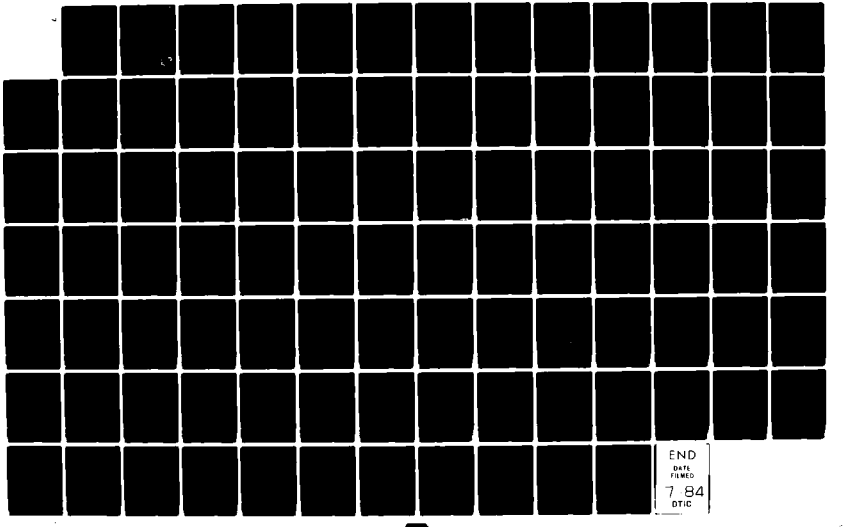
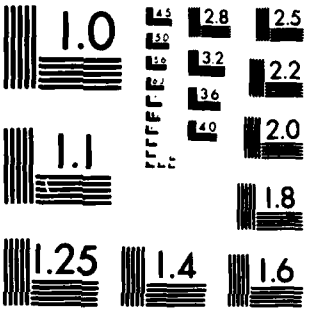
AD-A141 455 A MULTIPLE PROCESSING RESOURCE EXPLANATION OF THE SUBJECTIVE DIMENSIONS O..(U) ILLINOIS UNIV AT URBANA ENGINEERING-PSYCHOLOGY RESEARCH LAB W L DERRICK ET AL. 1/1

UNCLASSIFIED FEB 84 EPL-84-2/ONR-84-1 N00014-79-C-0658 F/G 5/8 NL



END
DATE FILMED
7-84
DTIC



ENGINEERING-PSYCHOLOGY RESEARCH LABORATORY

University of Illinois at Urbana-Champaign

TECHNICAL REPORT EPL-84-2/ONR-84-1

FEBRUARY 1984

A Multiple Processing Resource Explanation of the Subjective Dimensions of Operator Workload

William L. Derrick

Christopher D. Wickens

Prepared for
Office of Naval Research
Engineering Psychology Program
Contract No. N-000-14-79-C-0658
Work Unit No. NR 196-158

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER EPL-84-2/ONR-84-1	2. GOVT ACCESSION NO. AD-A144455	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Multiple Processing Resource Explanation of the Subjective Dimensions of Operator Workload		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) William L. Derrick Christopher D. Wickens		8. CONTRACT OR GRANT NUMBER(s) N000-14-79-C-0658
9. PERFORMING ORGANIZATION NAME AND ADDRESS Dept. of Psychology, University of Illinois 603 E. Daniel St. Champaign, IL 61820		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 196-158
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research, Eng. Psych. Program 800 N. Quincy St. Arlington, VA 22217		12. REPORT DATE February 1984 13. NUMBER OF PAGES 85
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Mental workload, task analysis, performance, physiological subjective ratings, multi-dimensional scaling, attention, multiple resources

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Multiple measures of operator workload may dissociate, or fail to agree, for a given task. The goal of this study was to determine which task difficulty (workload) as indexed by attentional resource demand could explain the attendant variance in a second index of workload, subjective ratings. A multiple resource model of processing resources (Wickens, 1980) guided construction of tasks of differential resource demand. These tasks were both performed by subjects and rated according to workload

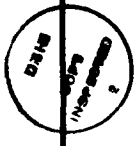
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

similarity. Scaling and clustering analyses of the similarity data produced subjective dimensions/clusters of workload that were explained in terms of resource demand, task structure, and task characteristics. Data collected to support this analysis - task performance, physiological measures of heart period variability, effort ratings - revealed three primary dissociations. These dissociations were explained by using the parameters of Wickens' multiple resource theory:

1) When contrasting subjective ratings with performance, the former was relatively more sensitive to the number of tasks performed concurrently, while the latter was relatively more sensitive to the difficulty of a single task, particularly if this difficulty was related to responding. 2) Subjective difficulty ratings did not discriminate a task performed concurrently with an identical task from a task time-shared with a different task. However, performance was reliably better in the second configuration. 3) While as noted in (2), time-sharing two different tasks, using separate resources lead to better performance, this condition also yielded higher cardiac measures of mental workload.



A Multiple Processing Resource Explanation of the Subjective Dimensions of Operator Workload

William L. Derrick
Christopher D. Wickens

Abstract

Multiple measures of operator workload may dissociate, or fail to agree, for a given task. The goal of this study was to determine which task difficulty (workload) as indexed by attentional resource demand could explain the attendant variance in a second index of workload, subjective ratings. A multiple resource model of processing resources (Wickens, 1980) guided construction of tasks of differential resource demand. These tasks were both performed by subjects and rated according to workload similarity. Scaling and clustering analyses of the similarity data produced subjective dimensions/clusters of workload that were explained in terms of resource demand, task structure, and task characteristics. Data collected to support this analysis - task performance, physiological measures of heart period variability, effort ratings - revealed three primary dissociations. These dissociations were explained by using the parameters of Wickens' multiple resource theory.

1) When contrasting subjective ratings with performance, the former was relatively more sensitive to the number of tasks performed concurrently, while the latter was relatively more sensitive to the difficulty of a single task, particularly if this difficulty was related to responding. 2) Subjective difficulty ratings did not discriminate a task performed concurrently with an identical task from a task time-shared with a different task. However, performance was reliably better in the second configuration. 3) While as noted in 2), time-sharing two different tasks, using separate resources lead to better performance, this condition also yielded higher cardiac measures of mental workload.

INTRODUCTION

Theoretical Overview

In designing and developing complex man-machine systems, system designers and evaluators need to determine if trained operators can adequately perform required tasks to achieve successful system performance. One aspect of this determination utilizes a construct referred to as operator workload. In its most basic usage, operator workload is simply how hard a person must work to satisfy a given set of task demands. Clearly, a system which requires excessive workload will lead to system failure.

To deal with this loose and very intuitive concept, numerous and independent operational definitions of operator workload have appeared. For example, workload was and still can be defined by the time tasks require divided by the time available (Brown, Stone, & Pearce, 1975), a decrease in heart rate variability as the cognitive processing demands of a task increase (Mulder, 1978), the number and level of inputs to an operator determined by task analysis (Gartner, Ereneta, & Donohue, 1967), the perceived magnitude of fatigue, tension, and difficulty gathered from a questionnaire (Jenney, Older, & Cameron, 1972), and the spare mental capacity or reserve attention of an occupied operator as determined by performance on a secondary task (Krause & Roscoe, 1972). As diverse as these approaches are, they represent only a small subset of the techniques that have been employed to assess operator workload.

To impose some order in workload assessment research, measurement classification systems have been proposed (Gartner & Murphy, 1976; Gerathewahl, 1976; Jahns, 1973; Sheridan & Stassen, 1979), workload conferences convened (Moray, 1979), literature reviews published (Gartner & Murphy, 1976; Hartman & McKenzie, 1979), and guides describing use of techniques written (Roscoe, 1978; Wierwille & Williges, 1978). The result of all this interest and activity is that no single widely acceptable definition of operator workload yet exists and no circumscribed set of assessment techniques can be labeled as the most valid. These failures can be attributed to numerous examples of technique dissociation; that is, when two or more techniques are utilized in the same study, one subset will indicate a task manipulation increases operator workload while another subset will indicate no change in workload (e.g., Gunning, 1978; Hicks & Wierwille, 1979; Krebs, Wingert, & Cunningham, 1977). These same seemingly insensitive techniques, however, have been shown to be sensitive to task difficulty manipulations in other studies (Wierwille & Williges, 1978).

Two responses to this state of affairs are evident. The first concentrates on methodology. Much inconsistency is attributed to crude and inappropriate use of a technique or poor scoring. The complexities of secondary task usage are now receiving a great deal of attention (Ogden, Levine, & Eisner, 1979; Pew, 1979) as are the use of various physiological measures (Wierwille, 1979). This kind of information

produces prescriptions of how techniques should be used (Wierwille & Williges, 1978).

The second response is more basic, and in the long run, more fruitful. It asks the question, "Do we measure what we want to measure with the usual measure of mental (operator) load?" (Sanders, 1979). This approach deals with the data inconsistency mentioned above by acknowledging that operator workload is a multidimensional construct. Controlling for methodological problems, this view accepts the fact that some techniques may produce evidence of changed workload when task difficulty is apparently manipulated while others may not. This simply means that no one number (result from a specific technique) represents the total workload of a task, and tasks cannot be rank ordered by difficulty with these single numbers. Thus, the workload of a task is not a scalar quantity but a vector quantity associated with some number of dimensions (Moray, Johannsen, Pew, Rasmussen, Sanders, & Wickens, 1979). To answer the question posed above, the dimensions of operator workload must be uncovered and metrics with limiting values associated with each dimension. A battery of measures, each sensitive to a given dimension, would then be employed to assess operator workload.

The nature of these dimensions is currently speculative, however. Disagreement among the results of workload studies is too great to generate a firm set of candidates. Moray et al. (1979) contend that the most system designers can currently do is generate task-specific dimensions and specify the vector that describes the resultant operator workload. Manipulatin of a task parameter may or may not affect the vector's length depending upon its relationship to the workload dimensions. Structurally similar tasks and task situations should produce comparable workload vectors; however, vector comparisons between different tasks can be done only with careful justification. Thus, the theoretical position that operator workload is a multidimensional construct may be supported by workload research results, but it will afford limited utility until the nature of the dimensions common to all tasks can be described.

One set of candidates for the dimensions of operator workload has been proposed by Wickens (1979; 1980; 1981). Arguing that the concept of human processing resources can be partitioned into separate and limited quantities (Navon & Gopher, 1979; Sanders, 1979), Wickens contends that these multiple resources underlie human performance in information processing tasks. These separate resources, therefore, can serve as the sought-after dimensions of operator workload.

The purpose of this research was to determine how these processing resource dimensions of workload are related to the much used but little understood workload technique of subjective assessment. Specifically, tasks whose resource demands were defined a priori were performed singly and in several dual task combinations. Subsequent proximity judgments of workload similarity were analyzed by multidimensional scaling and clustering procedures. Resultant solutions were interpreted with the use

of task performance data, unidimensional ratings, and physiological measures. The final result is an explanation of multifaceted workload opinion ratings in terms of processing resources.

Theoretical Analysis

Subjective opinions of operator workload are frequently gathered by designers and engineers during the evaluation of proposed system configurations. These opinions are expressed via rating scales, questionnaires, and interviews, instruments which tend to be very task specific. Operators may be asked to rate the "flyability" of a given aircraft configuration or comment on the adequacy of symbolic codes used to convey information. Since information of this type is relatively easy and inexpensive to obtain, it is often gathered in conjunction with other measures of workload including primary task performance, physiological measures, and secondary task techniques (Gartner & Murphy, 1976).

Although providing valuable clues on equipment and procedural modification that may improve system performance, opinion data has really provided little insight into how feelings of workload are produced or related to other assessment techniques. In spite of this fact, Sheridan (1980) has argued that "mental workload should be defined as a person's private subjective experience of his or her own cognitive effort" and that this experience can be measured most directly with rating scales. According to Sheridan, the remaining workload assessment techniques operationally measure the construct defined in these terms. Here, the multidimensionality of mental workload is related to the number of rated attributes, such as perceived risk, stress, and complexity, that covary with task manipulations and performance. Whatever these attributes, their assessment depends upon an operator's phenomenal awareness of this mental state during task performance.

The following review and analysis describes an alternative multidimensional view of workload and proposes that subjective ratings of workload can be explained by a non-phenomenal construct whose existence is determined empirically. Material is organized within three areas: 1) description of and evidence for multiple processing resource models of task performance, with emphasis on Wickens' (1980) three dimensional model; 2) review and evaluation of subjective workload assessment studies; and 3) description and application of multidimensional scaling and clustering techniques to uncovering the dimensionality of subjective workload ratings.

1) Processing resource models of task performance. Within the last few years increasing evidence has been provided that the "undifferentiated resource" theory of human attention in which only a single supply of processing resouces underlies task performance, cannot adequately account for much of the experimental data (Kantowitz & Knight, 1976; Navon & Gopher, 1979; Wickens, 1980; Wickens & Kessel, 1980). Much of the evidence for this view has been discussed in detail in Wickens (1981, 1984) and so will not be repeated here. Multiple resources theory

set forth as the alternative to undifferentiated capacity theory proposes that a series of sub-capacities underlie dual task performance. Two tasks demanding separate resources for their performance will be successfully time-shared. Furthermore, if the difficulty of one of two time-shared tasks is varied, and the manipulated parameter changes demands on resources not required by the concurrent task, then concurrent task performance will be unaffected. This accounts for what Wickens (1980, 1981, 1984) has labelled "difficulty insensitivity."

One possible problem with a multiple resource explanation of dual task performance is that the time-honored concept of attention, that phenomenal experience of which we are all aware, is now partitioned into many resources of which we cannot begin to be aware. The answer to this problem is that resources are not phenomenal; they are not the same thing as attention or even consciousness. As Navon and Gopher (1979) point out, we could "entertain our firm and phenomenally valid belief that we cannot attend to more than one thing at a time and our growing conviction that performance limitations are often unrelated with this basic fact, if we divorced the notions of attention and resources" (p. 235). Resources are thus "provisions" for information processing. Their existence will be determined from performance decrements in dual task studies, not introspection. Further, resource allocation is also an empirical question, answered by assigning different priorities to each of the time-shared tasks and examining the performance data.

A second potential problem with multiple resources theory is the possible proliferation of resources. There is a danger that the number of resources postulated to account for each piece of experimental data can grow so large that the theory loses all predictive value (Navon & Gopher, 1979; Wickens, 1980). Reviewing nearly 60 dual studies conducted over a 15 year period, Wickens (1980) concluded that three dichotomous dimensions -- processing stages, modalities, and codes -- could account for at least a large portion of the variance in time-sharing efficiency, and so could be labelled as candidates for multiple resources. The experimental evidence in support of these three dimensions has been described elsewhere (Wickens, 1980; 1981), and so they will only be briefly described here.

(1) Processing stages: The processes involved in perceptual evaluation and the central processing operations of memory rehearsal and transformations draw from a different pool of resources than those involved in the selection and execution of overt manual or vocal responses.

(2) Processing modalities: Perceptual processing of auditory stimuli draws upon separate resources from processing of visual stimuli.

(3) Processing codes: Verbal processing relies upon different

> resources from the processing of non-verbal material. This dichotomy apparently underlies the processes of perception, transformations in working memory, and responses. The latter defines the contrast between vocal and manual responses, assuming that vocal responses are usually verbal, and manual responses are normally spatially guided.

Described in this manner, the multiple resource model is a performance-based approach to workload assessment. Users would apply a battery of secondary tasks, validated according to resource demand, to determine the resource pattern underlying acceptable primary task performance. This position may be contrasted with that of Sheridan (1980) who stated that subjective perceptions of cognitive effort constitute the essence of workload. The next section of this paper contains a selective review of subjective assessments of operator workload and examines how these perceptions might be related to the three processing resource dimensions.

2) Subjective assessment of workload. Reviews of workload assessment techniques consistently conclude that operator opinions are valuable indices of workload (Moray, 1979; Hartman & McKenzie, 1979; Roscoe, 1978; Wierwille & Williges, 1978). The quote from Gartner and Murphy (1976) that "the pilot's direct perception or estimation of his feelings, exertion, or condition may provide the most sensitive and reliable indicators" of workload, is often repeated or paraphrased. Indeed, the last ten years have produced nearly 100 published studies in which operator opinions of workload, both direct and otherwise, have been reported. The opinions gathered, however, are almost always related directly to the equipment configuration under evaluation and not to any general task characteristics or other possible situational attributes that produce feelings of load. Thus, situation-specific opinions of operator workload are adjunct measures typically gathered in addition to the more "scientific" workload measures (Ellis, 1977).

For example, Spady (1978) looked at instrument scanning patterns of pilots during instrument landing approaches in two different system configurations under two conditins of atmospheric turbulence. The scan patterns differed reliably and pilots' opinions concerning instrument usage were found to be in agreement for the most important instruments. North and Graffunder (1979) attempted to assess the workload of simulated landings for vertical takeoff and landing aircraft. Multivariate discriminate functions were formed from both flight performance and visual response variables. Data from seven physiological variables and opinion scales were collected but only the physiologial data were used to predict the discriminant. The fate of the opinion data is not even discussed.

Despite the frequency and agreed upon utility of operator opinions, reviews of the subjective assessment technique note that this situation specificity of rating scales and questionnaires precludes rigorous scale development. Ordinal scale ratings may be treated as interval data, and validation and reliability studies are rarely done (see Reid, Shingledecker, & Eggemeier, 1981 for a notable exception using conjoint measurement techniques). Further, studies supporting subjective ratings almost never report controls for eliminating rating biases due to previous experience, for changing impressions as a rated system becomes more familiar, and for separating effects of physical task loading from mental task loading (Wierwille & Williges, 1978). It has also been argued (Gartner & Murphy, 1976) that during periods of intense concentration when operator workload should be very high, an operator may be least aware of the amount of effort he is investing in task performance. To these points can be added another serious flaw. Often the assessed workload construct is never defined for the rater, apparently because he is supposed to know what the term means. He is just asked to rate the level of workload or acceptability of workload for a particular system configuration (e.g., Gunning, 1978).

Because the previous paragraphs fairly depict the method as it is used, a comprehensive and fine-grained analysis of subjective opinion data, especially as it might relate to proposed dimensions of workload defined by processing resources, is not warranted. Specific conclusions would be overdrawn because of the methodological problems thus far associated with subjective workload studies. Rather, it will be more instructive to examine those few studies that have attempted to relate task characteristics or dimensions of subjective experience to perceptions of workload. The goal of these attempts was to discover what caused perception to vary, a goal that was not and cannot be fully realized because of reasons which will be discussed.

Several studies have varied parameters in analytical models of pilot controlling behavior to predict known pilot ratings of handling characteristics from actual systems. The ten-point Cooper-Harper scale (Cooper & Harper, 1969) or variants thereof have been used in these studies to gather subjective opinions of handling qualities or "flyability." This is the one widely used and validated workload scale, although to call it such one needs to assume that an aircraft which is rated difficult to fly also imposes a heavy cognitive load.

In an effort to determine what characteristics of manual control tasks, as modeled by analytical techniques, give rise to ratings of flyability, Hess (1977) proposed a rating hypothesis. Briefly, Hess hypothesized that if the index of performance in the optimal pilot model was representative of human pilot performance and if the variables in this index of performance were directly observable by the pilot, then the numerical value of this index can be related to a Cooper-Harper rating by a relatively simple function. Taking data from existing configuration-rating studies, Hess attempted to find the relationship

between several index of performance values and several ratings. The data revealed that the ratings could be predicted. Hess concluded that perceived operator workload, or flyability, was related to the number of separate variables whose deviations the pilot considers pertinent to the task, one of the terms in the equation for the index of performance. Stated in other optimal control theory terms, fraction of attention was the most important variable in predicting these ratings.

Wewerinke (1974) has pursued a similar line of research within the framework of the optimal control model. Using data from both a compensatory tracking task (Wewerinke, 1974) and later hover and navigation tasks in helicopters (Wewerinke, 1977), Wewerinke attempted to relate a parameter of the model indicative of attention allocation to subjective ratings of effort expended and to perceived task demands. This model parameter is based upon optimal allocation of attention among displays presenting different types and rates of information. In both studies he found that the "control effort model" could predict these ratings and concluded that it seems to provide a meaningful representation of workload.

A somewhat different conceptual approach is reported by Smith (1976). Contending that handling quality metrics are arbitrary and have given us no real insight into what produces flyability ratings, Smith proposes what he calls a unified theory of pilot opinion rating. Starting from a rather dubious physiological position, Smith hypothesizes that a pilot's subjective response originates from a particular area within his central nervous system and that this response is directly related to the strength of the neural signal in this location. A physiologial measure of an opinion rating would be a measure of the rate of nerve impulses in this region. Since such a measure is not possible, Smith proposes an analytical pilot model with a corresonding parameter related to pitch attitude rate. He attempts to correlate values of this parameter with data from existing studies where subjective handling qualities ratings have been gathered. Smith reports some success, but he concludes that it is difficult to quantify his rate parameter since existing analytical studies have no comparable parameter.

Other research within the domain of manual control has also identified task characteristics that seem to produce feelings of greater mental load. Jex and Clement (1979) identify two characteristics that reliably affect ratings on Cooper-Harper type scales. The first is the requirement to generate lead when controlling a plant of higher order dynamics. The more lead required, in the absence of augmentation displays, the greater the perceived effort required to control the system. The second characteristic is the stability of the controlled element. This characteristic has been extensively investigated with the use of the critical tracking task, a task with a transfer function which produces instability proportional to the value of the parameter lambda. The greater the control instability, the higher the perceived effort or attention required to perform the task.

A more elaborate approach to identifying those aspects of the manual control situation which produce perceptions of load has been suggested by Higgins (1979). Higgins argues that a correlation between engineering calculation procedures (ECP) and pilot rating procedures (PRP) provides the criterion for determining the effectiveness of a completed aircraft system evaluation. ECPs discussed by Higgins include the information content of displays, quantity and type of air-to-ground communications, and manipulation of aircraft handling characteristics. PRPs are composed of magnitude estimations of both cognitive and physical workload for a particular system configuration.

A unique system configuration is defined by quantified ECPs. This configuration is flown during a selected mission after which pilots render magnitude estimation ratings of workload. The two variables are then correlated with a high correlation indicating that the ratings are related to this particular set and level of ECPs. A low correlation would indicate that the workload ratings were a function of some other ECPs or the same ECPs set at different levels. If the correlation is low, new ECP levels are set and the process is repeated. Higgins reports a brief validation study of this method in which changes in control forces (ECP) were highly correlated ($r = 0.89$) with magnitude estimates of workload (PRP).

Outside the manual control literature, Borg and his colleagues have attemped to identify those task characteristics that produce feelings of increased workload (Borg, 1978). Although much of their work deals with perceptions of physical load, some work has been done with cognitive tasks such as visual search and answering items from intelligence tests. This type of work often reports high correlations between perceived difficulty (workload) and actual difficulty as indexed by reduced performance. The sources of these perceptions, according to Borg (1978), are task characteristics such as the number of decision alternatives, insufficient data, uncertainty of decisional outcomes, inadequate feedback, scarcity of time, and perceived probability of failure.

In a recent review of the subjective workload literature, Moray (1982) concludes that "little effort has been made explicitly to understand the origins of subjective feelings of load" (p. 37). His search for origins seems to be limited to the same task characteristic approach discussed above. He reviews the manual control and cognitive task literature, plus additional areas, for the purpose of presenting a list of characteristics (e.g., generating lead, the degree of precision in a response) that reliably produce perceptions of load. These characteristics are then integrated into four very general categories--physical exertion, rate of information processing, memory load, and subjective performance criterion--that Moray believes will be involved in any model of subjective mental load.

The efficacy of the task characteristic approach for understanding workload perceptions and helping designers predict operator opinions of proposed systems is limited by several factors. First, identification of task characteristics is not enough. Those identified thus far--generation of lead, instabiity, time pressure, etc.--could have been identified with common sense. Such characteristics would not be incorporated into systems unless factors such as technology limitations, cost, low performance criteria, and so forth mitigated otherwise. Second, the only task characteristics both identified and quantified as to their impact on perceptions of workload come from analytical pilot models. Dealing primarily with only one aspect of pilot-system performance, handling qualities, these models ignore much of the total flying task demands. Even their reported success leaves some doubt because of the difficulty in estimating accurate values for some of the model parameters. Hess (1977) points out that specifying the form of the index of performance in the optimal pilot model requires a good deal of experience on the part of the analyst.

A third limiting factor is that one may eventually find dozens of task characteristics that affect feelings of effort expenditure or workload. Higgins' (1979) methodology suggests many such characteristics varied over several levels in several different combinations might have to be examined to determine what drives subjective assessments. Not only would such an undertaking require a tremendous amount of time, previous workload research indicates that induced load would vary by performance criteria, mission profiles, maneuvers, phases of flight, and a host of other factors (e.g., Stackhouse, 1973). Thus, no one quantitative relationship between a task characteristic and perceived workload would exist. Fourth, and finally, this approach would enable predictions of workload, as rated by operators, only to the extent that similar task characteristics exist from situation to situation. Even then, if two load-inducing characteristics were combined in a novel way in a new system, the predictions might not hold. When task situations vary significantly, subjective assessments could be gathered only after simulation of the design decisions.

Another approach to understanding the causes and correlates of workload perceptions, related to the discussion above is the use of multiple scales to assess operator opinions. This approach implicitly acknowledges that any given task situation may have multiple and possibly unrelated effects on an operator's awareness of the effort expended to meet task demands. Like the proposed multidimensionality of processing resources, this view presumably considers cognitive effort to be multidimensional. As such, both views would claim that values on the relevant dimensions define what is meant when the construct of workload is termed multidimensional. In both views, one number or a scalar quantity should fail to specify operator workload.

Sheridan and Simpson (1979) have developed a subjective workload scale, proposed for use by pilots and based upon the Cooper-Harper handling qualities scale. During task performance, pilots are asked to rate cognitive effort or mental workload based upon three attributes claimed to contribute to subjective experience: 1) fraction of time busy; 2) level of problem solving or planning complexity; and 3) level of emotional stress. Unfortunately, neither the rationale for choosing these particular attributes nor data supporting the scaling have yet been published.

