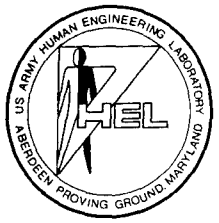
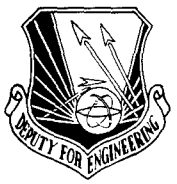




HANDBOOK OF PERCEPTION AND HUMAN PERFORMANCE

VOLUME II
Cognitive Processes and Performance

Editors:

KENNETH R. BOFF *Armstrong Aerospace Medical Research Laboratory*

LLOYD KAUFMAN *New York University*

JAMES P. THOMAS *University of California at Los Angeles*

A Wiley-Interscience Publication

JOHN WILEY AND SONS

New York • Chichester • Brisbane • Toronto • Singapore

CHAPTER 42

WORKLOAD ASSESSMENT METHODOLOGY

COLONEL ROBERT D. O'DONNELL

Human Engineering Division, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio

F. THOMAS EGGEMEIER

Wright State University, Dayton, Ohio, and Systems Research Laboratories, Inc., Dayton, Ohio

CONTENTS

Preparation of this chapter was supported in part by a contract to Wright State University from the Air Force Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio in conjunction with the Air Force Office of Scientific Research (Contract No. F33615-82-K-0522).

Colonel O'Donnell's present affiliation: Ergometrics Technology, Inc., Dayton, Ohio.

The human operator has a limited capacity to process and respond to information. Under most conditions, increases in task difficulty lead to increases in resource or capacity expenditure. Resources and capacity as used here are interchangeable terms and refer to limited processing facilities within the human system that enable task performance (e.g., Navon & Gopher, 1979; Norman & Bobrow, 1975; Wickens, 1984a). If the processing and response demands of a task exceed available capacity, the resulting overload can lead to decrements in operator performance. The term *workload* refers to that portion of the operator's limited capacity actually required to perform a particular task. The objective of workload measurement is to specify the amount of expended capacity. This quantification can be used to avoid existing or potential overloads to ensure adequate operator performance. This chapter reviews major categories of empirical workload measurement techniques and provides guidelines for the choice of appropriate assessment procedures for particular applications.

A large number of individual techniques have been proposed as workload assessment procedures. Wierwille and Williges (1978), for example, identified 28 specific techniques used to measure workload. In spite of this diversity, most empirical assessment procedures can be classified into one of three major categories: (1) *subjective measures*; (2) *performance-based measures*; and (3) *physiological measures*.

Subjective measures (discussed more extensively in Sectio 2) require operators to judge and report their own experienc of the workload imposed by performing a particular task. Ratin scales are a frequently used version of this technique.

Performance-based measures derive an index of workloa from some aspect of operator behavior or activity. There ar two major types of performance-based measures. *Primary tas measures* (discussed in Section 3) specify the adequacy of operato performance on the principal task or system function of intere (e.g., the number of errors made by a pilot while flying a aircraft). *Secondary task measures* (discussed in Section 4) pr vide an index of primary task workload based on the operator ability to perform an additional or secondary task (e.g., respon to a radio communications signal) concurrently with the primar task of interest (e.g., flying an aircraft).

Physiological measures infer the level of workload fron some aspect of the operator's physiological response to a task or system demand. These measures may include autonomi responses (e.g., pupillary reflex), central nervous system re sponses (e.g., event-related potentials), or peripheral measure (e.g., muscle activity or eye movements). These measures ar discussed more extensively in Section 5.

Given the variety of workload assessment technique available, care must be exercised in selecting appropriate tech niques for specific applications. Such choices should be guide by both theoretical considerations and practical constraints. Current data and theories (e.g., Navon & Gopher, 1979; Wickens 1980) provide some guidelines about the nature and utility o each measure, and these guidelines can then be related to the environment in which the workload measurement is to be taken. Section 1 discusses several key factors to be considered in choosing a workload assessment technique.

1. CRITERIA FOR SELECTION OF WORKLOAD ASSESSMENT TECHNIQUES

The categories of assessment techniques just described, as well as individual procedures within each category, vary along a number of dimensions that can serve as criteria for selection of a procedure for a given application. A procedure that is satisfactory for one application may not meet the measurement objectives and constraints of another application. Several authors (Chiles, 1982; Eggemeier, 1984; Gartner & Murphy, 1976; Shingledecker, 1983; Wickens, 1984b; Wierwille & Williges, 1978) have discussed major dimensions of workload assessment techniques that affect their applicability. Table 42.1 is adapted from these sources and lists five major criteria to be considered in selection of an assessment technique. This section reviews the status of each major category of assessment technique as it relates to the proposed criteria, and provides a basis for initial choice of the class(es) of techniques that should be considered for a particular application.

1.1. Sensitivity

Sensitivity refers to the capability of a technique to detect changes in the levels of workload imposed by task performance (e.g., Chiles, 1982; Gartner & Murphy, 1976; Wickens, 1984b). Techniques differ with respect to their sensitivity (e.g., Bahrick, Noble, & Fitts, 1954; Bell, 1978; Eggemeier, Crabtree, & LaPointe, 1983; Hicks & Wierwille, 1979; Isreal, Chesney, Wickens, & Donchin, 1980; Wickens & Yeh, 1983; Wierwille &

Table 42.1. Criteria for Selection of Workload Assessment Techniques

Criterion	Explanation
Sensitivity	Capability of a technique to discriminate significant variations in the workload imposed by a task or group of tasks.
Diagnosticity	Capability of a technique to discriminate the amount of workload imposed on different operator capacities or resources (e.g., perceptual versus central processing versus motor resources).
Intrusiveness	The tendency for a technique to cause degradations in ongoing primary task performance.
Implementation requirements	Factors related to the ease of implementing a particular technique. Examples include instrumentation requirements and any operator training that might be required.
Operator acceptance	Degree of willingness on the part of operators to follow instructions and actually utilize a particular technique.

