observation of successive radar scope patterns and responses to critical changes in the patterns.

A 6 by 6 matrix of lights, mounted in a four-inch, square array, displays the patterns, and the display is mounted in the lower left corner of the instrument panel. Any combination of lights from one to thirty-six can be presented simultaneously. The patterns are presented in pairs and the subject is required to indicate 'same' or 'different' for each pair of patterns presented.

COMPENSATORY TRACKING

Tracking has been used as a task in a great variety of studies of perceptual-motor performance. It is a task which is sensitive to changes in many conditions, such as control-display relations, environmental factors, and operator variables. It has been used extensively as a selection device, and appears eminently suitable for detecting performance effects of stress and fatigue over a wide range of conditions and work periods. Fatigue effects are most clearly shown in increased variability in performance over extended time periods. Compensatory tracking represents a kind of activity required for control of an aircraft under instrument flight conditions. Successful performance requires fairly precise perceptual-motor coordination.

The display chosen consisted of a standard cross-pointer indicator presentation housed in a standard circular instrument case approximately three inches in diameter, mounted in the upper center of the instrument panel. Movements of the horizontal and vertical pointers are programmed by separate random-signal generators. The task of the subject is to correct deviations, maintaining alignment of the pointers with the reference work.

SCALE POSITION MONITORING

Vigilance for the detection of out-of-tolerance or other non-normal indications on instruments was found by Mackworth to deteriorate over time. He found that the frequency of detection of 'jumps' of a modified clock-pointer decreased from approximately 8t percent to approximately 70 percent over a 4.5 hour watch period. Broadbent, on the other hand, employing a 90-minute watch period, observed that little or no decrement of performance occurred in watch-keeping involving up to 20 dials, some of which show extreme indications at infrequent intervals.

A task was designed to sample the kind of vigilance required for the detection of out-of-tolerance indications on instruments such as the airspeed indicator, altimeter, tachometer, and directional gyro. Such instruments are normally maintained within planned tolerances, and corrections are needed only at relatively infrequent intervals during a long cruise.

Four vertical moving-tape scales were located in the upper portion of the panel, two on either side of the compensatory tracking display. The scales were displayed in an open-window framework, three-quarters of an inch wide and two and one-quarter inches high. They resemble scales recently developed for airspeed and altitude registering instruments. A Gaussian distribution of settings presents a random sequence of movements and positions.

PERFORMANCE SCHEDULE

Since it is reasonable to suppose that missions will involve periods of high- and low-performance demand as well as periods of intermediate demand, the original task battery was evaluated with tasks presented in various combinations in order to determine the inter-correlations between performance on the various tasks. A basic 2-hour task program eventuated which gives flexibility in building work-rest cycle evaluations in 2-hour units. It would be feasible to modify such a schedule to provide 1-hour units if desirable. The 2-hour program, as presently constituted, is shown in Table 2.

TABLE 2

Basic two-hour performance task program

TASK	MINUTES								
	000	015	030	045	060	075	090	105	120
Warning Lights	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
Auditory Vigilance	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
Probability Monitoring	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
Code-Lock Solving			xxxxx	xxxxx	xxxxx	xxxxx	xxxxx		
Arithmetic Computation		xxxxx	xxxxx	xxxxx					
Target Identification						xxxxx	xxxxx	xxxxx	

PHYSIOLOGICAL MEASURES

Instrumentation was provided for obtaining such measures as skin conductance, skin temperature, heart rate and respiration rate. A biochemical laboratory was also constructed and equipment provided for the determination of biochemical indices. Utilization of these measures has been limited largely as a function of their cost.

ARITHMETIC COMPUTATION

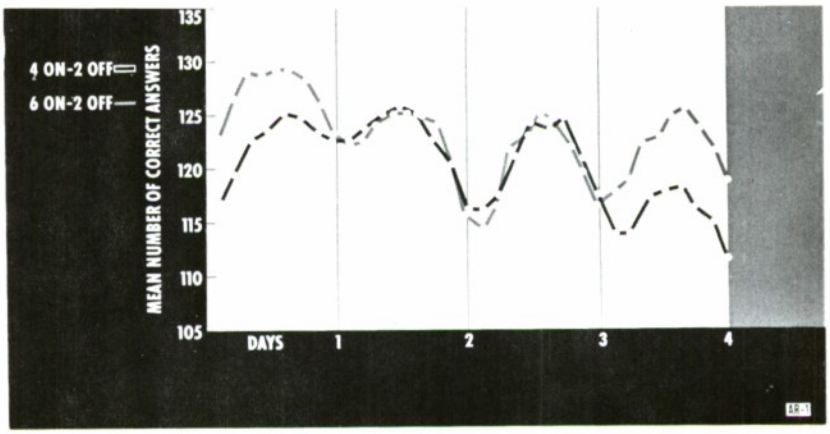

Figure 3. Comparison of the 4-2 and 6-2 schedule groups in terms of trends in mean performance levels for arithmetic computation.