Milord and Perry (1977), concerned with the perception of task overload, also proposed three attributes or dimensions that should be measured: 1) intensity (frequency of decisions); 2) diversity (complexity of decisions); and 3) patterning (coherence or continuity of the stimulus input). Three very different dual task situations were constructed which varied the level of each dimension. Supporting their predictions, a combined driving-memory task was rated the most distracting or overloading because the levels of stimulation were higher (intensity), the amount of variation in input and output was higher (diversity), and the aperiodicity of the distracting stimulation was lower (patterning). Again, however, the reason for measuring these particular dimensions to determine overload was not explained.

Many studies have used multiple scales to assess perceptions of workload. For example, Steininger (1977) gave commercial aviation pilots and copilots four attributes to rate on scales running from 1 (great or high) to 7 (little or low). The four attributes were: 1) overall workload, 2) preflight confidence, 3) perceived pressure in coping with errors and failures, and 4) pressure of responsibility. The ratings did dissociate somewhat, especially for copilots. Goerres (1977) asked several types of pilots to rate the "mental workload stressors" of multiple display monitoring, complex decision making, and system handling. The monitoring and handing ratings indicated little associated stress while the decision making ratings were somewhat higher.

In neither study, nor several similar studies that could be cited, is any rationale presented concerning the choice of dimensions or attributes that are rated. Nor are attempts made to relate these dimensions to a theoretical construct such as effort. Thus, the multiattribute scaling studies shed little light on the prediction of workload perceptions across task situations.

An additional problem with garnering meaningful information from subjective ratings of workload is that quite often the results are counterintuitive. Compared to system configuration A, system configuration B might yield equal performance, decreased performance, or increased performance. Corresponding subjective ratings of operator load could also be described by the three outcomes of no change, increased, or decreased. A matrix of nine outcome possibilities can be generated, each cell of which requires a different explanation of events. Workload

researchers normally expect a decrease in performance between configurations to be associated with an increase in perceived workload, assuming the operator population is motivated to meet the same performance criterion in both configurations. Quite often this outcome is reported.

Kopala (1979) examined two conditions of symbol coding--shape and shape plus redundant color coding--on a flight display in an A-7 aircraft. During simulated flights, pilots were able to use the redundant color-coded display in a threat recognition task both faster and more accurately than the shape-coded display, and their perceptions of display efficiency (workload) matched the performance results. Hicks and Wierwille (1979) manipulated workload in a driving simulator by moving the center of pressure of simulated wind gusts from near the front wheel (high workload) to 50 cm rearward (medium workload) and to 100 cm rearward (low workload). Performance, measured by steering reversals and lateral plus yaw deviations, and perceptions of workload, measured by ratings of attention, confirmed the differential disruptive effects of the wind gusts. Stone, Sanders, Glick, Wiley, and Kimball (1979) looked at pilot performance during various nighttime helicopter maneuvers using two bifocally configured night vision goggles. Both flight performance and subjective ratings indicated that one version of the goggles was superior.

If only one parameter differs between conditions where workload is assessed, then it is not surprising to find a covariation between measures. However, if two or more variables differ between conditions, and they may oppose each other's influence, then dissociation might well occur, because we hae no models of task characteristics, that determine which parameters will "drive" subjective measures, and which will "drive" performance. For example, Stackhouse (1973) examined helicopter pilot opinions of workload for several attributes such as ability to maintain altitude and fore/aft movement. These ratings were gathered for several maneuvers and alternative displays and were compared to performance measured by eight error scores. For the precision hover maneuver, Stackhouse found no correlation coefficient greater than 0.2 between actual performance and pilots' opinions of their performance. Results were similar for the sideways flight maneuver.

Wolf (1978) paired three levels of flight task difficulty, defined by gust level and flight control system mode, with two levels of memory search task difficulty, defined by size of the positive set (Sternberg, 1969). Several physiological and subjective measures were gathered as each of the six test conditions was flown. Using paired comparison judgments of relative task difficulty and an overall workload rating derived from a modified Cooper-Harper scale the following effects were obtained: (1) performance on tracking was influenced only by tracking difficulty; (2) performance on the memory search task was influenced both by tracking difficulty and by memory set size. Yet (3) subjective workload was influenced only by tracking and failed to reflect any characteristics of the memory task. From a system design point of view,

however, these results reveal the limitations of relying upon perceptions to determine what combinations of task situations improve performance and reduce workload.

A less theoretical example of the same problem was reported by Herron (1980). Noting that typical air-to-air and air-to-ground head-up displays (HUD) in the same aircraft often use very different symbology and that a pilot may have to transition rapidly from one HUD to another, Herron tested a "standardized" HUD that could be used across modes. An initial modified HUD was compared to the existing display in an A-7D aircraft. Flying air-to-ground missions, the candidate display was associaed with significantly worse bombing error scores, but pilot opinions of the display were favorable enough to encourage further development. A second candidate display was subsequently compared to the existing display, and again the candidate was associated with less accurate weapons delivery but very favorable operator opinions. Herron concluded that one should not rely too heavily on operator opinion in such evaluations. Similar sorts of results were obtained by Murphy et al. (1978) when they evaluated three candidate displays for V/STOL aircraft. The rank order of performance across the three was precisely opposite the rank order of pilot opinion data.

Changes in system configurations that alter operator performance have important consequences for system design, irrespective of how these changes are subjectively perceived. Workload ratings that match previously calibrated task difficulty manipulations (Hicks & Wierwille, 1979) do not validate the use of ratings for all systems and configurations. Although it may be interesting to explore the workload rating process for its own sake (e.g., Hart, 1982), the dissociation of performance outcomes and subjective ratings demands that ratings be explained by the more relevant performance data, not the other way around.

As a performance-based approach to workload assessment, the multiple resource view ties task difficulty to resource cost (Navon & Gopher, 1980). Any task performed to criterion will demand resources from some number of structures. Different tasks may have different compositions of resources. Any two tasks performed simultaneously may demand some number of joint resource structures and some number of disjoint resource structures (demanded by one task but not at all by the other). A task difficulty manipulation is one that requires the operator to utilize more resources from one or more structures to meet a performance criterion.

Within this framework, resource usage is determined by employing secondary or loading tasks. If the secondary task does not overlap in resource demand with the primary task of interst, very efficient time-sharing will result but primary task resource cost (workload) will remain undetermined. When the two tasks partially overlap in demand but the difficulty manipulation in the primary task taps a resource unneeded by the secondary task, workload will still go unmeasured. Only when both tasks demand the common resource needed by a difficulty manipulation will

the secondary task reflect the resource cost of the primary task. Since three separate dimensions of resources are thought to exist (Wickens, 1980), many secondary tasks may be required to accurately assess the resource cost of any primary task.

The goal of the research reported here was to map the three resource dimensions onto subjective ratings of workload. Stated another way, this study attempted to determine how task difficulty, defined by resource cost, could alter perceptions of workload. The method employed for this mapping and its rationale are described below. Unlike simple correlational analysis, it did not require an a priori specification of what categories of ratings to collect, nor did it require explanation of the complex resource concept to subjects. Further, instead of providing a matrix of correlations or a set of relationships, it provided one integrated mapping of a complex construct onto another.

3) Multidimensional scaling and clustering. Multidimensional scaling (MDS) is a set of mathematical techniques that enables the user to uncover what is often referred to as the "hidden structure" in a particular set of data (Kruskal & Wish, 1978). Using proximities among objects, where one proximity is a number indicating how similar or different two objects appear to be, scaling programs output geometric spatial representations of these objects in an n-dimensional space. Distances between objects in this space are proportional to the judged similarity of the objects as shown by the proximity values. Judgmental models implicit in scaling assume that these mathematical distances reflect "psychological distances" among the objects for the object raters. Interpretation of the scaling solution dimensions tells the user what attributes the rater is using to generate the initial proximities, although the rater may be unaware that he is using these attributes.

A large number and variety of methods and models of MDS exist, and they are used within many different fields. Studies of memory structure (Shepard, Kilpatric, & Cunningham, 1975; Shoben, 1976), perception and evaluation of auditory stimuli (Soli & Arabie, 1979; Wish & Carroll, 1974), interpersonal communication (Wish, 1979), consumer behavior (Olshavsky, MacKay, & Sentell, 1975) plus several other areas use variants of MDS techniques. These variants differ with respect to properties of the data analyzed and properties of the models employed. Carroll and Arabie (1980) provide a taxonomy of MDS techniques with respect to these two properties.

Many types of MDS use essentially only one matrix of proximities, a lower triangular matrix of each stimulus object compared to all others. With several of these matrices, one for each subject, these "two-way" MDS models treat differences among matrices as due to random error. Further, the coordinate solutions provided by two-way MDS programs are not generally susceptible to direct interpretation (Kruskal & Wish, 1978). Stimulus objects in this n-dimensional space are tied to the dimensions or axes by coordinate values, but if the total configuration is rotated these values change considerably. The reason is that this MDS solution

is based upon distances between stimulus objects, distances which do not change when the axes are rotated; however, this rotation changes the coordinate values. Finding the correct orientation of these axes and discovering their meaning require additional analytical techniques.

In contrast, individual differences scaling, or "three-way" MDS, assumes that different individuals perceive the stimulus in terms of a common set of dimensions or attributes, but that these attributes have differential importance (Carroll, 1972). The most common three-way MDS model is INDSCAL which stands for INdividual Differences SCALing. With INDSCAL, differences in proximity matrices among subjects are not treated as random error; rather, they are used to weight or modify the distances among the stimuli in what is termed the group stimulus space. In addition, INDSCAL produces a second space, called a subject space, which plots each rater along the same dimensions used to solve for the psychological distances among the stimuli. With this information, the user can reconstruct how each rater perceived the overall relationships among the stimuli.

In addition to recognition of individual differences in perception, INDSCAL solves for that particular orientation of axes that maximizes the amount of variance accounted for among proximity matrices. This unique orientation is produced because the subject weights which produce the group stimulus space are allowed to have their effect only along the coordinate axes (Kruskal & Wish, 1978). These axes or dimensions thus achieve a special status in INDSCAL, and "might be assumed to correspond to fundamental psychological processes that have different saliences for different individuals or under different experimental conditions" (Wish & Carroll, 1974, p. 452). Specifying the nature of these processes, however, may require additional data collection and analysis.

A related approach to finding "hidden structures" in a set of proximities data has used clustering analysis. Unlike the spatial-distance MDS models, clustering is essentially a non-spatial or discrete representation of stimulus objects grouped or clustered together because they have a property in common. Clustering models are employed because the assumption of continuous spatial representations of stimuli may not appear warranted with certain judgments or certain classes of stimuli. For example, Shepard and Arabie (1979) question the utility of spatial-distance models when applied to the perceived similarity of consonant phonemes, kin and other category-specific terms, social structures, and possibly even continuously variable stimuli that are not perceived as such.

Although the dominant mode of clustering analysis within psychology is hierarchical, Shepard and Arabie (1979) argue that the requirement of having clusters hierarchically nested is unduly limiting. Once one object is grouped with a second on the basis of a characteristic, it is not possible to separate the objects and identify one with a different category. They must now both be members of all the categories superior to them in the hierarchical solution. In response to this limitation,

Shepard and Arabie (1979) report a new method of nonhierarchical clustering called additive clustering (ADCLUS). The essential feature of their model is that proximities data can be represented by overlapping subsets of objects. Members of a subset share some discrete property and only those members share that property. However, any object may also be a member of another subset of objects which also share another unique property.

The model is additive because it is assumed that the discrete property contributes a fixed increment to the similarity between any two objects sharing that property, independently of the contributions of any and all other properties. Shepard and Arabie thus define object similarity as a sum of weights associated with just those subsets to which both objects belong. When one object belongs to a subset not shared by the other, the sum of weights is not incremented.

Additive clustering solutions generate a variable number of subsets with each subset containing some number of the rated objects. Each subset has an associated weight that indicates the importance of the subset in the final solution. The higher the subset weight, the greater the psychological salience of the property that clustered the objects into a subset. As with MDS solutions, however, the identification of that binding property must come from further analysis.

Both INDSCAL and ADCLUS can be used to confirm the known physical structure of a set of stimuli. For example, Wish and Carroll (1974) report a reanalysis of color judgment data with INDSCAL. In the initial experiment, subjects arranged three color chips into a triangle such that the lengths of the three sides were proportional to the psychological distances between the colors. INDSCAL produced a two-dimensional group stimulus space from these judgments in which the stimuli formed the familiar color circle. Rosler (1978) had subjects rate the dissimilarity of circle and triangle figures which were constructed according to relationships of size and form. The INDSCAL solution recovered these relationships although these attributes were never specified during the ratings. Shepard and Arabie (1979) report a reanalysis of social categorization data using ADCLUS. They found that the clustering model could adequately account for some of the seemingly contradictory aspects of the original data. In contrast, these models can also be used when one cannot specify the physical structure a priori, such as Wish and Carroll's (1974) scaling of perception of rhythm and accent in words and phrases.

Given the operation and previous applications of both the INDSCAL and ADCLUS models, they were chosen to explore the relationship between perceptions of operator workload and dimensions of processing resources. In general, the method required the construction of several tasks that utilized different patterns of resources. These tasks were then performed by subjects and subsequently served as the stimuli for all pairwise ratings of workload similarity. The ADCLUS and INDSCAL models were applied to the ratings to uncover the attributes used by subjects in

making these ratings. Identification of these attributes depended upon location of tasks in the MDS space, grouping of tasks in clusters, the performance data, and additional rating data collected on each task.

These models, at least theoretically, generate the so-called "hidden structure" or identify and weight of "psychological characteristics" in a set of ratings data. The position argued here is that whatever that structure might be for operator workload ratings, much of it should be explanable by the performance-based concept of processing resources. Obviously the explanation is not simple. Decrements in performance may be followed by either perceived increases or decreases in workload. But if workload is to be measured by resource cost and both resource cost and subjective data are to be collected, one must explain how the two covary together. Further, if one can predict the final resource cost of a task that currently exists only in a simple simulation mode, one would also like to predict the final operator reaction to that fully implemented task. The following sections of this paper present an experiment that explores the processing resource-workload rating relationship.

EXPERIMENTAL OVERVIEW

This research was designed to answer one primary question and explore answers to related questions. That one primary question asks how judgments of operator workload are related to performance as interpreted within the framework of the multiple resources model, where resources are defined by codes, modalities, and stages of processing. One related related question asks what might be the differences in perceptions of workload among tasks requiring few resources, tasks requiring large amounts of resources from few structures, and dual task pairs that are time-shared either efficiently or inefficiently. A second related queston asks what might be the relative importance of the resource cells in the Wickens' (1980) model depicted in Figure 1. For instance, do two tasks that demand processing of verbal codes influence subjective reactions more than two tasks that compete for manual response resources?

These questions were studied with a multidimensional scaling and clustering study in which the stimuli to be rated consisted of actual tasks which the subject performed. Using the resource model in Figure 1, four tasks were constructed that were assumed to demand resources from different structures. Based upon the premise that increasing task difficulty increases resource cost (Navon & Gopher, 1980), two versions of each task were created: one easy, one difficult. The difference between them was the manipulation of a task parameter that presumably increased resource cost. Each of the four easy tasks were then combined in all possible pairs to form ten dual task combinations, some of which presumably competed for the same resource structures while others did not. All possible pairs of the 18 tasks (4 easy, 4 difficult, and 10 dual tasks) were compared on workload similarity. These paired comparison ratings were then raw material for the INDSCAL and ADCLUS analyses.

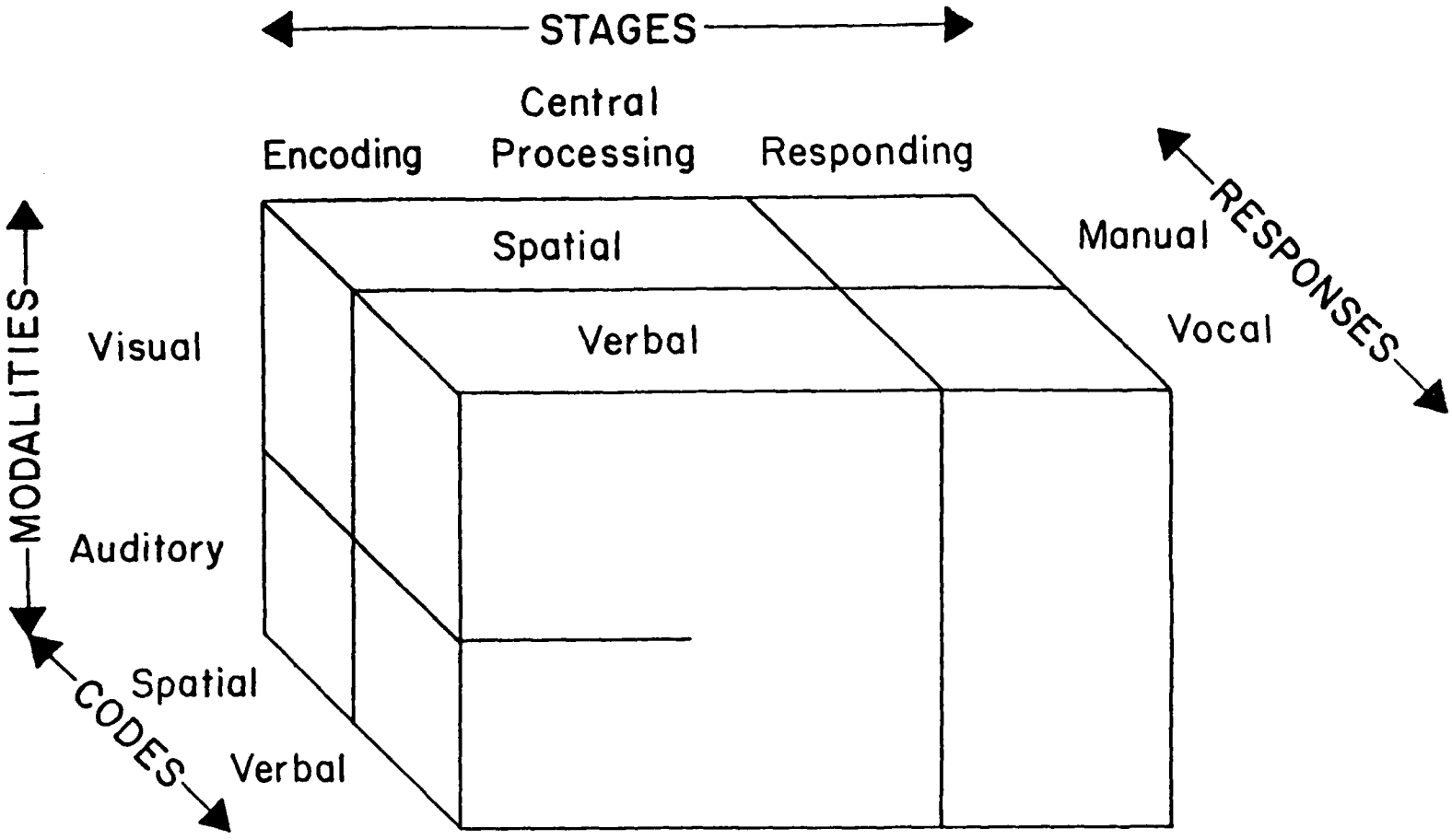


Figure 1: The proposed multiple structure of processing resources (after Wickens, 1983).

To interpret the meaning of MDS dimensions or ADCLUS clusters, several techniques were employed. First, the attributes of the tasks themselves were examined relative to their locations in the coordinate solutions. Second, the performance data for each task was compared to the scaling and clustering solutions to determine its impact on perceptions. Third, each task was rated on several unidimensional scales, and these ratings were correlated with the MDS scaling weights. High correlations implied that a unidimensional scale described a MDS dimension. Finally, physiological measures related to heart rate variability were collected while the subjects performed the tasks. This construct, which has been found to index sustained attentional demands and operator workload (Porges, 1981; Wierwille, 1979; Mulder & Mulder, 1980) was evaluated against subjective perceptions using correlational analysis. All four techniques were attempts to explain what the subjective dimensions of operator workload might be and how these dimensions are related to both physiological and performance-based measures of workload.

HYPOTHESES

Based upon the previous discussion, the following two research hypotheses can be formulated:

(1) Performance in the time-shared tasks will be related to competition for resource structures. When assigning equal task priorities, those tasks which compete for relatively few resource structures will suffer less severe performance decrements than tasks which do compete for common structures.

(2) The underlying structure of workload perceptions can be explained by task resource cost when resources are defined by stages, codes, and modalities. Specifically, one dimension of a scaling solution or some number of clusters from an additive clustering analysis should be clearly interpretable with respect to the number of resource structures demanded by the task and/or competed for by a task pair.

Hypothesis 1 is derived directly from the Wickens' (1980) model. Although increased resource demand and thus increased task difficulty can be produced in most single tasks, empirical evidence for increased resources can be obtained only from dual task interference studies where predicted changes in performance are supported based upon particular task combinations (Wickens & Derrick, 1981). Thus hypothesis 1 is limited to 10 of the 18 task configurations performed by subjects.

This hypothesis has received support from other studies (e.g., Wickens, Mountford, & Schreiner, 1981) and is therefore not original. However, given the relative recency of the multiple resource model and the small body of evidence supporting it, the hypothesis should be evaluated. Further, failure to find the predicted dual task performance effects would eliminate the need for any scaling or clustering analysis.

If this hypothesis were refuted, indicating that the resource composition of each task was indeterminate, then the primary question addressed in this research could not be answered.

Hypothesis 2 contains the argument that overlapping resource demand will be related to performance decrements and jointly be related to perceptions of increased workload. This relationship should not be perfect, however, given the previously discussed dissociation between performance and opinions of workload (e.g., Herron, 1980). If this effect did not emerge from a scaling or clustering analysis, one would have to question any other conclusion from that analysis.

What other scaling dimensions or clustering subsets might contain cannot be predicted. The ragged nature of the subjective workload data base coupled with this unique attempt to combine manipulations of resource cost with workload perceptions greatly limit the prediction process. One can hope, however, that additional scaling or clustering information will reveal how the resource dimensions of codes, stages, and modalities map onto subjective dimensions and thus influence workload perceptions.

EXTENSION OF TECHNIQUES

Typical scaling and clustering studies ask subjects to rate the similarity (or comparable concepts) of all possible stimulus pairs with few instructions as to what is meant by similarity (e.g., Wish & Carroll, 1974). One goal of these studies is to determine what attributes the subjects employ in making such judgments. Further, the stimuli being rated are typically faces, color chips, sounds, countries, and similar objects which can be presented rapidly to subjects. Rapid presentation enables the experimenter to collect pairwise ratings for a large number of stimuli. In contrast, the present study sought to determine the structure of workload ratings among a set of stimuli which were tasks that must actually be performed before ratings could be collected. These differences forced some departures from conventional techniques.

Tasks as Stimuli

The number of tasks used in this study was limited to four. As mentioned earlier, four tasks yield 18 task variants of interst. Eighteen task stimuli require 153 proximity ratings (N(N - 1)/2) to complete a triangular matrix of all possible pairs of stimuli. A further addition of one task would yield 25 task variants and require 300 proximity ratings. Since these tasks could not be rated until they had been performed at a stable level, the addition of even one task was deemed excessive of experimental time.

Task structure was based upon the resource cells in Figure 1. Four tasks - manual tracking, visual search, memory search, and tone judgment - were constructed so that theoretically they demanded resources from the three dimensions along which resources are defined. Although

restrictions to four tasks precluded all possible resource combinations in Figure 1, the four that were chosen differed to the maximum extent possible. The specific form of these tasks is discussed in a later section of this study.

Proximity Judgments

To derive the structure of workload perceptions, simple similarity judgments were not used. Because ten of the stimuli are actually very separable stimulus compounds of two tasks, similarity ratings between compounds would likely be affected by the physical structure of the tasks themselves. For example, if stimulus 1 were a tracking-visual search task pair and stimulus 2 a tracking-tone judgment task pair, any similarity rating would likely be dominated by the presence of tracking in both dual task pairs. The same logic applies if stimulus 1 were single task tracking. Scaling and clustering solutions of such data would likely produce dimensions related only to the surface form of the tasks.

Scaling and clustering solutions do permit more specific similarity judgments, however, such as political similarity or cultural similarity (Kruskal & Wish, 1978). In this study, subjects were asked to rate the *similarity of task difficulty* between all pairs of tasks. Here, "task difficulty" substituted for "operator workload" or "mental workload", although Moray (1982) correctly points out that difficulty and workload judgments have never been empirically equated.

For operators of actual systems, increasing task difficulty may not be equated with increasing workload if subjective and objective performance criteria do not coincide. In the laboratory environment of this study, however, this discrepancy was eliminated by using generous performance incentives and eliminating objective performance criteria. These precautions permitted the use of task difficulty ratings, ratings that did not force the experimenter to define the unfamiliar term "workload" and possibly bias the results. Further, task difficulty is a concept that naive subjects can easily deal with, yet it is rich enough to encompass all the suggested characteristics of subjective workload (e.g., stress, complexity, etc.).

Ideally, each of the 153 similarity of difficulty ratings would be collected after the two task stimuli which comprise the rated pair had been performed. Such a procedure is normally used in scaling and clustering studies (Kruskal & Wish, 1978) but was not used here primarily because continued practice with the tasks would alter perceptions of difficulty. The basis for similarity of difficulty ratings would change markedly and comparisons of ratings collected early in the experiment with those collected later would be inappropriate.

To circumvent these problems, similarity of difficulty ratings were gathered in one session. After an equal amount of practice on all tasks, subjects were given pictorial representations of the tasks and asked to generate 153 proximity values. This procedure is analogous to that employed by Nygen and Jones (1977) who gave their subjects names of political candidates and asked for ratings between the people those names represent. In both cases, subjects had to recall information prior to rating a stimulus pair.

