MIT

FLIGHT TRANSPORTATION LABORATORY
REPORT R 79-4

TOWARD THE DEFINITION AND MEASUREMENT OF THE MENTAL WORKLOAD OF TRANSPORT PILOTS

T.B. Sheridan

and

R.W. Simpson

January 1979

DEPARTMENT OF AERONAUTICS & ASTRONAUTICS

FLIGHT TRANSPORTATION LABORATORY
Cambridge, Mass. 02139

TOWARD THE DEFINITION AND MEASUREMENT OF THE MENTAL WORKLOAD OF TRANSPORT PILOTS

T.B. Sheridan
Man-Machine Laboratory

and

R.W. Simpson
Flight Transportation Laboratory

Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

FINAL REPORT

January, 1979

Program of University Research
Department of Transportation

Contract DOT-OS-70055

EXECUTIVE SUMMARY

This report describes work performed in the first year of a continuing research project aimed at developing useful methods for measuring the workload of pilots operating aircraft in the ATC system. Good methods of measuring mental workload of human operators are needed to evaluate the introduction of new technology and new procedures in the man-machine environment. The present research is concentrating on developing subjective assessment methods for any phase of an IFR (Instrument Flight Rules) flight and any crew station on the flight deck.

One of the results achieved in the first year is an expanded conceptual structure which allows a more precise definition of terms and assumptions in defining pilot mental workload in a multi-task environment. A second area of research has concentrated on reviewing the alternative approaches to developing a measurement scheme for workload, with some emphasis on the subjective assessment approach. A tentative result in this area is the generation of a prototype subjective rating method for IFR pilot workload modeled closely on the Cooper-Harper rating developed in 1969 to evaluate aircraft handling qualities. This scheme and others will be tested in a transport aircraft simulation during the coming year. If successful, it will be used in a variety of cockpit simulators at NASA research centers (Ames and Langley) and FAA NAFEC as part of a joint research program to evaluate cockpit display of traffic information.

TABLE OF CONTENTS

1. INTRODUCTION

In the last several years the subject of mental workload has received renewed attention by scientists and engineers.[1,2] A NATO/AGARD Specialists Aerospace Medical Panel on Pilot Workload convened in the spring of 1977 in Cologne, Germany,[3] followed by a NATO Scientific Affairs Division Workshop on Mental Workload in Mati, Greece that summer.[4] A Joint Services Committee on Workload has had several meetings. In July of 1978 the U.S. Air Line Pilots Association convened a meeting in Washington, D.C. on pilot workload, and in September 1978 a National Science Foundation teleconference cum electronic journal on mental workload began involving a transatlantic hookup between the U.S. and Europe.[5] Why all the interest in such a topic?

Increasing air traffic over metropolitan areas, more stringent fuel constraints and stricter noise regulations have militated in the direction of increased pilot workload. While increased automation and larger, more sophisticated aircraft have definitely reduced the workload of direct manual control, the airline pilot has increasingly become a "flight manager," with a wide range of monitoring responsibilities combined with expectations that he be able to take over manual control suddenly should the situation demand it. These same factors have changed the demands on the ATC controller.

A scientific basis for definition and measurement of mental workload is sought -- to help decide how many air crew are necessary for sufficient safety, to help decide on relative roles for pilot and ground controller, to help decide on work-rest schedules for both air and ground crews, and a host of cockpit design and procedure questions.

The subject of mental workload has similarly arisen in various other settings such as nuclear and chemical plants and factories where human operators are controlling expensive machines and where failure is not only extremely costly in dollars, but also in lives. In these other settings as well, automation is changing the operator's role from that of in-the-loop manual controller to that of machine supervisor, fault diagnostician, and emergency back-up.

Interest in human operator workload is not new. Workload was an active research area during the early days of "scientific management" in the factory, circa 1930, and led to various motion-time-methods and piece-work standards. But in those days skills seemed to be more easily defined and measured on a physical basis -- visual and manipulative events. As the tasks of interest have become less routine, and especially more recently as the computer has become a cooperative party in decision-making and control, the appropriate basis for defining workload has become more elusive. This is particularly true when we are talking about large transient workload demands which occur, for example, as a rather dull autopilot monitoring situation suddenly becomes a frantic effort to take over control of an unstable aircraft or to avert a midair collision.

2. THE DEFINITION OF PILOT MENTAL WORKLOAD

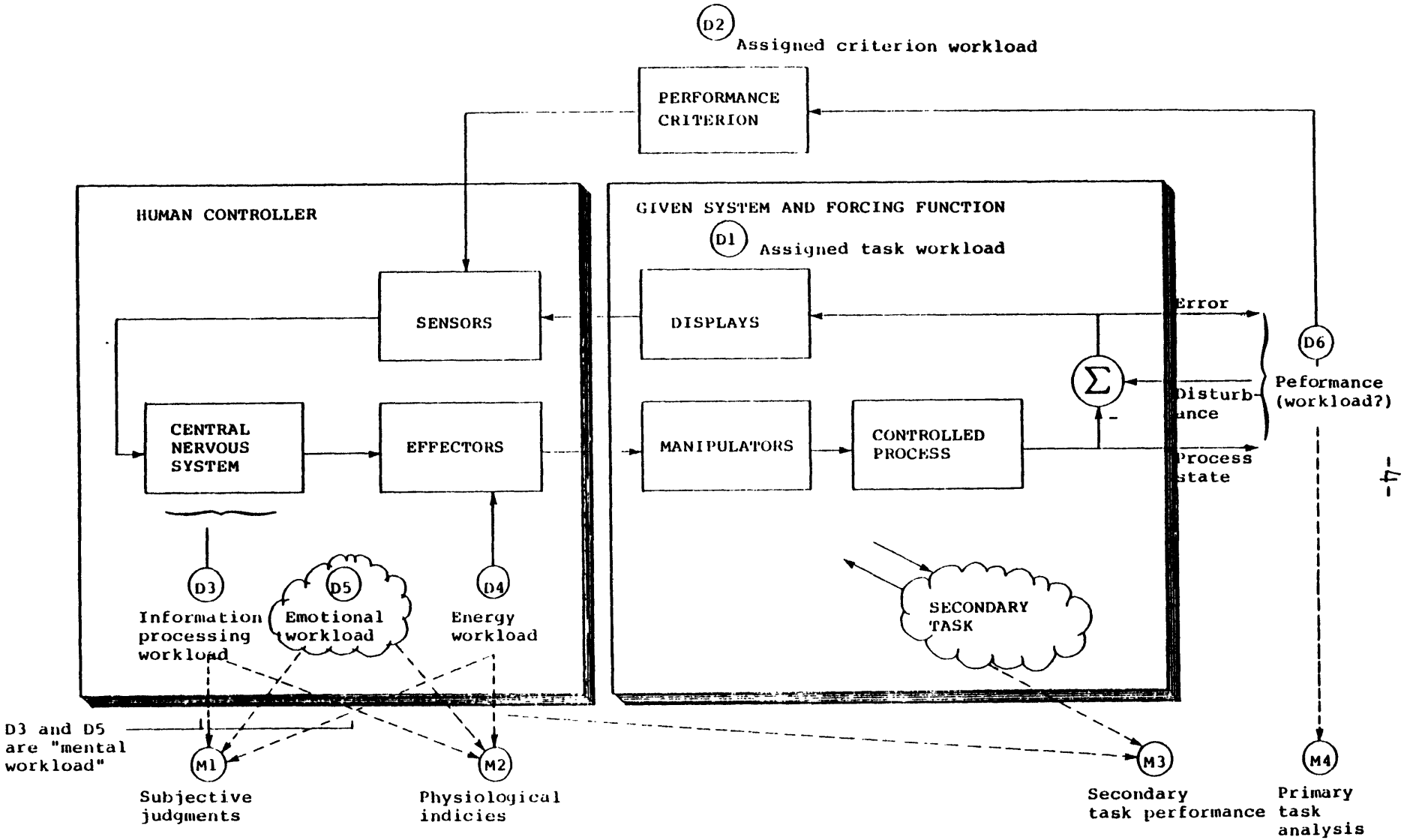
2.1 WHAT IS MENTAL WORKLOAD?

We are not particularly concerned here with physical work, which is easily measured in terms of changes in respiratory gases from O_2 to CO_2, heart rate, and other means. Clearly we are concerned with information processing and decision making -- events in the head (though the senses and the muscles are involved as input and output transducers).

There are other words closely associated with "mental workload," such as "information processing," "thinking," "attention," "stress," "emotion" and "fatigue." Are these all the same, or are they distinctly types of mental events which contribute to or relate to mental workload in different ways? Figure 1 illustrates in the context of a non-machine control loop the "loci" of some alternative definitions of mental workload, labeled D1-D6).

Probably a first assertion which must be made is that, in terms of experimental science, mental events, including mental workload, are "intervening variables," i.e., they are not directly measurable, even by electrodes in the skull. Mental events intervene between measurable stimuli and measurable responses. There is no way, for example, to verify that what one person sees as "red" is what another sees as "red." Mental workload can only be inferred, not directly measured.

One can evade this measurement problem by defining mental workload to be some measurable quantity. For example, in Figure 1 task demand, D1, can be defined for a given scenario in terms of task elements, their nominal time duration, and their schedule, i.e. a normative, detailed description of what has to be done. This is called "task timeline analysis" and is practiced by the manufacturers of transport aircraft.



(From Sheridan and Stassen Paper, reference 4)

FIGURE 1

Alternative Definitions of Workload and Performance Illustrated in Control Paradigm

Alternatively, task performance, D6, can be defined as workload in terms of accuracy, timeliness, etc., and compared to an established task criterion, D2, for performance. These quantities can often be measured, but the trouble with such definitions of mental workload is that they are nothing more than nominal activity or performance, or perhaps relative performance.

Clearly, when we speak of mental workload we normally mean something to do with a sense of mental effort, how hard one feels one is working, and it is obvious that one person can claim to have a feeling of great mental effort, while another can claim to be exerting no mental effort at all -- and both be performing at the same level. Thus, as the term is usually used, mental workload is not performance and it is not task demand. Mental workload would seem to be some combination of mental effort, information processing, D3, and emotion, D5, in response to task demand. We would expect that different persons will have different responses to the same task demands and task performance.

Thus, the measurement problem has caused a definitional problem. It is clear from reviewing the extensive literature by human factors researchers that there is some "semantic confusion" about workload. A conceptual structure for workload ideas is clearly necessary before we begin to devise experimental methods and new measures for mental workload appropriate for various pilot activities. We badly need a general theory, or conceptual structure, or common language to organize researchers in this area and to show the relationships between their individual efforts and ideas. This seems to be recognized by human factors researchers themselves and indeed was the central purpose of the workshop at Mati, Greece in September, 1977.[4]

2.2 A DESCRIPTION OF THE CURRENT TASK ENVIRONMENT FOR A TRANSPORT PILOT

Considerable effort is undertaken by the designers of the flight decks of transport aircraft in describing the nominal taskloads of crew members in all phases of a nominal IFR flight. In current task timeline analysis, pilot activities are described in what might be called macroscopic (or even microscopic) detail. For example, the simple actuation of a wing flap lever is broken down into verbal command and acknowledgment, grasp, actuate, hold, monitor flap extension, and report of actual flap position (see reference 11). For evaluating pilot mental workload, the information processing requirements for various macroscopic tasks (such as change of aircraft attitude, change of heading, etc.) can be determined in microscopic detail by looking at the information needed on bank angle, heading, heading rate, and how it must be processed during the macroscopic task for a given mode of the aircraft's flight control system. Perhaps this would be a promising direction for further research into task loading, and might lead to a methodology for establishing ideal task effort for various pilot tasks. Here we shall simply provide an overview of the various kinds of tasks performed by pilots in current ATC operations. We shall attempt to classify them in two different ways and to provide a small set of task characteristics.

2.2.1 A FUNCTIONAL CLASSIFICATION OF PILOT TASKS IN IFR FLIGHT

We may group the various tasks performed by a crew of an ATC aircraft into four main classes in decreasing order of significance:

1. Communication and Traffic Control
2. Navigation
3. Guidance (or Piloting or Steering)
4. Aircraft Systems Monitoring and Management

In the first class of tasks, one member of the crew usually acts as communicator by monitoring all radio messages, receiving and responding to messages for his aircraft, selecting the desired radio frequencies, and maintaining an awareness of the current ATC situation relative to his aircraft. These messages occur randomly, but to some degree can be expected in routine operations. They require immediate response and usually cannot be deferred. There may be garble or interference in obtaining the information from a given message. The response may require some information processing and decision making and internal communication amongst the crew. If future ATC systems use an automatic data link, the workload of the communicator will be quite different from today where a broadcast voice link is used.

The set of tasks in the Navigation class consist of operating the aircraft navigation systems to establish current aircraft position, and to make decisions relative to the nominal desired track of the aircraft in three (or four) dimensions. Today, there are a number of navigation charts, maps, etc., which accompany the use of navigation systems such as VOR, DME, Omega, Loran, Inertial, RNAV, etc. For radio navigation, there is the task of selecting and identifying ground stations.

Once the desired track and current position relative to it are established, the continuous task of aircraft guidance can be performed. We define this as piloting or steering the aircraft such as to follow the desired track with some

degree of precision. This may be performed manually, semi-automatically by controlling or instructing the autopilot, or completely automatically by inserting a detailed description of the desired track (and altitude) into an advanced flight control system. Notice that it is necessary to monitor the performance of these automated flight systems so that the continuous guidance task for the pilot is converted to intermittent discrete monitoring tasks.

The final set of tasks have been classified as Systems Monitoring and Management. These tasks consist of monitoring and managing all the aircraft subsystems such as fuel, air conditioning and pressurization, electrical, hydraulic, engines, etc. These tasks are usually very routine and deferrable over a short period except that unexpected, rare, emergency conditions can occur which may require instantaneous response. Cockpit checklists describe routine work in this class.

2.2.2 TASK CHARACTERISTICS

We now define some abstract characteristics of tasks which occur in IFR piloting which are important in determining workload. We will assume that we are dealing with discrete task events with a starting time, a duration, and an ending time. The continuous task of manual guidance does not constitute an important part of current taskloads for a transport crew since it occurs only in the takeoff and the final stages of approach and landing. Interference from other task events is minimal at these times. For some pilots in General Aviation, manual guidance may occur in all phases of instrument flight as a background continuous task, but we are still primarily interested in the discrete task events which arise from the ATC, Navigation, and Systems Management classes.

First, for discrete tasks there is the characteristic of "Task Arrival Randomness." Tasks may arrive randomly within a phase of flight, or in some irregular arrival pattern, or may be considered as rigidly scheduled. Even though a task is expected to occur, its timing may be random or unexpected. For example, ATC messages are expected but arrive at varying times within a flight phase. On the other hand, the landing gear is scheduled to be lowered as the outer marker is passed on landing approach.

Secondly, there is the characteristic of "Task Uncertainty." Tasks may be expected or predictable, or unexpected as to their content. For example, ATC messages may contain unexpected commands due to air traffic or weather situations which then cause a set of unexpected tasks to be carried out.

Thirdly, there is a characteristic "Task Priority". Some tasks are "deferrable," i.e. upon arrival they may be delayed while other tasks are executed. Other tasks are "pre-emptive," i.e. upon arrival they interrupt work on current tasks. Interrupted tasks may not be "partitionable" so that if interrupted they must be totally repeated. Non pre-emptive, but high priority tasks may simply go to the head of the line, allowing current tasks to be finished before they are initiated. All of these situations occur due to some degree of randomness and uncertainty which causes tasks to occur simultaneously and a task queue on backlog to be created. Workload during these peak or "busy" periods is high, and requires further supervisory decisionmaking in managing tasks and coordinating the sharing of tasks amongst the crew. The priority assigned to a task will depend upon its importance to the safety of flight.

Finally, there is the characteristic of "Task Sequence". Discrete tasks may have to be performed in a particular order where some tasks have precedence over others whose execution

is contingent upon their completion. For example, a reduction in airspeed in the terminal area may be contingent upon a further extension of wing flaps, or a reduction in altitude below 10,000 feet must follow a speed reduction to 250 knots.

2.2.3 TASK CATEGORIES

These task characteristics determine the occurrence of peak work rates, the size of task backlogs, etc., as a function of expected rates of arrival of tasks. In particular, the simultaneous arrival of two non-deferrable tasks means that the two tasks have to be worked on simultaneously, and the arrival of any non-deferrable task is likely to interrupt deferrable tasks in busy periods. The probabilities of working on two or more tasks simultaneously, of tasks being interrupted, of a backlog of size n of deferrable tasks, etc., are all good indicators of high taskload or workload conditions and increase the probabilities of pilot errors or omissions. Thus, it may be useful to categorize tasks depending upon the degree of their deferability as follows:

Category	Description
A	non-deferrable, or pre-emptive
B	deferrable for a short period (say, 60 seconds)
C	deferrable (for more than 60 seconds)

THe first category describes tasks which are of critical importance to the safety of flight and must be worked on when they arrive. For example, ATC commands must be monitored, received, acknowledged, and responded to as they occur, or changes in flight profile or aircraft configuration which must be made at required times. In general we may call this category "Operating Tasks."

