NASA-CR-166596

NASA CONTRACTOR REPORT 166596

NASA-CR-166596
19840026089

N84-34160

(NASA-CR-166596) THE WORKLOAD BOOK: ASSESSMENT OF OPERATOR WORKLOAD TO ENGINEERING SYSTEMS Final Report (Illinois Univ.) 24 p HC A02/MF A01 CSCL 05H G3/53

Unclas 22526

The Workload Book: Assessment of Operator Workload to Engineering Systems

D. Gopher

CONTRACT NAS2– NCC 2-233
November 1983

NASA

NASA CONTRACTOR REPORT 166596

The Workload Book: Assessment of Operator Workload to Engineering Systems

D. Gopher, Visiting Professor
Department of Psychology
University of Illinois
Urbana-Champaign, IL

Prepared for
Ames Research Center
under contract NCC 2-233

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

Final Report - NASA Grant NCC 2-233

The Workload Book: Assessment of Operator Workload to Engineering Systems.

Daniel Gopher,

Cognitive Psychophysiology Lab.

Univ. of Illinois at Urbana - Champaign.

The objective of this work is the writing of an integrative manuscript on the analysis and evaluation of workload in engineering systems. Assessment of the workload imposed on the operator in the performance of tasks has emerged as a major topic of concern in today's design and evaluation of the human interface with engineering systems. Modern airplane cockpits and air-traffic control units, among many other advanced engineering systems, are good examples for the significance of this issue in the operational environment. In the past three decades, intensive efforts have been dedicated both in basic research and in the applied domain to the study of the phenomenon of workload. However, little has been made to bridge the communication gap between these two lines of research. We believe that the main group that has suffered from this state of affairs is the professionals in the field who face the requirement to select their measurement techniques and justify their conclusions. By writing this manuscript, we hope to contribute towards narrowing the gap by providing a summary and integrative discussion of the main theoretical, methodological, and practical issues that have accompanied the development of the workload construct in human performance theory and research. A special attention is devoted to problem areas that bear direct relevance to human-factors engineering applications.

The manuscript is written by the author of this report in collaboration with Sandra Hart from the Man - Vehicle Research Division at NASA-Ames. A total period of 20 months has been requested for the completion of the writing, of which the first 8 months have been conducted at the Dept. of Psychology of the University of Illinois, and at NASA-Ames Research Center, during the sabbatical year of D. Gopher. The present report summarizes the major accomplishments during this period.

The first phase of the work was dedicated to the definition of the problem, development of an outline of the scope of the book, and a review of the literature. During this phase a major effort was made to compile, organize and review the main theoretical, methodological and applied references that have been published since the turn of the century. This work has led to the structuring of the chapters of the book, and to the establishment of a reference library on all aspects of the study of workload.

Along with the literature survey, writing has begun of the introductory chapter and the theoretical section. A first draft of these two sections has been almost completed during the 5-week summer visit of Dr. Gopher to NASA-Ames.

The introductory section discusses the main elements of the phenomena of workload, and lays the foundations of a general framework for a systematic treatment of this problem area. The theoretical section reviews the development of theory and research on the limitation of the human processing and response system. Special attention is given to those limitations that are assumed to result from the work of a central limited processing mechanism. The assumed existence of such a mechanism was the main trigger in the effort to model and quantify workload.

The chapter discusses the emergence of this concept from the early works on consciousness, through the information theory based models and the Post Second World War formulation of single channel capacity approaches. Structural and energy constraints on processing and response capabilities are contrasted, leading to a discussion of the present state of the art. Current notions of multiple resources, control and automatic processes, and functional organization of the processing system, serve to delineate a profile of dimensions that has to be pursued in the establishment of measurement procedures. The methodological section is generally divided into three sub-sections: Performance measures, physiological indices and subjective scales. A first draft of the main assumptions, advantages and disadvantages of each of these approaches has also been written during this period.

The summer stay at NASA has also been used to coordinate efforts with Sandra Hart. Special consideration was given to an outline of the main steps that has to be taken in order to render the book useful and appealing, both to the academic and the applied community. The strategy that was decided upon, is the pursuit throughout the book of several major practical examples, such as evaluation of driver workload as a result of increased speed. These examples are first introduced in colloquial terms, then task analysed and reformulated within the proposed theoretical framework. They will also be confronted with several measurement approaches, to demonstrate the use and expected outcome of each approach. In this way we hope to demonstrate the necessity of theoretical analysis for practical purposes on the one hand and, on the other hand, the importance of application for the enrichment and testing of theoretical thinking.

Another development during the first period of work on this grant has been the conduct of preliminary work to study a new approach to the

development of workload scales based upon subjective experience. Extensive effort has been directed in recent years to study the subjective experience of workload during the performance of tasks. This research is accompanied by several attempts to develop measurement scales to quantify the experience of workload, and relate it to task demands. In the review of this work for the purposes of the book, it became clear that although the general approach gains popularity quite rapidly, it is seriously lacking in its theoretical and measurement rational. Hence it was difficult to integrate it into the general framework of the book.