Interpretation of Scaling Dimensions

1) Rating scales. To interpret the INDSCAL dimensions reported in this study, unidimensional ratings of the tasks were correlated with the INDSCAL dimension weights (Wish, Deutsch, & Kaplan, 1976). Selection of those scales was based in part upon the work of Sheridan and Simpson (1979) who contend that fraction of time busy, level of problem solving complexity, and perceived stress contribute to the subjective experience of workload. Given the nonexistence of objective performance criteria and no real meaning of failure for these tasks, however, stress was not evaluated. For these same reasons, several task characteristics indentified by Borg (1978), such as the probability of failure and the adequacy of performance, were not included. Borg's work, however, did suggest the remainder of the scales. These scales are listed in Table 1.

In addition to the five rating scales, subjects rank ordered all tasks from easiest to most difficult. This ranking was included to validae the INDSCAL scaling solution. One should expect this ranking to correlate with all reported dimensions since the INDSCAL analysis was performed on similarity of task difficulty ratings.

2) Physiological measures. Two physiological measures, recorded while performing the task configuration in question were used to interpret the scaling and clustering solutions. The two measures used here were derived from physiological recordings of heart rate and respiration which were taken while the subjects were performing each of the 18 tasks. Starting with the early work of Kalsbeek and Ettema (1963), several researchers have reported decreases in heart rate variability (interbeat interval times become more uniform) under conditions of increased cognitive activity or sustained attention. Several investigators do not report such findings, however (e.g., Gaume & White, 1975). One cause for the discrepancy is the multitude of influences on heart rate variability such as blood pressure, temperature regulation, respiration, blood gasses, posture, movement, and hormones, plus the central neural influence of sustaining attention (Porges, Bohrer, Cheung, Drasgow, McCabe, & Keren, 1981). A second cause is the method chosen to quantify heart rate patterns, be it simple descriptive statistics or more sophisticated spectral analysis of heart beats over a period of time.

TABLE 1

Task Rating Scales in Order of Usage

Phase 1: Similarity of Difficulty Ratings Between Pairs

1	2	3	4	5	6	7	8
Extremely Similar		Similar	Slightly Similar	Slightly Different	Different		Extremely Different

Phase 2: Bipolar Rating Scales for Each Task

Input Complexity:

1	2	3	4	5	6	7	8
Extremely Simple		Simple			Complex		Extremely Complex

Response Complexity:

1	2	3	4	5	6	7	8
Extremely Simple		Simple			Complex		Extremely Complex

Effort Demands:

1	2	3	4	5	6	7	8
Extremely Undemanding		Undemanding			Demanding		Extremely Demanding

Feedback Adequacy:

1	2	3	4	5	6	7	8
Extremely Inadequate		Inadequate			Adequate		Extremely Adequate

Demands on Available Time:

1	2	3	4	5	6	7	8
Extremely Low		Low			High		Extremely High

Phase 3: Rank Order in Difficulty

Porges and his colleagues (Porges & Smith, 1980; Porges et al., 1981) have investigated the relationship between respiration and heart rate and have derived measures that describe this relationship and that may be related to sustained attention. During inspiration, stretch receptors in the lungs send information to the brain stem. This information signals the brain stem to inhibit (gate) the efferents in the vagus (a cranial nerve connecting brain and heart) and heart rate increases. During expiration, no gating occurs, the vagal efferents to the heart increase, and the time between successive heart beats is prolonged. This component of heart rate variability that is related to respiration is called respiratory-sinus arrhythmia (RSA).

Porges et al. (1981) use spectral analysis techniques to examine vagal activity during RSA. Heart rate and respiration rhythms are decomposed into constituent frequencies, the variance accounted for by each constituent frequency in the total heart rhythm is found, and the variances at heart period frequencies associated with normal respiration frequencies are summed. This sum is represented by V and thought to be an estimate of vagal tone. The greater vagal tone, the greater the difference in heart rate during expiration and inspiration.

Increased vagal tone was initially thought to reflect increased central nervous system influences on the heart and thus reflective of sustained attention. However, Porges, Bohrer, Keren, Cheung, Franks, and Drasgow (1981) report highly positive and reliable correlations between heart period variability (HPV) and $\hat{V}$ in a drug study on hyperactive children. Thus decreases in HPV were associated with decreases in V although the latter decreases were smaller in magnitude.

In evaluating the $\hat{V}$ measure, Porges (in press) notes that within normal populations, vagal tone does not consistently parallel individual differences in attention. Here, attention is viewed as a unitary construct for which increasing task demands require increasing amounts (Kahneman, 1973). Given the multiple resource view that increasing demand can be within a structure or distributed over structures, "sustained attention" takes on a more complex meaning. For this reason, it was thought that $\hat{V}$ might be sensitive to the mobilization of many resources (which may or may not yield poor performance) but less sensitive to the mobilization of a large amount of few resources (which will yield poor performance). If this differential sensitivity is found and if the $\hat{V}$ measure is related to the scaling solution dimensions in an interpretable way, it will shed some light on the often reported dissociation among subjective, performance, and physiological measures of workload (Moray, 1979).

The second physiological measure employed was the simple descriptive statistic of heart period variability. Porges (in press) has described it as a crude estimate of vagal tone, crude because of the host of potential unwanted influences on it in addition to the sometimes reported attentional effects. Nonetheless, it is considerably easier to measure

and calculate. Further, Porges et al. (1981) found that $\hat{V}$ was highly correlated with heart period (which was in turn highly correlated with HPV) and noted that $\hat{V}$ has yet to be shown as the better index of vagal tone.

EXPERIMENTAL INVESTIGATION

Apparatus

The task stimuli used in this study had voice, tone, and visual input and required voice, manual positioning, and discrete manual responses. The basic equipment included a 7.6 cm x 10.2 cm Hewlett-Packard Model 1300 CRT, two spring-centered, dual-axis tracking hand controls (both with index-finger triggers), a loclly fabricated voice activation key with high impedence microphone, stereo headphones, and a Bruel and Kjaer Type 7003 four-channel tape recorder. The CRT presented all the visual displays and the hand controls were used for manual responses. Auditory stimuli from the four-channel recorder plus post trial performance feedback were presented over the headphones while the voice key signaled vocal responses. A Raytheon 704 16-bit digital computer with 24K memory and A/D, D/A interfacing was used to generate inputs to the CRT and tone stimuli, and to process responses of the subjects.

Heart rate and respiration data were measured with a Grass Model 79 EEG/Polygraph. EKG signals were amplified by a Grass Model 7P5A Wide Band pre-amplifier. A rubber bellows, Grass Model PT5A volumetric pressure transducer, and a Grass Model 7P1A DC pre-amplifier interfaced with the polygraph to record respiration. Heart rate and respiration data were recorded on an Ampex SP7000 four channel FM tape recorder for later analysis. A Tektronix 016-0597 trigger generator sent pulses to a third channel of the FM tape to signal relevant portions of the recorded data.

Subjects sat in a sound and light attenuated booth approximately 80 cm from the CRT. The two hand controllers were mounted on each side of the subject's chair. The microphone was attached to the headphones which were worn at all times. All stimulus generation and physiological recording equipment were located outside of the booth.

Task Stimuli

Four criteria were used to select the tasks used in this study. First, the composite of resource structures thought necessary for task performance should vary considerably across the four tasks. Second, the tasks should have easily adjustable parameters to create easy and difficult versions. Third, the task should permit rapid learning and performance stabilization, including performance in dual task conditions. Finally, time-shared task pairs should not induce structural interference (Kahneman, 1973). In other words, peripheral structures such as eyes, ears, and limbs should not be required to perform incompatible operations

(e.g., rhythmic movements in two arms that mutually interfere).

With respect to the first criterion, selection of tasks that demand the resources identified by the Wickens' (1980) model was simultaneously straightforward and hypothetical. Two dimensions of the model--codes of central processing and modalities of input and output--are dichotomous; that is, a task either has visual stimuli or it does not or a voice response is required or it is not. For these dimensions, it is clear whether a task demands the specified resources.

The third dimension - stages of processing - uses an ordinal scale to specify resource demand. It is thus impossible to assert with absolute assurance that a task manipulation is exclusively perceptual/central or response in its locus, since this dichotomy cannot be defined by structural morphology in the same way as the modality dimension. Nevertheless, the stages dichotomy can readily be justified on the basis of research that has identified and named stages of processing with additive factors methodology (Sternberg, 1969).

To meet the second, third, and fourth criteria, initial forms of the tasks had to be created and performed by subjects. Response data were analyzed in conjunction with subject comments. Based upon these inputs, several task parameters were modified to avoid data-limited tasks, or tasks so easy that performance could not be improved with increased effort (Norman & Bobrow, 1975). The tasks were also fine-tuned to prevent excessively difficult dual task pairs, a situation that could have produced performance changes related to motivation, not the competition for resource predictions. The final form of each experimental task is described below.

1) Critical tracking task (CT). Developed by Jex, McDonnel, and Phatac (1966), this is a one dimensional compensatory tracking task with control dynamics of the form $Y = K\lambda/S - \lambda$. These dynamics form an unstable positive feedback loop that drove the error cursor to the edge of the display at a velocity proportional to the error and the parameter λ. Lambda was set at a value of 0.95 to create an easy version of the task and 1.90 to create a difficult version. The cursor, which travelled in a vertical plane, required fore-aft movements of the joystick to null the error. This orientation was chosen to reduce incompatibility with the other tasks. Control was with the left hand except when the task was time-shared with itself, in which case both hands were used. Panel A of Figure 2 depicts the single task version of the CT.

This task has been related to processing resource demands in several dual task situations (Jex & Clement, 1979). According to the multiple resource model, this task should demand resources dedicated to spatial codes, visual-input modalities and manual response modalities. For the easy task version, the perceptual/central and response execution resource demands were presumably low, although the Wickens and Kessel (1980) evidence suggests resources related to responding receive the greater demand. When lambda was increased to 1.90 and system control became more

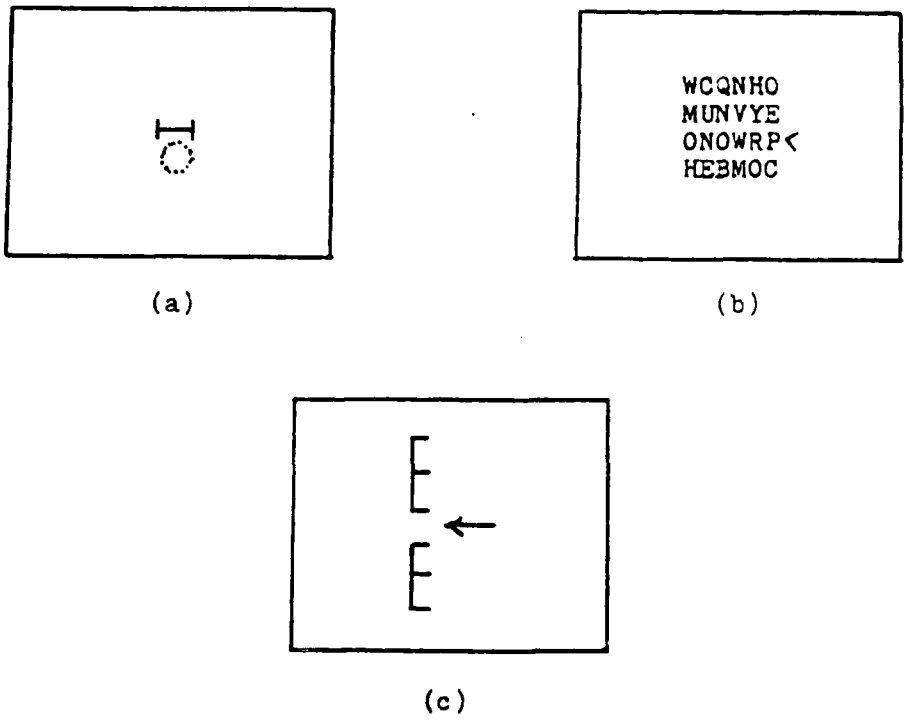


Figure 2. Visual displays for the experimental tasks in single task conditions: a) Critical Tracking Task, b) Visual Search Task, and c) Tone Judgment Task.

unstable, the resource locus of this manipulation presumably was response execution.

2) Visual search (VS). Adapted from a task used by McCarthy and Donchin (1981), this task required subjects to search for the target word NOW which was embedded in a 4 x 8 array of letters and non-letters. The array contained 15% non-letters for the easy task and all letters for the difficult task. Target probability was set at 35%. The display was continuously modified, one line of the display changed at a time. The line of change was random and occurred from 3.5 to 10.0 seconds after the previous change. When the target was detected, subjects responded by squeezing a trigger on the right joystick. This response placed an indicator (<) next to the line where the target occurred. A subsequent target could occur in the same line, in which case all eight characters would change and the indicator would disappear, or in a different line, in which case the previous target line would remain on the screen. When time-shared with itself, subjects searched for two targets, NOW and ONE, with ONE requiring a left hand response. In this condition, both targets could appear anywhere in the array. This display subtended 1.80 degrees of visual angle. The single task version of VS is depicted in Panel B of Figure 2.

The composite resource demands of this task should be defined by verbal codes of central processing, visual input and manual response modalities. The easy single task demands resources from both stages of processing, but McCarthy and Donchin (1981) have shown using evoked brain potentials that the demands from a similar task are clearly more perceptual in nature. Deleting the non-letter characters in the array to make the search task more difficult presumably increased resources from the perceptual/central structure.

3) Auditory Sternberg (AS). Using the basic Sternberg (1969) memory search paradigm, this task required subjects to retain in memory either two (easy task) or five (difficult task) memory set letters. Subsequent auditory presentations of letters to the right ear required a vocal YES if the heard letter came from the memory set, a vocal NO otherwise. Memory set letters occurred 50% of the time. Letter stimuli were presented at random intervals between three and seven seconds. When the subject responded, the word HEARD appeared on the CRT in front of him. This feedback indicated the response was loud enough to activate the voice key and measure the response latency. When the task was time-shared with itself, a unique memory set of n = 2 letters was presented to each ear. Letters were presented randomly between ears but they never overlapped to preclude auditory masking. A YES was the correct response if the heard letter was both from a memory set and it occurred in the ear assigned to that memory set. Unlike the other three tasks used in this study, the vocal response precluded true time-sharing of responses for the two AS tasks.

According to the multiple resource model, this task should demand resources dedicated to verbal codes of central processing and modalities of auditory input and vocal response. With respect to stage-related resources, the mental rehearsal and matching of letters would indicate that perceptual/central responses were utilized more than those needed for the simple response. Further, manipulation of memory set size in the Sternberg task has been shown to affect resource demands of intermediate stages of processing (Wickens, Derrick, Micalizzi, & Beringer, 1980). Thus, the difficult single task should demand more perceptual and central processing resources than the easy task.

4) Tone judgment (TJ). In this absolute judgment task a 1200 millisecond tone burst to the left ear required subjects to position an arrow on a vertical scale associated with that tone frequency. The easy task required absolute judgment of four discriminable tones (400 Hz, 480 Hz, 650 Hz, and 850 Hz) while the difficult task added a tone at each end of the scale (325 Hz and 1000 Hz). The response scale consisted of four (six) response brackets with an open neutral position in the center (see Panel C, Figure 2). Subjects heard each of the four or six tones prior to task performance. When each tone was presented during this familiarization step, an X was displayed in the bracket associated with that tone.

Tone stimuli were presented in random order during actual task performance. Left hand joystick manipulations of zero order dynamics positioned the response arrow. When positioned, subjects depressed a trigger on the stick to indicate their responses. The arrow subsequently returned to the neutral position and an X appeared in the correct response bracket. At zero, one, or two seconds after disappearance of this feedback (determined randomly), a new tone stimulus was presented.

Although zero order dynamics were employed, the stick had a predetermined gain parameter that enabled the arrow to continue travelling for a short distance after stick movement ceased. Further, excessive stick deflection could send the response arrow past the outer bracket positions. These two characteristics increased the difficulty of responding beyond a simple, unskilled control movement. When four bracket positions were employed the display subtended 2.5 degrees of visual angle. Six positions subtended 3.5 degrees.

When time-shared with itself, two sets of brackets, two arrows, and both joysticks were used. The same four tones were employed, presented randomly to either ear, but never concurrently. Here the subjects' task was to position the arrow associated with the ear of tone delivery. When time-shared with the AS task, both auditory stimuli were deliverd to separate ears. On some occasions, tones and letters overlapped. However, since these pure tones and the speech were at approximately the same sound pressure level, they did not mask each other (Deatherage, 1972).

Performance on this task should demand processing resources associated with spatial codes of central processing and modalities of auditory input and manual response. The perception and judgment portion of this task was not trivial. Kidd and VanCott (1972) state that the human limit of absolute judgment of frequencies when used as auditory signals is four or five. Therefore, the perception and memory processes used should demand perceptual/central processing resources. Increasing the difficulty of TJ by adding two tones should increase the demand for these resources. Six tones is still within the range of human ability (Deatherage, 1972) but is not recommended for systems using auditory displays. As described earlier, responding in this task was no trivial matter either. Therefore, it was posited that TJ required response execution resources.

Table 2 summarizes the composite resource demands of the four tasks in their "easy" configurations. The numbers represent ordinal values only. For the processing stage resources, the comparisons of 1 versus 2 reflect within-task comparisons only. Response demands for critical tracking, assigned a 2, should not be judged as equal to response demands for tone judgment, also assigned a 2. Equation of these demands would require the four tasks be time-shared with a validated battery of secondary tasks whose resource demands are known. These secondary tasks currently do not exist.

The entries in Table 2 were used to predict dual task performance decrements in support of the first hypothesis. The coding scheme we used for this prediction is shown in Table 3. The ordinal data limitation precluded predicting equal decrements for different task pairs. Rather, each task was considered as a "home" for referent task and all other tasks were paired with it. The numbers were then used to predict the number and relative amount of resources demanded by the four task pairs in which home was represented. These demands were rank-ordered from one to four, the greatest predicted demand receiving a one.

For the critical tracking task, the CT/CT task pair (time-shared with itself) should exhibit the greatest competition for resources and thus the greatest performance decrement. The CT/VS and CT/TJ pairs represent intermediate resource overlap with the CT/TJ being the more difficult because of the increased response demand. The CT/AS pair has almost no resource competition and should be time-shared very efficiently.

With respect to the visual search home task, the VS/VS task pair should have the greatest decrement in performance. Pairs VS/TJ and VS/CT were similar in resource demand except that the manual response for TJ was much harder for subjects in the pilot study to learn and execute than the manual response for CT. Further, although TJ had an auditory input, visual processing was required to monitor the accuracy of the manual response. Therefore, it was predicted VS/TJ would have the greatest performance decrement of the two. Finally, the VS/AS pair should be time-shared most efficiently of the four.

TABLE 2

Composite Resource Demands of the Four Tasks (Demand level 2 > 1 > 0)

	RESOURCE DEMAND							
	STAGES		MODALITIES				CODES	
Task	Perceptual/ Central	Response	Audi- tory	Visual	Manual	Vocal	Spatial	Verbal
Critical Tracking (T)	1	2	0	1	1	0	1	0
Visual Search (V)	2	1	0	1	1	0	0	1
Auditory Sternberg (S)	2	1	1	0	0	1	0	1
Tone Judgment (T)	2	1	1	1	1	0	1	0

TABLE 3

Composite Resource Demands of the Ten Dual Task Pairs

	RESOURCE DEMAND							
	STAGES		MODALITIES				CODES	
Task Pair	Perceptual/ Central	Response	Audi- tory	Visual	Manual	Vocal	Spatial	Verbal
TT	1,1	2,2	0	1,1	1,1	0	1,1	0
VV	2,2	1,1	0	1,1	1,1	0	0	1,1
SS	2,2	1,1	1,1	0	0	1,1	0	1,1
JJ	2,2	1,1	1,1	1,1	1,1	0	1,1	0
TJ	1,2	2,1	1	1,1	1,1	0	1,1	0
TV	1,2	2,1	0	1,1	1,1	0	1	1
TS	1,2	2,1	1	1	1	1	1	1
VS	2,2	1,1	1	1	1	1	0	1,1
VJ	2,2	1,1	1	1,1	1,1	0	1	1
SJ	2,2	1,1	1,1	1	1	1	1	1

T = Critical Tracking, V = Visual Search, S = Auditory Sternberg, J = Tone Judgment

Unlike the preceding two home tasks, the AS/AS pair was not predicted to have the greatest performance decrement for the auditory Sternberg tasks. The primary reason is that vocal responses cannot be time-shared, thus easing the burden of resource demand. Further, to avoid masking, the speech input was sequential and not concurrent. These limitations make it impossible to argue that AS/AS is time-shared to the same extent as are the other three tasks. Therefore, the AS/TJ pair, which had concurrent input and responses, should exhibit the greatest decrement followed by AS/AS. Based upon the numbers in Table 2, AS/VS should have the third largest decrement and AS/CT the smallest.

For the tone judgment task, the TJ/TJ pair represents the greatest competition for resources even though the tone inputs never overlapped. Tone judgment and critical tracking (TJ/CT) should be the next most difficult task to perform. The interference produced by TJ/VS and TJ/AS should be very similar. AS overlaps with TJ on modality of input but not response; VS overlaps with TJ on modality of response but not input. In all other respects, VS and AS are very similar. Therefore, one should expect these two tasks to produce similar performance decrements when paired with tone judgment.

METHOD

Subjects

Twenty-two right-handed male subjects were recruited to participate in this study. All reported normal or corrected to normal vision and normal hearing. Because musical training might have influenced performance in the tone judgment task, these subjects were screened for vocal or instrumental competence. All claimed little or no musical background. Further, since forward movement of the stick in the critical tracking task moved the error cursor down (opposite to the movement of the artificial horizon in an aircraft attitude indicator), subjects with flying experience were eliminated due to potential negative transfer. Subjects ranged in age from 18 to 22. Each was paid $3.50 per hour plus performance bonuses.

Design

All 22 subjects performed each of the 18 tasks, 4 easy single task, 4 difficult single tasks, and 10 dual task pairs formed by all pairwise combinations of the easy tasks. Each task trial lasted two minutes. These tasks were performed during three one-hour sessions, two practice and one experimental. Task order during the practice sessions was fixed for the first 45 minutes with the remaining time devoted to tasks on which subjects needed additional training. For the experimental session, task order was unique for each subject and determined by a Latin Square procedure. One standard form of an 18 x 18 square was first randomized by rows and then columns with the restriction that the final square contained each task in each of the 18 serial positions. Task order for

subjects 1-18 followed the consecutive rows of the square; subjects 19-22 repeated the order dictated in rows 1-4. Physiological recordings of heart rate and respiration were taken during this third experimental session.

The three task performance sessions were run on different days. At the end of the third session, task ratings were collected. First, similarity of task difficulty ratings were gathered for the 153 pairs of tasks. These pairs were presented in the same order for all subjects by means of an ordering developed by Ross (1934) which reduces presentation bias in paired comparison ratings. Twenty task pairs selected at random were repeated as pairs 154-173 for all subjects to provide estimates of rater reliability. Following these ratings, each of the 18 tasks was rated on the unidimensional scales listed in Table 1 and then ranked according to difficulty.

Procedure

1) Bonus payments. A bonus system was used to encourage good performance and equal concern for both tasks during dual task trials. During sessions 2 and 3, the previous day's single task (easy) performance was used to guage performance during dual task trials. If performance of both tasks in a dual task trial met or exceeded the respective single task performances of the previous day, a 15 cent bonus was paid for that trial. If one task met this criterion and the other showed a decrement of less than one standard deviation of the single task level, a 10 cent bonus was paid. Two dual task decrements of less than one standard deviation each earned 5 cents, while any decrements greater than these earned no money. With this incentive system, subjects could earn a maximum of $1.50 per hour above their $3.50 per hour base pay.

2) Practice: Session 1. After reporting to the lab, subjects were given a general explanation of the study and the demands that would be placed upon them. Subsequent to signing a consent form, subjects were introduced to the tasks and began practicing. All single task trials were performed first followed by the dual task combinations. Prior to each dual task trial, subjects were told to give equal priority to both tasks. The success of this instruction was evaluated and conveyed to them by referencing dual task performance to respective single task performances. The latter part of the session was tailored to each subject's learning difficulties. No bonuses were paid during this session.

3) Practice: Session 2. Task order was again fixed for all subjects but this time began with the difficult single tasks. Performance feedback was provided after each trial and bonuses were paid for dual task performance where appropriate. Again, the last 15 minutes of the session was devoted to extra practice on tasks which proved to be difficult.

4) Experimental: Session 3. During this session heart rate and respiration data were collected while the subjects performed the tasks. Prior to beginning the session, each subject's cardiac activity was detected by three Ag/AgC/electrodes secured in the following locations: left temple, right calf, and left calf, the latter serving as a ground. A rubber bellows was secured below the rib cage to measure respiration and connected to the equipment mentioned earlier. As the subject sat at rest in the experimental booth, the controls on the polygraph were adjusted to clean up and amplify the incoming signals.

To both warm up on the tasks and ensure that the signals would be measured and recorded properly, subjects performed each of the four difficult single tasks. Each then performed the 18 tasks in the order discussed in the design section. Each task was preceded by a one minute baseline or complete rest period. The onset of this period was indicated by sending a discrete pulse to the FM recorder where the analog signals were being recorded. At the end of one minute, the task began. During task performance, subjects were instructed to make no unnecessary movements, including posture changes, and engage in no irrelevant behavior. When the task was complete, subjects were allowed to stretch, take deep breaths, and so forth in preparation for the next three minutes of limited movement.