The second category describes tasks which are of critical importance to the safety of flight, but which can be deferred for some short period of time. Examples of these tasks are the monitoring and cross-checking of flight instruments, aircraft systems and engine instruments, the visual scan for other aircraft and the execution of checklists. We can define the short period as one or two minutes, or alternatively, in terms of task durations. In general terms we may call these tasks "Monitoring Tasks".

Finally, we have a category of tasks which are deferrable over longer periods. While they are necessary for the safe conduct of flight, they are not time-critical. Examples in this category are tasks in monitoring and planning of navigation, tasks in preparing and planning for ATC procedures, tasks which obtain flight information (altimeter setting, weather, ATIS), and tasks in monitoring aircraft performance (cruise control, determining landing weights and approach speeds). We may define these tasks as "Planning Tasks."

A subjective rating scheme for pilot workload based upon these categories of tasks is given later in the report. By observing the occurrence of simultaneous operating tasks, interruptions of monitoring and planning tasks, etc., we have a basis for constructing a subjective workload assessment which might be used by pilots or trained observers.

2.3 A CONCEPTUAL FRAMEWORK FOR DEFINING MENTAL WORKLOAD

2.3.1 A QUALITATIVE PARADIGM FOR THEORETICAL ANALYSIS

Before attempting a detailed analysis, let us consider qualitatively the paradigm shown in Figure 2. Discrete tasks arrive on the left side (circles), are selected for processing and when finished, leave on the right side. However, we have not restrained ourselves, as strict behaviorists might, to a single "black-box," stimulus-response paradigm, but instead

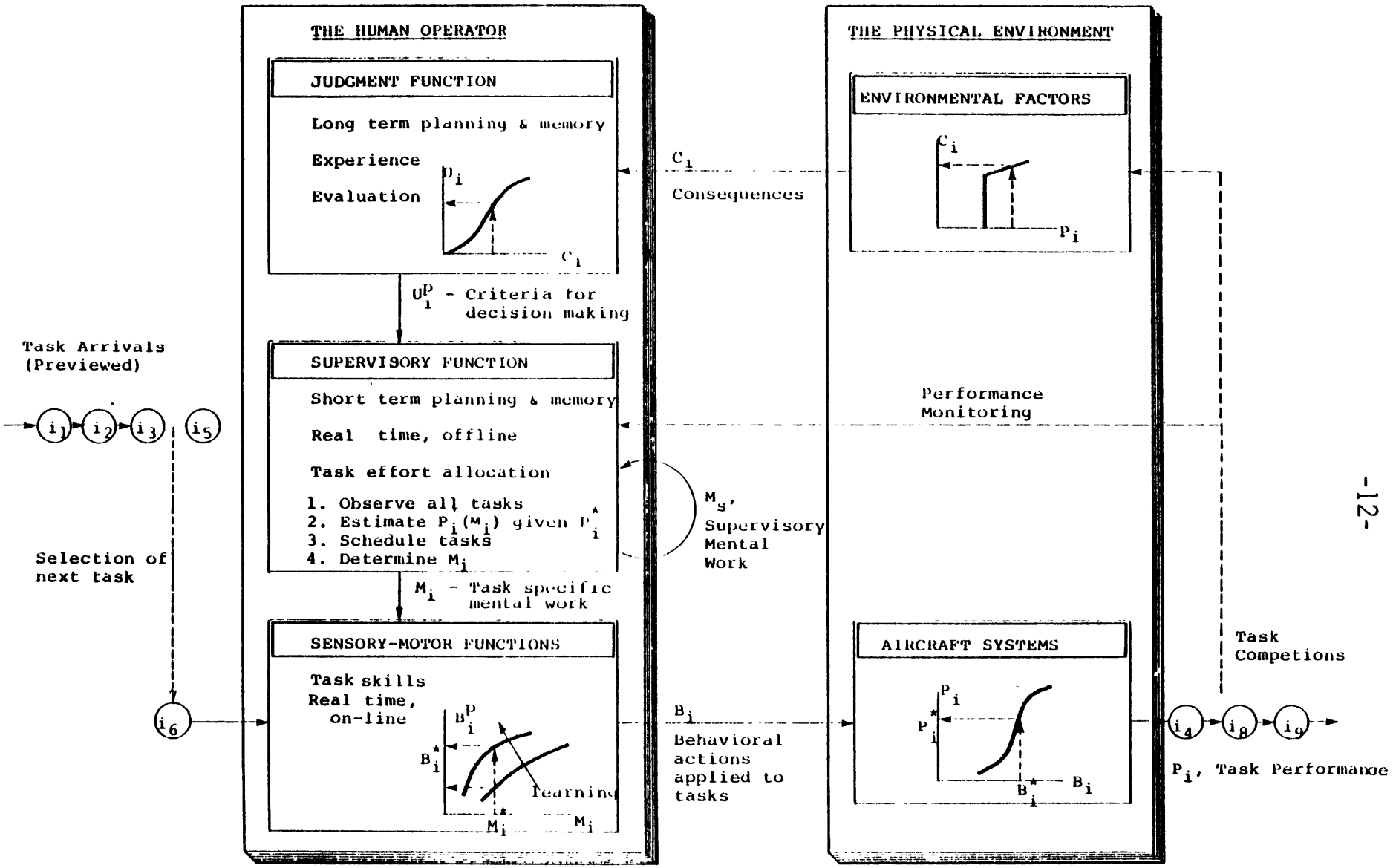


FIGURE 2

Qualitiative Paradigm for Pilot Mental Workload

have postulated three hypothetical levels of information processing for the human operator, which, we believe, accord with common anecdotal experience. As well, we have created two physical blocks, one for the aircraft and another for the physical factors external to an IFR flight.

The three left hand blocks represent information processing functions of the pilot at three levels:

(1) Sensory-Motor Function

At the lowest level are those functions which constitute real-time and on-line controlling (sensory-motor skills), which are learned by the pilot, and which, when learned, do not in themselves involve much conscious thought. There is an observable behavioral activity, B_i, associated with each task i which has some physical effect on the aircraft and its systems. Depending upon the skill level of the pilot, there is some mental work, M_i, associated with B_i.

(2) Supervising Function

At the intermediate level is the supervisory function, i.e., short term planning and decision making to schedule and allocate effort amongst the arriving tasks. From experience, the pilot has an estimate of the tradeoff between the quality of task performance, P_i, and his behavioral and mental work, B_i and M_i. Based on a set of utilities, U_i, for achieving P_i, and a set of evident task parameters (arrival time, duration, deadline), he must allocate a level of mental work M_i consistent with his mental resources. Notice that $M_i = 0$ corresponds to a decision not to select task i to be worked on at the present time. Solving this task allocation problem imposes "real time," non-task-specific, mental work. We shall call this "supervisory" mental work, M_S.

(3) Judgmental Function

At the highest level, the pilot has a set of longer term planning functions, and the job of retaining in long term memory experiential data concerning task performance, mental effort expended, and the consequences of performance levels. This experiental data is used in evaluating the consequences, C_i, of task performance so that a set of values called "utilities," U_i, are established for performance of current and future tasks. These U_i are then used by the Supervisory function in allocating mental effort as explained above.

Now, the two right-hand blocks of Figure 2 are processes or physical activities external to the pilot:

(4) Aircraft Systems

These represent the aircraft and its display and control systems. The pilot inputs of behavioral activity B_i result in a system performance, P_i. We assume a monotonic relationship where an increasing input of B_i will cause higher quality task performance, P_i.

(5) Environmental Factors

These represent factors external to the pilot and aircraft such as weather turbulence, air traffic control, other aircraft, etc. Given these external factors, there is a consequence C_i of an aircraft system performance, P_i.

There are two decision-control loops in our paradigm. A larger, judgment or experience loop which is off-line, and feeds back P_i, and $C_i(P_i)$ to establish U_i in the judgmental function. The smaller, intermediate loop monitors the

performance P_i in real time to affect the decisions made by supervisory control in allocating mental effort to obtain desired performance.

We would wish that all the blocks of Figure 2 were stationary. However, we know they all change (learning affects motor skill at the lowest level, learning and experience affects short term supervisory capability at the intermediate level, and experience causes improved judgment at the highest level). Similarly, the Aircraft Systems and Environmental Factor blocks can change over time and/or circumstances.

We note that the triad of human operator blocks is only our way of describing the situation, i.e., a model. There is no way to measure explicitly the signals which flow between these blocks. We have created M_i and M_s as components of mental work. They are "intervening" variables which we can only infer, not measure. The behavioral activities, B_i, can be measured experimentally, although there is little agreement on standard methods.

2.3.2 A FRAMEWORK OF DEFINITIONS AND ASSUMPTIONS

We now extend our paradigm with some definitions and assumptions, using more precise semantics and symbology. We should emphasize the hypothetical nature of our model and its needs for validation.

(1) Description of Tasks

We assume that discrete tasks arrive with some degree of uncertainty or randomness within a given phase of flight. Thus, they have a task arrival time, t_{oi}, a task duration,

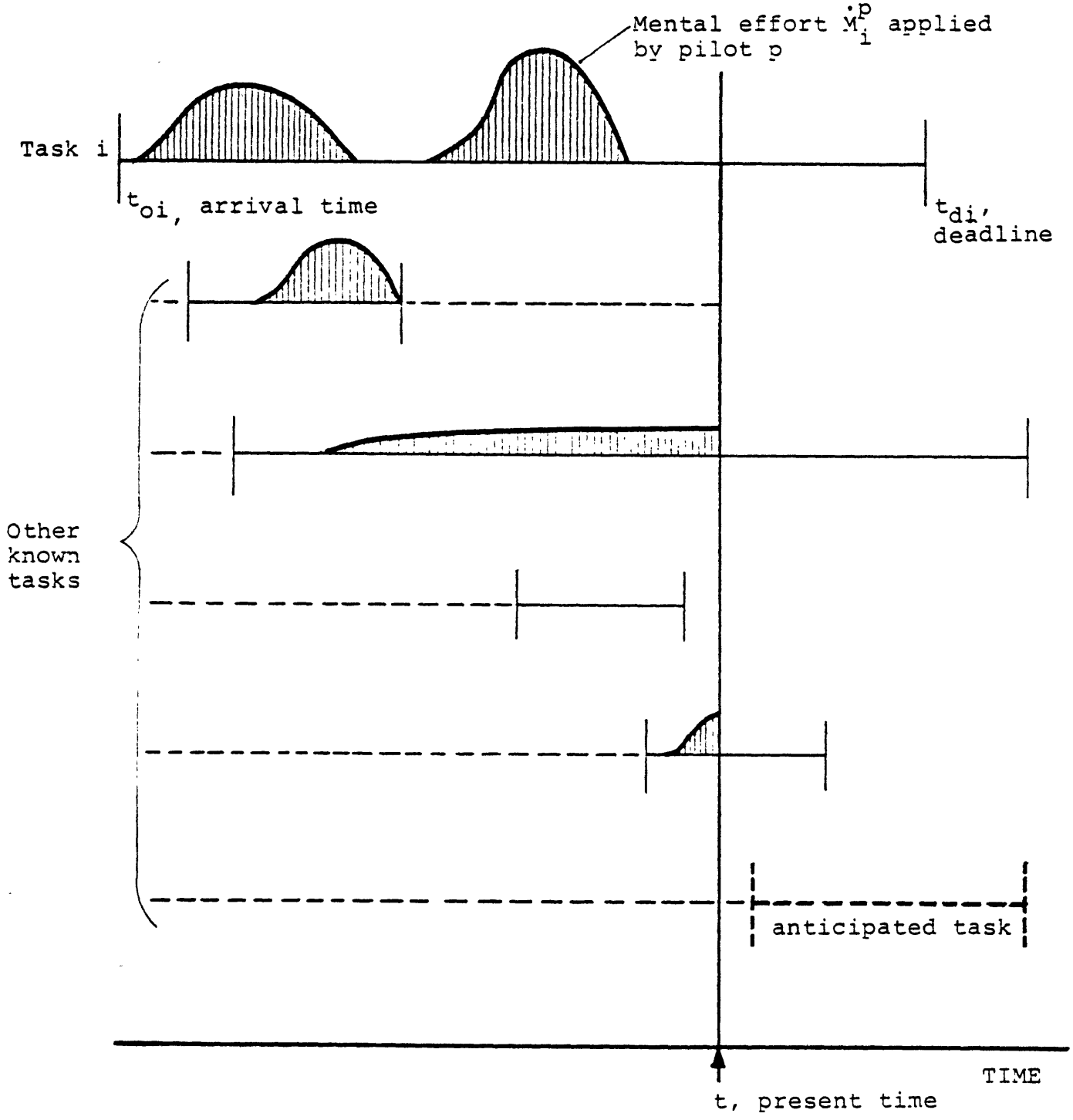


FIGURE 3

Task arrivals, deadlines and mental effort allocation functions

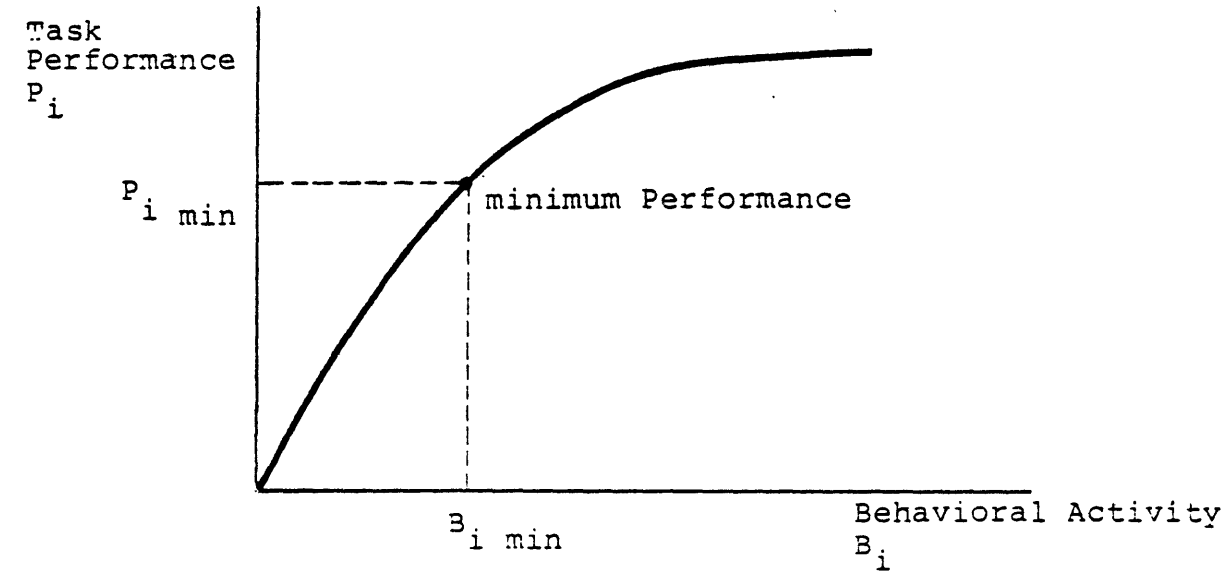
and a task completion time. Also, we assume that a task has a deadline time, t_{di}, after which there is no benefit from working on it. Figure 3 displays this multiple task situation in a format similar to that used by Tulga.[8] On each task line, the rate of expenditure of mental effort, $\dot{M}_i^p$ (defined later), can be plotted versus time, so that the area under the curve represents mental work applied to that task. Note that we assume that more than one task can be worked on at any given time.