Our experimental work addressed this gap. We have proposed to model and treat the issue of constructing subjective scales of workload within the general problem area of psychophysical scaling. This approach can be defended on theoretical grounds and also benefits from the rich theoretical and methodological knowledge that has been accumulated in this research. To support our claims, test the feasability of the approach and examine its predictions, we conducted an experimental study. This study was performed at the Engineering Psychology Laboratory, in collaboration with Christopher Wickens and Rolfe Broune. It was also supported by a grant from the U.S. office of Naval Research. The results of this study provided an overwhelming support to our approach. They have been summarized in a paper that was presented at the Second Symposium on Aviation Psychology, which was held in April 1983, at Ohio State University, Columbia, Ohio. The paper was also published in the proceedings of this conference. A copy is enclosed with the present report, as a part of the final report.

In summary, the progress of work on the Workload book has met all the objectives that were outlined in the statement of work of the proposal.

In addition, experimental work has been conducted on a topic that was identified to be of prime importance to the study of workload. This work yielded encouraging results, and was received with interest by the scientific community.

Proceedings of the 2nd Aviation Psychology Symposium, Ohio State University, April 1983

ON THE PSYCHOPHYSICS OF WORKLOAD: WHY BOTHER WITH SUBJECTIVE MEASURES?

Daniel Gopher & Rolf Braune
University of Illinois

ABSTRACT

Psychophysical functions describe the relationship between variations in the amplitude of a defined physical quantity and the psychological perception of these changes. Examples are brightness, loudness, and pain. The regularities of these relationships have been recognized since the early days of experimental psychology, and have been formulated into psychophysical laws. The measurement methodology of psychophysical scaling has been refined by the Harvard group led by S. S. Stevens, who proposed a power function as a general form for such laws. The main argument of the present paper is that a similar scaling approach can be adapted to the measurement of workload and task demands based upon subjective estimates given by subjects. The rationale is that these estimates, like other psychophysical judgments, express the individual's perception of the demands imposed on him by the surrounding environment. This approach was successfully applied to the assessment of 21 experimental conditions given to a group of 60 subjects. The paper discusses the main results of this effort and their implication to theory and application in human performance.

The measurement of workload has emerged as a central topic of interest in current human performance theory. In addition to a general theoretical interest in this issue within the domain of cognitive psychology, it is also of much relevance to many applied problems in the domain of human factors engineering.

The hypothetical construct associated with the notion of workload has been employed as a generic term in a variety of situations to explain the inability of a human operator to cope with the requirements of a task that he is given to perform. In such instances, the task is argued to impose high "workload" in reference to the underlying processing and response capabilities of the human processing system. Within this framework, a workload measurement procedure is one in which an attempt is made to characterize the conditions under which task demands can or cannot be met by the performer. A workload measure is one by which the latter differences are expressed in relation to the overall ability of the human processing system to process information and generate responses. It is generally assumed that performance on tasks depends upon the deployment of processing facilities, and that there are upper limits on the rate at which the system can recruit its resources to accommodate task demands (Navon and Gopher, 1979, 1980).

Three main types of measurement approaches have been developed to evaluate workload. Within one approach, demands are expressed in terms of the

objective parameters of tasks (e.g., signal quality, information rates, number of response alternatives, etc.). A second class relies primarily on measures of response (either behavioral or physiological). Finally, in recent years an extensive effort has been dedicated to the development of a third measurement technique based upon the subjective appraisal given by the performer to the load experienced by him during task performance (for a review of the three approaches, see Moray, 1979, 1982; Williges and Weirwille, 1979; Ogden, Levine, and Eisner, 1979).

In principle, all three approaches represent alternative paradigms to the study of the same phenomenon, i.e., the relationship between the demands imposed on the human by the task (the environment), and his ability to cope with them. In practice, however, there is only sparse knowledge on the way in which measures obtained by one method are related to those obtained by another. Furthermore, considerable disagreement appears to exist between proponents of each method as to which provides a "better," or a more "valid" estimate of the underlying limitations.

The above brief introduction serves to place the discussion of subjective measures, which is the topic of interest in the present paper, within the general perspective of workload research. Subjective measures represent the conscious judgment of the performer of the difficulties encountered by him in the performance of the evaluated task. They are easy to obtain and have a very high face validity. This validity is, indeed, so compelling that it appears to have led several researchers to argue that "If the person tells you that he is loaded and effortful, he is loaded and effortful whatever the behavioral and performance measures may show" (Moray, Johanssen, Pew, Rasmussen, Sanders, and Wickens, 1979, p105), thus subordinating all other measures to the mundane truism of this statement.

With few variations, in recent years, this general philosophy has guided the development of several subjective measurement scales. Sheridan and Simpson have developed a general workload assessment version of the old Cooper-Harper rating scale which was originally developed for the description of the handling qualities of flight vehicles (Sheridan and Simpson, 1979). Hart, Childress, and Bartolussi (1981, 1982) and Bird (1981) have experimented with a variety of bipolar rating techniques. Wickens and Yeh (1982) and Derrick (1981) explored a different type of rating scale. Reid, Shingledecker, and Nygren (1981), Reid, Singledecker, and Eggemeier (1981), at the Wright-Patterson Air Force Laboratories, have devoted a considerable effort to the development of a Subjective Workload Assessment Technique (SWAT), based upon a conjoint measurement approach.

The overall outcome of these efforts is confusing, and, to some extent, disappointing. Human subjects appear to have no difficulty in assigning numerical values to their experience. However, the experimenter has the burden of selecting the appropriate dimensions for rating. In the absence of a formal theory of workload, informal intuitions have led experimenters to select different rating dimensions, a fact that greatly complicates any comparison between studies. In addition, techniques vary in their initial measurement assumptions and their resultant constraints on subjects' freedom in rating. Another disturbing outcome is that while variations within tasks

produced consistent changes in the subjective load profiles, no such consistency was found between tasks (e.g., Hart et al, 1981).