5) Task ratings: Session 4. Session 4 followed Session 3 on the same day. Subjects sat at a table where 18 - 3" x 5" cards were displayed, each card pictorially depicting a task that had just been performd. Each card was also numbered from 1 to 18. Below the task cards the subjects saw the Similarity of Difficulty scale printed on a 5" x 8" card. The scale ran from 1 (extremely similar) to 8 (extremely different) with a verbal anchor for each number. The experimenter then explained the rating procedure to the subject. Several task pairs were selected at random and the subject assigned ratings to the pairs. When the subject understood the use of the scale, the ratings were collected. The experimenter would call out the pair number from the Ross ordering (e.g., 6, 13), the subject would locate these cards (sometimes placing them next to each other to aid concentration) and then verbally give a rating. These ratings were recorded by the experimenter.

After a rest break, subjects were given the unidimensional scales and a response sheet for each task and asked to assign these ratings. Upon completion, the tasks were rank ordered by perceived difficulty on another sheet. The pictorial cards were available as prompts during these ratings.

Although all 22 subjects provided data for all tasks and conditions, subsequent analyses of the heart rate variability data revealed that these data were incomplete for three subjects. Because much of the analysis that follows attempts to relate the multiple measures provided by each subject, all data for these three subjects were excluded from further consideration.

RESULTS

Task Performance

To test Hypothesis Two, the predicted dual task performance decrements listed in Table 4 had to be confirmed. Before these predictions could be evaluated, however, several transformations of the raw task performance data were needed. These transformations are discussed in the following sections. Table 5 presents the performance variables selected for each task along with their means and standard deviations. The method of selection is presented below.

1) Selection of Performance Variables. Performance on the Critical Tracking task was measured by root mean squre (RMS) error. Van Cott and Kinkade (1972) conclude that this error measure is a most adequate measure of performance in this type of tracking task.

Two measures were collected on the Auditory Sternberg task: Reaction time (RT) to correct responses and percent correct responses. To check for a possible speed/accuracy tradeoff, RTs to all S tasks during Session 3 (single-easy, single-hard, 4 dual task pairs) were compared to percent correct responses for the same tasks. The following mean RTs (in milliseconds) and percents were found: 494 ms, 98%; 606 ms, 96%; 562 ms, 98%; 566 ms, 97%; 571 ms, 96%; and 611 ms, 97%. Although the reaction times increased, no substantive difference in percent correct occurred. Therefore, reaction time was chosen as the single performance measure for the Auditory Sternberg task.

Since the Visual Search task was analogous to a signal detection or inspection task, a measure of detection sensitivty was deemed appropriate. However, since the d' distribution assumptions from signal detection theory could not be met in such short task trials, the distribution-free sensitivity measure A' (Craig, 1979) was computed. A' is calculated from the hit and false alarm data and ranges from .5 (chance) to 1.0 (perfect performance). A second task measure, RT to hits, was also taken.

Both measures changed considerably from task combination to task combination but in no correlated fashion. Thus, both measures offered useful information about subject performance and both were utilized. To combine A' and RT into one performance measure, the respective standard deviations were calculated for all S tasks. Each subject's A' and RT were then divided by these standard deviations, and the resultant quotients were subtracted one from another to produce a linear composite score. This division by standard deviation had the effect of equating .029 units of A' to 155 milliseconds of latency. Each composite score then reflected the equal contribution of A' and RT for that one trial.

Two performance measures were collected for the Tone Judgment (J) task: RT for corrects and percent corrects. A check for the

TABLE 4

Predicted Rank Order of Performance Decrements for Each Home (Referent) Task

HOME TASK

Critical Tracking (T)		Visual Search (V)		Auditory Sternberg (S)		Tone Judgment	
Task Pair	Rank	Task Pair	Rank	Task Pair	Rank	Task Pair	Ra
TT	1	VV	1	SJ	1	JJ	1
TJ	2	VJ	2	SS	2	JT	2
TV	3	VT	3	SV	3	JV	3
TS	4	VS	4	ST	4	JS	4

TABLE 5

Task Performance Measures: Means (Standard Deviations)

Task	Single-Easy	Single-Hard	With Critical Tracking	With Visual Search	With Auditory Sternberg	With Tone Judgment
Critical Tracking (RMS error)	0.06 (0.02)	0.17 (0.04)	0.15 (0.06)	0.10 (0.03)	0.07 (0.03)	0.10 (0.03)
Visual Search (A', RT composite)	1.88 (0.89)	3.05 (1.56)	2.91 (2.10)	4.79 (1.62)	2.37 (1.07)	3.89 (1.46)
Auditory Sternberg (RT in milli-seconds)	494 (91)	566 (92)	505 (71)	562 (103)	571 (83)	611 (122)
Tone Judgment (% correct, RT composite)	3.42 (1.11)	6.61 (1.50)	4.02 (1.25)	5.74 (1.51)	5.18 (1.75)	6.11 (1.34)

speed/accuracy tradeoff revealed that increasing reaction times across all J tasks were accompanied by increasing error rates, suggesting that the tradeoff did not occur. However, the error rates were so high (ranging from 11% to 25%) that the RT data alone could not project a complete picture of subject performance. As with the Visual Search task, the RT correct data and the percent correct data were combined into one performance measure. Standard deviations were calculated for both tasks and used to divide the separate scores for each subject. Resultant quotients were subtracted to produce a linear composite score. Here 9.40 units of percent correct equated to 162 milliseconds of latency and, as with the VS task, each composite score reflected the equal contribution of reaction time and percent correct.

2) Normalized Performance Decrements. To analyze performance in the dual task trials, further transformations on the performance data were necessary. Table 5 presents four very different types of performance measures that cannot be equated in their existing state. Yet these measures must somehow be combined to produce a measure of performance that reflects the joint consequence of performing two tasks simultaneously. Further, that measure must be comparable across all dual task pairs.

Adopting a procedure reported by Wickens, Mountford, and Schreiner (1981), task performance was normalized or converted to standard scores. The four single easy tasks (Te, Ve, Se, Je) were considered to be baseline or reference tasks. Performance on all other versions were compared to the baseline (e.g., Th, TT, TJ, TS, TV), and the differences computed. These difference scores, or performance decrements, were then analyzed with a one-way ANOVA with repeated measures. The goal of the ANOVA was to obtain the Mean Square error term, or that residual variance in the scores that was not associated with different task types (the systematic variance). The square root of the error variance was then used in the denominator of the standard score transform equation. Placed in the numerator were the original baseline score and the comparison score. For example, one such equation was $Th - Te/\sqrt{MS\ error}$. This type of equation produces a "normalized" performance decrement.

The main assumption underlying this procedure is that if little error variance is found in the performance decrement scores, "a given change in performance from single to dual-task conditions represents proportionately greater loss in efficiency than for tasks that are highly variable" (Wickens et al., 1981). When this procedure was used with all four types of tasks, it permitted all four performance measures to be cast in terms of their performance variability. Thus, one unit of performance decrement for the Critical Tracking task was equated to one unit of performance decrement for the other three tasks.

Table 6 presents the means and standard deviations of these normalized performance decrement scores for each task. Note the high degree of variability associated with many of the means. This occurred because on any given dual task trial, a subject may not have assigned

TABLE 6

Normalized Performance Decrement Scores: Means (Standard Deviations)

Task	Single-Hard	With Critical Tracking	With Visual Search	With Auditory Sternberg	With Tone Judgment
Critical Tracking	3.35 (1.14)	2.93 (1.68)	1.14 (0.60)	0.37 (0.58)	1.35 (0.71)
Visual Search	0.88 (1.05)	0.80 (1.15)	2.10 (0.85)	0.37 (0.61)	1.44 (1.24)
Auditory Sternberg	1.28 (1.52)	0.21 (0.96)	1.25 (1.03)	1.41 (1.01)	2.18 (1.41)
Tone Judgment	2.85 (0.80)	0.50 (0.83)	1.95 (1.29)	1.49 (1.14)	2.26 (0.73)

equal priority to both tasks. Thus, the normalized decrement for one of the two tasks for that subject may have been very small (or may have even shown an improvement over the single task baseline) whereas the other task bore the main cost of concurrent performance (had a large decrement). A second subject may have exhibited a reversed priority effect while yet a third managed to follow the equal priority instructions. As such, the performance decrements for individual tasks presented in Table 6 only roughly indicate what happened in the dual task trials. Although a subsequent analysis used the individual data points that are averaged in Table 6, a more comprehensive view of the dual task performance measures is valuable at this point.

To create one performance decrement score per subject per dual task trial, the individual task performance decrements were averaged. For example, when Subject #4 performed the Visual Search and Tone Judgment tasks together, his V decrement score was .572 and his J decrement was 1.537. Based upon the normalization procedure, his J task performance was nearly three times worse than his V task performance, and taken together, Subject 4's dual task performance decrement was scored as 1.055. Following this procedure, all dual task decrement scores were found. In a format similar to Table 4, the means and standard deviations of these decrements are reported for each Home task in Table 7.

When the predicted order of performance decrements from Table 4 is compared to the means listed in Table 7, there is a good deal of agreement. In general, increased resource competition between tasks is associated with poorer dual task performance. The one major exception to the predictions appears to be the joint performance of the Tone Judgment and Critical Tracking tasks. When both the T and J tasks are listed as Home, the TJ (or JT) pair reveals much better dual task performance than predicted. To confirm if the impressions drawn from the comparison of Tables 4 and 7 have any validity, the next section reports a statistical analysis of the dual task performance data.

3) Analysis of Performance Data. Performance analyses focused on the joint performance decrements that are summarized in Table 6. First, these decrements were visually represented in Performance Operating Characteristic (POC) spaces (Norman & Bobrow, 1975). Figure 3 depicts a hypothetical POC that contains the joint performance of tasks A and B. The two axes of the POC are scaled in normalized decrements, permitting performance comparisons. Had tasks A and B been time-shared perfectly, their joint performance would have been plotted at point P. Any point on the positive diagonal lower than P indicates that both tasks suffered equal decrements in performance. Points on either side of the diagonal reflect unequal priorities during dual task performance. The point plotted in Figure 3 reveals a relatively greater dual task performance decrement for task A than for task B.

The four POCs relevant to the performance analysis are presented in Figure 4. Identified above each panel is the Home task whose performance decrements are plotted on the y-axis. The decrements for the three other

TABLE 7

Dual Task Performance Decrements for Each Home Task[1]

HOME TASK

Critical Tracking (T)		Visual Search (V)	
Task Pair	Mean (Standard Deviation)	Task Pair	Mean (Standard Deviation)
TT	2.99 (1.68)	VV	2.16 (0.88)
TV	1.02 (0.67)	VJ	1.66 (1.05)
TJ	0.92 (0.62)	VT	1.02 (0.67)
TS	0.38 (0.50)	VS	0.81 (0.63)

HOME TASK

Auditory Sternberg (S)		Tone Judgment (J)	
Task Pair	Mean (Standard Deviation)	Task Pair	Mean (Standard Deviation)
SJ	1.75 (0.93)	JJ	2.27 (0.77)
SS	1.63 (0.97)	JS	1.75 (0.93)
SV	0.81 (0.63)	JV	1.66 (1.05)
ST	0.38 (0.50)	JT	0.92 (0.62)

[1]Larger numbers represent greater performance decrements.

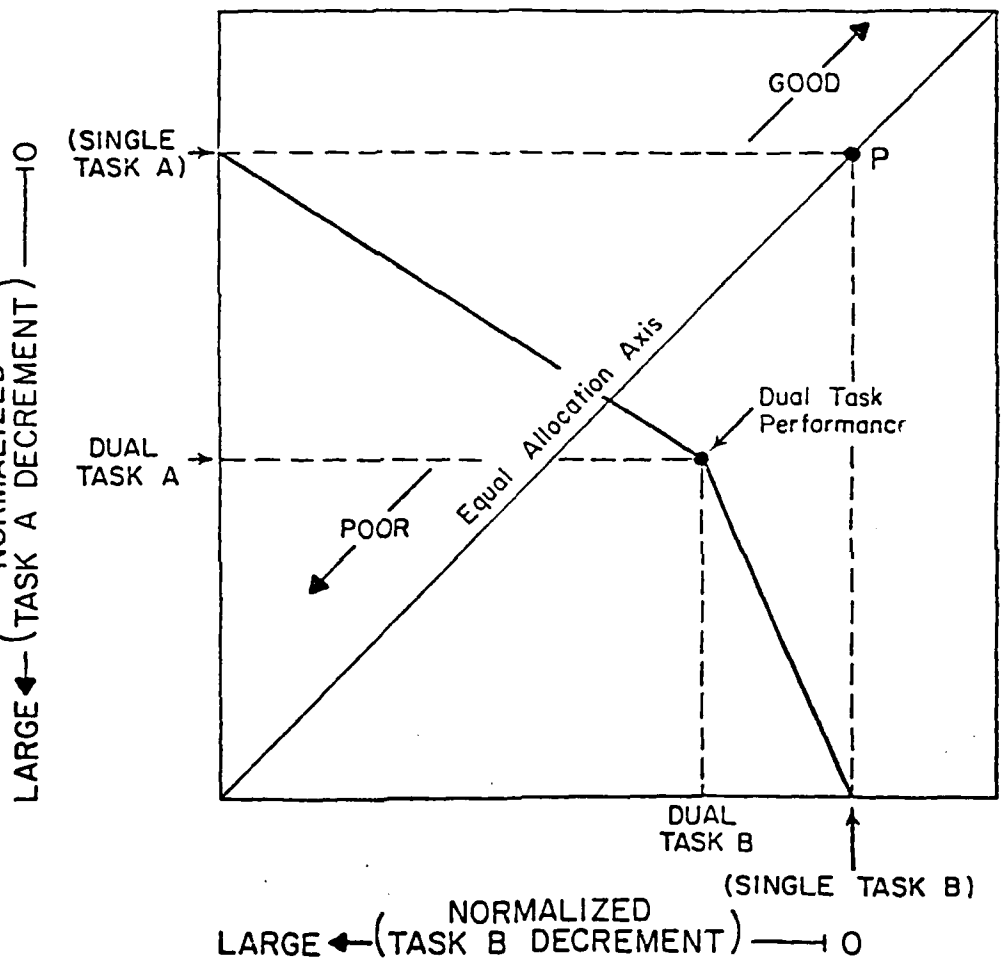


Figure 3. Hypothetical Performance Operating Characteristic (POC) with equated performance scales.

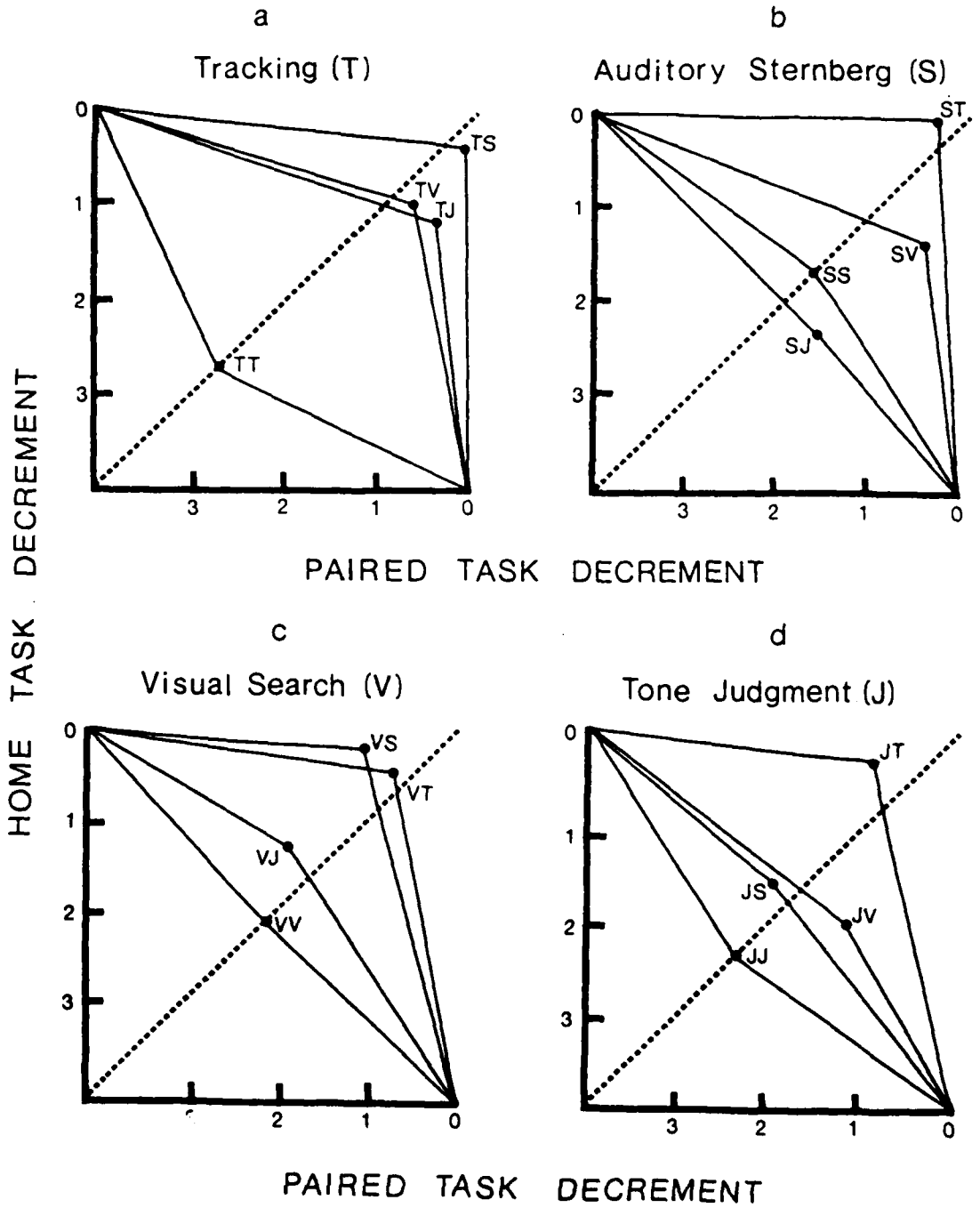


Figure 4: POC representation of the dual task performance decrements for the four experimental tasks.

tasks time-shared with the Home task are plotted on the x-axis. Zero represents single task performance; numbers 1 - 3 represent increasing performance decrements. The dashed line represents the equal allocation or equal priority axis.

To determine if these points differed from one another in each of the four POCs, a repeated-measures multivariate analysis of variance (MANOVA) was run on the data from each panel. Each subject was treated as a bivariate observation whose two dimensions consisted of the two normalized dual task decrement scores. The following Wilks' Lambda and approximate F values were obtained from the four MANOVAs: Critical Tracking--Lambda = .446, $F = 13.77$ ($p < .01$); Visual Search--Lambda = .638, $F = 6.97$ ($p < .01$); Auditory Sternberg--Lambda = .572, $F = 8.92$ ($p < .01$); and Tone Judgment--Lambda = .619, $F = 7.51$ ($p < .01$). Each F calculated was evaluated against an F critical with 6 and 146 degrees of freedom and all four were statistically reliable.

With the rejection of the four null hypotheses--the points in each panel of Figure 4 did differ--the next step was to discover the sources of the differences. According to Harris (1975, p. 104), the critical values for the multivariate planned comparisons depend greatly upon two factors: Was the linear combination of outcome factors selected on a priori grounds, and was the particular contrast among groups selected a priori? For both questions, the answer here is yes because only two outcome factors were used and because the comparisons were intended to verify the decrement order of Table 4. Therefore, these contrasts reduce to simple Hotelling T^2 analyses for differences between selected mean vectors on the dependent variables (Harris, 1975, p. 105).

For the data plotted in the Critical Tracking panel of Figure 4, the TT task pair produced reliably worse performance than the TJ pair ($T^2 = 45.52$, $p < .01$). Pairs TJ and TV did not differ from each other ($T^2 = 3.35$, $p > .05$); however, both differed from the TS pair ($T^2 = 22.17$ and 22.42, respectively; both $p < .01$).

With respect to the predicted performance decrements listed in Table 4 and the associated discussion, these data support the predictions derived from Wickens' multiple resource model. For the extreme tasks (TT, TS), the model specified the performance decrements correctly. For the intermediate tasks (TJ, TV), the model suggested that the TJ pair should be the more difficult of the two, but only slightly. In fact, no difference in performance was found. One surprise to note here was how well the TJ pair was performed. Based upon the resource competition scheme presented in Table 3, one would expect to find the joint plot of this task pair closer to the origin of the POC. Possible reasons for this occurrence will be discussed later.

In the planned comparisons for the Auditory Sternberg data (Figure 4, panel b), SS and SJ did not differ from each other in performance ($T^2 = 5.70$, $p > .05$), but both produced reliably worse performance than SV ($T^2 = 17.49$, $p < .01$ and $T^2 = 23.03$, $p < .01$, respectively). Further,

SV in turn was associated with worse performance than ST (T^2 = 18.67, $p < .01$).

These results also support the performance decrement predictions. Since SS was not truly time-shared (no simultaneous auditory input; no simultaneous vocal response), it was no more difficult to perform than the second task pair of the four (SJ). Further, the other two tasks decreased in competition for resources and thus task difficulty as the classification system in Table 3 suggested.

For the data plotted in the Visual Search panel of Figure 4, the VV task pair produced reliably worse performance than the VJ pair (T^2 = 8.74, $p < .05$) which in turn produced reliably worse performance than the VT pair (T^2 = 9.14, $p < .05$). However, task pairs VT and VS produced no difference in performance (T^2 = 2.19, $p > .10$).

Again, the predictions from the multiple resource model were generally supported. The greatest competition for common resources (VV) was associated with the worst dual task performance while lesser competition was associated with increasingly better performance. Task pair VS, however, was predicted to be time-shared the best of the four Visual Search Home tasks. In fact, it was time-shared reliably better than VV and VJ but had no better performance than VT.

Finally, for the Tone Judgment data in Figure 4, the dual task pair JJ was time-shared more poorly than either JS (T^2 = 9.70, $p < .05$) or JV (T^2 = 8.70, $p < .05$) but neither of these latter two tasks differed from each other (T^2 = 6.22, $p > .05$). The last task in the panel, JT, was time-shared with the best of all, differing from JS (T^2 = 15.39, $p < .01$) and JV (T^2 = 16.17, $p < .01$).

This is the one panel of the four where the predictions derived from the multiple resource model were not supported. The primary culprit was the Tone Judgment-Critical Tracking task pair, on which the two tasks were time-shared very effectively, even when the classification scheme of Table 3 suggested that this would be a difficult combination of tasks to perform. When this task pair was discussed in panel a (Critical Tracking), its apparent relative ease of performance failed to differentiate it from the TV pair. In panel d, it was predicted to be the second most difficult task pair; in fact, it was the easiest. Considering the other two task pairs here--JS and JV--the JV pair was predicted to be the more poorly performed of the two, but only slightly so. In fact, their performance decrements were equal.

To summarize, the predictions derived from the Wickens' resource model were generally supported. Task decrements within each Home task set generally lined up as expected. Three small deviations from these predictions and one larger one were encountered. Differences, albeit small ones, were predicted between TJ and TV, between VS and VT, and between JV and JS but none occurred. Further, JT was time-shared much more efficiently than predicted. Had its associated performance been

worse, the TJ versus TV equality would have been removed.

The most likely cause of these reversals is the rather simple classification system depicted in Table 3. Because a task's resource demands can be specified only on an ordinal scale, and since ordinal values cannot be added, one can only roughly guage the level of resource demand for a dual task pair. Thus, the predictions of small differences in performance between different sets of dual tasks perhaps are better labeled as guesses. Finally, although this procedure may be on questionable ground because of the ordinal scales used to rate tasks, a procedure was derived to predict the total amount of resource competition for each of the dual task pairs. When these values, for each task pair in Table 3, were correlated with the final combined performance decrement in Table 7, the final Pearson correlation value was +0.84, indicating the successful prediction by the model.

While noting exceptions, the conclusion to be drawn from this section of the study is that Hypothesis One is supported. For this set of tasks, increasing levels of resource competition do lead to poorer dual task performance. With the validation of these tasks, the next portion of the study will report how this concept of resource competition maps onto the subjective dimensions of task workload.

Multidimensional Scaling

To determine the "hidden structure" or psychological dimensions that produced the similarity of task difficulty ratings, these ratings were analyzed by multidimensional scaling techniques. The resultant dimensions were interpreted by the location of the task stimuli along the dimensions and by correlations of dimension weights with performance decrements, unidimensional ratings, and heart period variability scores. Derivation of these heart period variability scores is described below.

1) Heart Period Variability Measures. To derive Vagal Tone ($\hat{V}$) and Heart Period Variability (HPV), the heart rate and respiration data recorded on the FM tape were analyzed with cross-spectral techniques. Since this type of analysis assumes that the data are sequences of events equally spaced in time, the beat-to-beat measures were converted into time-based data (Porges et al., 1981). The heart period was sampled from the tape every 250 milliseconds and computed as the sum of each heart period that partially or wholly occupied the 250 millisecond interval. Respiration amplitude was sampled on each R-wave and linearly interpolated to provide estimates on successive 250 millisecond intervals. These time series for heart period and respiration were then prestationed by removing linear trends using the technique of calculating successive differences (Porges et al., 1981). Assuming that resting adults take between 6 and 24 breaths per minute, the cross-spectral analysis was then conducted on the data sets between .1 Hz and .4 Hz. The resultant variables, $\hat{V}$ (the amount of variance of the heart period process occurring between .1 and .4 Hz) and HPV, the simple measure of heart period variability, were then transformed to their natural

logarithms to normalize the distributions of these estimates (Porges et al., 1981).

Initially, each subject had three $\hat{V}$ and three HPV scores per task: one baseline or pretask resting score, a score for the first minute of the task, and a score for the second minute. The two on-task measures were compared to determine if continued performance altered these physiological responses. For the $\hat{V}$ measure, minute one to minute two comparisons for all tasks produced very high correlations (median r = 0.81) and no reliable differences between mean scores. The HPV data were very similar. The median correlation was r = 0.68 and only two means were reliably different. Therefore, the data for both minutes were combined and averaged for both variables.