(2) Task Performance

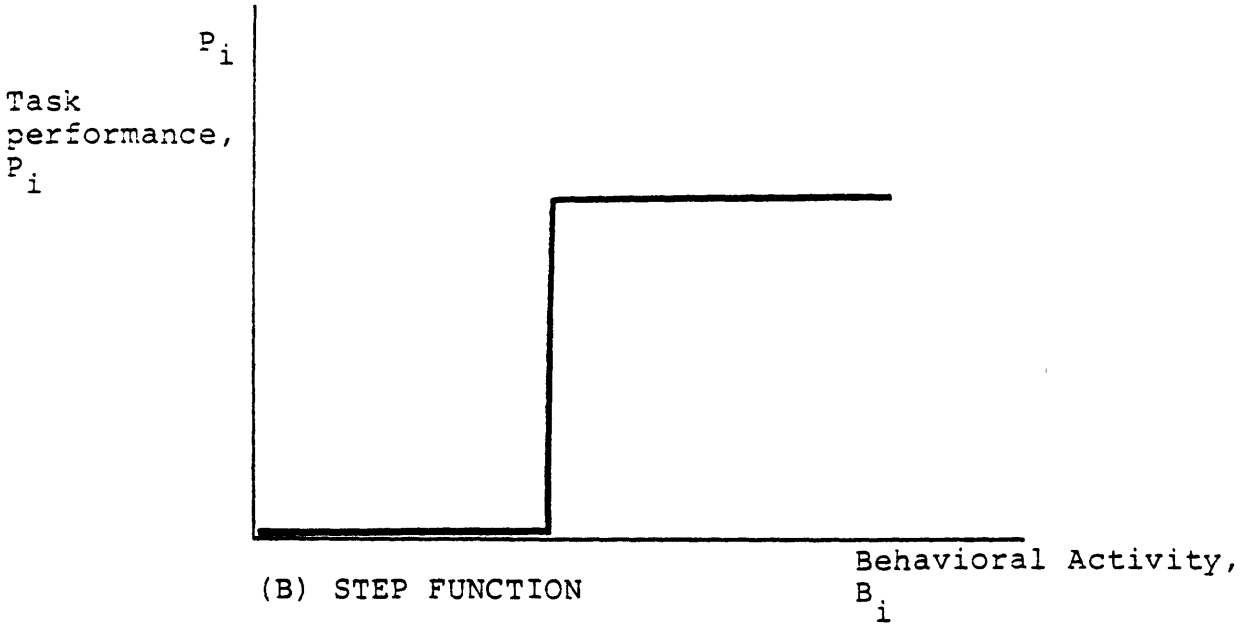
In general, we assume that a single valued task performance function, $P_i(B_i)$, exists such that an increasing expenditure of behavioral activity will result in an increasing quality of task performance (see Figure 4a). The task must be completely defined in terms of control system, display, air turbulence, etc. This assumes the task has some quality dimension such as accuracy, timeliness, etc. Note that some minimum requirement of task performance could be specified which implies a minimum requirement of behavioral input. Also for certain other tasks, the task is either "done" or "not done" and may require a fixed amount of input as shown in Figure 4b. This is commonly assumed in task timeline analysis.

(3) Value of Task Performance -- Utility

In our paradigm, we assume that there is a functional relationship between task performance and the consequences of task performance, $C_i(P_i)$. For each pilot, p, there is a value placed on these consequences, called a utility, $U_i^p(C_i)$. This value is developed by experience and judgment. We assume



(A) GENERAL FUNCTION



(B) STEP FUNCTION

FIGURE 4

Task performance functions, $P_i(B_i)$

that a single valued transformation between U_i and P_i exists, $U_i(P_i)$, as a result of these two functional relationships. Thus, the pilot can place a value on increasing task performance under given environmental circumstances.

(4) Mental Work, M_i, Associated with Task Behavioral Activity, B_i

We assume that there is a monotonic relationship between the amount of work expended by a given pilot, p, and a given amount of behavioral activity. We denote this as $B_i^p(M_i)$. This relationship is a function of learning and practice, where a "higher skill level" can be achieved by the pilot such that he can produce the same behavioral output, with lesser amounts of mental input, as indicated in Figure 5.

In Figure 3, we assumed that the mental work associated with task i can be spread out in time. Thus, we define instantaneous functions of time, $\dot{M}_i^p(t)$, $\dot{B}_i^p(t)$, called "mental effort," and "behavioral effort," respectively. By analogy to physical work and effort, the time integral of mental effort is mental work,

$$M_i^p(t) = \int_0^t \dot{M}_i^p(t) \cdot dt$$

(6) Mental Effort Level, Behavioral Effort Level

We have assumed that the pilot may do more than one task simultaneously, where we mean within some short period of time, τ. Whether he "time shares" within this period, or is really working instantaneously on more than one task is not important. We can average the expenditure of mental or behavioral work over the time period τ, and also over the tasks, i. To indicate this averaging process, we shall call

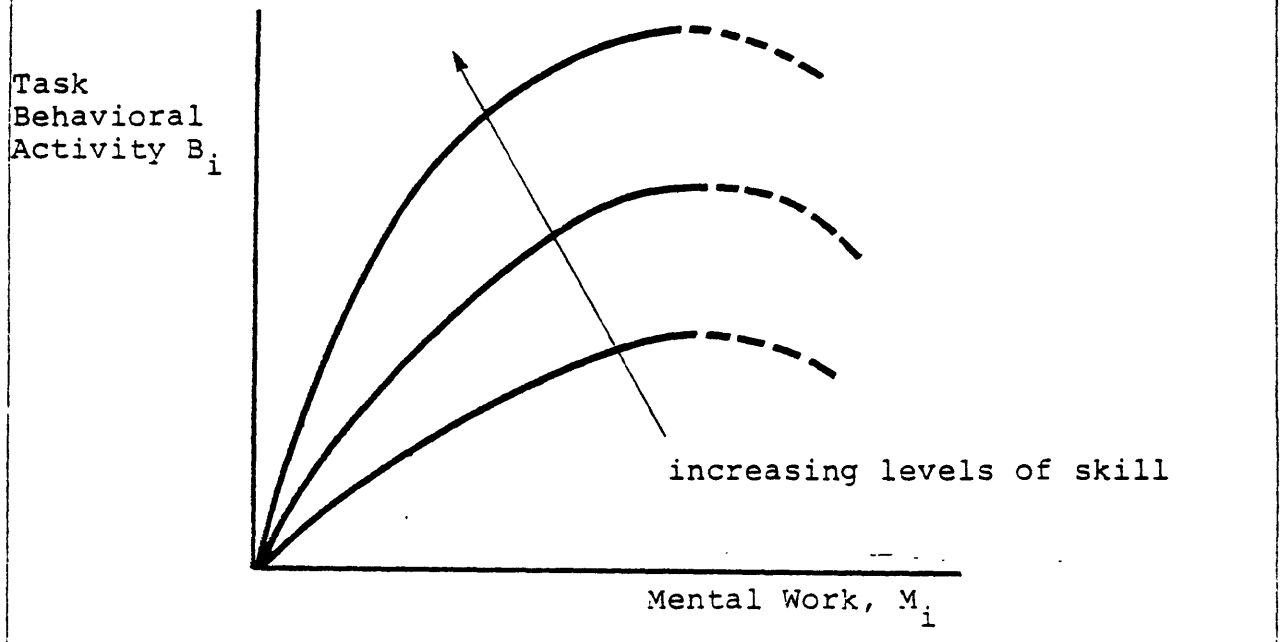


FIGURE 5

Behavioral Activity as a Function of Mental Work

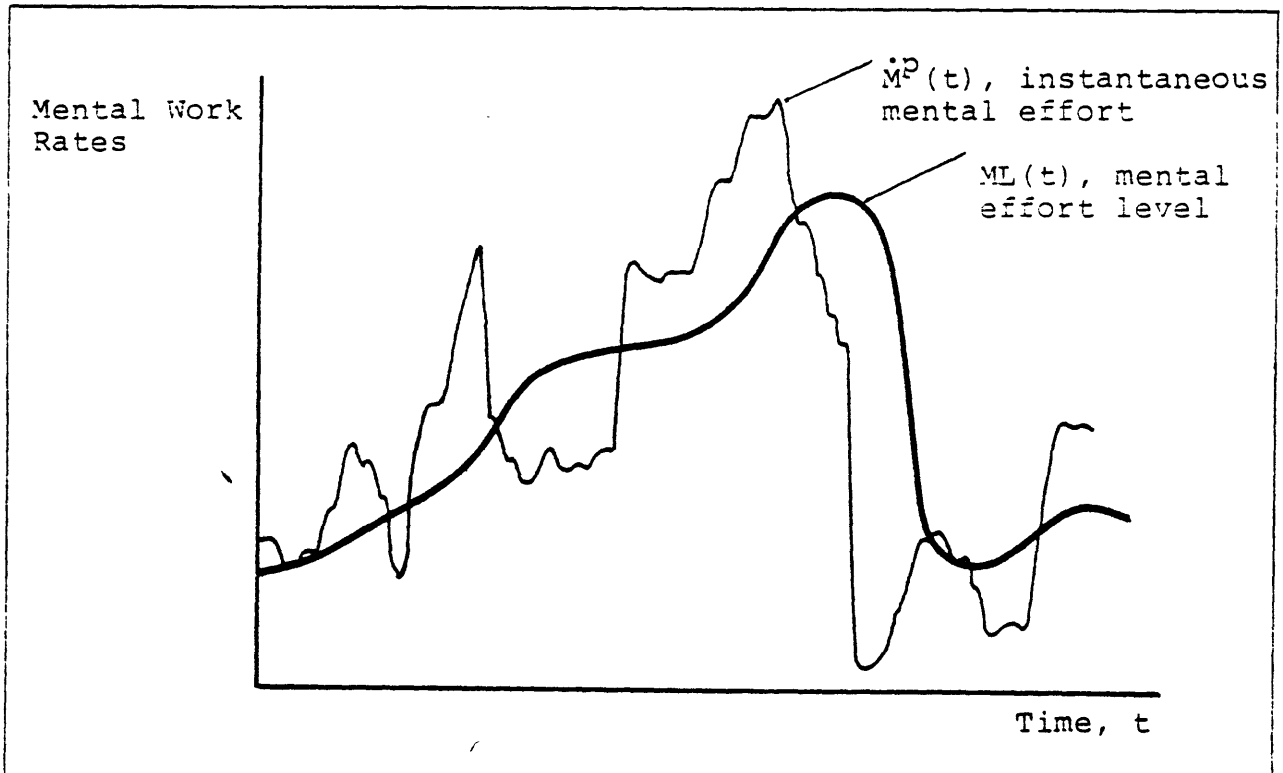


FIGURE 6

Mental Effort Level and Instantaneous Mental Effort

these quantities "effort levels," ML or BL.

Total mental effort level over all tasks and a moving time window of duration τ

$$ML^p(t) = \frac{1}{\tau}\left[\sum_i \int_{t-\tau}^{t} \dot{M}_i^p(t)\cdot dt + \int_{t-\tau}^{t} \dot{M}_s(t)\cdot dt\right]$$

The effects of time averaging are shown in Figure 6. Notice that peak instantaneous work rates will exceed the effort level at times. Also, the value of effort level at time t will depend on the value of τ used in the time averaging process.

2.3.3 CAPACITIES AND RESOURCE TRADE-OFFS IN SKILLED BEHAVIOR

It is assumed above that the level of skilled behavior applied to task i is a simple, monotonically increasing function of mental work expended upon that task. Now we shall elaborate our notions of the $B_i^p(M_i^p)$ relationship and suggest that, in combination with other "simultaneous" tasks, and when particular limiting conditions are reached, skilled performance may deteriorate as workload increases still further.

We propose that the lower level psychomotor skills in the sensory-motor block of the pilot be thought of as a set of j distinguishable but interconnectable elements, such as are illustrated in Figure 7. These may be defined either on an anatomical basis (sensory elements such as vision, hearing, vestibular senses, or motor elements such as left hand, right hand, left foot, head) or on the basis of seemingly separate cognitive behaviors (memory, learning of particular procedure, etc. -- we will not be fussy about specific categories at this point).

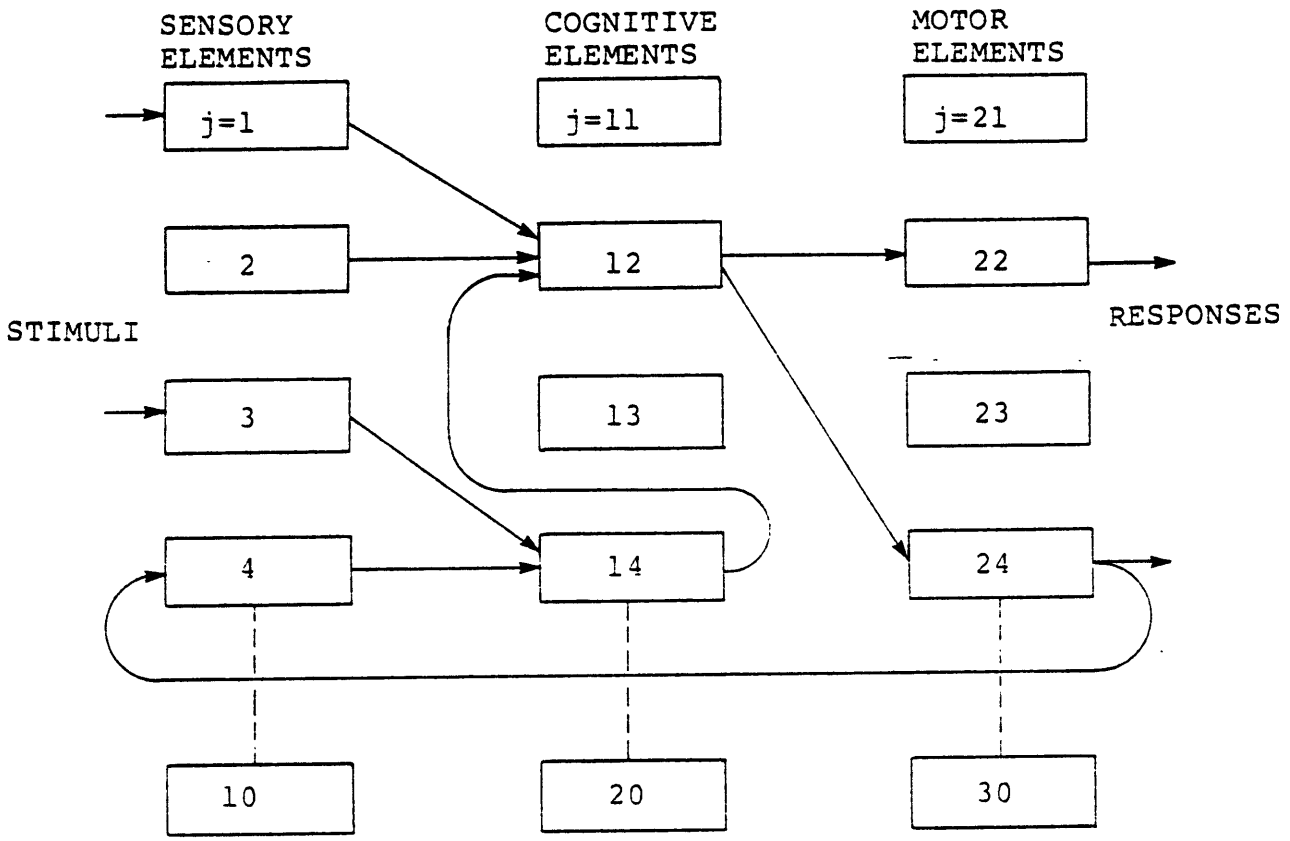


FIGURE 7

Hypothetical Instantaneous Interconnection of Sensory, Cognitive and Motor Elements for a Given Task