However, most annoying of all are the recurrent findings of dissociation between subjective estimates and objective measures of task performance. That is to say that in some instances a strong correspondence is found, while in others no relationships are revealed. Furthermore, in most cases reliable but low correlations are obtained (e.g., Wickens et al, 1982). It should be recognized that the main theoretical justification for instituting the workload concept, in the first place, was the desire to improve the ability to predict performance, given task conditions. What is the sense and what is the value of developing a workload measure that does not correspond or is only weakly related to the actual behavior of subjects?

The problems with subjective measures should be considered both from a theoretical and a methodological viewpoint. From a theoretical perspective, the issue is the nature and content of the conscious experience and its relationship to attention, information processing, and performance. Does the conscious experience, and hence the subjective measure, incorporate all the phenomenon of interest included in the notion of workload? A detailed discussion of this topic is beyond the scope of the present article. On the methodological level, the challenge is the development of a scaling approach that would impose few a priori constraints, that would enable quantification of subjects' experience, and that would allow comparisons to be made within and across tasks.

We propose to examine the psychophysical measurement theory as developed by S. S. Stevens (1957, 1966) as a viable candidate for such a scaling approach. We argue that if there is any basis to the assumption that the human information processing system invest or commits processing facilities to enable the performance of tasks, then subjective measures can be thought to represent the perceived magnitude of this investment, in much the same way that the perception of brightness is changed with manipulations of light intensity, and loudness reflects variations in sound pressure levels. The importance of this analogy is that it places the efforts to construct a subjective workload scale in the center of a rich body of methods, theory, and data that have a long history of scientific excellence.

In general, it was found that all psychophysical functions can be adapted to a power function of the form:

$$1.\quad P = KI^e$$

or in its log form

$$2.\quad \log P = \log K + e \log I$$

Where: P = psychological estimate
I = the physical quantity
e = the exponent
k = a unit scalar that depends on the range of numbers used.

Different physical functions were shown to change only the size of the exponent. In addition, subjects were shown to be able to compare units across modalities, with a resultant exponent for the comparative power function that is a proportion of the within modalities exponents. In a psychophysical experiment, it is conventional to employ one quantity as a modulus or a reference point, and ask subjects to assign values to other quantities relative to this point. Subjects are not restricted in their selection of values and are free to select any number that best represents the differences between two conditions.

We applied this approach to the study of workload estimates given to 21 experimental conditions. A one-dimensional second-order tracking task was employed as a reference task. This experiment was conducted as part of a more general study aimed at investigating individual differences and age effects on performance.

METHOD

Tasks and experimental conditions.

Subjects were given a total of 21 experimental conditions composed of 14 single and 7 dual task conditions. The selection of tasks was guided by the multiple resource paradigm suggested by Wickens (1980). They were designed to represent the following facets: perceptual motor control, short term memory, verbal and spatial abilities, selective and divided attention, and time sharing capabilities. The battery of tests included: (1) a one-dimensional 2nd order compensatory tracking, (2) a critical tracking task, (3) a delayed digit recall (2 back), (4) seven variations of the Sternberg memory search task. These were auditory stimulus-verbal response with memory set size of 2 and of 4, visual-verbal 2 and 4, visual-spatial 2 and 4, and auditory-spatial 2, (5) card mental rotation, (6) hidden figures, (7) maze tracing, (8) Gopher dichotic listening task. The seven dual task conditions were combinations of the tracking task with the first six variations of the Sternberg task and the delayed digit recall. For a detailed review of these tasks, see Braune and Wickens (in preparation).

Apparatus

The experiment was performed at the Engineering Psychology Research Laboratory of the University of Illinois. A PDP 11/40 minicomputer was used to generate the stimuli and record the subjects' objective performance. The computer was interfaced with a Hewlett-Packard display generator and a Measurement System, Inc. Model 521 control stick. Auditory stimuli were generated by a Centegram Corporation Mike-2 unit, interfaced to the PDP 11/40. Subjects sat in a sound and light attenuated booth approximately 90 cm from a Hewlett-Packard Model 1300 CRT. The CRT was used to present all of the visual stimuli to the subjects. The only task that was not computer-generated was the Dichotic Listening Task. This task had previously been recorded in a professional recording studio. It was copied onto a stereo-cassette and was played to the subjects via a stereo-cassette player. The subjects received the messages through a headset with different messages sent to the left and right ear simultaneously. The subjects had to record their responses on a recording sheet.

Procedure

The 21 experimental conditions were administered 3 times in a single experimental session that lasted about 4 hours. It was divided into halves, with a 45-minute break between halves. During the first half, subjects were given 1 minute practice on all tasks, followed by a two-minute test. On the dichotic listening task, the first half of the test was performed. In the second half of the meeting, each condition was performed for 3 minutes, and the second half of the dichotic listening task was given.