For each subject and each task, this on-task average score was subtracted from the immediately preceding resting state, or baseline, score. The goal was to determine if the attentional demands of the tasks could be indexed by decreases in HPV, $\hat{V}$, or both. Table 8 presents the descriptive statistics associated with these baseline-minus-task comparisons. These are the means that were used in the subsequent analysis of the scaling dimenions.

2) INDSCAL Solution. The similarity of difficulty paired comparison ratings for all 18 tasks produced a matrix of 153 ratings for each subject. Technically, these are dissimilarity ratings since the highest value on the rating scale--8--represented "Extremely different in task difficulty." In addition to this full set of ratings, each of the 19 subjects provided repeated ratings on the first 15 tasks specified by the Ross ordering (1934). These repeated ratings were used to estimate the subjects' reliabilities. Using a minimum reliability criterion of r = 0.50 (Nygren & Jones, 1977), all estimates of reliability exceeded this value; therefore, the ratings from all 19 subjects were used in the INDSCAL analysis.

The 19 matrices of ratings, or proximities, were analyzed by the Symmetric INDSCAL computer program. Scaling solutions based upon the combined data were computed in two, three, four, and five dimensions. The variance accounted for (VAF) by these solutions was 48.3%, 55.0%, 57.9%, and 60.6%, respectively. The corresponding median correlations between the subjects' data (scalar products) and the four scaling solutions were 0.70, 0.74, 0.75, and 0.78, respectively.

Both statistical and interpretation criteria were used to select the final n-dimensional solution. The correlational measure of fit increased by .05 or more for seven subjects going from two to three dimensions. For three of these subjects, the increase was .15 or more. No increase of this magnitude occurred going from three to four or four to five dimensions. Further, higher dimension solutions are nearly impossible to interpret, so the higher order solutions were rejected.

Based upon statistical criteria, the choice between two and three

dimensions was not completely clear, however. A VAF increase of 6.7% was not considered a good tradeoff for one additional dimension to interpret. The three dimensional solution was selected, however, as subsequent analyses of the dimensions by the unidimensional ratings, performance data, and physiological measures revealed that the third dimension contained unique information that was not reflected in the other two. The correlations among the three chosen dimensions (0.42, 0.44, and 0.39) were not reliable ($p > .05$), suggesting that the dimensions were indeed independent. Figure 5, panels a, b, and c, present the three dimensional INDSCAL solution.

3) Interpretation of Dimensions. To explain the nature of these three dimensions used by subjects to compare tasks on similarity of task difficulty, the location of the 18 tasks in the MDS space was examined. Comparing Dimension 1 weights for these tasks with the single-hard and dual task performance decrements (Tables 6 and 7, respectively), it was apparent that Dimension 1 was associated with processing resource cost. In general, the most positive Dimension 1 weights were associated with the single-easy tasks, scaled as zero performance decrements (Te, Se, Ve), and the easily time-shared task pairs (TS, VS, TV). The one exception was the single-easy Tone Judgment task (Je) which was earlier identified as the most demanding single task because of its joint high levels of perceptual and response load. In addition, two of the single-hard tasks (Vh, Sh), whose performance decrements were moderate, also had fairly positive Dimension 1 weights. Toward the middle of Dimension 1 were those tasks that proved to be more difficult to perform because of resource demand (e.g., TJ, Th, VJ), and at the far left with the most negative weights were those tasks that were predicted to have and did have the greatest performance decrements (e.g., TT, JJ, VV).

Task location on Dimension 2 indicated that workload perceptions were influenced by the input modality required. Positive Dimension 2 weights were associated with tasks such as Vh, TT, and VV that used primarily visual input. The main exception was the single-easy Auditory Sternberg task. This task had a positive weight on Dimension 2, but its input was auditory. Slightly negative Dimension 2 weights occurred with mixed auditory and visual input tasks such as VJ and TJ. Tasks with the most negative weights along Dimension 2 were associated with primarily auditory input (e.g., SS, Jh, SJ).

Task location along Dimension 3 provided no information to aid in interpretation. None of the resource structures in the Wickens' model nor their degree of use suggested why the tasks were arranged in this particular order.

To help label or explain the characteristic that defined Dimension 3 and to substantiate the interpretation given to the other two dimensions, three data sets were correlated with task weights derived from the scaling analysis. Mean performance decrements (Tables 6 and 7), mean heart period variability measures (Table 8), and the mean ratings of the tasks on the five unidimensional scales and difficulty ranking (Table 9)

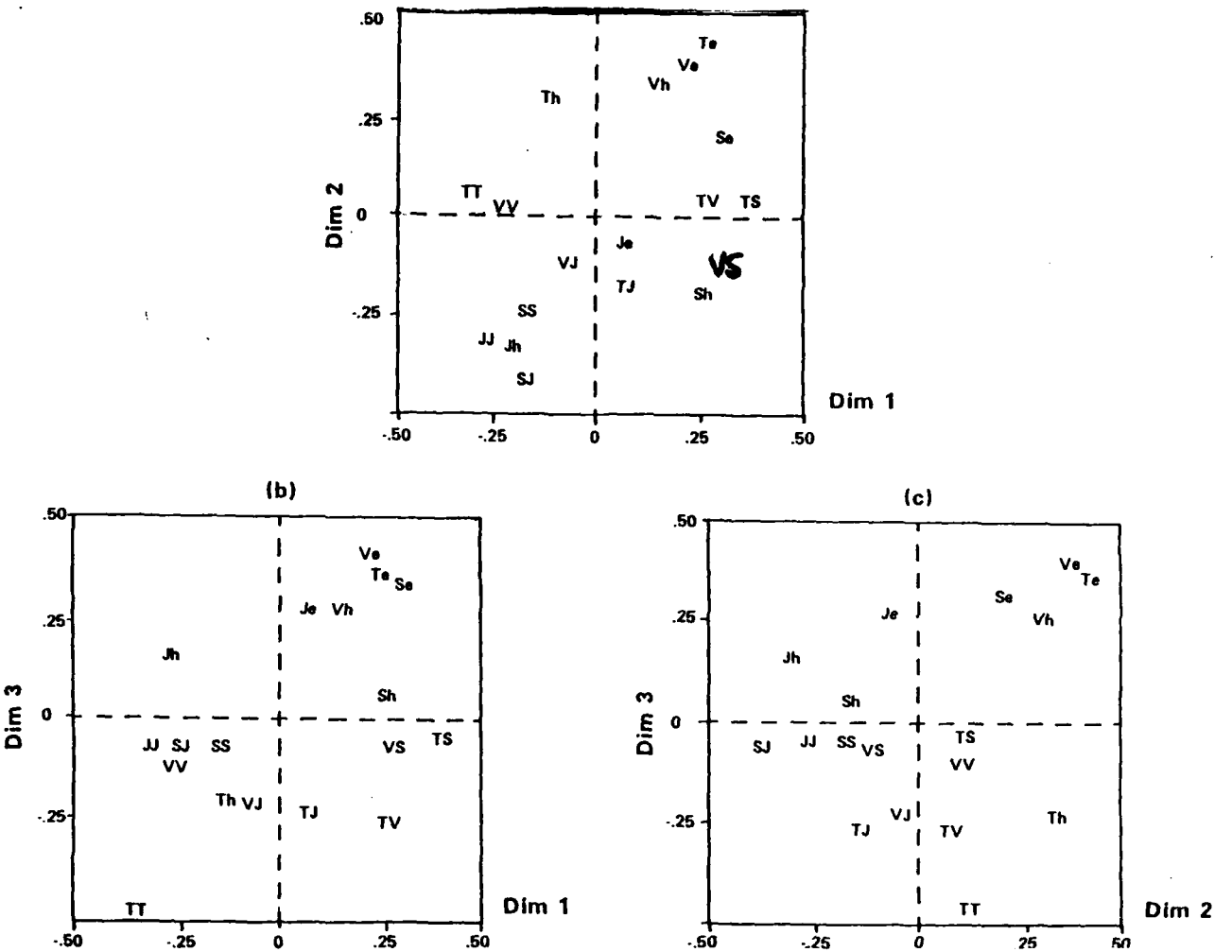


Figure 5: Three dimensional INDSCAL solution for similarity of task difficulty ratings (e = easy single task; h = hard task, two capital letters are dual task pairs).

TABLE 8

Vagal Tone ($\hat{V}$) and Heart Period Variability (HPV) Scores (Baseline - Task): Means (Standard Deviations)

Task	HPV	$\hat{V}$	Task	HPV	$\hat{V}$
Te	.409 (.29)	1.15 (.87)	Th	.558 (.37)	1.08 (1.1)
Ve	.222 (.38)	.452 (.88)	Vh	.354 (.30)	.726 (.65)
Se	.172 (.38)	.199 (1.1)	Sh	.266 (.40)	.183 (1.1)
Je	.523 (.41)	.624 (1.6)	Jh	.247 (.52)	.043 (1.3)

Task	HPV	$\hat{V}$
TV	.539 (.31)	1.04 (.68)
TS	.493 (.39)	.662 (.81)
TJ	.562 (.32)	.974 (.79)
TT	.540 (.41)	1.61 (1.1)
VS	.412 (.30)	.685 (.75)
VV	.417 (.20)	.719 (.76)
VJ	.588 (.49)	1.19 (1.1)
SJ	.511 (.31)	.782 (.85)
SS	.169 (.32)	-.229 (.92)
JJ	.504 (.32)	.663 (.86)

T = Critical Tracking, V = Visual Search, S = Auditory Sternberg, J = Tone Judgment, e = easy single task, h = hard single task, two capital letters = dual task pair.

TABLE 9

Unidimensional Ratings and Difficulty Ranking by Task: Means (Standard Deviations)

	UNIDIMENSIONAL SCALE					
Task	Input Complexity	Response Complexity	Effort Demands	Feedback Adequacy	Time Demands	Difficulty Ranking
Te	1.74 (0.81)	2.84 (1.12)	2.84 (1.92)	6.89 (0.94)	3.89 (2.11)	2.37 (1.30)
Ve	2.63 (1.26)	2.21 (0.92)	2.74 (1.24)	6.10 (1.41)	4.42 (1.87)	2.74 (1.19)
Se	1.74 (0.65)	2.00 (0.88)	2.94 (1.47)	5.05 (1.78)	4.00 (1.89)	1.95 (0.91)
Je	2.89 (1.10)	4.00 (1.67)	4.21 (1.23)	5.21 (1.65)	3.68 (1.34)	4.95 (2.34)
Th	2.68 (1.00)	5.26 (1.33)	5.26 (1.52)	6.79 (1.27)	6.21 (1.36)	9.05 (3.41)
Vh	3.32 (1.25)	2.84 (1.21)	4.11 (1.45)	6.11 (1.10)	5.32 (1.97)	4.95 (2.04)
Sh	3.72 (1.53)	3.17 (1.69)	4.78 (1.80)	4.94 (1.83)	6.06 (1.83)	6.06 (3.19)
Jh	4.17 (1.65)	5.39 (1.54)	5.44 (1.50)	5.11 (1.75)	4.50 (1.38)	11.61 (3.55)
TV	4.53 (1.50)	4.74 (1.69)	5.11 (1.56)	6.16 (1.26)	5.79 (1.93)	9.95 (2.55)
TS	3.42 (1.35)	4.11 (1.24)	4.32 (1.49)	5.89 (1.27)	5.74 (1.45)	8.32 (2.14)
TJ	4.37 (1.34)	5.05 (1.54)	5.16 (1.21)	5.37 (1.21)	5.59 (1.64)	13.00 (2.29)
TT	3.95 (1.61)	6.16 (1.57)	6.47 (1.78)	6.63 (1.38)	7.05 (1.22)	14.89 (3.83)
VS	4.32 (1.42)	3.89 (1.10)	4.68 (1.38)	5.47 (1.31)	5.63 (1.57)	11.26 (3.00)
VV	4.00 (1.80)	3.26 (1.52)	4.47 (0.96)	5.68 (1.60)	5.95 (1.81)	12.89 (2.02)
VJ	4.94 (2.01)	5.17 (1.62)	5.56 (0.98)	5.39 (1.46)	6.17 (1.50)	13.94 (1.92)
SJ	4.53 (1.87)	5.42 (1.71)	5.95 (1.51)	4.89 (1.41)	6.26 (1.41)	15.52 (2.57)
SS	4.11 (1.88)	3.50 (1.65)	5.39 (1.94)	4.83 (1.86)	6.28 (1.87)	13.39 (3.40)
JJ	4.37 (1.71)	5.63 (2.11)	5.47 (1.90)	5.00 (1.76)	5.16 (1.83)	14.26 (3.77)

T = Critical Tracking, V = Visual Search, S = Auditory Sternberg, J = Tone Judgment, e = easy single task h = hard single task, two capital letters = dual task pair.

were correlated with the INDSCAL dimension weights. Since the four single-easy tasks had no performance decrement scores, Table 10 presents the correlations of the dimension weights for all 18 tasks with the ratings and variability scores only. Table 11 presents the correlations for 14 tasks (no single-easy) between the dimension weights and all three data sets.

When comparing Table 10 to 11, note that the correlations decrease as the task set drops from 18 to 14 tasks. However, the general pattern of correlations remains the same with two exceptions. In Table 10, Response Complexity is reliably correlated with Dimension 1 and Effort Demands is reliably correlated with Dimension 2. Neither is the case in Table 11.

In addition to the rationale described earlier, the choice of a three-dimensional scaling solution was further validated by these correlations. The average difficulty ranking was reliably correlated with all three INDSCAL dimensions, suggesting that the initial paired comparison rating of similarity of task difficulty was both understood by the subjects and underlies the entire scaling solution. Further, several items were correlated with Dimension 3 and not the other two dimensions, suggesting that this dimension or attribute of the ratings set should be interpreted. Finally, the high Effort Demand correlations with all three dimensions suggested that this global concept was tied to any conceptualization of task difficulty and ultimately workload.

In addition to the Effort Demand and Difficulty Ranking correlations with Dimension 1, two other variables--Performance Decrement ($r = -0.80$) and Response Complexity ($r = -0.61$)--demonstrated reliable correlations. The fact that performance decrement scores were not correlated with either of the other two dimensions reinforced the earlier conclusion that Dimension 1 was related to the resource demand or resource cost concept inherent in the multiple resources model.

Single tasks with relatively small resource demands or dual tasks where the resource demands were spread across several structures (in both cases suggesting fairly good performance) had positive Dimension 1 weights. As task resource demand was increased by loading a smaller and smaller number of structures (suggesting poorer performance), the Dimension 1 weights became more negative. In conjunction with this change, perceptions of response complexity went from low to high. Based upon these two correlations, the actual task location along Dimension 1, and the lack of differentiation of Effort Demands and Difficulty Ranking among all dimensions, Dimension 1 was labeled "Complex, Resource-expensive versus Simple, Resource-inexpensive." Tasks that were complex and resource expensive were seen as very effort demanding and very difficult; the simple and resource inexpensive tasks were perceived as easier to perform and required less effort.

These results support the predictions of Hypothesis 2. At least for these tasks, the data suggest that people rate tasks as similar in

TABLE 10

Correlations of INDSCAL Dimension Weights with Unidimensional Scale Means and Physiological Measure Means (all 18 tasks)

	Dimension		
Variable	1	2	3
Unidimensional Scales:			
Input Complexity	-.45	-.75	-.51
Response Complexity	-.61	-.53	-.75
Effort Demands	-.69	-.66	-.83
Feedback Adequacy	.22	.82	-.07
Time Demands	-.38	-.27	-.87
Difficulty Ranking	-.74	-.72	-.81
Physiological Measures:			
Heart Period Variability (HPV)	-.20	-.07	-.61
Vagal Tone ($\hat{V}$)	-.11	.32	-.56

Note. The following critical values can be used to evaluate the correlations reported in the table: $r(16) = .590$, $p < .01$; $r(16) = .542$, $p < .05$.

TABLE 11

Correlations of INDSCAL Dimension Weights with Unidimensional Scale Means, Average Performance Decrements, and Physiological Measure Means (10 dual plus 4 single-hard tasks)

	Dimension		
Variable	1	2	3
Unidimensional Scales:			
Input Complexity	-.17	-.62	-.28
Response Complexity	-.47	-.22	-.56
Effort Demands	-.67	-.32	-.61
Feedback Adequacy	.17	.92	-.38
Time Demands	-.07	.36	-.72
Difficulty Ranking	-.70	-.54	-.55
Performance Decrement	-.80	.07	-.23
Physiological Measures:			
Heart Period Variability (HPV)	-.01	.29	-.64
Vagal Tone ($\hat{V}$)	-.07	.48	-.68

Note. The following critical values can be used to evaluate the correlations reported in the table: $r(12) = .661$, $p < .01$; $r(12) = .532$, $p < .05$.

difficulty, in part, because of resource cost. This resource cost concept is further associated with perceptions of effort, difficulty, and response complexity.

Two additional INDSCAL dimensions indicated that factors other than resource cost, however, influenced similarity of task difficulty ratings. For Dimension 2, previously discussed in terms of input modality, two new variables correlated with the dimension weights: Input Complexity ($r = -0.75$) and Feedback Adequacy ($r = 0.82$). As the tasks changed from primarily visual input (positive weights) to primarily auditory input (negative weights), ratings of Input Complexity shifted from low to high and Feedback Adequacy demonstrated the opposite effect. Thus, the auditory input tasks were perceived as having a complex stimulus input and providing little performance feedback.

Putting together the information on task location with these correlations, Dimension 2 was labeled "Visually simple, good feedback versus Auditorily complex, poor feedback." Tasks that were seen as visually simple with good feedback were also seen as easy and requiring little effort; the complex auditory tasks with poor feedback were perceived as difficult and effortful.

These results are also consistent with Hypothesis Two which stated that some of the dissociation in workload measures could be explained by the multiple resource model. Here the performance-based measure of workload, resource cost, is related to only one of the subjective dimensions of workload: a dissociation. However, the concept of input modality can help explain the nature of the second subjective dimension. It is, of course, impossible to determine from these data which of the three attributes, input modality, feedback, or complexity, is the true driving force behind this dimension of subjective load.

Task location along Dimension 3 could not provide an interpretation of the dimension, but the correlations of Tables 10 and 11 were suggestive. This was the only INDSCAL dimension where both the physiological variables and the Time Demand variable were correlated with the dimension weights. Tasks with negative Dimension 3 weights such as TT and TV were generally associated with high ratings of time demand and response complexity but little heart period variability. Conversely, tasks such as Ve and Je with positive Dimension 3 weights were associated with little time stress, simple responses, and large amounts of on-task heart period variability.

One interpretation of these data is that this third dimension or aspect of task difficulty ratings is associated with time stress and response complexity which are in turn associated with changes in heart period variability. Since Reid et al. (1981) have argued that time load is a relevant variable in subjective perceptions of workload and since it is unlikely that operators "read" the state of their heart rhythms to render workload judgments, Dimension 3 was labeled as "Simple, time-free versus Complex, time-committed." The simple-response tasks that did not

demand all the available time were seen as less difficult and less effortful; the complex and time-demanding tasks were associated with high ratings of effort and difficulty.

In summary, the MDS analysis supported Hypothesis Two. Three subjective dimensions of task difficulty were uncovered in the proximities data. Two dimensions were explained by the multiple resource model. Task location plus correlational analyses suggested the labels "Complex, resource-expensive versus Simple, resource-inexpensive" for Dimension 1 and "Visually simple, good feedback versus Auditorily complex, poor feedback" for Dimension 2. Dimension 3, not explained by the model, was given the label "Simple, time-free versus Complex, time-committed." In addition, a dissociation between measures was found. Of the three subjective dimensions, only one was related to processing resource cost while another was related to heart period variability. Finally, measures of effort and difficulty were related to all subjective dimensions and thus were not sensitive enough to discriminate between difficulty as indexed by resource cost or difficulty as indexed by heart period variability.

Additive Clustering

An additive clustering analysis (Shepard & Arabie, 1979) was run on the same proximity data to determine if a non-spatial or discrete representation of the stimulus objects (tasks) could provide more information concerning the relationship of the difficulty ratings to the other measures collected. The 19 proximity matrices of 153 paired-comparison ratings were averaged to produce one matrix. This matrix was analyzed by MAPCLUS (Arabie & Carroll, 1980), a computer program for fitting the ADCLUS model.

Multiple runs of MAPCLUS produced several proposed solutions to the proximity data. These solutions differed by the number of clusters in the solution, the weights for each cluster (reflecting the importance of the property that produced the grouping), the variance accounted for by the entire solution, and the "density" or size of the clusters in each solution. Of the several possible solutions, the one chosen accounted for 68.4% of the variance using 9 clusters and had a relatively low associated density (.24). Although the VAF fell short of the recommended standard of 80% (Shepard & Arabie, 1979), this particular combination of VAF, cluster number, and density produced the most interpretable solution. As with the MDS solutions, this criterion of interpretation must be considered along with the statistical criteria when selecting a MAPCLUS solution (Arabie & Carroll, 1980). Table 12 presents the clustering solution.

Determination of the properties that produced the task clusters was aided by embedding the clusters within an MDS representation and drawing closed curves around each cluster. Figure 6 presents the data from Table 12 in a two-dimensional KYST scaling space (Kruskal, Young, & Seery, 1973). This spatial representation readily depicts the extensive overlap

TABLE 12

MAPCLUS Solution to Judged Similarities of Task Difficulty for 18 Tasks

Rank	Cluster Number	Weight	Elements					
1	4	.256	Th	TT	TV			
2	1	.250	Te	Ve	Se	Vh		
3	7	.242	Je	Sh	TV	TS	TJ	VS
4	6	.200	TJ	VJ	SJ	SS		
5	8	.199	Te	Ve	Se	Je	Vh	Sh
6	5	.186	Jh	TT	SJ	SS	JJ	
7	9	.173	Th	TJ	VV	VJ		
8	2	.129	Sh	TS	VS	TV		
9	3	.098	JJ	Jh	SS	SJ		

Note. VAF = 68.4%, density = .24. Two capital letters are dual task pairs, e = easy single task, h = hard single task.

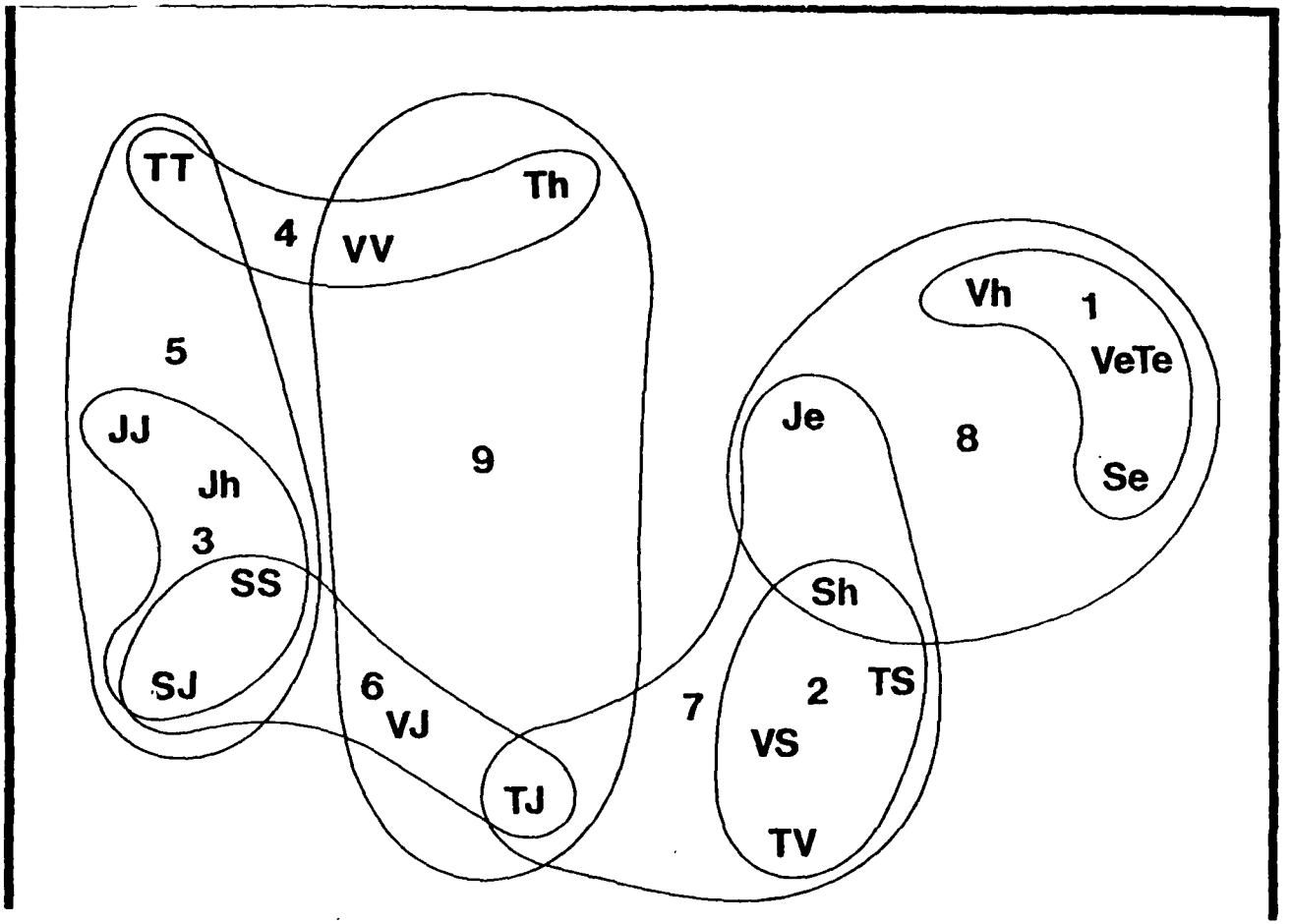


Figure 6: The nine ADCLOS subsets for the 18 tasks embedded in a two-dimensional scaling representation.

of the nine clusters. For example, the easy single Auditory Sternberg task (Se) is a member of cluster number 1 but also a member of cluster number 8 while the hard single Auditory Sternberg (Sh) belongs to Cluster 2 but also to Clusters 7 and 8. The asymmetry of these curves should be contrasted with the less complex figures of Shepard and Arabie (1979). The task stimuli used here, however, are more complex than the stimuli analyzed by these authors.