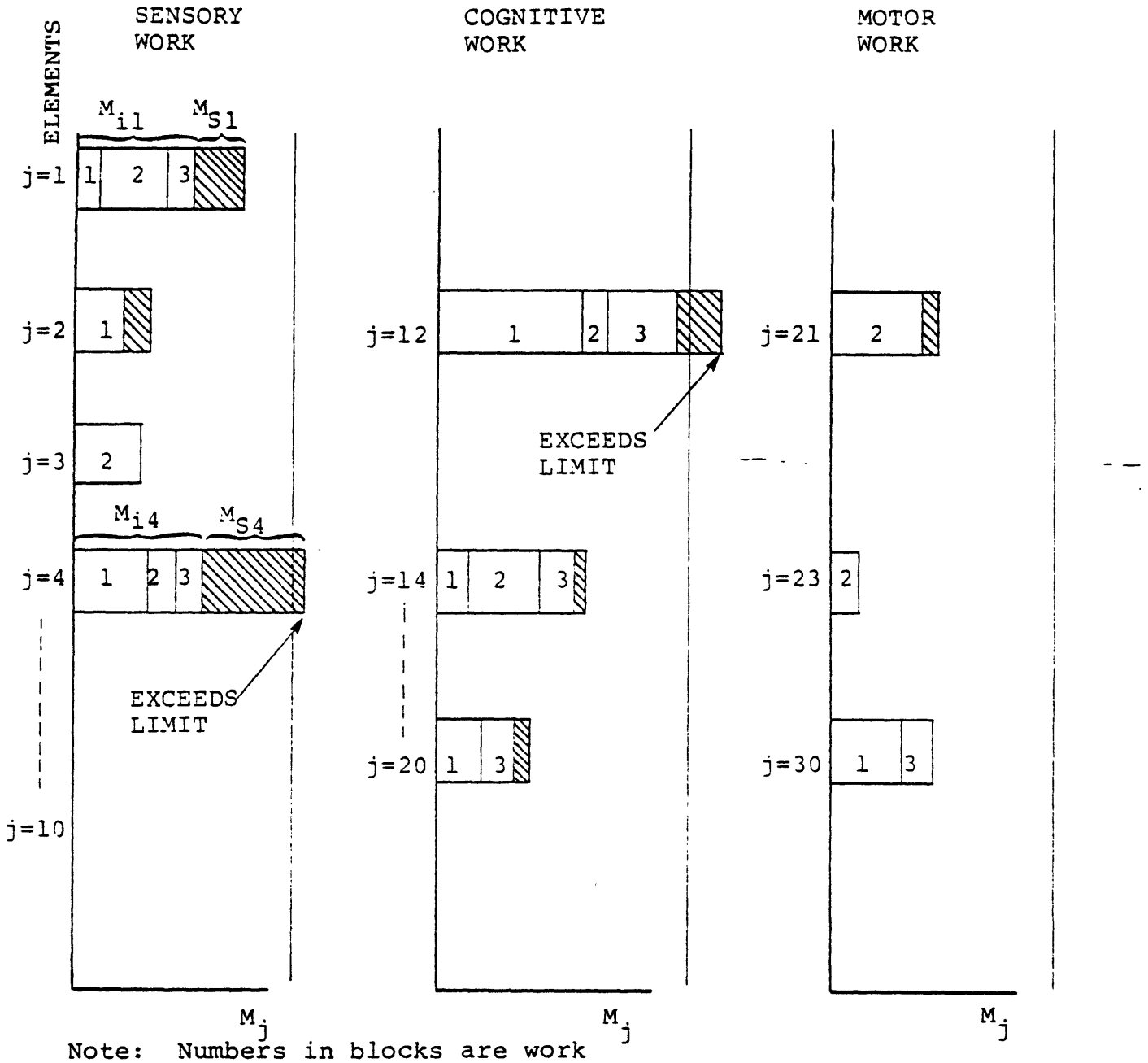
The implication is that whether elements are sensory, motor or "purely" cognitive, all require some mental effort. At any instant of time the doing of a given task i requires that some quantity of mental effort $\dot{M}^D_{ij}$ be allocated to element j, and that the allocation of mental effort to the various elements be in some fixed proportion, such that whatever mental effort is applied is distributed to these elements according to this proportion. At another instant, the task may require that some of these elements be switched, so that the mental effort becomes allocated in different proportions. However, we assume that at any one time a given element can be assigned to no more than one task.

The switching activity itself is assumed to use up an additional quantity of mental effort, specific to the element being switched to, as well as the given task. Thus, over some short time period τ, the amount of mental work applied to each sensory, motor or cognitive element by task i, including mental work expended to switch to each element, is

$$M^D_{ij} = \int_0^{\tau} \dot{M}^D_{ij}(t)\, dt$$

The set of M^D_{ij} may be allocated such as is shown in Figure 8. By dividing through by τ we determine a "mental effort level" expended over τ, which can also be calculated as a "moving window" time function as was done above.

In conjunction with this "continual reallocation of mental effort among elements" model of skill, we make the following additional assumptions:



Note: Numbers in blocks are work allocations to tasks i=1,2, and 3. Shaded area is non-task-specific supervisory mental work.

FIGURE 8

Hypothetical Allocation of Mental Work During Time Window to Sensory, Cognitive and Motor Elements

(1) M^p components add linearly across tasks and across elements for the period , i.e.

(a) $M_i^p = \sum_j M_{ij}^p$, total mental work for a given task

(b) $M_j^p = \sum_i M_{ij}^p$, total mental work for a given function

(c) $M^p = \sum_{ij} M_{ij}^p$, total mental work

(2) Associated with each psychomotor element, j, is a maximum work level or "capacity," $MCAP_j^p$, for a given pilot, p. We can speculate that these capacities may be a function of training or practice, of innate physical ability, of physical condition, and of psychological or emotional state. Fatigue could be defined as a degradation of these work capacities due to a high level of sustained mental effort.

We define these "capacities" in terms of work level rather than instantaneous work rate following the usual practice in queueing theory. Thus, it is possible for an instantaneous, work rate to exceed these capacities for short periods of time.

Now, since there are limits on the levels of mental work which can be performed by a psychomotor element, there will be limits on the levels of behavioral activity for tasks which require that element, and consequently the overall levels of performance on tasks will be limited. The choice is to do a good job on some of the arriving tasks, or to ration the available mental resources across all tasks and accept lowered levels of task performance.

(3) We may now define a "workload ratio" for each element

for a given pilot. This is the ratio of that element's effort level during τ to its capacity.

$$\rho_j^p = \frac{ML_j^p}{MCAP_j^p}, \quad 0 \leq \rho \leq 1$$

This gives us an index of the "loading" or "utilization" of the capacity of that element. We have a set of such ratios, one for each psychomotor element, instead of a single overall value for the operator. For example, the visual element might be saturated with $\rho = 1.0$ at the same time other elements are at low values of utilization.

Thus we imagine the pilot to be continually reconnecting his sensory, motor and purely cognitive elements of Figure 7 and that this is done on a time scale too fast for us to do much bookkeeping on mental workload within a period τ. The pilot's motivation, presumably, is to use his elemental resources in such a way as to (1) minimize lost time in switching, (2) avoid mental workload limits and (3) get the necessary tasks done before their respective deadlines.

When task deadlines give him enough leeway, he may delay working on some tasks or spread his allocation of mental effort out to avoid the overload condition. Or, when loading becomes too great or deadlines too pressing, he may intentionally drop some tasks which are of little importance.

When stringent task demands pose a situation which is not easily resolved, the non-task-specific short-term planning or "supervisory" component of mental work increases, with concomitant sense of uncertainty, anxiety and generalized stress. Under such circumstances we may expect

that some of the limiting values of some sensory, motor or cognitive functions may be reached, and the pilot's skilled behavior will be compromised. We may also expect that deadlines of some important tasks may be reached with those tasks not yet completed.

(4) We can now introduce the concept of "net subjective utility," NSU^p, for each pilot. It is useful to describe the task allocation, or supervisory problem, in terms of the subjective utility, U_i^p, which the pilot places on task performance and its consequences, and the subjective cost, $\bar{U}_i^p$, which the pilot places upon his mental or behavioral work. We could postulate that in some period, τ, the pilot will maximize his net subjective utility

$$\text{Max}(NSU^p) = \text{Max}\left[\sum_i U_i^p(P_i^p) - \bar{U}_i^p(M_i^p)\right]$$

He accomplishes this optimization under the following constraints:

(a) The mental effort level, for every psychomotor element, is less than its capacity

$$ML_j^p \leq MCAP_j^p \quad \text{for every element } j$$

(b) The performance on a task is above the minimum performance requirements, if such exist.

$$P_i \leq P_{mini} \quad \text{for any task } i$$

(c) The behavioral activity for task i is a function of the mental work expended on that task.

(d) THe performance achieved on task i is a function of behavioral activity expended on that task.

This is the supervisory or task allocation problem stated as a static, deterministic optimization problem for the period τ. A lazy pilot would have high subjective costs and low utilities on task performance. As a result, he would allocate a minimum amount of work such that minimum performance requirements are just met. An ambitious pilot has high utilities and low costs, and will maximize his overall performance with his workload ratios equal to unity. A complete range of intermediate solutions in terms of task work allocations, M_i^p, workload ratios ρ_j^p, task performances, P_i, can be generated by assuming various values for subjective utilities and costs.

But this static, deterministic model of the supervisory problem may be too simple to describe the rather dynamic, stochastic work allocation problem which the pilot usually faces. The tasks are not all present at the same time, do not arrive at expected times, etc., so that the pilot has a much more difficult problem to solve (see Tulga[8] for a more dynamic model of this supervisory problem, where such a static optimization process must be iterated as each new task arrives).

Notice that this work allocation model emphasizes the importance of briefing subject pilots in experimental workload simulations with a view to establishing the values of U_i and $\bar{U}_i$. We can expect different solutions from lazy or ambitious pilots. Further, in the real world the subjective utilities of high performance on certain tasks may

be considerably different than those found in the safety of an aircraft simulator. This reality brings into question the validity of simulator experiments which attempt to study pilot workload. Pilot behavior and work allocations could be different in the aircraft where the consequences of poor task performance are real.

2.4 RELATED THEORY IN MENTAL WORKLOAD

The ideas described above seem to us in accord with several current theories of skilled performance. Figure 9 illustrates how, during a period τ, in the process of doing two simultaneous tasks, the capacity limits of one sensory element and two motor elements may create tradeoffs under mutually conflicting demands. Notice that there is a "simplex" of constraint boundaries rather than the simple concept of a single boundary often used in connection with secondary or auxiliary task concepts. With non-linear task performance curves, this simplex is transformed into a "performance boundary" which would seem to accord with theories espoused by Norman and Bobrow[7] on the conflicting use of behavioral resources, wherein behavioral performance tradeoff curves (such as Figure 10) will result. Verplank[8] has also discussed this problem.

The theory presented here is also in accord with the work of Tulga,[9] who experimented with human operators and devised a normative model for allocation of effort among multiple conflicting task demands. Tulga suggested some ways in which subjective sense of mental workload accords with amount of short-term planning required to complete tasks before deadlines are reached. Tulga's model operator had only one "channel" (or "element" in terms of the present paradigm), but

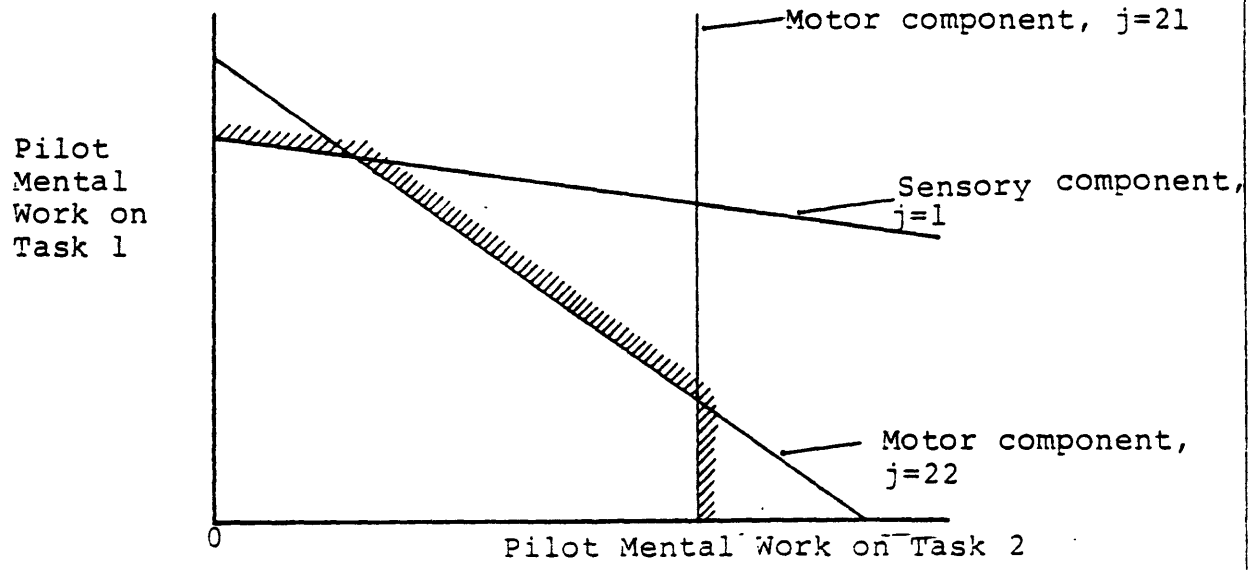


FIGURE 9

Mental Work Capacities for Two Simultaneous Tasks

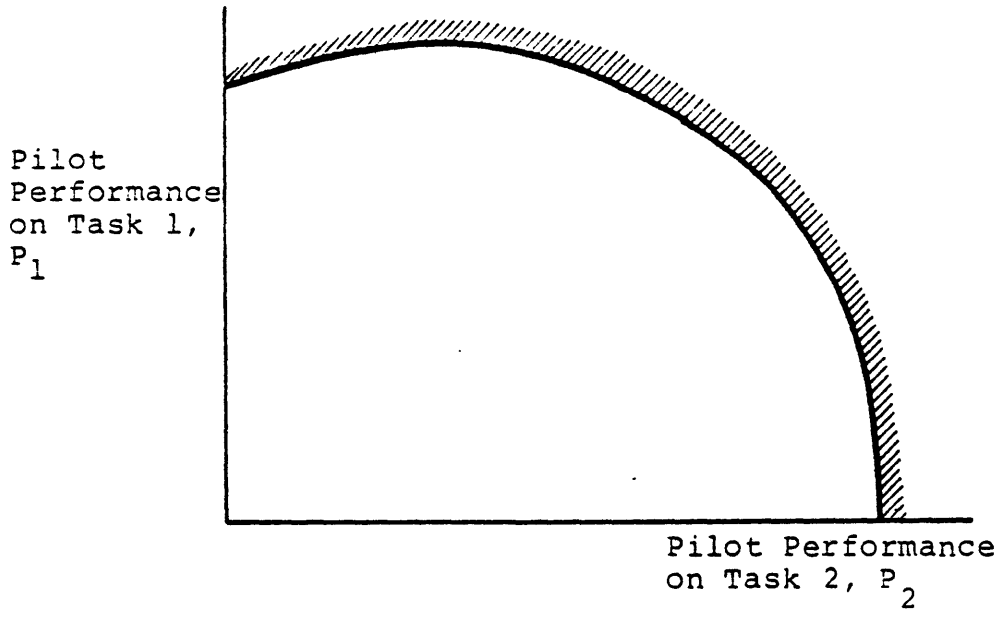


FIGURE 10

Performance Level Tradeoff for Two Simultaneous Tasks

we feel his general decision-theoretic approach can be extended to accommodate not only dynamic task demands with deadlines but also the allocation of multiple elements with each having a workload capacity.

A third theoretical direction is that of queueing theory which seemingly could contribute to our understanding of how pilots serve task demands which queue up. It should be clear that stochastic variations of mental effort, $\dot{M}^p(t)$, in response to the stochastic variations in arrival times for tasks, are important in determining overall performance by a pilot. As ρ_j^p values approach unity, the probability that instantaneous mental work values, $\dot{M}_j^p(t)$, will exceed the capacity levels, $MCAP_j^p$, during short periods, rapidly increases. This probability also depends upon the degree of randomness in arrival times for discrete tasks.

We expect that task performance will degrade because of peaking effects. For simultaneous continuous tasks, it occurs because the capacity limits force work levels on some or all tasks backward along the $P_i(B_i)$ curves during these periods. For discrete simultaneous tasks, it is possible to delay working on some tasks, i.e. to create a backlog or queue of tasks. The degradation in performance occurs because of delays in accomplishing tasks, the omission of some tasks because the queue size has grown beyond the capacity of the operator's short term memory, or the increased probability of task interruption due to the arrival of a task of higher priority. The expected degradation of task performance as a function of ρ and the randomness of task arrivals is schematically shown in Figure 11, where we have moved to some overall measure of task performance across all tasks, and an overall measure of workload ratio.

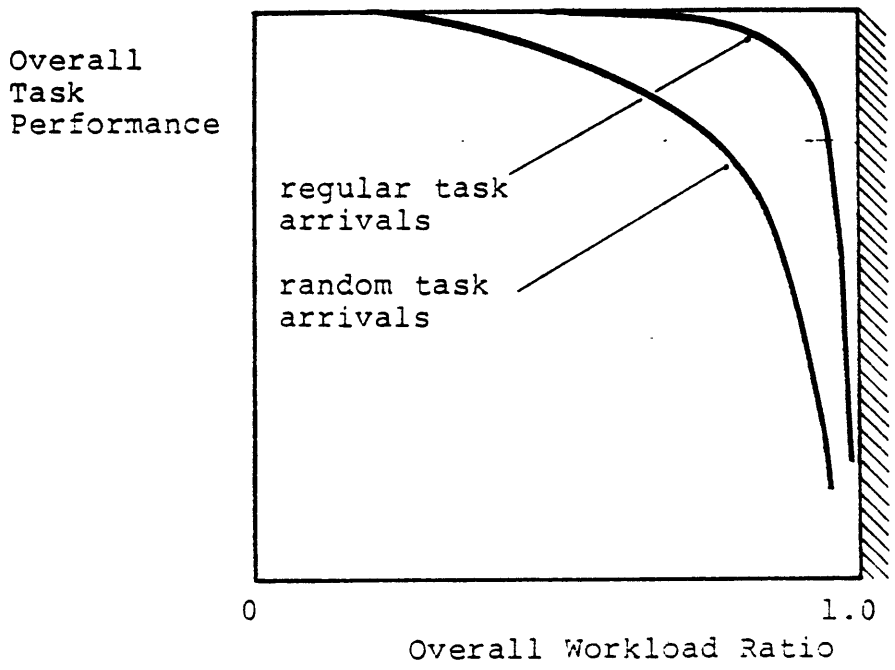


FIGURE 11

Overall Task Performance versus Workload Ratio

2.5 SUMMARY

In this section, we have postulated a rather complex structure for defining concepts about pilot mental workload which is consistent with the variety of concepts, ideas, and definitions contained in current literature. We do not contend that our structure is valid, only that it is necessary. Further refinement of such definitional structures is needed to guide future research on pilot workload in theoretical and experimental areas. Results from future research will determine the validity of such a structure, and will contribute to its further development or modification.