Following the performance of each task, subjects were required to give a number that would express the load or the demand imposed on them by the task. In accordance with the conventions of psychophysical scaling methodology, the single dimension tracking task was identified as the modulus or reference task. It was assigned the value of 100. Subjects were thus required to estimate the load of the currently experienced task relative to the value that was given to tracking. The experimenter strongly emphasized their liberty to select any number and range of values that would best represent their judgment. Any further clarification questions on the nature of the required judgment were skillfully evaded. In addition to these estimates that were given following the actual performance of each task, at the end of the first and second halves, subjects were also asked to reevaluate all of the tasks in one instance. We, therefore, obtained 5 ratings for each experimental condition. When giving their estimates, subjects were not allowed to see their former ratings.

Subjects

Sixty males from the Champaign-Urbana community, between the ages of 20 and 60, served as subjects. The subjects were all volunteers that had responded to ads in local newspapers. All reported to be in good health with 20/20 corrected vision and normal hearing. Each subject was paid $3.00 per hour.

RESULTS

In this study, as in previous studies with subjective measures, subjects had no difficulty in complying with the request to assign a numerical value to the perceived demands imposed by tasks. However, because they were given the freedom to select their own numbers (aside from the reference task which was given the value of 100), they varied widely in their values and ranges. Some subjects limited their estimates to a range between 70 and 150, while others employed the whole range between 0 and 800. To facilitate the comparison between subjects and tasks, the values given to tasks in each of the 5 rating instances were rescaled separately within the range of scores given by each subject in each rating instance, based upon the following formula:

$$3.\quad Xt(i) = \frac{X(i) - (Xmin-1)}{Range}$$

Where: $Xt(i)$ = transformed score of task i
$X(i)$ = raw score i
$Xmin$ = minimum value given to a task

Range = overall range of the 21 tasks

Note that this transformation rescales all the values given by subjects within a range of 0-1, without changing the original distances between tasks. In all of our preliminary analyses, we employed both raw and transformed scores, thus having a total of 10 scores for each subject on each test.

Average perceived load of the 21 tasks.

In the present paper, we limit the description and discussion of results to the average values of perceived load obtained for each of the 21 single and dual task conditions. Hence, unless otherwise indicated, the scores of individual tests always represent an average across the whole sample. Table 1 summarizes these averages (raw and transformed scores) for all tasks in the first and second test periods, in which subjects were required to give their load estimate immediately following task performance. These two periods correspond to the beginning and the end of the whole experimental session. Table 2 presents the intercorrelation matrix between the 10 subjective estimates of the tasks. As can be observed, the correlation

TABLE 1 -- Average Raw and Transformed Scores for the 21 Experimental Conditions

Experimental Task	Test I		Test 2	
	Raw Sc.	Trans-formed	Raw Sc.	Trans-formed
1. Sternberg, Visual-Verbal 2	54.5	.057	59.0	.052
2. Stern. Auditory-Verbal 2	60.5	.081	62.3	.069
3. Stern. Visual-Verbal 4	65.9	.111	79.3	.166
4. Stern. Auditory-Verbal 4	75.0	.139	80.7	.172
5. Stern. Visual-Spatial 2	77.5	.150	83.8	.184
6. Hidden Pattern	95.2	.223	84.3	.180
7. Card Rotation	118.8	.315	97.8	.242
8. 1-dimens. compens. tracking	100*	.283	100*	.302
9. Maze Tracing	120.4	.325	113.6	.329
10. Stern. Auditory-Spatial 2	148.1	.415	115.6	.344
11. Stern. Visual-Spatial 4	118.4	.334	118.1	.368
12. Critical Tracking	109.2	.290	122.8	.424
13. Dual, trac. & Aud.-Verb.2	163.8	.488	132.4	.431
14. Dual, trac. & Vis.-Verb.2	156.9	.481	132.8	.442
15. Delayed digit recall	156.6	.465	141.5	.473
16. Dichotic Listening	175.0	.505	146.7	.482
17. Track & Vis-Verb.4	170.1	.523	146.0	.512
18. Dual, Trac. & Aud.-Verb.4	170.4	.514	150.6	.516
19. Dual, Trac. & Vis.-Spat.2	184.6	.576	144.6	.490
20. Dual, Trac. & Vis.-Spat.4	212.5	.686	176.7	.667
21. Dual, Trac. & Delayed Digit	280.9	.921	243.8	.894

N = 55

* the reference task

Table 2

Intercorrelation Matrix between the Average Subjective Scores of the 21 Conditions, Based Upon the 5 Evaluations and Their Transformation

		1	2	3	4	5	6	7	8	9	10
Practice	1	.X									
Test 1	2	.98	X								
Reevaluation 1	3	.95	.98	X							
Test 2	4	.95	.98	.99	X						
Reevaluation 2	5	.92	.96	.98	.99	X					
Trans. Practice	6	1.00	.99	.96	.95	.92	X				
Trans. Test 1	7	.97	1.00	.98	.98	.96	.98	X			
Trans. Reev. 1	8	.95	.98	1.00	.98	.98	.96	.98	X		
Trans. Test 2	9	.94	.97	.99	.99	.98	.95	.97	.99	X	
Trans. Reev. 2	10	.90	.95	.98	.98	.99	.91	.95	.98	.98	X

N = 55

coefficients between all evaluation instances are very high. They do not differentiate between those instances in which the subjects were given the opportunity to estimate all the 21 tasks together (reevaluation 1, 2) to those in which an estimate was given following the actual task performance (practice, test 1, 2). Similar levels of correlations are revealed for raw estimates and transformed scores, despite the large individual variability in the selected range of numbers. Note that these correlations do not imply that subjects did not change their estimates from one instance to the other, but rather that if such changes occurred they preserved the relative position of the different tasks.