HEART RATE LEVEL

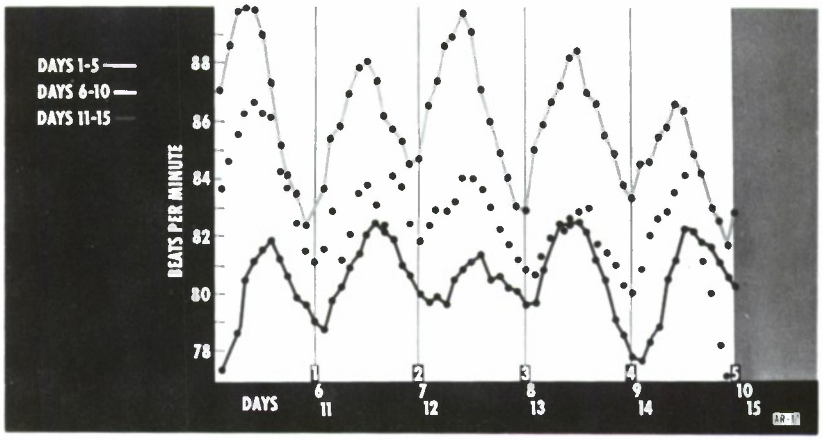

Figure 4. Mean level and mean fluctuation for heart rate over a 15-day period.

ARITHMETIC COMPUTATION

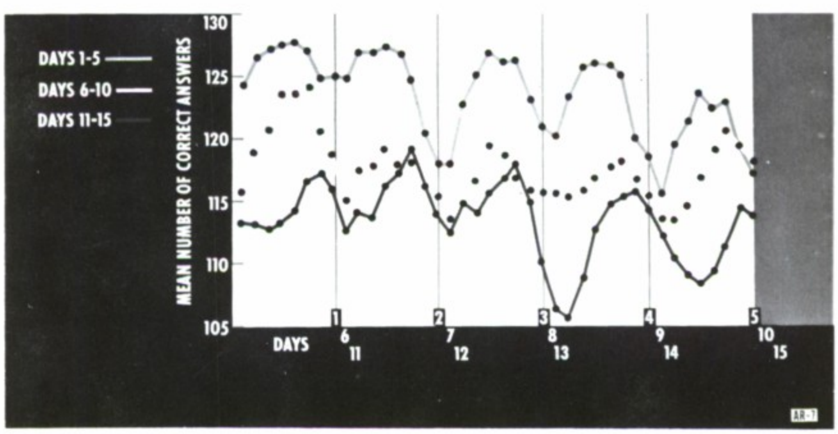

Figure 5. Mean levels achieved in arithmetic computation during 8 presentations of the 2-hour task program per subject per day.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Lockheed-Georgia Company, Marietta, Georgia and Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP N/A

3. REPORT TITLE

USE OF THE EXPERIMENTAL METHOD FOR EVALUATIONS OF PERFORMANCE IN MULTI-MAN SYSTEMS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (Last name, first name, initial)

Passey, George E., PhD
Alluisi, Earl A., PhD
Chiles, W. Dean, PhD

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1966	26	

8a. CONTRACT OR GRANT NO. Prepared in part under AF 33(657)-10506 b. PROJECT NO. 1710 c. Task No. 171002 d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AMRL-TR-66-121

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
Originally published as AMRL Memorandum P-67, February 1964.	Aerospace Medical Research Laboratories, Aerospace Medical Div., Air Force Systems Command, Wright-Patterson AFB, Ohio

13. ABSTRACT

This report discusses the use of the experimental method as a technique for arriving at solutions to human factors engineering problems encountered in the design of multi-man systems. Of specific concern are the methodological decisions that must be made in the design of the research. Factors that are likely to influence these decisions are considered as well as the implications of these decisions with respect to the validity and generality of the data thus obtained. These various decision points are illustrated through use of data on group performance during long-term confinement.

DD FORM 1 JAN 64 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Multi-man systems Human group performance Performance assessment Methodology Human factors engineering Long-term confinement						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
____________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
____________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
____________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.