3. THE MEASUREMENT OF PILOT WORKLOAD

3.1 WHY MEASURE PILOT MENTAL WORKLOAD?

The primary reason to measure pilot workload is to predict performance. There may be secondary reasons, such that, when performance is not in question, the work may be too great for comfort, etc. But the primary reason is to ensure satisfactory performance of the pilot and safety of the aircraft and its passengers.

Then why not measure performance directly? Especially since performance is relatively easily measured compared to mental workload? The answer is that measured performance can maintain a high level as task demand and (what appears to be) resultant mental workload gradually increases -- right up to the point where the pilot is obviously complaining about mental workload. At this point measured performance, as a function of task demand, would then predict continued satisfactory performance as task demand increases further.

But we know from experience that this is not always the case. Presumably the pilot can "struggle" harder, i.e., apply more "mental effort" by sheer will, to compensate for increasing task demand and maintain performance, but only up to a point. Beyond this point his "capacity" is exceeded and his performance deteriorates. This phenomenon was indicated by the curves of Figure 11 where a sudden drop in performance occurs at high values of ρ.

Thus, it is believed that if suitable measures of mental workload can be found, at relatively high levels of task demand they will be better predictors of what <u>would</u> happen <u>if</u> task demand <u>were increased still further</u> than would measures of performance per se. That is, mental workload,

it is claimed, is a more sensitive predictor of response to increased demands than is performance.

Such a claim for the use of mental workload may be made not only in predicting pilot response to increased task demand, but also in predicting pilot response to sustained task demand at the same level, i.e., fatigue.

3.2 ALTERNATIVE MEASURES OF PILOT MENTAL WORKLOAD

Referring again to Figure 1 we see four types of (categories of) measures of mental workload, M1 through M4, associated with the six definitions, D1 through D6, discussed above.

3.2.1 SUBJECTIVE JUDGMENT

Subjective judgment of mental workload, either by the pilot himself or by an observer, is the most widely accepted method of measurement. Subjective judgment is also the intuitively valid reference for the meaning of "mental workload." It provides the basis on which other measures are validated, i.e., to answer questions such as "Is this a good measure of pilot mental workload?" More is said about subjective judgment below. For comparison we cite the other mental workload measures commonly used:

A special kind of subjective judgment measure does not require the pilot to judge workload directly, but rather to make subjective judgment of other variables in the situation. One new and promising example is subjective time estimation, e.g., the pilot estimates when ten seconds has elapsed. (Results suggest that as task demand is increased the actual

measured interval, corresponding to what he claims to be ten seconds, consistently increases by 4 seconds and becomes more variable -- but only up to a point which Hart calls "critical mass."[10] Beyond that the measured interval shifts to suddenly back to ten seconds and diminishes as task demands further increases.)

3.2.2 PHYSIOLOGICAL INDICES

Physiological indices are many. All measure something in a very scientific way. The question is whether what they measure is correlated with what we think of as mental workload (i.e., what we make subjective judgments about when asked about "mental workload" in a given situation). Some physiological measures used with reference to mental workload are:

-- heart rate variability, heart rate
-- respiration rate
-- galvanic skin response
-- pupillary diameter
-- biochemical changes in blood and urine
-- electroencephalogram changes
-- changes in frequency spectrum of voice

All of these measures exhibit a high degree of variability and may be affected by other conditions independent of mental workload (environmental physiological stress, diet, etc.). Further, most must be individually calibrated. Finally, some are sufficiently distracting to make them invalid for use on pilots in situ, i.e., while flying simulator or actual missions.

A special kind of physiological measure, which is a common measure of one aspect of pilot workload, is eye movement recording. That is, number of eye transitions to a given

instrument suggests a sampling rate, which translates (supposedly) into information rate. Visual scans outside the nominal task stimuli or to side tasks indicate visual "free time."[11]

3.2.3 SECONDARY TASK PERFORMANCE

Secondary task performance measures assume that the primary task takes a fixed percentage of time, and what time remains can be relegated to a "secondary" task. The more time the pilot spends on the secondary task, presumably, the higher his score on it. Thus, the greater the time which must be spent on the primary task the less the score on the secondary task. This theory is impeccable, with one exception: namely, that the pilot actually spends as much time as is needed on the primary task, and will then spend what remains on the secondary task. Too often in laboratory experiments experimental subjects get the idea that the secondary task "really counts" and will make that the primary task or divide their attention evenly. Too often in operational flights pilots simply ignore secondary tasks altogether.

Note that our framework for mental workload agreed with the non-linear performance tradeoff curves of Norman and Bobrow. It is difficult to describe another task as "secondary," in the multi-task environment such as experienced by a transport pilot.

3.2.4 PRIMARY TASK PERFORMANCE

Primary task performance may be said to be indicative of mental workload, but normally shows little change until task load (and accordingly mental workload) is so extreme that performance deteriorates. But at that point one is not normally interested in mental workload anyway, since the threshold of deterioration of primary task performance is the very thing the workload measure was meant to predict.

3.3 CRITERIA FOR A USEFUL MEASURE OF MENTAL WORKLOAD

Ephrath[12] has suggested five conditions for a useful measure of pilot workload. They are worth repeating here:

- -- minimal interference with primary task
- -- applies uniformly to various types of piloting tasks
- -- applies uniformly to various types of information channels internal to the pilot which "become loaded"
- -- applies uniformly to various pilots, in various states of training, fatigue, etc.
- -- is capable of operation in actual flights as well as in a simulator

We might add:

- -- must be feasible economically
- -- should be capable of measuring changes in mental workload which occur as a function of major changes in flight phase or conditions (i.e., a new measurement should be possible every five minutes or so)

Observations of air crews at work leads us to comment that well-learned visual scanning and manipulative or closed-loop manual control activity would seem to make for relatively little mental workload. As the pilots tend to put it, "We have no trouble flying the airplane itself." Except for take-off and final-approach/landing, these routing motor skill operations seem to take no large fraction of the pilot's attention. Yet these are the events which are most easily amenable to eye movement analyses and other visual-motor measurements.

Mental workload of a more troubling kind occurs when entering a terminal control area, communicating with ground controllers (and sometimes disagreeing with the advisability of the instructions the controllers give as to speed, holding pattern, descent profile, etc.). It also occurs when an instrument fails, or there is conflicting information from different sources, etc., and the pilot must make diagnostic and management decisions which involve tradeoffs between probabilities and risks and consequent anxiety. These are the workloads we wish to get at. We have a suspicion they are associated with levels of "supervisory" mental workload.

3.4 PROS AND CONS OF SUBJECTIVE MEASURES OF PILOT MENTAL WORKLOAD

Subjective measures would seem to fit the Ephrath criteria stated in Section 3.2.2 above about as well as any measure. They may be the only measures which can also yield independent data points at five-minute intervals. Insofar as cost is concerned, <u>if</u> it were to prove practical for pilots to self-administer workload judgment tasks the cost would be minimal. If an observer is required, it will be more expensive. But

then an observer with a pad of paper is cheaper than an observer plus expensive instrumentation.

It was stated above that subjective judgment measures of mental workload have inherent validity. This is what is formally called "external validity" -- whether the measure measures what it is supposed to measure. There is also "internal validity" -- or what is also called "reliability" (independence of spurious factors, which boils down to consistency, precision, or variability).[13] Obviously both inter- and intra-judge variability will depend in part on how carefully the subjective scaling exercise is planned and implemented. It is important to insure reliability, to:

(a) make sure that the events to be scaled, i.e., the experience of a given mental workload situation (whether actually experienced as mental workload or imagined from a description) be administered with sufficient clarity of detail that there be no ambiguity about what the situation is.

(b) make sure the procedure which the judge is to use for assigning numbers to the experienced events also is clear, i.e., that the judge understands what the numbers at each end and in the middle of the scale mean ("anchoring") and what is the relation between any two numbers on the scale. (For example, does 3 relative to 1 mean 3 times as much workload, or a "workload distance" of 2 units difference comparable to 2 units elsewhere on the scale, or are 3 and 1 just ordered categories?)

Happily, when care is taken, internal validity of subjective measures, both within and between judges, can be quite good, as attested to by various researchers (Stevens,[14] Borg,[15] Jenny[16]). Probably the reliability is as good as objective physiological or secondary task performance measures.

The charge is sometimes made that subjective scales can never be more than ordinal scales, or further, that it is never appropriate to perform quantitative analysis on subjective data. Both charges are quite fallacious, as the rich and sophisticated literature in econometric and psychometric research will attest. Subjectively-based interval ratio scales are quite obtainable, depending upon the procedure used to get and process the data.

Another question (and sometimes a criticism) about subjective measures of mental workload concerns whether it is appropriate for the pilot to judge his own mental workload during the event itself. As already noted, one way out of this dilemma is to have an observer riding in the jumpseat, and have the observer do the judging. The latter makes little sense unless the observer is himself an experienced pilot and therefore can empathize (even then in some situations he may miss some "mental loading" event). Perhaps a better practice is to have the pilot rate his own mental workload of a given phase shortly after he has completed it (and during a flight period for which mental workload will not be assessed). Alternatively, or in addition to the above, an assessment can be made during a debriefing on the ground. If the latter is done it is useful to have a video and sound recording of cockpit events to remind the pilot/judge what occurred during the period in question.

What kinds of pilots should do the judging? It has been asserted (e.g. by Gartner and Murphy[1]) that the best results are obtained only if experienced subjective assessors (test pilots) do the judging. Others disagree, claiming that an experienced test pilot's judgments of workload are just not the same as a novice pilot's, that data from both types of pilots is necessary and is likely to constitute an important

differentiating factor. That is, mental workload is not simply a function of aircraft and procedures; mental workload is also a function of pilot.

A further question which can arise with subjective judgment scales is whether they are absolute or relative. The answer must be that subjective judgment is always relative, to some extent. It's like when A says to B, "How's your wife?" and B responds "Compared to what?" The point is that what we call absolute judgment is necessarily based on comparisons with experienced events. The experimenter, through well-defined and unambiguous (as possible) reference on baseline events can attempt to "anchor" one or both ends of a judgment scale. But then, the judge's interpretation of (the words of instruction used to explain) the scaling procedures depends on his experience. There really is no absolute judgment.

One way the relativism of mental workload judgment might arise is as follows. Suppose the pilot is asked to scale two several-minute phases of flight: 1) entry into a terminal control area; 2) the landing itself. In the former the pilot's task is to be familiar with the charts, to communicate with the ground and get the proper information, and occasionally to update and monitor the autopilot for course and altitude. There are uncertainties, but the visual and manipulative demands are <u>normally</u> mild. In the actual landing, on the other hand, nearly 100% of visual and manipulative capacity is <u>normally</u> required and this is considered satisfactory. Will the pilot reference his workload estimate to what is <u>normally</u> required, or will he use a more absolute scale? And what about the fact that phase (1) is more of a planning and communicating workload, and phase (2) is more of a visual-motor skill workload? Comparing the two phases is like comparing cows and potatoes. The scaling procedure must be amenable to scaling both. And the pilot/

judge must have some guidance on whether the scale is to be referenced as "normal" or not.

It is sometimes claimed that, given a well-defined and repeatable procedure for asking the questions, the organism's judgmental apparatus is a "noisy filter," easily biased by attraction or aversion to the experimenter, or by the desire to please or aggravate, or by what he ate for breakfast, etc. This is true to an extent, and cannot necessarily be eliminated by averaging. One simply tries to minimize such bias.

Finally, there is the claim that pilots "may not be in touch with their own bodies," that they may feel quite able, and even judge their own workload to be modest, but nevertheless be on the verge of performance breakdown. Should this be true, and should there be physiological measures which reveal the "verge" better and more reliably than subjective ones, we don't know what they are. Only hard and costly evidence will tell.

3.5 FOUR CANDIDATE PROCEDURES FOR OBTAINING SINGLE-DIMENSIONAL SUBJECTIVE SCALES

3.5.1 CATEGORY SCALES

The Cooper-Harper scale[17] is the outstanding example, but also in other fields there have been well-accepted category scales. Usually the category rating scale consists of five to nine discrete categories, each identified by a phrase which the judge can understand with a minimum of confusion as to the meaning of the words. Sometimes (like the "revised" form of the Cooper-Harper scale, Figure 12) there is a decision tree, or hierarchy of subdivisions which help the judge by a succession of decision stages to select a final category.

FIGURE 12

Cooper-Harper Scale

Controllable Capable of being controlled or managed in context of mission, with available pilot attention.	**Acceptable** may have deficiencies which warrant improvement, but adequate for mission. Pilot compensation, if required to achieve acceptable performance is feasible.	**Satisfactory** Meets all requirements and expectations, good enough without improvement. Clearly adequate for mission.	Excellent, highly desirable	A1
			Good, pleasant, well behaved	A2
			Fair. Some mildly unpleasant characteristics. Good enough for mission without improvement.	A3
		Unsatisfactory Reluctantly acceptable. Deficiencies which warrant improvement. Performance adequate for mission with feasible pilot compensation.	Some minor but annoying deficiencies. Improvement is requested. Effect on performance is easily compensated for by pilot.	A4
			Moderately objectionable deficiencies. Improvement is needed. Reasonable performance requires considerable pilot compensation.	A5
			Very objectionable deficiencies. Major improvements are needed. Requires best available pilot compensation to achieve acceptable performance.	A6
	Unacceptable Deficiencies which require mandatory improvement. Inadequate performance for mission even with maximum feasible pilot compensation.		Major deficiencies which require mandatory improvement for acceptance. Controllable. Performance inadequate for mission, or pilot compensation required for minimum acceptable performance in mission is too high.	U7
			Controllable with difficulty. Requires substantial pilot skill and attention to retain control and continue mission.	U8
			Marginally controllable in mission. Requires maximum available pilot skill and attention to retain control.	U9
Uncontrollable Control will be lost during some portion of mission.			Uncontrollable in mission.	U10

Category scales have the advantage of simplicity. Everyone, including test pilots, understands them with a minimum of explanation. Since they have been accepted by pilots for use in rating handling qualities,[17] it does not appear to be a large step to employ such a category scale for mental workload. However, mental workload is not the same as handling quality (e.g. mental workload will decrease with experience, handling quality will not). So a new set of category identifiers must be developed.

Some words may be similar, e.g. an initial subdivision of mental workload into "below level where performance will necessarily break down and control of aircraft be lost," and "above" same, and within the first category some elaboration on phrases like "acceptable," and "unacceptable." The words should emphasize "mental effort," and possibly refer to normal and well-known levels of mental effort such as landing under some "standard" conditions. They might refer also to fraction of attention required, or, equivalently, what types of additional tasks could be accomplished (specific examples would be given) at the same time. The identifying words might also refer to how long that mental workload level could be sustained (minutes, hours, etc.). Finally, there might be different sets of words to provide equivalencies for mental workload due to sudden transients vs. that due to long term vigilance, mental workload due to visual-motor skill vs. that due to planning, etc.

Pilots accustomed to judging "handling quality" would have to be reprogrammed to think "mental workload." But the problems of devising and implementing such category scales are not insuperable.

3.5.2 CROSS-MODALITY MATCHING

This is a procedure developed by Stevens[14] whereby a knob (on the other control) is adjusted until some artificial stimulus (like the amplitude of a vibration on the skin) "psychologically matches" the thing being scaled. Stevens has obtained consistent cross-modality matches between handgrip force produced by the judge and various psychophysical stimuli such as loudness of a tone, brightness of a light, etc. Cross-modality matching has the advantage that it is simple. But it remains to be seen whether it could yield anything useful for "mental workload."