To further test the consistency of the emerging task profile, a split half reliability test was conducted on the task averages. The correlations between task scores obtained in the two halves of the sample are presented in Table 3. None of them is lower than 0.96, indicating the high reliability of the obtained estimates. Given this consistency, it was decided to concentrate on the data obtained in tests 1 and 2 (Table 1), which represent the most direct estimates of task experience early and late in the experiment. Several aspects are worth noting in Table 1. One is the previously indicated consistency in the order of tasks across the two tests. It is evident in both the raw and the transformed scores, but it is more apparent in the latter. All estimates of dual task conditions are higher than their single task components. On the transformed scores, the actual values of dual task conditions are higher than the additive values of their components. These relationships are not maintained in the raw averages. The raw averages are biased in favor of subjects who used a wide range of numbers. The reference tracking task is located on both tests at about the 30% point of the range.

There is an interesting shrinkage of range between the two rating instances in the load values given to tasks. This reduction is primarily due

TABLE 3

Split Half Reliability for the
Average Scores of the 21 Conditions *

Raw Scores		Transformed	
Practice	.98	Practice	.98
Test 1	.96	Test 1	.98
Reevaluation 1	.97	Reevaluation 1	.98
Test 2	.97	Test 2	.98
Reevaluation 2	.97	Reevaluation 2	.97

(N first half = 29; N second half = 26)

* Correlations computed between the average scores obtained for the 21 conditions in each of the two halves of the sample, on each of the subjective measures.

to the lower values given to dual task conditions, although there is also some elevation of tasks at the lower end. It is as though dual task conditions were perceived as less demanding at the end of the experiment. To test the power and reliability of this observation in the performance of individual raters, paired comparison tests were conducted on the lower, upper and range values employed by subjects on the two tests. This analysis is summarized in Table 4. All the t tests are highly significant, but the shrinkage from the top is greater by a factor of 5 from the elevation from the bottom.

TABLE 4

Changes in the Minimum and Maximum
Values, and the Range of Scores Used by
Subjects Early and Late in the Session

(Test 1 vs Test 2)

		Test 1	Test 2	Diff.	t(59)	P
Minimum	$\bar{x}$	41.38	49.05	+ 7.67	3.366	< .0013
	SD	27.1	30.3			
Maximum	$\bar{x}$	298.68	260.42	-38.27	5.644	< .00005
	SD	146.13	146.79			
Range	$\bar{x}$	257.3	211.37	-45.93	6.236	< .00005
	SD	145.86	143.34			

To summarize this section, we have shown that consistent profiles of the perceived load of tasks can be obtained despite the freedom given to subjects in the selection of values and logic of rating, the multitude of tasks, and their vast heterogeneity. In addition, a consistent differentiation was found between single and dual task conditions. Also, there was clearly a practice effect.

Derivation of the psychophysical power function.

We can now turn to consider the feasibility and merit of an attempt to construct a psychophysical function (as in equation 1) from the present data. This attempt is based upon the transformed scores of the tasks. In a regular psychophysical experiment, the experimenter has a knowledge of both the physical quantity and its perceived value. He can then derive the values of the exponent and the scalar.

In the present experiment, we do not know the values of the underlying driving function. We only know the perceived load of each task. The hypothetical amount of invested resources (I), the exponent (e), and the scalar (k), are unknown. Altogether, we have a set of 21 equations, composed of 14 single tasks and 7 dual tasks that are combinations of tracking and seven of the single tasks. It is also assumed that the values of k and e are the same in all equations and that only the value of I is changing.

To be able to solve the equations for the values of k and e, one more assumption was made. The Sternberg auditory-verbal memory search tasks, with set size 2 and 4, were assumed to be one unit apart on the hypothetical resource investment function (and hence were given the actual values of 1 and 2). The rationale for this decision was that these are highly familiar and compatible tasks that were consistently located by all subjects at the lower end of the scale. The selection of the actual numerical values was influenced by the information theoretic analysis that would assign to these tasks 1 and 2 bits of information respectively.

Once the above assumption has been made, the equations for the two auditory tasks on the two tests could be solved for the values of k and e. These in turn were used to derive the resource units of the other single tasks. The resultant functions for tests 1 and 2 (in log-log units) are depicted in Figure 1. The differences in slopes reflect the presumed change due to practice between these two evaluation instances.

One question is, of course, what is the merit of these equations above and beyond the consistencies that have already been shown. A second question is can we test the validity of the derived values (and by that support the underlying assumptions)? A conclusive answer to these questions should await the collection of more experimental evidence. However, suggestive data has been obtained in several analyses that are described below. The question posed in the first analysis was, can the values derived for single tasks, based upon these equations, be used to predict the perceived load of dual task conditions that are composed of these single task conditions. There are several possible combination rules for I units derived from the single tasks. We started with the simplest and most powerful con-