Following the procedure published by Shepard and Arabie (1979), the unidimensional ratings, performance decrements, and heart period variability scores were then examined in the context of these clusters. A graphic representation of these results is presented in Figure 7. The labeled dashed lines segment the clusters by three dimensions: Performance, Input Complexity, and Effort Demands. Of the nine pieces of information per task, these are the three that could reliably explain why tasks are grouped into some clusters and not others.

The average performance decrement for the tasks classified by Worse Performance was 2.40 while the comparable decrement for the Better Performance tasks was 0.88 (excluding the four easy single tasks where the performance decrement would be zero). Looking simply at the rank order of performance decrements, the average rank order for the Worse Performance tasks was 14.5 while it was 5.5 for the other group. For the High, Moderate, and Low Effort tasks, the average Effort Demand ratings were 5.71, 4.75, and 3.16 while the average Effort Demand rankings were 15.5, 8.5, and 2.5, respectively. High Input Complexity tasks produced mean complexity ratings of 4.41 while Low Input Complexity tasks had a mean rating of 3.01; the complexity rankings were 14.5 and 5.5, respectively.

Based upon the segmentation, the cluster properties were interpreted in the following manner. Tasks group into Cluster 3 (JJ, Jh, SS, SJ) because they were poorly performed, demanded high effort, and were seen as having high input complexity. Task TT was associated with these four because of shared effort and performance properties, but it was also a member of Cluster 4 (TT, VV, Th) with which it shared the properties of low input complexity and poor performance. Task VV and Th did not share the input complexity property with tasks in Cluster 3 so there was no overlap between them and Cluster 3.

Interpretation of all clusters proceeded in this fashion. The one failure of the analysis was found in the comparison of Cluster 2 (Sh, TS, VS, TV) and Cluster 7 (these four plus tasks Je and TJ). Tasks were grouped into Cluster 2 because of better performance and moderate effort, the same properties that supposedly created the larger Cluster 7. Thus the properties were not unique.

This additive clustering analysis lent additional support to Hypothesis Two. Again, the concept of resource cost derived from the multiple resource model and its implications for task performance played a major part in affecting perceptions of task workload. Unlike the

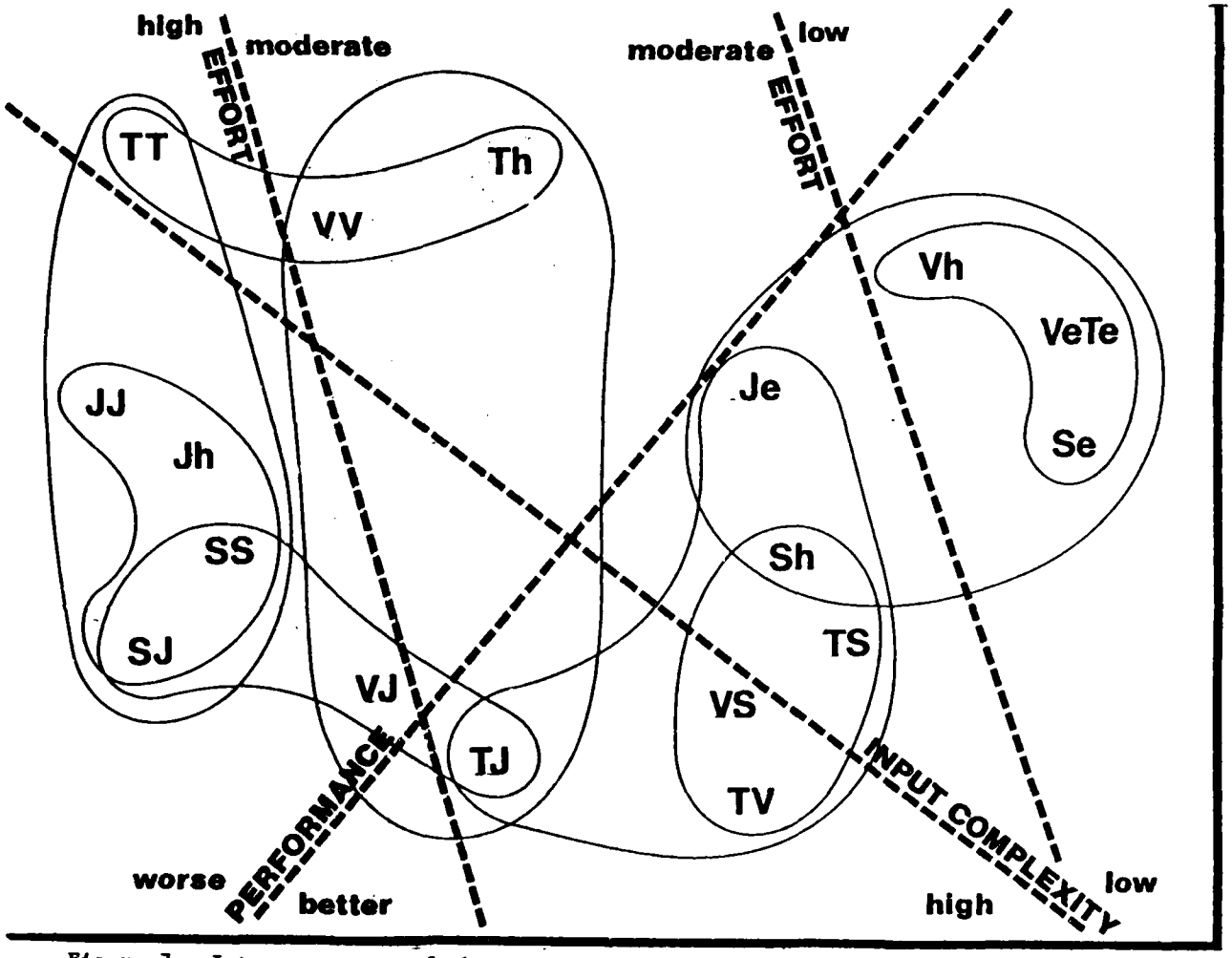


Figure 7: Interpretation of the ADCLOS subsets by the properties of task performance, effort demands, and input complexity.

INDSCAL analysis, however, no other component or prediction from the model could explain why these particular task clusters were formed. Further, many of the additional variables that were helpful in explaining the INDSCAL dimensions were of no use here. As such, the MDS spatial-distance model proved to be a better way to map predictions and parameters of the resource model onto dimensions of subjective workload.

As with the INDSCAL dimensions, however, one striking feature of the MAPCLUS solution and interpretation was the apparent dissociation of performance-based and subjective measures of workload. Segmenting clusters by performance (resource cost) clearly demonstrates why workload among five groups of tasks was perceived as similar yet different from four other groups. Segmenting by effort produced slightly different results (note tasks VV and Th which were lowest in terms of performance but only moderate in terms of effort), and segmenting by input complexity produced results nearly orthogonal to performance. Thus, the subjects in this study saw the single-hard version of Critical Tracking (Th) as similar in workload to the Critical Tracking-Tone Judgment pair (TJ), yet Th was also seen as having lower input complexity than TJ and Th also produced reliably worse performance.

The measure dissociations thus far reported in this study have all been analyzed within the framework of scaling or clustering solutions. The final part of this Results section discusses dissociation when each measure is treated as a dependent variable.

Measure Dissociation

The data collected to interpret the scaling and clustering solutions were also examined from a more traditional workload research perspective (Moray, 1979). Each of the 18 tasks had four pieces of information that are often collected in workload research: a performance decrement score (zero for single-easy versions) which, as described in an earlier section, was predicted from a resource demand/resource competition model (Wickens, 1980, 1983); two measures of heart period variability (Mulder & Mulder, 1981); and a global measure of task effort (Reid et al., 1982). Since the scaling and clustering analyses suggested that these measures do not covary to drive perceptions of workload (a dissociation), the analyses reported below focused on the interrelationships of these variables outside of the scaling and clustering framework.

Figure 8 depicts all four variables plotted on a common scale. For each variable, higher levels indicate greater workload and/or larger performance decrements. The points in the figure were averaged over all 19 subjects for all tasks that fit each of the configurations along the abscissa. The configuration Dual-Self refers to a task time-shared with itself (e.g., TT) while Dual-Others refers to a task time-shared with each other task (e.g., TV, TS, SJ).

To generate the Dual-Others data points, the three scores that were associated with "others" were combined and averaged. For example, for

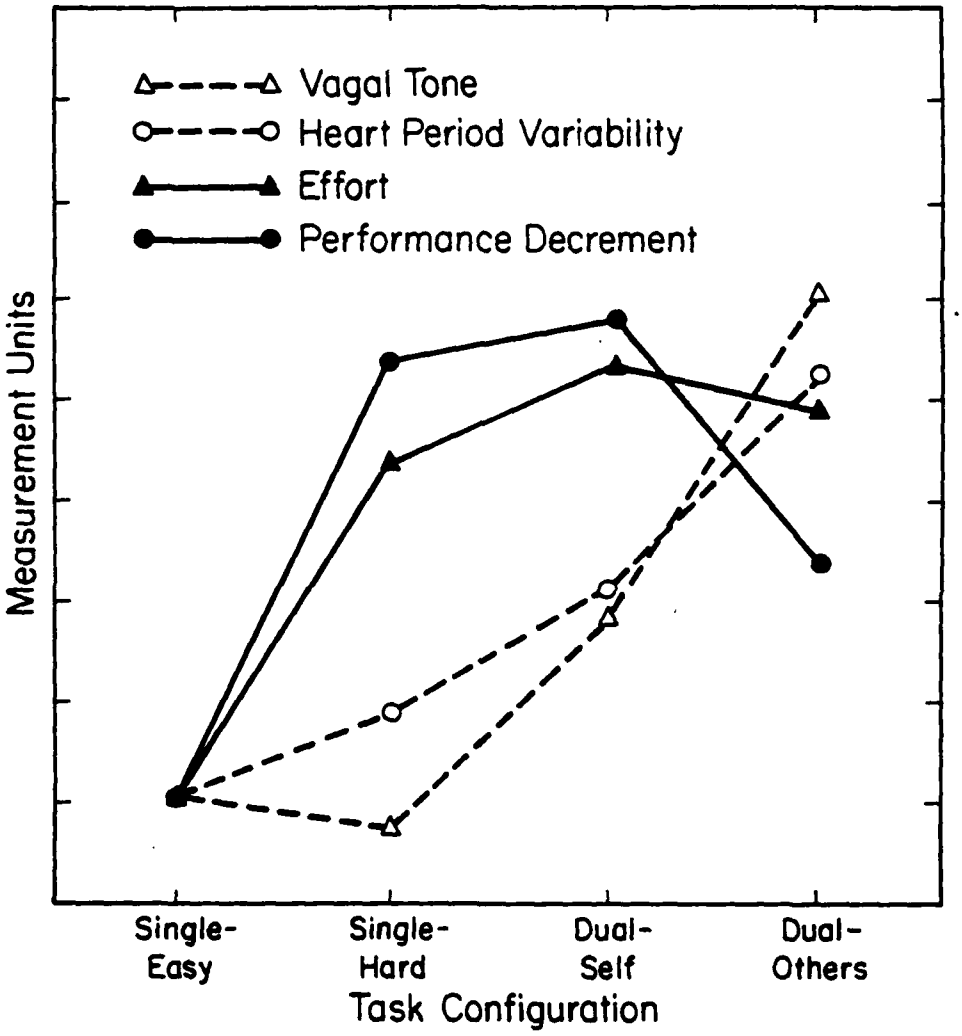


Figure 8: The four workload variables by task configuration averaged over all tasks.

Subject #1, his effort ratings for TV, TS, and TJ were 5, 5, and 4, respectively. Thus, his Critical Tracking-Others score was 4.67. When done for all 19 subjects, the mean Critical Tracking-Others score was 4.86. When done for the remaining three core tasks and combined with Critical Tracking-Others, the grand Dual-Others mean for effort was 5.13. This latter average is the point Dual-Others for the effort function in Figure 8.

The functions plotted in Figure 8 do demonstrate some association or convergence among the four workload measures. Compared to the Single-Easy task configuration, both dual task configurations produced greater workload, be it indexed by worse performance, lower heart period variability, or higher effort ratings. Further, both the effort ratings and the performance data indicated that the difficult single tasks were associated with greater workload than the easier single tasks. Thus, in some comparisons, the four diverse measures did distinguish levels of workload in harmony as would be expected in multi-measure workload study.

Measure dissociation apparently did occur, however. Starting with the baseline Single-Easy tasks, the performance decrement function rose to the Single-Hard configuration, was roughly equivalent to the Dual-Self tasks, and then fell (performance improved) for the Dual-Others configuration. Effort ratings also increased from Single-Easy to Single-Hard, indicating that the latter tasks were rated as more effort demanding. However, the effort function continued to rise to Dual-Self (performance was unchanged) but failed to drop substantially from Dual-Self to Dual-Others (performance did). This divergence suggested two dissociations between effort and performance measures. Apparently effort measures were sensitive only to the difference between single and dual task combinations whereas performance decrements were differentiated by focused (Single-Hard, Dual-Self) versus distributed (Dual-Others) resource demand.

Both measures of heart period variability, $\hat{V}$ and HPV, generally increased across the four task configurations. This increase represented greater baseline-minus-task differences in variability or decreased heart period variability while performing the task. Although no changes were evident in the performance decrements between Single-Hard and Dual-Self, the actual on-task variability measures declined suggesting greater workload. Further, while performance improved from Dual-Self to Dual-Others, HPV and $\hat{V}$ actually suggested the greatest workload (lowest variability) in the Dual-Others configuration. Again, a dissociation between measures appeared to exist.

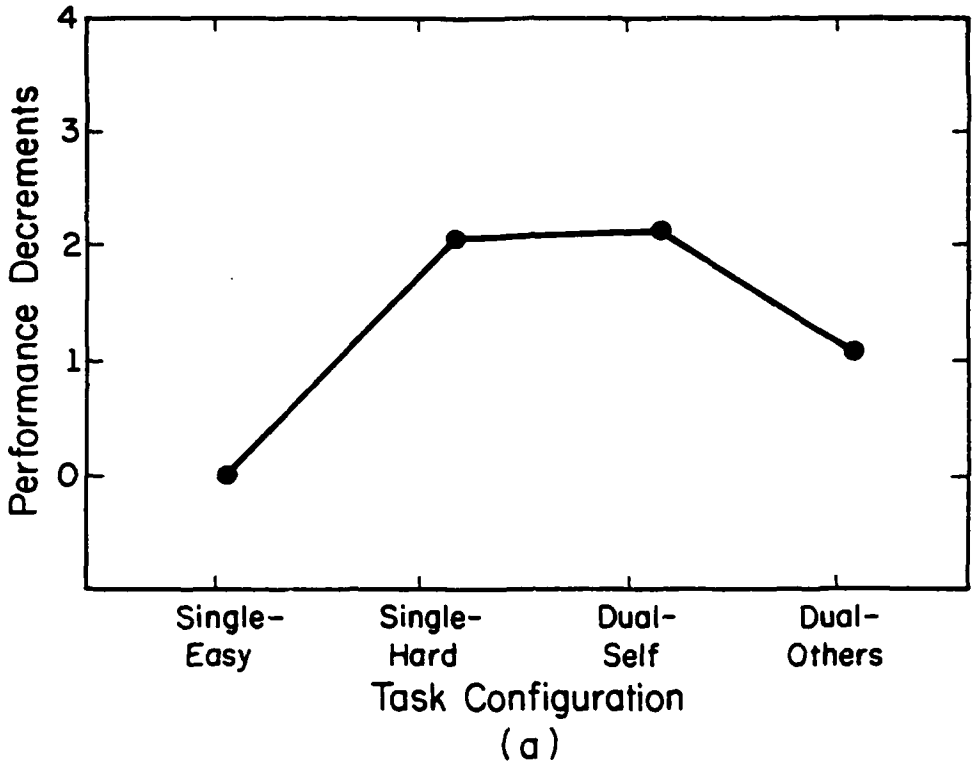
To confirm the dissociations suggested by Figure 8, 4 (Task Type: I, V, S, J) x 4 (Task Configuration: Single-Easy, Single-Hard, Dual-Self, Dual-Others) ANOVAs with repeated measures on both factors (Keppel, 1973) were run on the effort, $\hat{V}$, and HPV data. A 4 x 3 (Single-Easy dropped) ANOVA with repeated measures on both factors was run on the performance data. The primary purpose of these analyses was to test for a main

effect for Task Configuration. Planned comparisons were then evaluated on the points mentioned above. The effect of Task Type was also evaluated to determine if the functions plotted in Figure 8 accurately described the data for each of the four task types and so could be considered "general," or represented the combination of divergent trends.

The performance data that were analyzed are plotted in Figure 9. Panel a is the overall performance function previously plotted in Figure 8 while panel b presents the separate task means. Note that the task means are very similar for the Dual-Others configuration. Subsequent figures using the remaining variables demonstrated the same effect. This similarity occurred because the same dual task performance score was used more than once. For example, to calculate the Tone Judgment mean, the JT, JS, and JV scores were used, and to calculate the Critical Tracking mean, the TJ, TS, and TV scores were used, but JT and TJ are the same task. Although this procedure would invalidate any comparison among tasks for Dual-Other performance, it does not invalidate the comparison of means for Dual-Self versus Dual-Other.

For these performance data, the effects of Task Configuration, $F(2,36) = 65.83$, Task Type, $F(3,54) = 15.69$, and the interaction, $F(6,108) = 12.31$, were all reliable (all $p < .01$). To examine the first set of dissociations, planned comparisons revealed that performance decrements did not differ between Single-Hard and Dual-Self, $F(1,18) = 1.32$, $p > .10$, but did differ reliably between Dual-Self and Dual-Others, $F(1,18) = 87.74$, $p < .01$. Conversely, the effort ratings data (Figure 10) which also had a reliable main effect for Task Configuration, $F(3,54) = 60.66$, $p < .01$, demonstrated a reliable increase in means from Single-Hard to Dual-Self, $F(1,18) = 9.14$, $p < .01$), but no reliable difference between Dual-Self and Dual-Others, $F(1,18) = 2.70$, $p > .10$. With respect to the overall Task Configuration means, this analysis supported the dissociations proposed earlier: Effort was perceived to be greater in dual as opposed to single task conditions, but it did not differentiate between dual task conditions. Performance, on the other hand, was clearly better when separate tasks were time-shared in a dual task condition.

Examination of panel b in Figure 9, however, reveals that the essential equivalence of performance means between Single-Hard and Dual-Self was a function of two divergent trends. Performance of the two manual analog tasks, Tone Judgment and Critical Tracking, actually seemed to improve from Single-Hard to Dual-Self while performance of the two cognitive decision-making tasks, Auditory Sternberg and Visual Search, seemed to deteriorate between the same points. Conversely, the effort ratings data from panel b, Figure 10, all seemed to increase between Single-Hard and Dual-Self. Thus, between these two points and between these two groups of tasks, another measure dissociation was suggested. For Critical Tracking and Tone Judgment, performance and effort dissociated but for Auditory Sternberg and Visual Search, performance and effort associated.



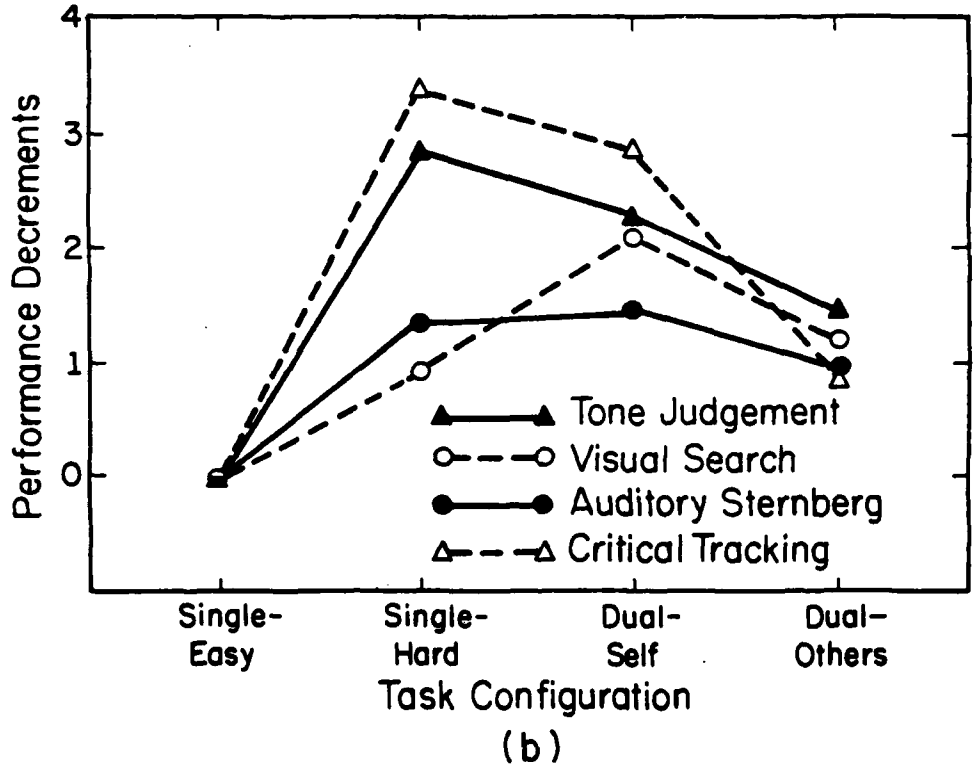
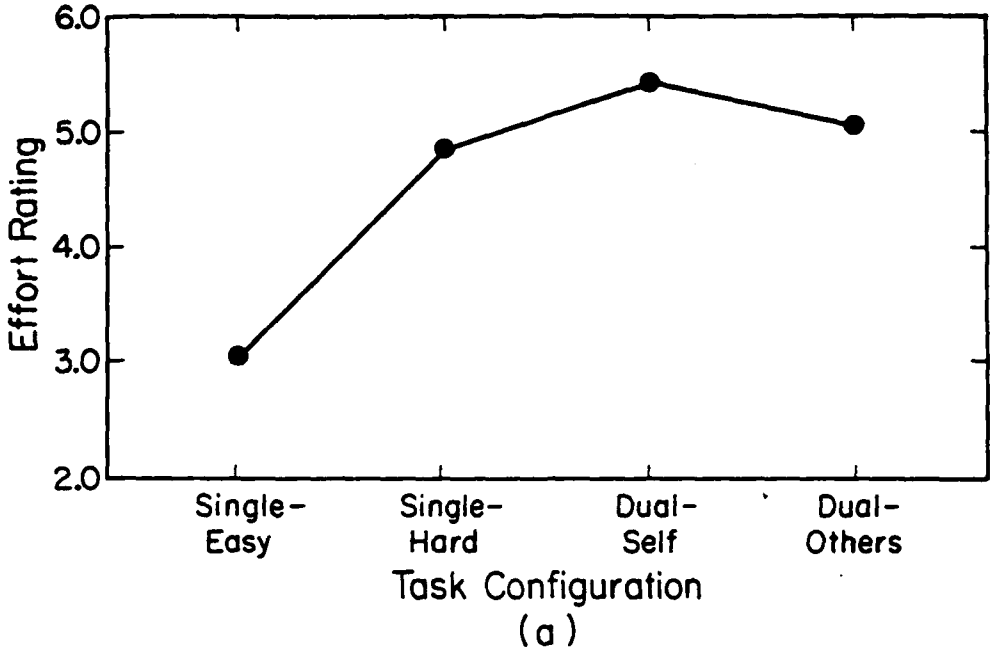


Figure 9: Performance decrement scores by task configuration averaged over tasks (a) and by task type (b).



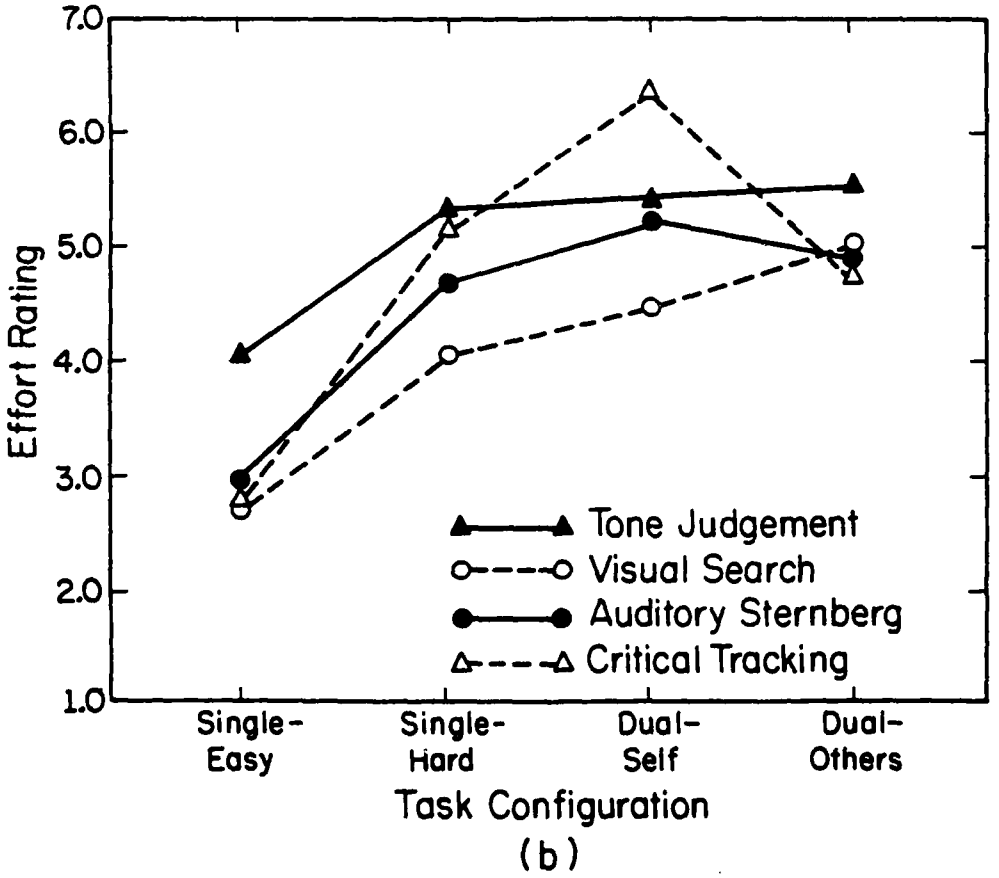


Figure 10: Effort ratings by task configuration averaged over tasks (a) and by task type (b).