3.5.3 THURSTONIAN OR POIKILETIC SCALES

Thurstone proposed[19] that an equal-interval subjective scale could be based upon the degree of confusion, or variability, or scatter of judgments. Situations (work-load events) to be judged are considered in pairs, and the consistencies of ordering which is the greater (mental workload) determines the distance between them. If the order is consistently in one direction the distance is large; if it is randomly either way there is no discrimination and hence no distance on the psychometric scale. Naturally, any one pair is not presented to the same judge very often; there are quantitative means to aggregate and infer average distances based on all pairings. This method may also be thought of as related to the law of Fechner, i.e., scale distances between stimuli are based on number of "just-noticeable differences." Figure 13 gives the general idea.

McDonnell[20] used a poikiletic technique to scale a set of words often used in rating scales. McDonnell's scale is shown in Figure 14.

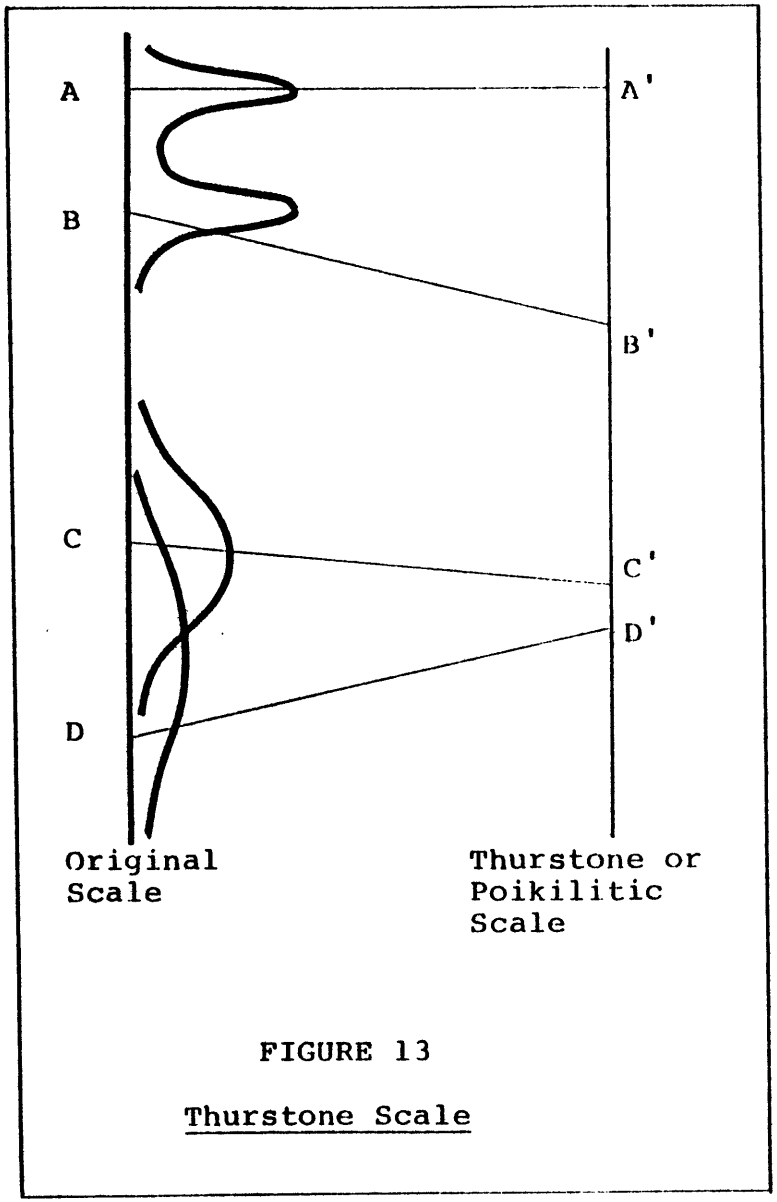


FIGURE 13

Thurstone Scale

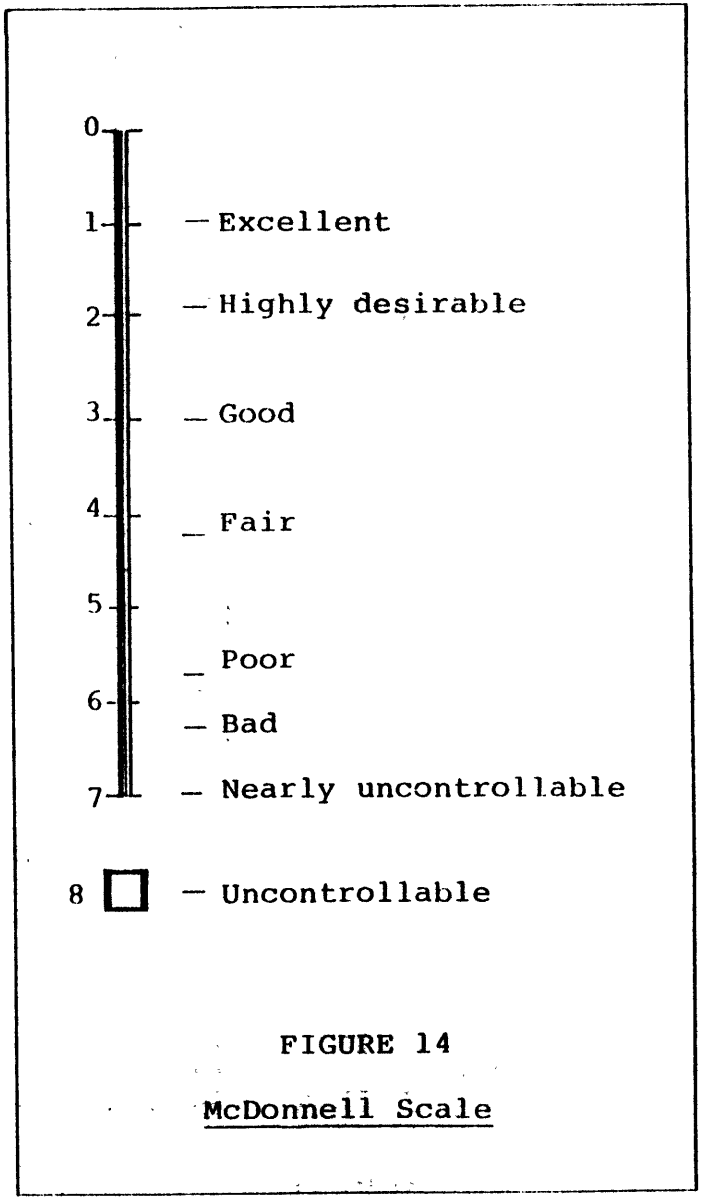


FIGURE 14

McDonnell Scale

3.5.4 DIRECT MAGNITUDE ESTIMATION

Direct ratio scaling assumes that a judge, given a standard or reference level of the workloading events being judged, can specify whether some other level is 1/2 of the reference, or 3.5 times it, or whatever. This method has long been used by Stevens and others in psychophysical scaling.[14]

Lottery-based magnitude scaling assumes that the judge can specify a point of indifference between a test situation (set of events) and a lottery consisting of a different situation having probability p and another situation having probability (1-p). Figure 15 indicates the indifference and the equivalent equation, which serves as the Von Neumann definition of "utility," and is widely employed in decision theory.[9] Theoretically either the situations can be changed or the probabilities can be changed until the judge says he is indifferent. In pilot mental workload measurement the hypothetical probabilities are the easiest to change. Thus the pilot would decide on a probability p, for example, such that he is indifferent between a test workload x happening 100% of the time and a lottery consisting of workload y happening p% of the time (say y were some maximum tolerable) and workload z (say none at all) happening if y didn't. From a series of such judgments a mathematically defensible interval scale (and some might say a ratio scale) can be developed. Such a proecedure yields a scale of utilities U for all workload events i, and is capable of predicting an important quantitative relationship: For any of the workloading events on the scale, when it is estimated to occur with any given probability p_i, a "subjectively expected utility" (SEU) results as the product of U_i and p_i. Then the pilot is predicted to prefer taking the risk of (U_i, p_i) to any other (U_j, p_j) which has a smaller SEU. One might criticize this procedure by claiming that "utility of workload" and "workload" are not the same. Perhaps not, and it would

be interesting to compare mental workload scales derived in the "utility" manner with those derived by direct magnitude estimation.

3.6 FOUR CANDIDATE PROCEDURES FOR OBTAINING MULTI-DIMENSIONAL SUBJECTIVE SCALES OF MENTAL WORKLOAD

3.6.1 SEPARATE SCALES OF DIFFERENT ASPECTS OF MENTAL WORKLOAD

Conceptually the simplest way to obtain a multidimensional scale is to use separate scales. That is, by any of the methods described in Section 3.5, subjective judgments are obtained for the given events, using some words to describe various levels of some one aspect of mental workload, say "mental workload involved in handling the aircraft." Then another set of judgments is obtained relative to the same events but using a different set of words to describe various levels of a different aspect of mental workload, say "mental workload involved in communications and navigation." The same can be done for several different aspects. The whole set of workloading events can be run through with each new aspect, or, alternatively, for each event in turn, judgments about the different aspects can be run through.

There are various ways to categorize the different aspects of mental workload. One, say, is by piloting task (as already suggested: aircraft handling, communications, navigation, systems monitoring, attendance to passengers and staff, etc.). Another cut is by bodily activity (e.g., speech communication, control manipulation, planning and decision, memory and information accessing, etc.). Still a third is by transient events vs. steady state (fatiguing) events.

Judged indifference: Situation x for sure $\approx$ p: Situation y; 1-p: Situation z (lottery)

Equality (definition): $U_x = p\,V_y + (1 - p)\,U_z$

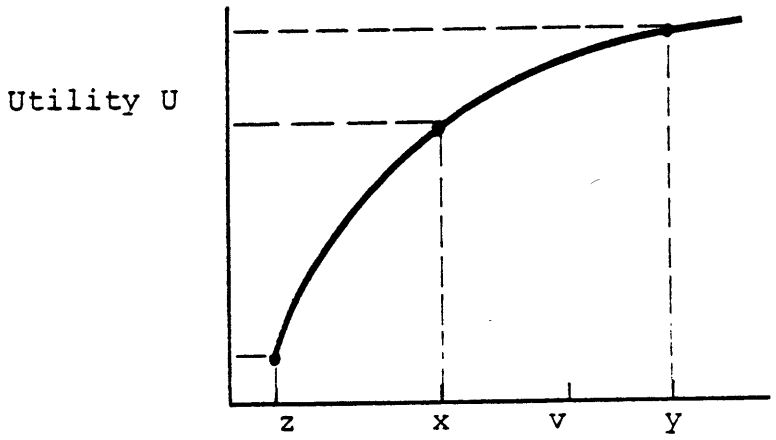


FIGURE 15

"Utility" Judgment

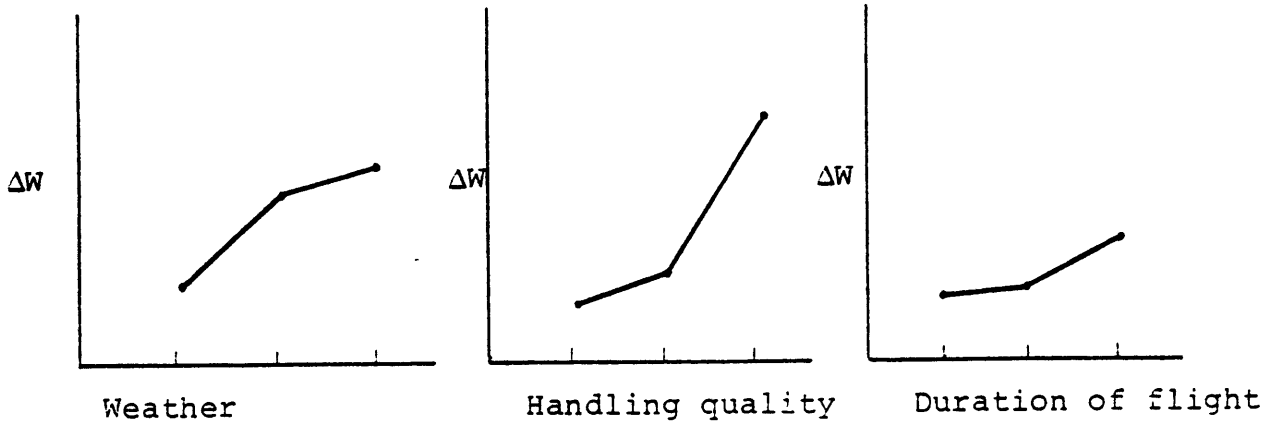


FIGURE 16

Policy Capturing Results

Clearly the separate scales obtained are not going to be independent, so that a multi-dimensional scale with true properties of orthogonality will not result from such a procedure. With proper care, however, and provided only a small number of separate scales is used (the pilot won't stand for many!) some interesting differentiations might be obtained.

3.6.2 JUDGING CONDITIONED DIFFERENCES

After a judgment has been made, say about the mental workload of a particular five-minute period of events just experienced, it may be appropriate to ask the pilot to make some additional conditional judgments. For example, he might be asked what the mental workload would be if one factor of the situation changed in a certain way (say, in the weather, procedures, etc.). This can be thought of as providing a "partial derivative" of mental workload with respect to that factor around a given operating point (set of events experienced and initially judged). This "piggy backing" technique could provide several times the original amount of workload judgment data, per unit of flight time.

3.6.3 POLICY CAPTURING AND MULTI-ATTRIBUTE UTILITY

These are commonly considered different techniques but will be commented upon together because of their similarity.

In policy capturing, each of a small discrete number of given attributes which might affect workload would have several different levels. For example, the attribute "weather" might have levels ("zero-zero," several levels combining ceiling and visibility and wind may be specified, and finally "CAVU"). An attribute "handling quality" might have level ratings which are standard Cooper-Harper ratings. And

so on for "cumulative duration of flight," "condition of aircraft," and so on. Then different bundles of attributes (i.e., one particular level of each attribute employed) are separately rated for mental workload. Ideally these would constitute separate simulation runs. Less ideally they would be described to the pilot/judge verbally. He would make a direct mental workload magnitude estimation in each case. After the whole exercise the data, based on many combinations of various levels of all attributes, would be turned over to a computer, and a "policy capturing" algorithm based on non-linear regression, is run. It produces, for each attribute, a piecewise function showing how, on the average, various levels of that attribute contribute to (weight) the overall judgment of "mental workload," Figure 16. Hammond and his colleagues[21] have employed these methods in a variety of applications.

Multi-attribute utility is an extension of single-dimensional (single-attribute) utility scaling, and has been written about extensively by Raiffa and Keeney[22] and their colleagues. One form of it is called "interpolation between corners," and is explained in Figure 17. Assume three attributes, each of which has several different "levels", as with policy capturing. For each separate attribute the pilot is asked to assign a number from 0 to 1 for each level, corresponding to the judged mental workload of that level, and such that he is indifferent between mental workload of 0.5 and a 50-50 chance of either mental workload of 1 and mental workload of zero. In other words, the judgement is based on a lottery, as before, and the underlined words convey the meaning of the numbers. Our pilot does this for each of three attributes, then does it again for the eight combinations of the best and worst levels of each attribute (the "corners" of the cube). Then, by a simple data aggregation and interpolation procedure similar to that used for policy

capturing, a three-dimensional utility function is obtained (i.e., with the non-metric-ordered levels of the three attributes as the arguments). Yntema, Torgerson and Klem[23] successfully used this method to obtain preference functions for instructor-pilots landing aircraft under various combinations of ceiling, visibility and fuel remaining. Instead of the "corners" procedure, multi-attribute lotteries can be judged using the more mathematically precise (but perhaps more psychologically confusing) procedure described earlier in Section 3.5.

3.6.4 MULTI-DIMENSIONAL SCALING

This is a technique pioneered by Shepard[24] at the University of Pennsylvania and Carroll[25] at Bell Laboratories and now commonly employed in marketing and attitude research to infer what are the attributes of situations with respect to which people judge those situations to be different. In other words, one does not start with definite attributes and definite ordered levels for each attribute. One ends up there.

One starts simply with different situations, experienced by the pilot in our case, or described to him in words. He considers all pairs of such situations and judges how dissimilar each pair is, one from the other. (This is a magnitude scale of dissimilarity). That is all the judge does. In the present context the pilot/judge would be instructed that "dissimilarity with respect to mental workload, otherwise considering any and all factors which affect mental workload."

The data are then turned over to one of a number of available computer algorithms, and a two- or three-dimensional plot is generated, plotting each of the situations in that cartesian space.