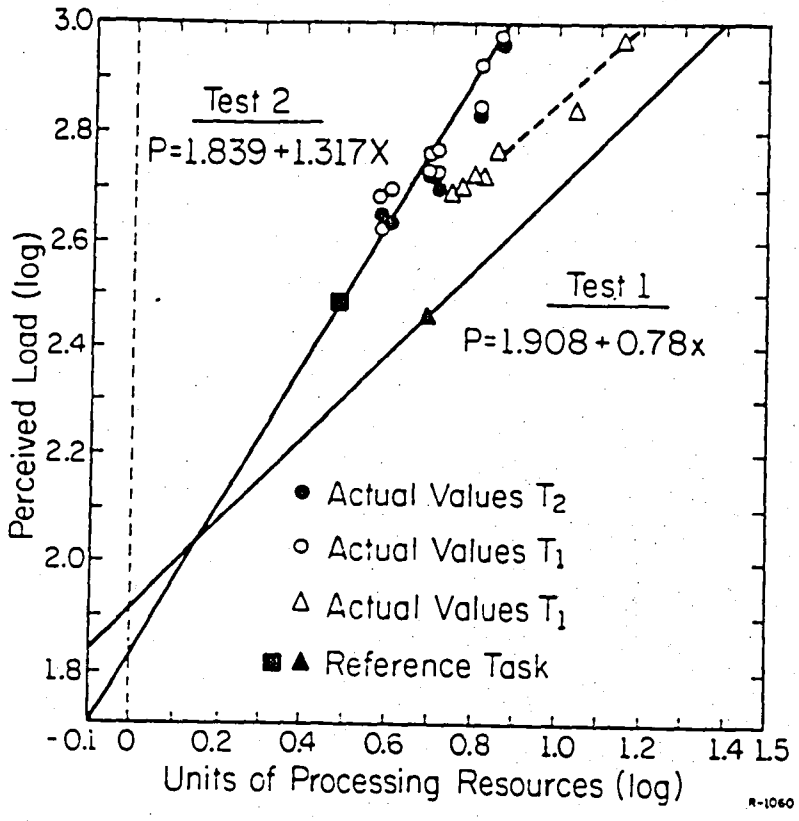

Fig. I - Power functions describing the relationship between perceived load (transformed scores) and the underlying resource function, for the 21 conditions.

straint, the additive rule. For example, the derived units of processing for tracking in test 2 was 3.07, and for delayed digit recall, 4.31. The equation for the dual task condition in which they were combined was then written as follows:

$$4.\quad \log P' = 1.839+1.317\log(3.07+4.31)$$

$$\log P' = 2.982$$

The predicted value obtained from equation 4 can now be compared with the actual value obtained from the estimates given by the subjects. This value was 2.951. Note that the predicted value was obtained by simply adding the values of the single tasks. This procedure was repeated for the seven dual task conditions of the two tests. The respective correlations between actual and predicted load estimates were 0.98 for test 1 and 0.95 for test 2 ($P < .01$ in both cases). The actual data points are plotted in Figure 1 along with the solid function lines that represent the predicted values at that point.

The actual values obtained in test 1 are plotted twice, once as related to the test 1 equation (empty triangles), and once in relation to the test 2 equation (empty circles). It was done to illustrate the finding that although a high correlation was found between actual and predicted measures based upon the test 1 equation, all actual values were shifted upwards by a constant, as though there was an added perceived constant cost of concurrence for dual task performance. These extra costs disappeared in the estimates obtained in test 2 and were eliminated if test 1 results were predicted from the equation of test 2.

Relation to task parameters and performance measures.

To assess the relationship between the average load profile of the 21 tasks based upon subjective estimates and measures of task demands based upon task characteristics or subjects' performance, two preliminary comparisons were conducted. In the first, subjective measures were correlated with task scores obtained from an index of difficulty suggested by C. D. Wickens from the University of Illinois. In the second, these scores were correlated with the average response times of all tasks that had a reaction time score as their prime performance measure.

Task difficulty scores based upon Wickens index of difficulty represented an unweighted sum of scores on four dimensions: (1) Familiarity of stimuli (0=letters, 1=spatial dot patterns, tracking cursor), (2) Number of concurrent tasks (0=single, 1=dual), (3) Task difficulty (0=memory set size 2, 1=set size 4, 2nd order tracking, delayed recall), and (4) Resource competition (0=no competition, 1=competition for either encoding or control processes, 2=competition for both). These dimensions were used to score 15 of the 21 tasks (all of the single tasks that were also performed in dual task conditions, as well as their dual combination, see table 1).

Pearson product moment correlation coefficients were computed between these 15 scores and test 2 subjective measures. The correlation with the average transformed scores was 0.90, and 0.93 with the units derived from the power function (81 and 86 percent variance accounted for respectively, in both cases $p < .001$). Figure 2 depicts the relationship between the index of difficulty scores and the inferred units of resource demands for the 15 tasks.

In contrast to the high correlations with the index of difficulty, the correlations with reaction time measures were considerably lower. The coefficients were 0.29 with the transformed scores of test 2 and 0.30 with the scores derived from the Power function of this test (both correlations are non significant). These correlations, however, were computed only for the 11 single task conditions in which response time was the main dependent measure. They should therefore be treated with caution, and at best regarded as a first rough exploration of the relationship between these two types of measures of task demands.