Wickens and Yeh (1983) have presented a theory of workload measure dissociation that predicts the latter finding. Within the framework of the multiple resource model, they argued that tasks which are increased in difficulty by adding to the perceptual/central processing resource load will be readily perceived as more difficult or more effortful. However, tasks made more difficult by increasing the resource demands of responding will not be as accurately recognized or correctly interpreted. Using tasks and methods different from this study, Wickens and Yeh found evidence for this dissociation.

To determine if the increase in difficulty (performance decrements) from Single-Hard to Dual-Self for the S and V tasks were accompanied by increased effort ratings but the decreased difficulty of the T and J tasks were misperceived by subjects, the performance and effort data were examined only for the Single-Hard and Dual-Self configurations. A 2 (Configuration) x 4 (Task Type) ANOVA with repeated measures on both factors was run on each variable. As expected, no effect of Configuration was found for the performance data, $F(1,18) = 1.41$, $p > .10$, but the interaction between Configuration and Task Type was reliable, $F(3,54) = 6.48$, $p < .01$. Thus, the performance difference between Sh and Vh versus Th and Jh was considerably different than the difference between SS and VV versus TT and JJ. Conversely, a main effect for Configuration was obtained for the effort data, $F(1,18) = 9.15$, $p < .01$, indicating that a reliable increase in effort ratings did occur for all tasks. Further, the Task Type x Configuration interaction was not relible, $F(3,54) = 2.76$, $p > .05$, and, as evident in Figure 10b, all tasks showed a difference in the same direction. Thus, changes in task difficulty for the primarily cognitive tasks were accurately "read" by the subjects, but for the analog manual tasks, these changes were read inaccurately.

When examining the task-specific data that comprise the second dissciation discussed above (from Dual-Self to Dual-Others, performance improves, effort is unchanged), it is clear from panel b of Figure 9 that all tasks improved in performance. Since some of the tasks that comprise the "others" category were nearly as resource demanding as the Dual-Self tasks, this improvement would have been much greater had all the Other tasks truly had distributed resource demands. Nonetheless, the difference in performance was highly reliable ($p < .01$). Thus, the trend of the means in Figure 9 clearly reflects the trend of the components. On the other hand, the same cannot be said for the effort data. Not all tasks demonstrated zero slopes between Dual-Self and Dual-Others. Ratings dropped considerably for the Critical Tracking task, but not enough to change the overall null effect.

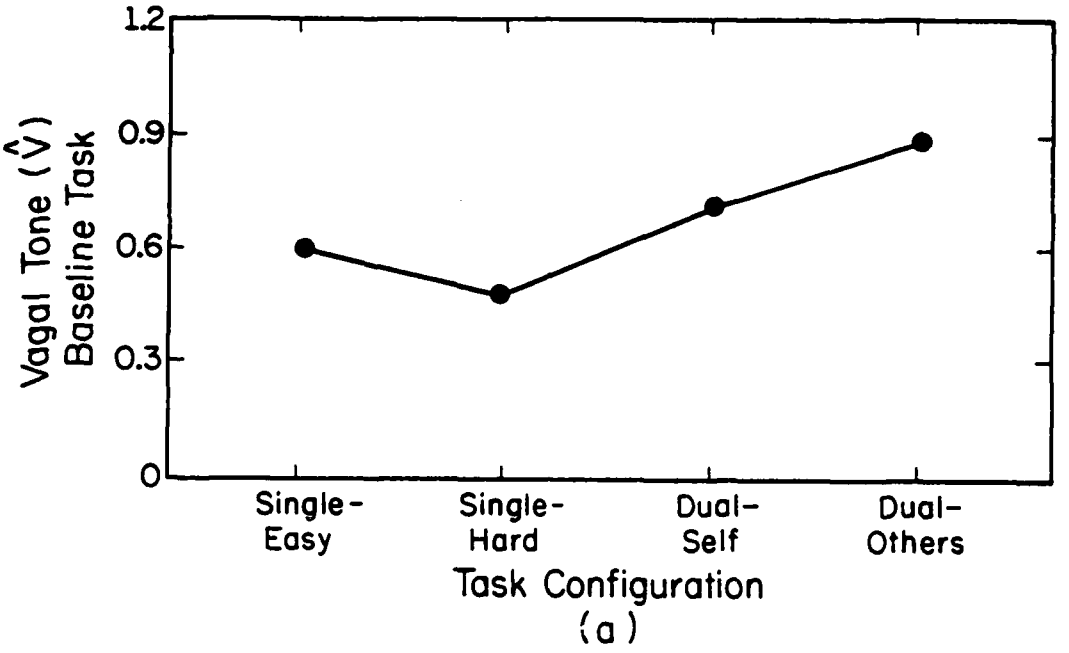
Wickens and Yeh (1983) have also explored this type of dissociation. They proposed that performance will be influenced by whether time-shared tasks compete for common or separate resources, but the subjective experience of workload will not be affected by this manipulation. Using three tasks that were representative of piloting activities, they

discovered this performance-ratings divergence in several different experimental comparisons. When time-shared task performance improved, effort ratings were insensitive to the change. The data from this experiment replicates that finding.

Panel a of Figures 11 and 12 present the heart period variability data. Both $\hat{V}$ and HPV functions look quite similar; in fact, they were highly correlated across all tasks ($r = 0.83$, $p < .01$). Using the same 4 x 4 repeated measures ANOVAs described above, the effect of Task Configuration was reliable for HPV, $F(3,54) = 7.80$, $p < .01$, and for $\hat{V}$, $F(3,54) = 7.99$, $p < .01$. Since these two measures should index workload in a fashion that parallels performance changes, the planned comparisons focused on the difference between Dual-Self and Dual-Others where performance improved but the variability measures suggested increased workload. For HPV, this difference was reliable, $F(1,18) = 5.13$, $p < .05$, but for $\hat{V}$ the difference was not, $F(1,18) = 3.93$, $.10 > p > .05$, although it approached the critical value because of its correlation with HPV. Based upon overall means, these results suggested a dissociation between performance and the simple measure of heart period variability: The former improved when different rather than identical tasks were time-shared, while the latter indicated greater workload.

Panel b of Figure 12, however, suggests that the overall effect was a function of disparate trends. The main effect for Task Type was reliable, $F(3,54) = 8.59$, $p < .01$, as was the Configuration x Task Type interaction, $F(9,162) = 2.02$, $P < .05$. The primary cause of the interaction was the erratic behavior of the Tone Judgment function between Single-Easy and Single-Hard. Although not pertinent to the dissociation under discussion, this large and unexplainable change reduced confidence in the HPV measure. Between the Dual-Self and Dual-Other points, HPV scores for S, J, and V did increase, while they decreased slightly for T. This decrease, however, was not enough to prevent the overall planned comparison from demonstrating a reliable increase. Contrasted with the more pronounced decrease of the same function in Figure 11, it is evident why the $\hat{V}$ score failed to demonstrate an overall change from Dual-Self to Dual-Others.

Within the framework of the multiple resource model, one explanation for this dissociaton may be related to distribution of resource demand. Table 3 indicated that the Dual-Self tasks had their resource demands focused within five structures or cells. The Dual-Other tasks had their demands spread over six, seven, or eight cells. Possibly the HPV measure was sensitive to the total resource demand of the processing system, irrespective of the fact that when a larger number of cells were employed, dual task performance actually improved. As noted earlier, Porges (1981) speculated that Vagal Tone, or the HPV estimate of Vagal Tone, should index increases in "sustained attention." Attention, however, was modeled on the Kahneman (1973) concept of undifferentiated capacity. Possibly more "attention" and more resources impact the physiological measures in the same fashion.



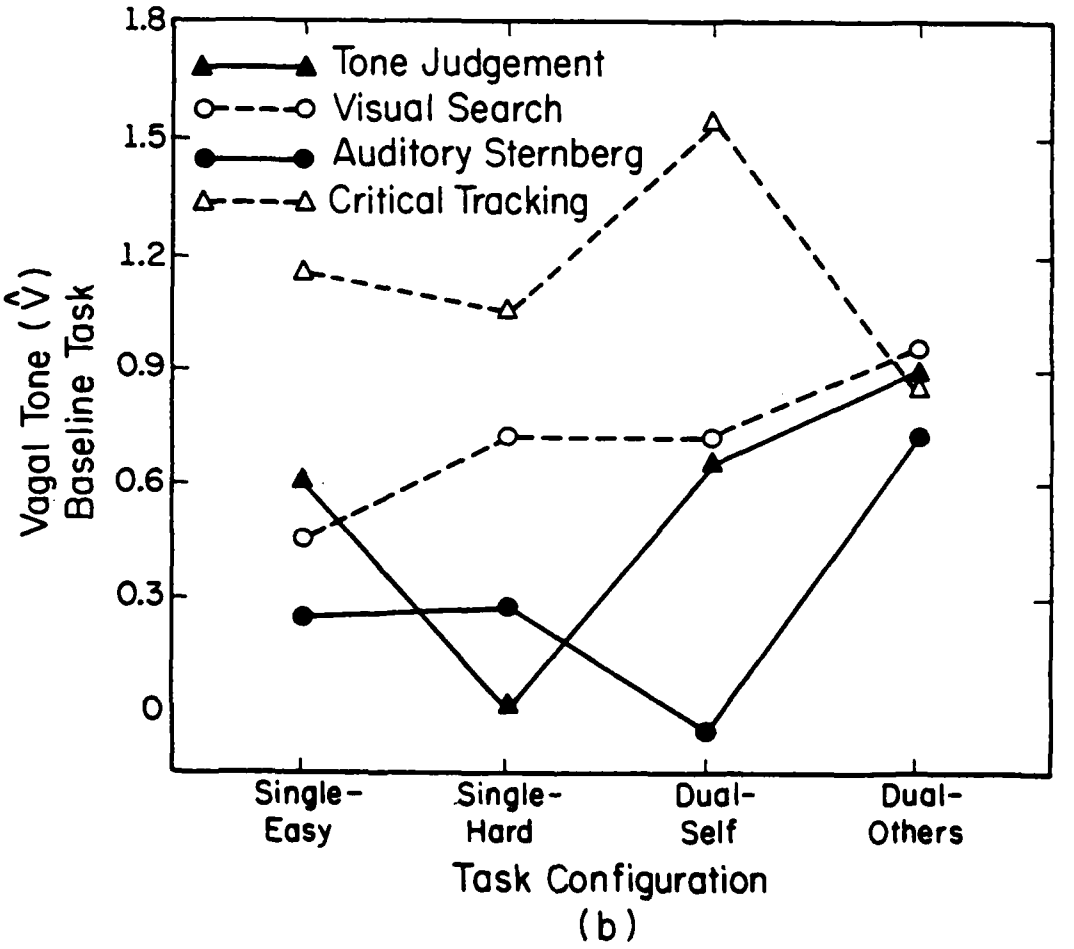
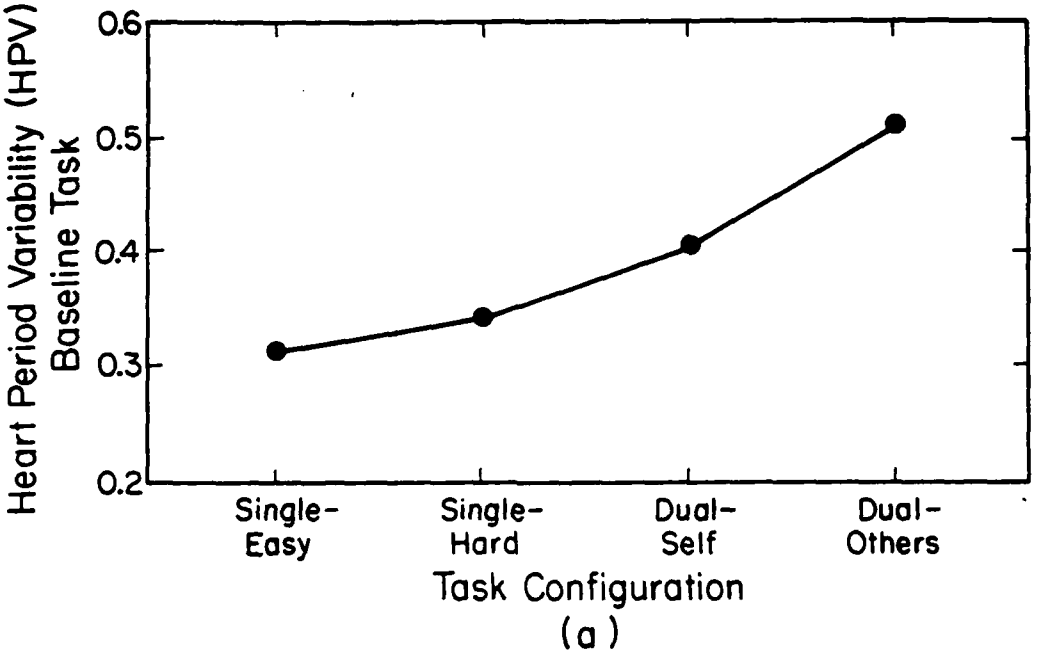


Figure 11: Vagal tone scores by task configuration averaged over tasks (a) and by task type (b).



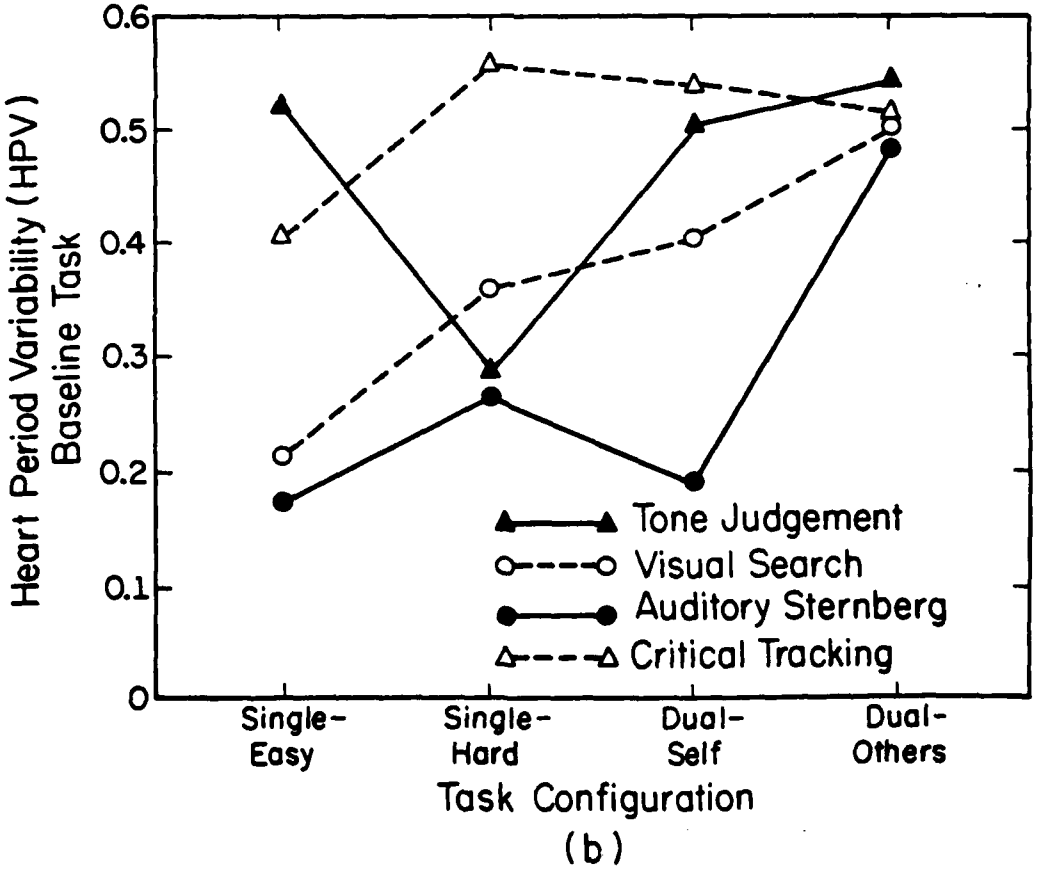


Figure 12: Heart period variability scores by task configuration averaged over tasks (a) and by task type (b).

In summary, analyses conducted on the performance, effort ratings, and heart period variability data revealed three dissociations. First, when task configuration changed from Single-Hard to Dual-Self, performance remained unchanged but perceived effort increased. This dissociation was manifested primarily in the manual analog tasks. Second, when comparing Dual-Self to Dual-Others, performance improved but effort ratings failed to covary. Third, for these same configuration points, workload as indexed by heart period variability increased while performance improved. The first two dissociations were found to support existing research while a tentative explanation was offered for the third.

DISCUSSION

This study set out to explore the multidimensional nature of the operator workload construct from the perspective of Wickens' multiple resource model (1980, 1983). For many, what makes this construct multidimensional is that no single number or no result from one workload measure can completely specify the workload of a task (Moray et al., 1979). The genesis of this view was the plethora of workload measures (primary and secondary tasks, time-line analysis, opinion data, physiological indices, and others) that have all been used successfully at one time or another to index workload but that have all failed to covary on occasion when used together. To help explain these successes and failures, and possibly even predict them, the framework of the multiple resource model was employed.

The primary workload measure investigated in this study was subjective ratings. A highly popular measure (Wierwille & Williges, 1978), its ease of collection has outweighed a host of problems, such as unreliability and rater bias, that should have reduced its usefulness (Gartner & Murphy, 1976). Some investigators (e.g., Hart et al., 1982; Sheridan, 1980) assert that subjective ratings capture the true essence of workload. Further, a great deal of research has been conducted on what aspects of workload should be rated and how these aspects should be combined (Reid et al., 1982) and what task characteristics consistently produce elevated ratings of workload (Borg, 1978; Moray, 1982).

What has not been done with subjective workload measures, however, is the formulation of any theoretically-based explanation of what underlies an operator's judgment of why one system configuration produces more workload than another configuration. The existing work on selection and refinement of rating scales has been atheoretical: No substantial rationale has been produced describing why the selected two, three, or more scales are the necessary and sufficient scales to collect operator opinions. The task characteristic work has also been atheoretical. Although some characteristics that increase load have been identified, these are all primarily related to single-task characteristics (i.e., generating lead). As such, there are no models to explain the combined effects of these characteristics in complex, multi-task environments. These, of course, are the very environments where workload

prediction/measurement is the most important.

With the use of the multiple resource model, this study attempted to go beyond the existing data to discover the complexity or dimensionality of workload ratings and explain how this dimensionality is tied to the resource model. Use of the model permitted a manipulation of task workload not by task characteristics but by a more abstract and powerful concept--resource cost. This cost resulted either from the excessive demand of a single task on a set of resources or by the competition by two tasks for common resources. Resource cost had its empirical verification in the differing levels of performance of the 18 tasks or activities that the subjects performed. In other words, this study attempted to explain the dimensionality of ratings in terms of task performance. Given that a designer's ultimate concern must be with system performance (Wickens & Yeh, 1983), the complex relationship between these two variables--sometimes converging, sometimes dissociating--must be understood.

Turning to the actual data, the multidimensional scaling and clustering solutions demonstrated that several factors drove people's perceptions of task difficulty and workload. These data suggested that people react to performance changes that can be explained by the multiple resource model. As the resource demand scheme in Table 3 indicates, task pairs such as TS and VS that have their resource demands widely distributed are not difficult, are time-shared fairly well, and load at one end of a scaling dimension. Conversely, task pairs such as TT and VV that compete for common resources are difficult, are time-shared poorly, and load at the other end of the dimension. In this instance, resource and opinion measures of workload associated.

In addition, people also reacted to the input modality of a task when they rendered workload judgments. Thus a structural characteristic of the multiple resource model influenced perceptions. People reacted to other aspects of the task stimuli, however, that were not related to the model, so that concepts such as time demand, complexity, and feedback apparently influenced ratings.

The single most important dimension to result from the multidimensional scaling analysis, based upon variance accounted for, was Dimension 1, the resource cost or competition dimension. Tasks with similar weights on that dimension, and thus seen as similar in workload, differed substantially in many characteristics. For example, consider the task pair VV compared to the single-hard Jh task. The first can be characterized by multiple tasking, intensive perceptual search, forced-pacing, no accuracy feedback, and simple discrete responses; the second by single tasking, auditory absolute judgment, forced-pacing, accurate feedback, and complex motor responses. The multiple resource model equated these diverse tasks on resource cost, permitting predictions of equivalent performance and an equivalent impact on subjective perceptions of workload. These predictions were supported. No task characteristic model exists that can put these very different

characteristics together and predict equivalent workload ratings.

Based upon task location in the MDS space, Dimension 2 could be considered a task characteristic dimension. Modality of input did differentiate perceptions of task difficulty. Here again, however, since the task characteristic literature has identified only the obvious "load inducing" characteristics, no previous work has suggested that this charcteristic could influence opinions of workload. Further, the other task characteristics that did correlate with and help identify the three subjective dimensions of workload--complexity, feedback, time demands--did so in an unpredictable fashion. The task characteristic literature suggests that more complexity, less feedback, and greater time pressure should combine to increase percpetions of workload. The patterns of correlations from Tables 10 and 11, however, suggest that these measures do not combine in such a simple fashion. They demonstrate differential effects on the overall perceptions of workload.

The scaling and clustering techniques presented in this stuy also offer an alternative approach to the multiple subjective scale studies mentioned above and reviewed earlier. Here the scales or dimensions were not given to subjects for workload ratings; rather, they were uncovered with analytical techniques. A theory guided construction of task stimuli and two methods helped uncover the dimensions or properties that produced the workload ratings. Since only 19 undergraduate subjects performed only 18 laboratory tasks, no claim is made that the three dimension labels named earlier are the labels for subjective dimensions of workload. However, the methods used produced interpretable results. Further, it should be noted that the most popular multiscale procedure (Reid et al., 1982) requires ratings on time, stress, and effort load. The results obtained in this study suggest that perceived effort, since it correlated with all three INDSCAL dimensions, can serve as a global measure of subjective workload but adds nothing unique to the understanding of workload.

Outside of the scaling and clustering solutions, workload measure dissocation was also found. This is not an unusual occurrence when multiple measures are collected, but investigators in this area rarely make an attempt to explain why the dissociation may have occurred. Typically, well-studied laboratory tasks whose parameters have been thoroughly investigated are pitted against a battery of potential workload measures. Those potential measures that fail to covary with task difficulty are simply dismissed as insensitive (Hicks & Wierwille, 1979; Wierwille & Connor, 1983).

Sheridan and Stassen (1979) discuss sensitivity as one criterion for an effective workload measure. For them, a measure is sensitive if the workload of a task is truly increased and the measure changes accordingly. Conversely, a measure is said to be selective if it does not change when some aspect of the task, other than difficulty, changes. Within the framework of the multiple resource model, a sensitive measure is one that changes when more resources are applied to the task (Wickens

& Derrick, 1981). More resources may not mean poorer performance, however. If the resources come from several cells in the model, performance may be unaffected. On the other hand, if resources come from just a few structures, performance will be degraded. The sensitivity of a measure is thus decoupled from performance.

Another criterion proposed by Sheridan and Stassen (1979) is diagnosticity. A measure is said to be diagnostic if it can specify the exact cause of the increased workload. A diagnostic resource measure of workload then is a battery of secondary tasks that indicate which resource structure(s) is more heavily utilized when the task has become more difficult (Wickens & Derrick, 1981). Diagnostic measures do reflect changes in performance.

The evidence from this study suggests that different workload measures may be sensitive but not diagnostic or vice versa. Comparing the Single-Hard to Dual-Self task configurations, the ratings of effort increased (see Figure 8). The resource demand scheme of Table 3 suggested that more resources were utilized in the Dual-Self configuration, so effort apparently was sensitive to this change but not diagnostic of its nature. These ratings gave no clue as to what resource changes were producing changes in perceived effort. Overall, however, performance remained unchanged between these points. There was no reason to believe that a large resource demand in one or two structures (Single-Hard) would produce any worse performance than Dual-Self. Thus, the resource model was insensitive to this change in effort.

When comparing Dual-Self to Dual-Others, effort ratings did not change, thus remaining sensitive to large resource demands. Performance did improve, however, and the cause of this improvement was the distribution of resource demand from few to many cells. This assertion was supported by the dual task performance scores which were diagnostic by pointing to the cause of the change.

Using these definitions of sensitivity and diagnosticity, the heart period variability measures faired poorly on both criteria. Variability continued to decrease from Single-Hard to Dual-Self to Dual-Others. As such, the measures simply appeared to reflect the total number of resource structures involved, irrespective of the amount of resources utilized or the consequences for performance.

Measures that are both sensitive and diagnostic throughout the range of task difficulty manipulations are probably rare. A secondary task that does not share a resource demand with a primary task whose workload is being measured will be labeled insensitive and dismissed. However, it is still diagnostic, at least in part, because its performance tells the investigator what resources are not important for primary task performance. Thus the data from this study suggest that one way to conceptualize and explain dissociation among workload measures is to focus on resource demand and determine if the measure is sensitive, diagnostic, or neither.