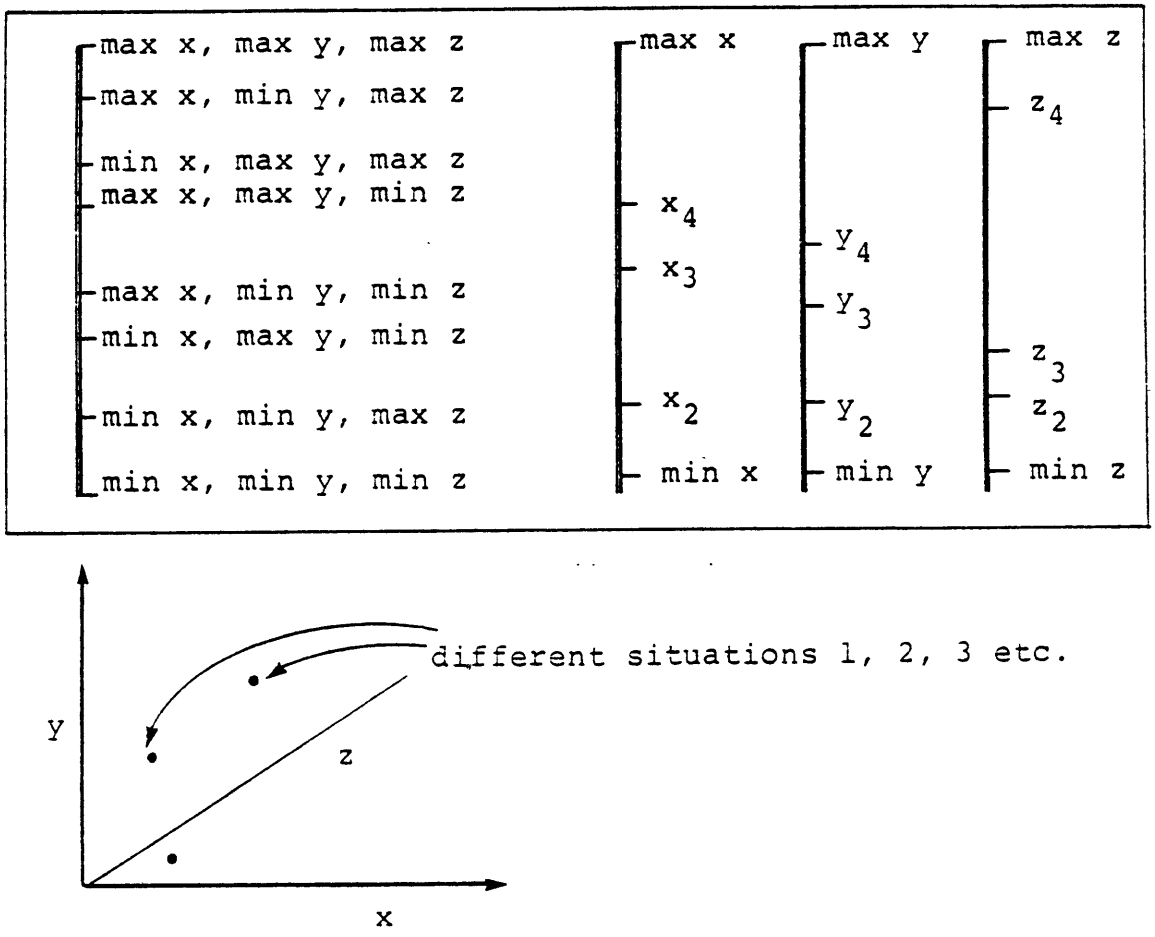


FIGURE 17

Interpolation Between Corners

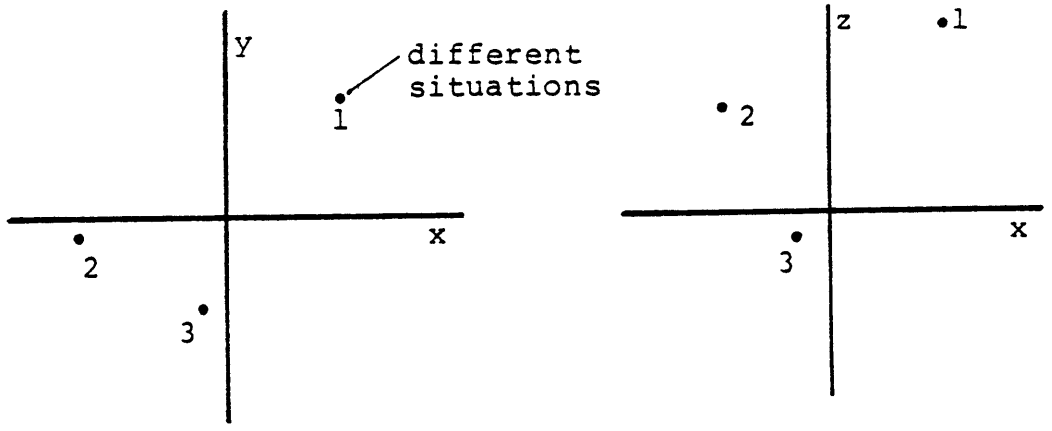


FIGURE 18

Results of Multi-dimensional Scaling

On what basis is such a plot generated and what does it mean? It turns out that the best fit of data for all pair-wise judged dissimilarities between N situations requires an N-1 dimensional space. For example, judging dissimilarities between four events creates a tetrahedron, and requires a three-dimensional plot. Because many workloading situations are likely to be used, a best-fit plot would require a plot of too many dimensions to be of any use.

The real question is, can a two- or three-dimensional plot provide a pretty good fit, i.e., account for satisfactorily large fractions of the "variance." There are various criteria for finding the best fit, and for assessing the degree of residual variance. One method is called "monotone regression" (Kruskal) and Carmone)[28] and determines a set of M(M is chosen arbitrarily) orthogonal axes such that the ranks (order) of interpoint dissimilarity distances is always preserved. A function is minimized, where d_{ij} denotes the dissimilarity

$$S = \left| \frac{\sum_{i \neq j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i \neq j}^{n} d_{ij}^2} \right| \quad 1/2$$

distance between i and j in the reduced space M and $\hat{d}_{ij}$ denotes a set of values chosen to be as close to the d_{ij} as possible subject to being monotone with the observed original dissimilarity distances (i.e., $\hat{d}_{ij} \leq \hat{d}_{kl}$ whenever δ_{ij} is judged less than δ_{kl} where δ denotes original judged distance). The mathematics is included here just to convey some sense of what the computer is asked to do.

One then looks at the resulting plot (Figure 18) of situations in M-space and tries to puzzle out what each of the

M dimensions really means. If these are the dimensions with respect to which piloting situations are most perceived to be dissimilar from one another (in mental workload) the plots should convey a sense of what the most important dimensions of judged mental workload are.

Some programs, like Carroll's INDSCAL,[25] not only provide a given judge's plot of the given situations in M space, but also, if a number of judges are used, plot all judges in the same M space to represent the relative weights each judge used relative to each dimension. If all judges plot in a tight bundle, there is high consistency between judges. If one judge plots far out along one dimension relative to the others, it means that judge weighted that dimension more heavily in making his dissimilarity judgments.

3.7 PROPOSED RESEARCH ON SUBJECTIVE MEASURES

3.7.1 A COOPER-HARPER TYPE CATEGORY SCALE FOR MENTAL WORKLOAD

In the face of immediate need by the FAA and NASA to evaluate CDTI (cockpit display of traffic information) configurations, as well as other near-term applications of subjective scaling of pilot mental workload, it is proposed that a category scale patterned directly on the Cooper-Harper type scale be developed. A first attempt is described in the next section. It will require much more effort to produce a useful and commonly accepted scale.

3.7.2 EXPERIMENTS WITH OTHER SINGLE AND MULTI-DIMENSIONAL SCALING PROCEDURES APPLIED TO PILOT MENTAL WORKLOAD

The recent developments in subjective scaling would seem to offer many interesting possibilities. Careful thought and data collection is required to prove or disprove the validity of these various methods and the usefulness of the resulting scales if they are shown to be valid.

It is suggested that, in conjunction with other persons, several of the other scaling techniques described above be experimented with, including at least one multidimensional technique, and using both simulator runs and armchair methods (verbal descriptions) to provide pilot subjects realistic or vicarious working experiences.

4. A PROTOTYPE SUBJECTIVE RATING SCHEME FOR IFR PILOT WORKLOAD

4.1 INTRODUCTION

In this section we attempt to construct a subjective method of assessment for IFR pilot workload which is similar to the Cooper-Harper rating scheme for evaluating aircraft handling qualities[17], i.e. we will use a decision tree to direct the assessment process and use a set of words and ideas easily understood by skilled pilots. This rating scheme is intended to be used by subject pilots after flying a real or simulated phase of IFR flight. It can also be used by trained observers (normally other pilots) present on the flight deck, or alternatively, by subjects or observers watching a video replay of the flight phase.

The indicators of workload to be observed are the occurrences of more than one task being worked on simultaneously, the percentage of tasks interrupted by the arrival of a priority task, and the existence of idle periods between tasks. Notice that these are not indicators of task performance such as task errors or omissions, or delayed responses to tasks although these could also be observed. We want to observe directly the more frequent indications of the underlying level of workload causing the degraded task performances.

This workload rating scheme is called a prototype because it will undoubtedly be modified or completely revised as experience with its usage is obtained, and suggestions from subject pilots begin to be gathered. The Cooper-Harper rating has a genealogy which we expect must be repeated in

developing a workload rating scheme. Acceptance of the rating amongst a wide community of instrument-rated pilots is a necessary condition in establishing its usefulness.

4.2 TASK SCENARIO FOR IFR WORKLOAD ASSESSMENT

Flight Phase -- the assessment is intended to apply to some phase of a complete IFR mission, i.e., to some segment of an IFR flight such as a departure from a particular airport or some portion of that departure of at least 5 minutes duration.

Crew Position -- the assessment applies to an established crew position and set of normal procedures and practice which describe the taskload for that position. For a two- or three-man crew, assessments can be made of each position simultaneously.

Flight Systems -- the assessment is intended to apply to normal and abnormal configurations of aircraft flight controls, displays, and other aircraft systems. It is not intended to assess workload under unexpected emergency operations of short duration.

Flight Environment -- the flight conditions for weather and traffic are assumed to be specified; i.e., weather factors such as day/night, external visibility, and wind speed and gustiness; traffic factors such as proximity of other aircraft, message rate and percentage utilization of the communications channel.

4.3 TASK CATEGORIES

The categorization of discrete tasks into Operating, Monitoring, and Planning tasks described in Section 2.2 of this report must be understood by the assessor. A task analysis of the crew procedures for the flight phase should be undertaken by the assessor before experiments are performed wherein each expected task is categorized. While the terms suggest the characteristic of tasks in each category, the actual basis of categorization is "deferrability," i.e., the possibility of deferring the arriving task over some period of time.

Reviewing the definitions: "operating tasks" are non-deferrable and must be done as they arrive to maintain the expected flight performance and ensure flight safety. The category of "monitoring" tasks describes tasks which can be deferred for a short period such as one minute. All of the monitoring tasks fall in this category, but there are other tasks such as check lists which can also be included. The last category is called "planning" and describes all those tasks deferrable by more than a few minutes. As with the previous category, it includes all the planning and preparation tasks of IFR flight, but may also include others.

4.4 A SUBJECTIVE RATING SCHEME

One proposed prototype rating scheme is shown in Figure 19. The decision tree divides the ratings into four groups:

1) impossible
2) unacceptable
3) unsatisfactory, but acceptable
4) satisfactory

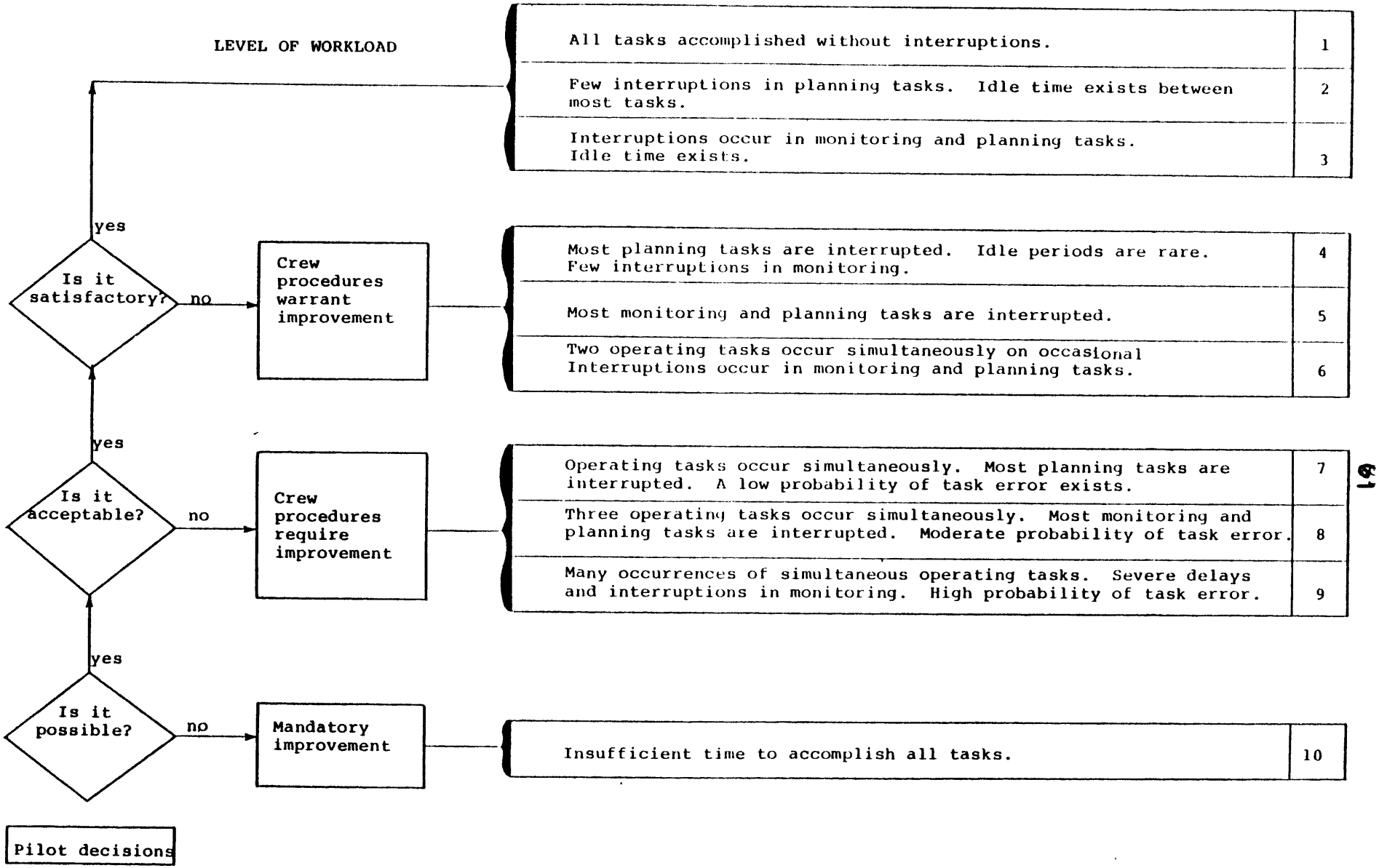


FIGURE 19

IFR Pilot Workload Rating Scale

If the rating is "impossible," then there is insufficient time to accomplish all the necessary tasks, and a rating of 10 is given. The group of ratings called "unacceptable" indicate the necessity to improve crew procedures and taskloads to ensure safety of operation since there is a probability of task errors and omissions due to task interruptions and crew loading. The next group of ratings are called "unsatisfactory" and describe workload levels which, although acceptable, are at such a level that many task interruptions are still occurring. Crew procedures should be improved if possible so that a better rating is achieved. The final category is called "satisfactory" and provides three ratings for pilot selection. In all three, some degree of idle periods should be observable.

For each rating, there is a brief description of the workload indicators to guide the pilot selection. The pilot or observer should be familiar with these indicators, and be trained to note their occurrence. It may be expecting too much of a pilot to have him perform, observe, and remember the occurrences of these indicators in a post-phase assessment without the use of a video-replay. A trained observer must be so familiar with the procedures (and perhaps the subject pilot) so that he knows the tasks being worked on at a given point in time. It might be feasible to have the subject pilot "think aloud" for the benefit of the observer. However the indicators are gathered, the description for each subjective rating selected should match the pilot's or observer's evaluation of the workload situation.

This rating scheme, and the words which make up the category descriptors, depends heavily on the idea of "interruption." There are other words and connotations upon which a subjective scale may be based. The next section discusses some alternatives.