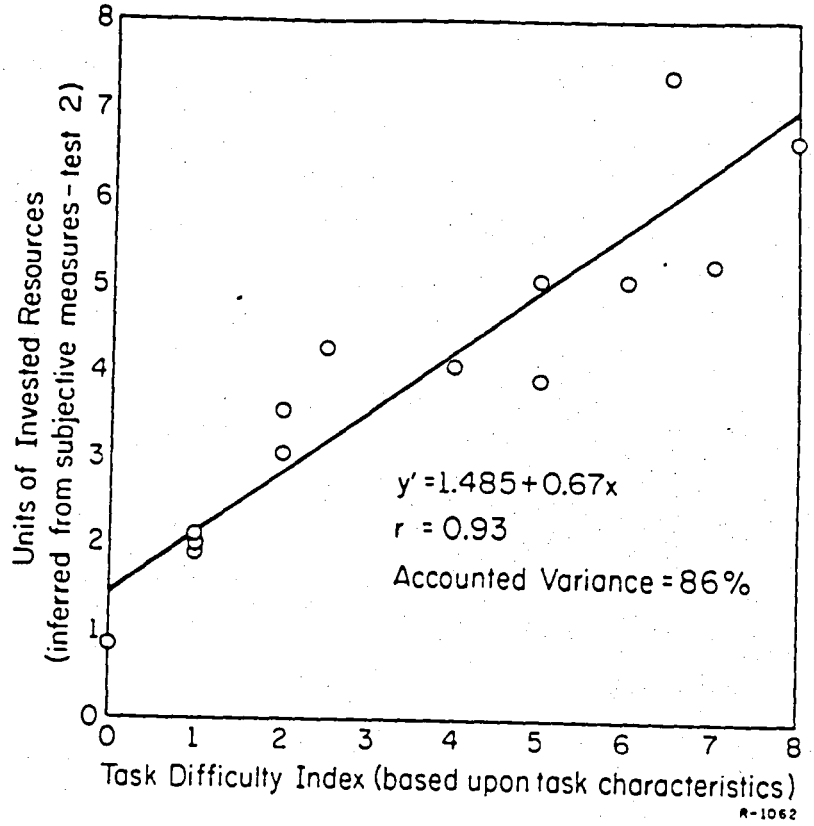

Fig. II - The relationship between the index of task difficulty and units of invested resources derived from subjective estimates.

DISCUSSION

Undoubtedly, the most important finding of the present experiment is the highly consistent profile of tasks that has emerged in response to the very loose requirement to express task demands by a numerical value. The reliability of this outcome has been underlined by the high correlations between rating instances, and the results of the split half reliability test. Its significance is accentuated in light of the diversity of the task battery, and the heterogeneity of the subject population. Regardless of the criteria employed by subjects, or the theoretical framework selected to interpret them, and irrespective of individual uniqueness in the range or method of assigning numbers, the discernable consistency argues for the presence of a coherent and potent psychological attribute. This attribute can be quantified in many levels, and is related to the characteristics of the experienced environment.

We shall name this attribute "resource requirements" which at this point have no meaning beyond the fact that the respective numerical values were generated in response to a request to evaluate task demands. We have observed the following phenomena in the assignment of values to tasks:

a) Subjects do not experience any difficulty in estimating tasks despite the differences in stimulus and response modes, and the variability in mental operations and transformation requirements.

b) Dual task conditions were rated as higher than the additive values of their components.

c) Replications and practice had ordered effects on the rating, such that difficult tasks were decreased, while easy tasks remained largely unchanged. These changes occurred as a unified constant factor across all difficult task conditions, and appeared primarily in dual task situations.

d) By constructing a psychophysical power function, it was possible to predict dual tasks perceived loads from the derived scores of the single tasks by applying a simple additive rule. A constant cost of concurrence was present early in training, but disappeared in later estimates.

e) Estimates of resource requirements correlated highly with an index based upon the processing characteristics of tasks.

f) Initial analysis showed that these estimates only showed low correlation to measures of task performance.

Taken together, the present results support the initial argument that psychophysical scaling may provide an adequate measurement approach to quantify the subjective experience of workload. It is usually the average load profile of tests that is of main interest to the applied researcher, and it was shown that tasks can be meaningfully compared across diverse parameters and conditions. Such a comparison was a major problem in previous subjective measurement approaches. Furthermore, with a psychophysical function one is allowed to take one more step to compare the relative distances between tasks, a comparison that could not be justified by other measurement approaches. An important finding was the ability to predict the perceived load of dual task conditions from single task values based upon the same power function. Along the same vein was the demonstrated sensitivity of the subjective magnitude estimate to the effects of practice. Finally, the higher correlation that was obtained between Wickens' index of difficulty and the inferred unit, as compared with the direct correlation of the perceived load values, supports the "power law" correction that was introduced in relating this index of difficulty to the underlying units of invested resources.

While the general merit of the methodological approach has been demonstrated in several ways, the details of the actual technique are still unclear. For example, do we have here a single driving function, such that all tasks map to the same function, or does the observed function represent the joint resolution of many cross modality matchings? What happens if a different reference task is selected? What are the results of selecting

different pairs of tasks to help the derivation of the exponent and the constant? These are some of the obvious questions that should be approached in future research. At the same time, it appears that based upon the present results this research can be looked upon as a refinement of method rather than an exploration of basic feasibility.