One final point should be made about how these workload measures dissociated. The effort-performance dissociation presents a real problem for designers who might collect these measures and get contradictory results. If a designer selects an equipment configuration to minimize subjective perceptions of workload, he will be biased toward single activities and away from multiple concurrent tasks (McCloy, Derrick, & Wickens, 1983; Wickens & Yeh, 1983). This would be a valid decision, however, only when the single activities were very easy or the concurrent tasks competed for common resources. Time-sharing tasks that require few common resources should produce very acceptable performance but unfortunately also produce high ratings of workload. Since system performance should be the ultimate criterion against which different systems are judged (Wickens & Yeh, 1983), people who employ these biases should be made aware of them.

In conclusion, the results from this study suggest the following generalizations. First, if a designer chooses to predict or assess operator workload with a performance-based measure, resource demand specified by the multiple resource model (Wickens, 1983) offers a very powerful, diagnostic alternative. In prediction, a task analysis based upon the more general concept of resource cost, not the specific concept of task characteristics, will specify the workload of the complex task being proposed. In an actual system, complex task workload should be measured by a battery of secondary tasks that tap different resource structures. Here, task difficulty would be specified empirically by resource cost, not from an incomplete and superficial task taxonomy.

Second, if operator opinion data are to be predicted, the designer should realize that subjective reactions will be related to workload as measured by resource cost, but the two will not covary completely. Subjective opinions will be driven by both the resource demand of the task and the number of simultaneous inputs and responses that must be dealt with (time-sharing). The resource cost approach, on the other hand, will discriminate between resource competition versus resource distribution during time-sharing.

Third, if operator opinion data are to be collected for existing systems, the designer should realize that operator responses will be driven by independent aspects of the task environment and questions should be constructed to get at those aspects. Characteristics such as task complexity (be it associated with stimulus input or required response), time demands, and feedback each contribute separately to the total percept of task difficulty. Perceived effort, on the other hand, can serve as a global measure of subjective difficulty, but its use masks important subtleties in the ratings data.

Fourth, and finally, a designer who chooses to measure workload with heart period variability scores should do so with caution. The measures can be sensitive to resource demand and thus performance differences for tasks which differ greatly in difficulty, but greater degrees of

sensitivity are unlikely. Further, these measures may index the degree of resource structure mobilization, but this index is unrelated to the resource cost, performance, or subjective opinion approaches to workload. As such, heart period variability scores appear to have a rather limited utility as workload measures at this time.

Over the last several years, workload researchers have provided the design community with a great deal of data, some guiding principles, and a lot of contradictory findings. What has been missing from all of this output is any appeal to theoretical models that can prescribe how workload should be defined, how it should be measured, and how the measures should relate both to system performance and to each other. This study has been an attempt to move workload research in that direction.

References

Arabie, P., & Carroll, J. D. (1980). How to use MAPCLUS, a computer program for fitting the ADCLUS model. Bell Laboratories.

Baddeley, A. D., & Lieberman, K. (1980). Spatial working memory and imagery mnemonics. In R. Nickerson & R. Pew (Eds.), Attention and performance VIII. Englewood Cliffs, NJ: Erlbaum.

Borg, G. (1978). Subjective aspects of physical and mental load. Ergonomics, 21, 215-220.

Boyd, S. P., (1982). The use of conjoint analysis in the interval subjective scaling of mental workload. Proceedings of the 8th Psychology in the Department of Defense Symposium, 334-339.

Brooks, L. R. (1968). Spatial and verbal components of the act of recall. Canadian Journal of Psychology/Review of Canadian Psychology, 22, 349-368.

Brown, E. L., Stone, G., & Pearce, W. E. (1975). Improving cockpits through flight crew workload measurement. Paper presented at the Second Advanced Aircrew Display Symposium, U.S. Naval Air Test Center, Patuxent River, MD.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. Shepard, A. Romney, & S. Nerlove (Eds.), Multidimensional scaling: Theory and applications in the behavioral sciences, Vol I. New York: Seminary Press.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. Annual Review of Psychology, 31, 607-649.

Cooper, G. E., & Harper, R. P. (1969). The use of pilot rating in the evaluation of aircraft handling qualities. Moffett Field, CA: NASA Ames Research Center (NASA TN-D-5153).

Craig, A. (1979). Nonparametric measures of sensory efficiency for sustained monitoring tasks. Human Factors, 21, 69-78.

Deatherage, B. H. (1972). Auditory and other sensory forms of information presentation. In H. Van Cott & R. Kinkade (Eds.), Human engineering guide to equipment design. Washington, D. C.: U.S. Government Printing Office.

Ellis, G. A. (1978). Subjective assessment. In A. Roscoe (Ed.), Assessing pilot workload (AGARD-AG-233).

Friedman, A., & Polson, M. C. (1981). Hemispheres as independent resource systems: Limited capacity processing and cerebral specialization. Journal of Experimental Psychology: Human Perception

and Performance, 5, 1031-1058.

Gartner, W. B., Ereneta, W., & Donohue, V. (1967). A full mission simulation scenario in support of SST crew factors research. Moffett Field, CA: NASA Ames Research Center (NASA CR-73096).

Gartner, W. B., & Murphy, M. R. (1976). Pilot workload and fatigue: A critical survey of concepts and assessment techniques. Moffett Field, CA: NASA Ames Research Center (NASA TN D-8356).

Gaume, J. G., & White, R. T. (1975). Mental workload assessment II. Physiological correlates of mental workload: Report of three preliminary laboratory tests. St. Louis, MO: McDonnell Douglass Corporation (MDC J7023/ 01).

Gerathewohl, S. J. (1976). Definition and measurement of perceptual and mental workload in aircrews and operators of Air Force weapon systems: A status report. In B. Hartman (Ed.), Higher mental functioning in operator environments. AGARD Conference Proceedings No. 181.

Goerres, H. P. (1977). Subjective stress assessment as a criterion for measuring the psychophysical workload on pilots. Proceedings of the AGARD Conference on Studies on Pilot Workload (AGARD-CP-217).

Gunning, D. (1978). Time estimation as a technique to measure workload. Proceedings of the 22nd Annual Meeting of the Human Factors Society, 41-45.

Harris, R. J. (1975). A primer of multivariate statistics. New York: Academic Press.

Hart, S. G., Childress, M. E., & Hauser, J. R. (1982). Individual definitions of the term "workload." Proceedings of the 8th Psychology in the Department of Defense Symposium, 334-339.

Hartman, B. O., & McKenzie, R. A. (Eds.). (1979). Survey of methods to assess workload (AGARD-AG-246).

Herron, S. (1980). A case for early objective evaluation of candidate display formats. Proceedings of the 24th Annual Meeting of the Human Factors Society, 13-16.

Hess, R. A. (1977). Prediction of pilot opinion ratings using an optimal pilot model. Human Factors, 19, 459-475.

Hicks, T. G., & Wierwille, W. W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. Human Factors, 21, 129-143.

Higgins, T. H. (1979). A systems engineering evaluation method for piloted aircraft and other man-operated vehicles and machines. U.S.

Department of Transportation, Federal Aviation Administration Report No. FAA-RD-78-78.

Israel, J. B., Wickens, C. D., & Donchin, E. (1979). The event-related brain potential as a selective index of display load. Proceedings of the 23rd Annual Meeting of the Human Factors Society, 558-562.

Jahns, D. W. (1973). Operator workload: What is it and how should it be measures? In K. Cross & J. McGrath (Eds.), Crew system design. Santa Barbara, CA: Anacapa Sciences.

Jenney, L. L., Older, H. J., & Cameron, B. J. (1972). Measurement of operator workload in an information processing task. Washington, D.C.: NASA Contractor's Report (CR-2150).

Jex, H. R., & Clement, W. F. (1979). Defining and measuring perceptual-motor workload in manual control tasks. In N. Moray (Ed.), Mental workload: Its theory and measurement. New York: Plenum Press.

Jex, H. R., McDonnell, J. D., & Phatak, A. V. (1966). A "critical" tracking task for man-machine research related to operator's effective time delay. Proceedings of the 2nd Annual Conference on Manual Control, Massachusetts Institute of Technology, 361-377.

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

Kalsbeek, J. W. H., & Ettema, J. H. (1963). Scored regularity of the heart rate pattern and the measurement of perceptual or mental load. Ergonomics, 6, 306.

Kantowitz, B. H., & Knight, J. L. (1976). Testing tapping time-sharing, II: Auditory secondary tasks. Acta Psychologica, 40, 342-362.

Keppel, G. (1973). Design and analysis: A researcher's handbook. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Kerr, B. (1973). Processing demands during mental operations. Memory and Cognition, 1, 401-412.

Kidd, J. S., & Van Cott, H. P. (1972). System and human engineering analyses. In H. Van Cott & R. Kinkade (Eds.), Human engineering guide to equipment design. Washington, D.C.: U.S. Government Printing Office.

Kinsbourne, M., & Hicks, R. (1978). Functional cerebral space. In J. Requin (Ed.), Attention and performance VII. New York: Academic Press.

Kopala, C. J. (1979). The use of color-coded symbols in a highly dense

situation display. Proceedings of the 23rd Annual Meeting of the Human Factors Society, 397-401.

Krause, E. F., & Roscoe, S. N. (1972). Reorganization of airplane manual flight control dynamics. Proceedings of the 16th Annual Meeting of the Human Factors Society, 117-126.

Krebs, M. J., Wingert, J. W., & Cunningham, T. (1977). Exploration of an oculometer-based model of pilot workload. Washington, D.C.: (NASA Report CR-145153).

Kruskal, J. B., & Wish, M. (1978). Multidimensional scaling. Beverly Hills, CA: Sage Publications.

Kruskal, J. B., Young, F. W., & Seery, J. B. (1973). How to use KYST: A very flexible program to do multidimensional scaling and unfolding. Murray Hill, NJ: Bell Laboratories Technical Report.

Lockhead, G. R. (1973). Choosing a response. In S. Kornblum (Ed.), Attention and performance IV. New York: Academic Press.

McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. Science, 211, 77-79.

McCloy, T. M., Derrick, W. L., & Wickens, C. D. (1983). Workload assessment metrics - What happens when they dissociate. SAE Second Aerospace Behavioral Engineering Technology Conference. Society of Automotive Engineers.

McLeod, P. (1977). A dual task response modality effect: Support for multiprocessor models of attention. Quarterly Journal of Experimental Psychology, 29, 651-667.

McLeod, P. (1978). Does probe RT measure central processing demand? Quarterly Journal of Experimental Psychology, 30, 83-89.

Milord, J. T., & Perry, R. P. (1977). A methodological study of overload. The Journal of General Psychology, 97, 131-137.

Moray, N. (1967). Where is capacity limited? A survey and a model. Acta Psychologica, 27, 84-92.

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

Moray, N. (1982). Subjective mental load. Human Factors, 23, 25-40.

Moray, N., Johannsen, G., Pew, R. W., Rasmussen, J., Sanders, A. F., & Wickens, C. D. (1979). Final report of the experimental psychology group. In N. Moray (Ed.), Mental workload: Its theory and measurement. New York: Plenum Press.

Mulder, G. (1978). The heart of mental effort. Unpublished doctoral thesis, University of Groninger.

Mulder, G., & Mulder, L. J. M. (1981). Information processing and cardiovascular control. Psychophysiology, 18, 392-402.

Murphy, M. R., McGee, L. A., Palmer, E. A., Paulk, C. H., & Wempe, T. E. (1978). Simulator evaluation of three situation and guidance displays for V/STOL aircraft zero-zero landing approaches. IEEE Transactions on Systems, Man, & Cybernetics, 18, 563-571.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system: A model of multiple capacity. Psychological Review, 86, 214-255.

Navon, D., & Gopher. D. (1980). Task difficulty, resources, and dual-task performance. In R. Nickerson & R. Pew (Eds.), Attention and performance VIII. Englewood Cliffs, NJ: Erlbaum.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. Cognitive Psychology, 7, 44-64.

North, R. A. (1977). Task components and demands as factors in dual-task performance. Savoy, IL: University of Illinois Aviation Research Laboratory Technical Report 77-2.

North, R. A., & Graffunder,, K. (1979). Evaluation of a pilot workload metric for simulated VTOL landing tasks. Proceedings of the 23rd Annual Meeting of the Human Factors Society, 357-361.

Nygren, T. E., & Jones, L. E. (1977). Individual differences in perceptions and preferences for political candidates. Journal of Experimental Social Psychology, 13, 182-197.

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

Olshavsky, R. W., MacKay, D. B., & Sentell, G. (1975). Perceptual maps of supermarket locations. Journal of Applied Psychology, 60, 80-86.

Pew, R. W. (1979). Secondary tasks and workload measurement. In N. Moray (Ed.), Mental workload: Its theory and measurement. New York: Plenum Press.

Porges, S. W. (1981). Individual differences in attention: A possible physiological substrate. In B. Keogh (Ed.), Advances in special education. Greenwich, CT: JAI Press.

Porges, S. W., Bohrer, R. E., Cheung, M. N., Drasgow, F., McCabe, P. M., & Keren, G. (1980). New time-series statistic for detecting rythmic

co-occurrence in the freqency domain: The weighted coherence and its application to psychophysiological research. Psychological Bulletin, 88, 580-587.

Porges, S. W., Bohrer, R. E., Keren, G., Cheung, M. N., Franks, G. J., & Drasgow, F. (1981). The influence of methylphenidate on spontaneous autonomic activity and behavior in children diagnosed as hyperactive. Psychophysiology, 18, 42-48.

Porges, S. W., & Smith, K. M. (1980). Defining hyperactivity: Psychophysiological and behavioral strategies. In C. Whalen & B. Henker (Eds.), Hyperactive children: The social ecology of identification and treatment. New York: Academic Press.

Reid, G. B., Shindledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the 25th Annual Meeting of the Human Factors Society, 522-526.

Reid, G. B., Eggemeier, F. T., & Nygren, T. E. (1982). An individual differences approach to SWAT scale development. Proceedings of the 26th Annual Meeting of the Human Factors Society, 639-652.

Rosler, F. (1978). Cortical potential correlates of selective attention in multidimensional scaling. Biological Psychology,7, 223-238.

Roscoe, A. H. (Ed.). (1978). Assessing pilot workload. AGARD-AG-233.

Ross, R. (1934). Optimum orders for the presentation of pairs in the method of paired comparisons. Journal of Educational Psychology, 25, 375-382.

Sanders, A. F. (1979). Some remarks on mental load. In N. Moray (Ed.), Mental workload: Its theory and measurement. New York: Plenum Press.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. Psychological Review, 86, 87-123.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. Cognitive Psychology, 7, 82-138.

Sheridan, T. B. (1980). Mental workload--What is it? Why bother with it? Human Factors Society Bulletin, 23, 1-2.

Sheridan, T. B., & Simpson, R. (1979). IFR pilot mental workload rating. Massachusetts Institute of Technology Flight

Transportation and Man-Machine Systems Laboratory Report.

Sheridan, T. B., & Stassen, H. G. (1979). Definitions, models, and measures of human workload. In N. Moray (Ed.), Mental workload: Its theory and measurement. New York: Plenum Press.

Shoben, E. J. (1976). The verification of semantic relations in a same-different paradigm: An asymmetry in semantic memory. Journal of Verbal Learning and Verbal Behavior, 15, 365-379.

Smith, R. (1976). A unified theory of pilot opinion rating. Proceedings of the 12th Annual Conference on Manual Control, University of Illinois, 542-554.

Soli, S. D., & Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. Journal of the Acoustal Society of America, 66, 46-59.

Spady, A. A., Jr. (1978). Airline pilot scan patterns during simulated ILS approaches. Hampton, VA: NASA Langley Research Center (NASA-TP-1250).

Stackhouse, S. P. (1973). Workload evaluation of LINO display. Minneapolis, MN: Honeywell (7201-3408).

Steininger, K. (1977). Subjective ratings of flying qualities and pilot workload in the operation of a short haul jet transport aircraft. Proceedings of AGARD Conference of Studies of Pilot Workload (AGARD- CPP-217) (B10-1-B10-12).

Sternberg, S. (1969). The discovery of processing stages: An extension of Donders' method. Acta Psychologica, 30, 276-315.

Stone, L. W., Sanders, M. G., Glick, D. D., Wiley, R., & Kimball, K. A. (1979). A human performance/workload evaluation of the AN/PVS-5 bifocal night vision goggles. Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory Report No. 79-11.

Triesman, A. M., & Davies, M. (1973). Divided attention between eye and ear. In S. Kornblum (Ed.), Attention and performance IV. New York: Academic Press.

Van Cott, H., & Kinkade, R. (Eds.). (1972). Human engineering guide to equipment design. Washington, D.C.: U.S. Government Printing Office.

Wewerinke, P. H. (1974). Human operator workload for various control situations. Proceedings of the 10th Annual Conference on Manual Control, Wright-Patterson AFB, OH, 167-192.

Wewerinke, P. H. (1977). Performance and workload analysis of

in-flight helicopter tasks. Proceedings of the 13th Annual Conference on Manual Control, Massachusetts Institute of Technology, 106-117.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson & R. Pew (Eds.), Attention and performance VIII. Englewood Cliffs, NJ: Erlbaum.

Wickens, C. D. (1983). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), Varieties of attention. New York: Academic Press.

Wickens, C. D., & Derrick, W. L. (1981). Workload measurement and multiple resources. Proceedings, 1981 IEEE International Conference on Cybernetics and Society, 600-603.

Wickens, C. D., Derrick, W. L., Micalizzi, J., & Beringer, D. (1980). The structure of processing resources: Implications for task configuration and workload. Proceedings of the 24th Annual Meeting of the Human Factors Society, 253-256.

Wickens, C. D., & Kessel, C. (1980). Processing resource demands of failure detection in dynamic systems. Journal of Experimental Psychology: Human Perception and Performance, 6, 564-577.

Wickens, C. D., Mountford, S. J., & Schreiner, W. S. (1981). Time-sharing efficiency: Evidence for multiple resources, task-hemispheric integrity, and against a general ability. Human Factors, 23, 211-229.

Wickens, C. D., Sandry, D. L., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. Human Factors, 25, 227-248.

Wickens, C. D., & Yeh, Y. (1983). The dissociation between subjective workload and performance: A multiple resource approach. Proceedings of the 27th Annual Meeting of the Human Factors Society, 244-247.

Wierwille, W. W. (1979). Physiological measures of aircrew mental workload. Human Factors, 21, 575-593.

Wierwille, W. W., & Connor, S. A. (1983). Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. Human Factors, 25, 1-16.

Wierwille, W. W., & Williges, R. C. (1978). Survey and analysis of operator workload assessment techniques. Blacksburg, VA: Systemetrics, Inc. Report No. S-78-101.

Winer, B. J. (1971). Statistical principles in experimental design.

New York: McGraw-Hill.

Wish, M. (1979). Dimensions of dyadic communication. In S. Weitz (Ed.), Nonverbal communication. New York: Oxford University Press.

Wish, M., & Carroll, J. D. (1974). Applications of individual differences scaling to studies of human perception and judgment. In E. Carterette & M. Friedman (Eds.), Handbook of perception, Vol. 2. New York: Academic Press.

Wish, M., Deutsch, M., & Kaplan, S. J. (1976). Perceived dimensions of interpersonal relations. Journal of Personality and Social Psychology, 33, 409-420.

Wolf, J. D. (1978). Crew workload assessment--development of a measure of operator workload. Wright-Patterson AFB, OH: Flight Dynamics Laboratory Final Technical Report (AFFDL-TR-78-165).

OFFICE OF NAVAL RESEARCH

Engineering Psychology Group

TECHNICAL REPORTS DISTRIBUTION LIST

CAPT Paul R. Chatelier
Office of the Deputy Under Secretary
OUSDRE (E & LS)
Pentagon, Room 3D129
Washington, D.C. 20301

CDR Paul Girard
Code 250
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Dr. Edward H. Huff
Man-Vehicle Systems Research Division
NASA Ames Research Center
Moffett Field, CA 94035

Mathematics Group
Code 411-MA
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Special Assistant for Marine Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

CDR James Offutt, Officer-in-Charge
ONR Detachment
1030 E. Green St.
Pasadena, CA 91106

Human Factors Department
Code N-71
Naval Training Equipment Center
Orlando, FL 32813

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, D. C. 20375

Engineering Psychology Group
Office of Naval Research
Code 442EP
800 N. Quincy St.
Arlington, VA 22217 (3 cys)

Physiology Program
Office of Naval Research
Code 441NP
800 N. Quincy St.
Arlington, VA 22217

Manpower, Personnel & Training Program
Code 270
Office of Naval Research
Arlington, VA 22217

CDR Kent S. Hull
Helicopter/VTOL Human Factors Office
NASA-Ames Research Center
MS 239-21
Moffett Field, CA 94035

Dr. Robert G. Smith
Office of the Chief of Naval Operations, OP987H
Personnel Logistics Plans
Washington, D. C. 20350

Combat Control Systems Dept.
Code 35
Naval Underwater Systems Center
Newport, RI 02840

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D.C. 20375

CDR Normal E. Lane
Code N-7A
Naval Training Equipment Center
Orlando, FL 32813

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Human Factors Engineering
Code 8231
Naval Ocean Systems Center
San Diego, CA 92152

Mr. Philip Andrews
Naval Sea Systems Command
NAVSEA 61R
Washington, D.C. 20362

Dr. L. Chmura
Naval Research Laboratory
Code 7592
Computer Sciences & Systems
Washington, D.C. 20375

Office of the Chief of Naval
Operations (OP-115)
Washington, D. C. 20350

Human Factors Technology Administrator
Office of Naval Technology
Code MAT 0722
800 N. Quincy St.
Arlington, VA 22217

Dr. Arthur Bachrach
Behavioral Sciences Dept.
Naval Medical Research Institute
Bethesda, MD 20014

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Commanding Officer
Naval Health Research Center
San Diego, CA 92152

Naval Training Equipment Center
Attn: Technical Library
Orlando, FL 32813

Dr. Ross Pepper
Naval Ocean Systems Center
Hawaii Laboratory, P.O. Box 997
Kailua, HI 96734

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D. C. 20380

Larry Olmstead
Naval Surface Weapons Center
NSWC/DL
Code N-32
Dahlgren, VA 22448

CDR C. Hutchins
Code 55
Naval Postgraduate School
Monterey, CA 93940

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 334A
Washington, D. C. 20361

CAPT Robert Biersner
Naval Medical R & D Command
Code 44
Naval Medical Center
Bethesda, MD 20014

Head
Aerospace Psychology Dept.
Code L5
Naval Aerospace Medical Res.
Lab.
Pensacola, FL 32508

Dr. W. Moroney
Human Factors Section
Systems Engineering Test
Directorate
U.S. Naval Air Test Center
Patuxent River, MD 20670

Dr. Harry Crisp
Code N 51
Combat Systems Department
Naval Surface Weapons Center
Dahlgren, VA 22448

Mr. Stephen Merriman
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch
Code 3152
Naval Weapons Center
China Lake, CA 93555

Director, Organizations and Systems
Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Director, Human Factors Wing
Defense & Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsview, Ontario M3M 3B9
Canada

Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Dr. Clinton Kelly
Defense Advanced Research Projects
Agency
1400 Wilson Blvd.
Arlington, VA 22209

Dr. M. D. Montemerlo
Human Factors & Simulation
Technology, RTE-6
NASA HQS
Washington, D. C. 20546

Dr. Harry Synder
Dept. of Industrial Engineering
Virginia Polytechnic Institute
& State University
Blacksburg, VA 24061

Dr. Robert T. Hennessy
NAS-National Research Council (COHF)
2101 Constitution Ave., N.W.
Washington, D.C. 20418

Dr. Robert Blanchard
Navy Personnel Research and
Development Center
Command and Support Systems
San Diego, CA 92152

Dr. Edgar M. Johnson
Technical Director
U.S. Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Technical Director
U.S. Army Human Engineering
Labs.
Aberdeen Proving Ground, MD
21005

Human Factors Engineering
Branch
Code 4023
Pacific Missile Test Center
Point Mugu, CA 93042

U.S. Air Force Office of
Scientific Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D. C. 20332

Defense Technical Information
Center
Cameron Station, Bldg. 5
Alexandria, VA 22314 (12 cys)

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks AFB, TX 78235

Dr. Robert R. Mackie
Human Factors Research Div.
Canyon Research Group
5775 Dawson Ave.
Goleta, CA 93017

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, CA 94305

Dr. Amos Freedy
Perceptronics, Inc.
6271 Variel Ave.
Woodland Hills, CA 91364

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

Dr. Paul E. Lehner
PAR Technology Corporation
P.O. Box 2005
Reston, VA 22090

Dr. Babur M. Pulat
Department of Industrial Engineering
North Carolina A & T State Univ.
Greensboro, NC 27411

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Executive Park
Burlington, MA 01803

Dr. Robert Wherry
Analytics, Inc.
2500 Maryland Rd.
Willow Grove, PA 19090

Dr. William B. Rouse
School of Industrial & Systems Engr.
Georgia Institute of Technology
Atlanta, GA 30332

Dr. Richard Pew
Bolt Beranek & Newman, Inc.
50 Moulton St.
Cambridge, MA 02238

Dr. Baruch Fischhoff
Decision Research
1201 Oak St.
Eugene, OR 97401

Dr. T. B. Sheridan
Dept. of Mech. Engr.
Massachusetts Inst. of Tech.
Cambridge, MA 01239

Dr. James H. Howard, Jr.
Department of Psychology
Catholic University
Washington, D. C. 20064

Dr. Stanley N. Roscoe
New Mexico State University
Box 5095
Las Cruces, NM 88003

Dr. Marvin Cohen
Decision Science Consortium
Suite 721, 7700 Leesburg Pike
Falls Church, VA 22043

Professor Michael Athans
Room 35-406
MIT
Cambridge, MA 02139

Dr. Edward R. Jones, Chief
Human Factors Engineering
McDonnell-Douglas Astronautics
Box 516
St. Louis, MO 63166

Dr. Douglas Towne
Univ. of Southern California
Behavioral Technology Lab
3716 S. Hope St.
Los Angeles, CA 90007