4.5 ALTERNATIVE SUBJECTIVE RATING SCALES

In thinking about words associated with "mental workload" which might constitute category descriptors (especially for the lowest-level category descriptors on the right side of Figure 19), it appears that there are three kinds of words:

1. those associated with task time constraints, i.e., with behavioral activity
   a. fraction of the total time available which is required
   b. number of interruptions
   c. nearness of deadlines

2. those associated with decision making or supervision, task uncertainty and complexity of planning, e.g.,
   a. uncertainty about what the tasks are
   b. uncertainty as to "nominal" values of behavioral activity, performance and consequences associated with given task (B_i, P_i, C)
   c. uncertainty with respect to importance of consequences U(C)
   d. uncertainty with respect to what consequences will probably result from aircraft performance, C(P)
   e. uncertainty with respect to how pilot behavior will affect aircraft performance, P(B)
   f. uncertainty with respect to how mental work will affect one's own (pilot) behavior, B(M)
   g. the amount of planning required
   h. how far ahead planning must be done
   i. number of tasks from which to select

3. those associated with psychological stress, e.g.,
   a. level of bodily risk to passengers and crew
   b. level of social embarrassment, or intimidation or anger relative to crew, or passenger, or controller

c. psychological frustration of efforts to plan, decide and execute tasks
d. physiological impairment
e. generalized confusion
f. generalized anxiety

The above suggests a three-dimensional rating scale, since the phrases within each of the three groups seem closely related, while it is clear that the three dimensions -- fraction of time required, uncertainty and complexity of supervision, and psychological stress -- can occur in very different proportions.

Thus, in combination with the subjective scale proposed above, which is predicated primarily on the idea of time constraint, individual scales can be devised based on the other two dimensions (Figures 20 and 21).

Possibly these three, or some refinement on them, can be used in combination to produce a three-dimensional rating scale. Alternatively, one scale can be devised where the descriptor at each level is a combination of descriptors at the corresponding level of the three separate scales (e.g., with "or" connecting the three phrases in each case). In the latter case the user must decide for himself which descriptor dimension is most appropriate, since if he felt each of the three to be both appropriate and independent, he would not be able to make his mental workload judgments on a one-dimensional scale.

In the rating scales described above, the pilot assigns ratings based on the interpretation or connotation of the verbal descriptor of each separate category. A complete departure from this is the "utility" type scale described in Section 3.5D and 3.6C wherein a workload at some point along the scale is defined as equivalent to a 50-50 lottery

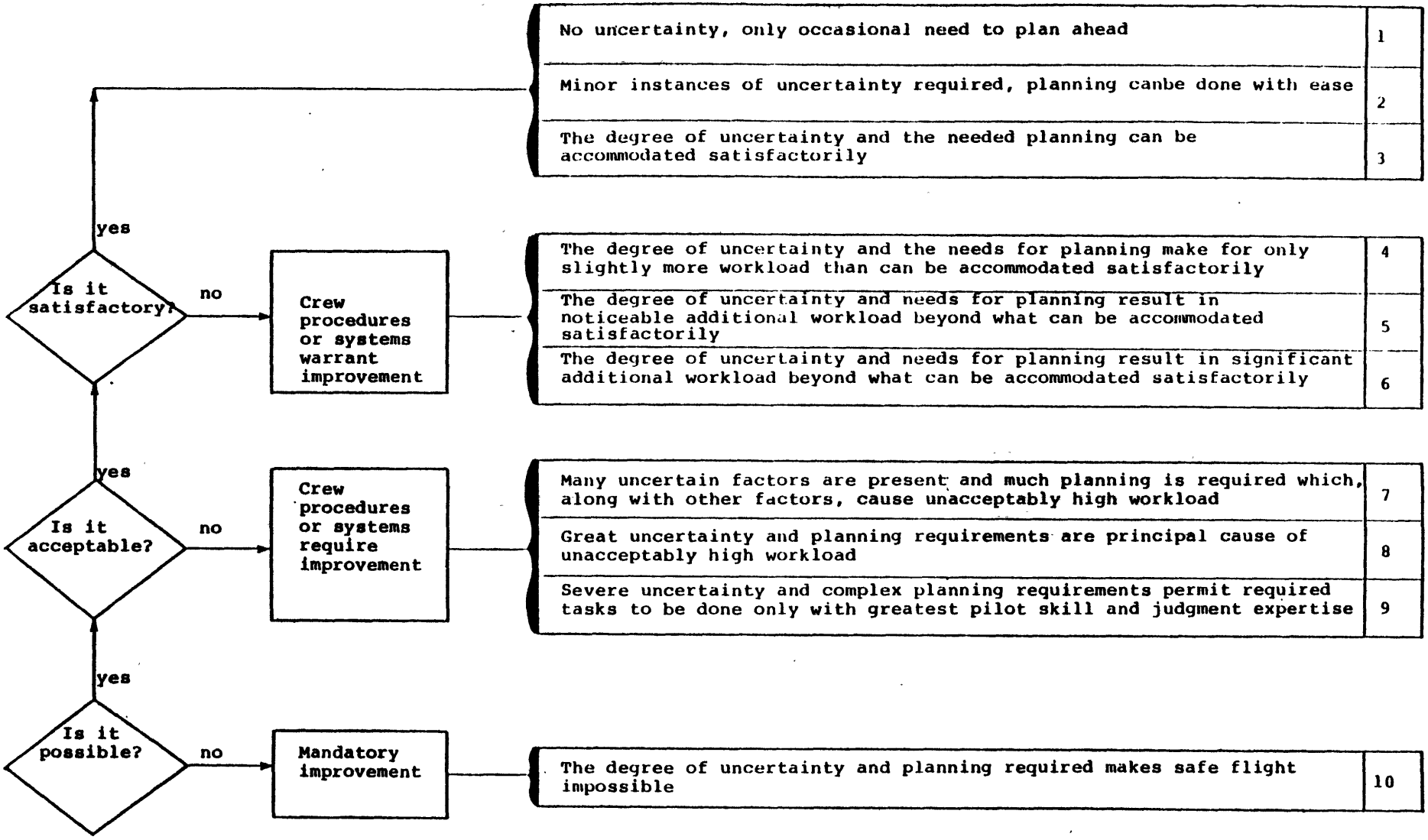


FIGURE 20

Rating Scale for Uncertainty and Planning Load

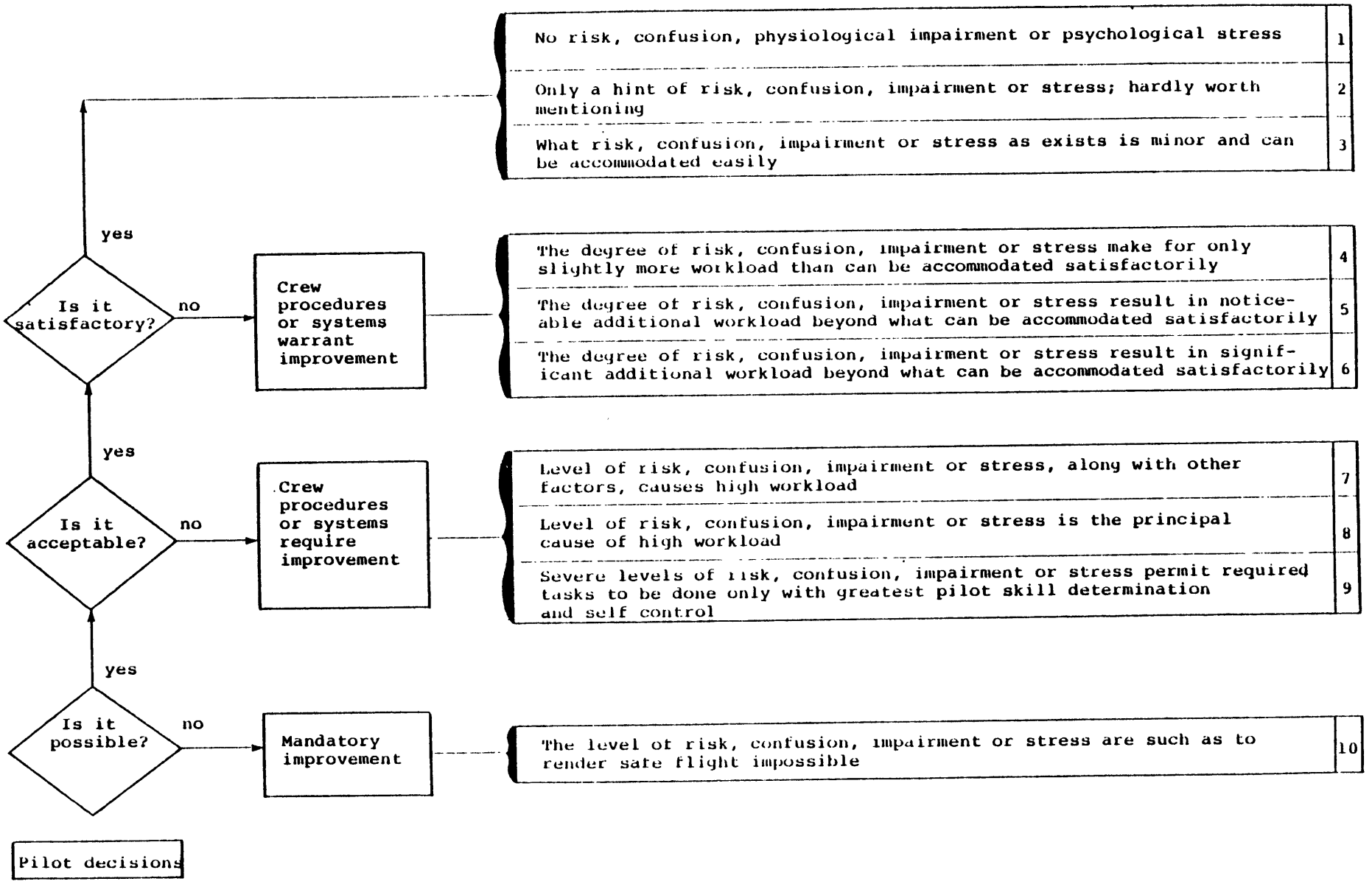


FIGURE 21

Rating Scale for Mental Stess Load

of a workload some number of levels up the scale and a workload at the same number of levels down the scale. That is, the pilot would be indifferent to flying continuously at 5 or flying half the time at 1 and half the time at 9. This method has the disadvantage of requiring the pilot to understand a somewhat more sophisticated technique. It has the advantage of being better "anchored," i.e., the numbers have more meaning.

In one-dimensional form it might be presented to the pilot in a form such as that shown in Figure 22. This, of course, could be expanded to the multi-dimensional form where, in lieu of "mental workload," would be "time load," "uncertainty and planning load" and "mental stress load", respectively. The three scales would be used independently to rate the various procedures and flight situations. Such "utility" ratings are most easily done after a whole battery of test flights is completed, since the rater must make comparisons between flight experiences.

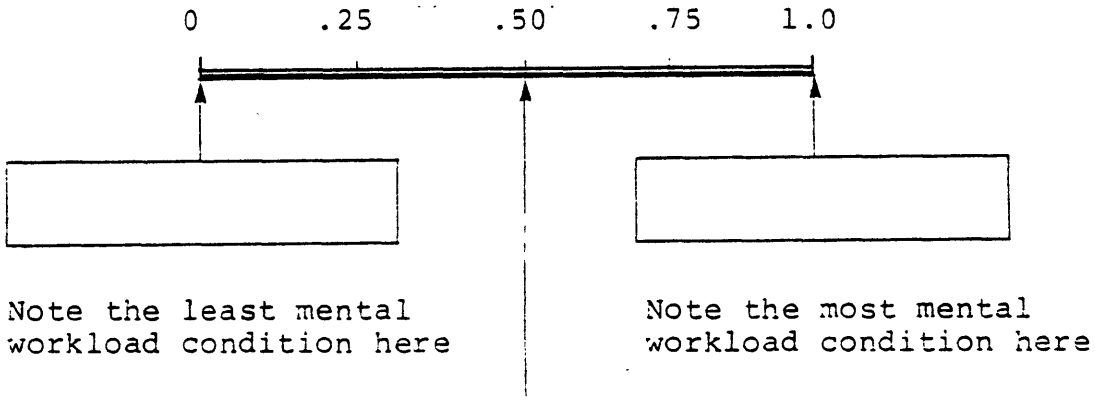


FIGURE 22

A Utility Scale for Pilot Workload

REFERENCES

1. Gartner, W.B. and Murphy, M.R., Pilot Workload and Fatigue: A Critical Survey of Concepts and Assessment Techniques. Adex Systems, San Jose, CA (no date).

2. Wierwille, W.W. and Williges, R.C., Survey and Analysis of Operator Workload Assessment Techniques. Report S-78-101, Systemetrics, Inc., Blacksburg, VA, Sept. 1978. See also Clement, W.F., Annotated Bibliography of Procedures which assess Primary Task Performance in Some Manner as the Basic Element of a Workload Measurement Procedure, Systems Technology Inc., Hawthorne, CA TR1104-2, Jan. 1978.

3. Review Manuscript of 1977 NATO/AGARD meeting in Cologne, F.R. Germany, to be published.

4. Moray, N., Editor, Mental Workload, Theory and Measurement, Plenum, N.Y., 1979 (Proceedings of 1977 NATO/Scientific Affairs Div. meeting in Mati, Greece).

5. (MIT Mental Workload Electronic Teleconference/ Journal Project.)

6. Norman, D.A. and Bobrow, D.G., "On Data-Limited and Resource-Limited Processes," Cognitive Psychology 7, 44-64, 1975.

7. Verplank, W.L., Is There An Optimal Work-Load in Manual Control?, Ph.D thesis, MIT, Dept. of Mechanical Engineering, August 1977.

8. Tulga, M.K., Dynamic Decision Making in Multi-Task Supervisory Control: Comparison of an Optimal Algorithm

to Human Behavior, Sc.D. thesis, MIT, Dept. of Mechanical Engineering, Sept. 1978.

9. Von Neumann, J. and Morganstern, O., Theory of Games and Economic Behavior, Princeton University Press, 1947.

10. Hart, S.G. and McPherson, D., "Airline Pilot Time Estimation During Concurrent Activity Including Simulated Flight," paper presented at 47th Annual Meeting of Aerospace Medical Ass'n, Bal Harbour, FLA, May 1976.

11. Gerathewohl, S.J., Brown, E.L., Burke, J.E., et al., "Inflight Measurement of Pilot Workload: A Panel Discussion," Aviation and Space Environmental Medicine, 49 (6), 810-822, 1978.

12. Personal communication, A. Ephrath, University of Connecticut, Storrs, CT.

13. Chiles, W.D., Chapter 5, "Objective Methods", in Assessing Pilot Workload, manuscript of 1977 AGARD meeting to be published.

14. Stevens, S.S., Psychophysics and Social Scaling (monograph), General Learning Press, Morristown, N.J., 1972.

15. Borg, G., "Psychological and Physiological Studies of Physical Work," in Singleton, W.T. et al., Measurement of Man at Work , London, Taylor and Francis, 1971, pp. 121-128.

16. Jenney, L.L., Older, H.J. and Cameron, B.J., Measurement of Operator Workload in an Information Processing Task, NASA CR-2150, Dec. 1972.

17. Cooper, G.E. and Harper, R.P., Jr., The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities, NASA TN-D-5153, April 1969.

18. Westbrook, C.B., Anderson, R.O., and Pietrzak, P.E., Handling Qualities and Pilot Workload, Wright-Patterson AFB Flight Dynamics Lab FDCC TM-66-5, Sept. 1966.

19. Thurstone, L.L., "A Law of Comparative Judgment," Psychol. Rev., 34, 1927, 273-286.

20. McDonnell, J.D., "Pilot Rating Techniques for the Estimation and Evaluation of Handling Qualities," Wright-Patterson AFB, Flight Dynamics Lab, AFF-DL-TR-68-76, Dec. 1968.

21. Stewart, T.R., West, R.E., Hammond, K.R., and Kreith, F., "Improving Human Judgment in Technology Assessment," J. Int'l. Soc. for Technology Assessment, June 1975, pp. 37-43.

22. Keeney, R.L., "Utility Functions for Multi-Attributed Consequences," Management Science, 18, 276-287, 1972.

23. Yntema, D.B. and Klem, L.,"Telling a Computer How to Evaluate Multi-Dimensional Situations," IEEE Trans. Human Factors in Electronics, HFE-6, 3-13.

24. Shepard, R.W., Romney, A.K. and Nerlove, S.B., Multidimensional Scaling, N.Y. Seminar Press, Vol. 1, 1972.

25. Carroll, J.C. and Chavy, J.J.,"Analysis of Individual Differences in Multi-dimensional Scaling via an N-Way Generalization of Eckert-Young Decomposition," Psychometrika, 1, 211-218, 1936.

26. Kruskal, J.B. and Carmone, F.J., How to Use M-D- SCAL (Version 5M) and Other Useful Information , Bell Telephone Labs, Murray Hill, N.J., 1969.