From a theoretical perspective, the present results raise several issues that cannot be addressed in detail in this article. Briefly, the subjective estimates data argues very strongly for the most strict model of a single undifferentiated pool of resources (e.g., Kahneman, 1973). Subjects appear to be able to use a single scale to evaluate all tasks, despite their huge diversity in modalities, mental operations, and response modes. In addition, the simplest additive model was sufficient to predict dual task conditions from single task units, and concurrence costs were a constant additive factor across all task combinations. No indications were found to show that different task combinations interact or parallel more or less with one another, nor was there evidence to argue that some tasks may compete with each other for common resources while others do not. The difficulty of the individual tasks was all that matters, and practice operated as a single unified factor across the whole difficult tasks domain.

Can one conclude from these results that the information processing system, in toto, behaves like a single capacity mechanism? Not at all. It is the conscious apparatus that appears to follow the pattern of a single channel. Throughout the history of experimental psychology there seems to be a recurrent confusion between the constructs of attention and consciousness (see Underwood, Geoffry, and Stevens, 1979; Posner, 1980, for a review). These two should be distinguished. Do we want to argue that all the phenomena of interest in the information processing mechanism which is linked with the notion of workload is also admitted to consciousness? Our consciousness would be eternally overloaded if bombarded by all the details of our mental life that are so rich and diverse. Clearly we are only aware of a part of this activity. In the case of workload, the nature of this part is hinted to a certain extent by the high correlation between the subjective measures and the index of task difficulty based upon the general characteristics of the performed tasks. The incomplete coverage of the phenomena of interest was exemplified by the low correlation between these estimates and the response time measures. One should always bear in mind that a better prediction of actual behavior from knowledge of task conditions is the ultimate objective of psychological modelling. The task of gaining a better understanding of the nature of the different parts of the information processing mechanism, their significance, and their relationship to each other, remains a topic for future investigation. However, the present study emphasizes the crucial importance of such research for the understanding of the basic phenomena, and provides a strong guide to the direction of productive research.

In conclusion, as in many other experiments, we have raised more questions than we have answered. We still do not know why one should bother with subjective measures, but we have very clearly shown they are well worth the bother.

ACKNOWLEDGMENT

The present work was supported in part by a NASA grant, NCC 2-233, Ms. S. Hart at Ames Research Center is the scientific monitor of this grant and in part by contract N00204-82-C-0113 from the Naval Aerospace Medical Research Laboratory at Pensacola. D. J. Owens is the scientific monitor of this contract. The authors are indebted to Dr. C. Wickens and Ms. S. Hart for their invaluable suggestions and helpful comments.

REFERENCES

Bird, K. L. Subjective rating scales as a workload assessment technique. Proceedings of the 18th Annual Conference on Manual Control, Los Angeles, 1981.

Childress, M. E., Hart, S. G., & Bertolussi,M. R. The reliability and validity of flight task workload rating. Proceedings of the 26th Annual Meeting of the Human Factors Society, 1982.

Derrick, W. The relationship between processing resource and subjective dimensions of operator workload. Proceedings of the 25th Annual Meeting of the Human Factors Society, 1981.

Hart, S. G., Childress, M. E., & Bartolussi, M. Defining the subjective experience of workload, Proceedings of the 25th Annual Meeting of the Human Factors Society, 1981.

Kahneman, D. Attention and Effort, Prentice-Hall, 1973.

Moray, N. (Ed.), Mental Workload: Its Theory and Application. New York: Plenum Press, 1979.

Moray, N. Subjective mental workload, Human Factors, 1982, 24, 25-40.

Moray, Johanssen, Pew, Rasmussen, Sanders, & Wickes. Report of the experimental psychology group. In Moray, N. (Ed.), Mental Workload: Its Theory and and Measurement. New York: Plenum Press, 1979.

Ogden, G. D., Levine, J. M., & Eisner, E., Measurement of workload by secondary tasks, Human Factors, 1979, 21, 529-548.

Posner, M. I., Mental chronometry and the problem of consciousness. In Jusczyk, D. W. & Klein, R. W. (Eds.), The Nature of Thought: Essays in the Honor of D. O. Hebb. Hillsdale, N. J.: Erlbaum Associates, Inc., 1980.

Reid, G. B., Shingledecker, C. A., & Eggemeier, T. F., Applications of conjoint measurement to workload scale development, Proceedings of the 25th Annual Meeting of the Human Factors Society, 1981.

Reid, G. B., Shingledecker, C. A., Neggren, T. E., & Eggemeier, T. E., Development of a multidimensional subjective measure of workload, Proceedings of the International Conference on Cybernetics and Society, Atlanta, Georgia, 1981.

Sheridan, T. B. & Simpson, R. W., Towards the definition and measurement of the mental workload of transport pilots, MIT, Cambridge, MA, Man-Machine System Laboratory, 1979.

Stevens, S. S., On the psychophysical law, Psychological Review, 1957, 64, 153-181.

Stevens, S. S., On the operation known as judgment, American Scientist, 1966, 54, 385-401.

Underwood, Geoffrey & Stevens, Aspects of Consciousness. Vol. 1, Psychological Issues. Academic Press, 1979.

Wickens, C. D., The structure of attentional resources, In R. Nickerson (Ed.), Attention and Performance VIII, Hillsdale, N. J.: Erlbaum Associates, Inc., 1980.

Wickens, C. S. & Yeh, Y. V., The dissociation of subjective rating and performance. IEEE Trans. of Sys. Man and Cyber., 1982, 600-603.

Williges, R. D. & Wierwille, W. W., Behavioral measures of aircrew mental workload, Human Factors, 1979, 210, 549-574.

End of Document