Categories of workload assessment techniques vary along a number of dimensions that can serve as criteria for selection of techniques for particular applications. When considered in conjunction with the measurement objectives and practical constraints of a given application, the capability to meet the listed criteria can provide an initial basis for choice of an appropriate technique or techniques. See text and Table 42.2 (Section 1.6) for further detail and information regarding the current status of each class of technique with respect to the listed criteria. The criteria are adapted from Chiles (1982), Eggemeier (1984), Gartner and Murphy (1976), Shingledecker (1983), Wickens (1984b), and Wierwille and Williges (1978).

Casali, 1983a), and the sensitivity of a procedure should be matched with the requirements of a given application. In some cases a relatively insensitive measure may be sufficient, such as in identifying areas of extreme workload or "choke points" during system operation. In other cases fine discriminations between two or more proposed system elements such as control/display design options, operational procedures, operator duty allocations, or crew composition must be made. Matching the sensitivity of the workload measurement procedure with the sensitivity required to meet assessment objectives, therefore, constitutes a basic decision in selecting a workload measure.

General guidelines for matching the sensitivity of the procedure to the application can be provided on the basis of the theoretical relationship between performance and workload. Figure 42.1 illustrates a hypothetical workload/performance relationship adapted from several sources (Meister, 1976; North, Stackhouse, & Graffunder, 1979; Tole, Stephens, Harris, & Eprath, 1982). Three regions are identified by the relative levels of workload imposed on the operator.

Region A includes low to moderate levels of workload and is characterized by adequate operator performance. In this region, increases in workload are not accompanied by variations in performance since the operator has sufficient spare information-processing capacity or resources to compensate for a workload increase. Consequently, by working harder, the operator is able to maintain adequate performance.

In this general region, the principal workload assessment question deals with determining the amount of reserve or spare information-processing capacity afforded by performance of a task. By knowing how close an operator is to a performance decrement, the potential for degraded performance can be assessed. Primary task measures of performance are insensitive to load variations in this region, since performance would be adequate in all cases. However, even under conditions of adequate performance, it may still be critical to determine, for instance, which of two display options (e.g., alphanumeric versus pictorial) imposes the lower workload and affords the greatest spare capacity. This would be the case when it was anticipated that other information-processing and response requirements (e.g., emergency situations, monitoring other displays) in the operational environment might be sufficient to overload the operator and lead to degraded performance. Subjective, physiological, or secondary task workload metrics can be more sensitive than primary task measures in this region and would, therefore, be more appropriate to meet the objective of identifying potential overloads. In particular, secondary task methodology is specifically designed to assess the spare or reserve processing capacity that remains after sufficient resources have been allocated to perform the primary task. In effect, the concurrent performance of the primary and secondary tasks moves the entire workload into Region B of Figure 42.1, thus permitting secondary task performance measures to reflect variations in primary task load.

Region B in Figure 42.1 represents higher levels of workload that exceed the capability of the operator to compensate. As a consequence, primary task performance decrements occur, and a monotonic relationship exists between workload and performance. Many investigators (e.g., Meister, 1976; Norman & Bobrow, 1975; North et al., 1979; Tole et al., 1982) assume that degradations in performance will be gradual during the initial stages of overload, and that catastrophic or total failure will

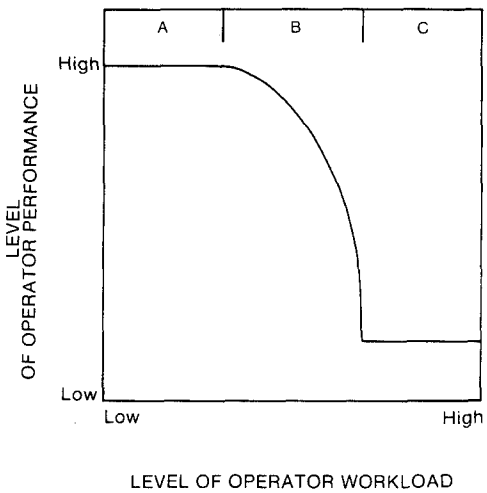


Figure 42.1. Hypothetical relationship between workload and operator performance proposed to depend upon the relative level of operator workload. There are three distinct regions in this relationship. Under low to moderate levels of operator load (Region A), increases in workload are not accompanied by variations in performance. It is assumed that in this region the operator has sufficient spare processing capacity or resources to compensate for increased levels of load and can therefore maintain adequate performance. In Region B higher levels of workload exceed the capability of the operator to compensate, and performance decrements occur. In this region a monotonic relationship exists between workload and performance. Under extremely high levels of load (Region C), very low levels of performance are assumed to result from the operator's lack of capacity to deal with the workload being imposed. (Adapted from D. Meister, *Behavioral foundations of system development*. Copyright 1976 by John Wiley & Sons, Inc. Reprinted with permission. From R. A. North, S. P. Stackhouse, & K. Graffunder, *Performance, physiological, and oculometer evaluations of VTOL landing displays* (TR 3171) NASA/Langley, 1979. Reprinted with permission. From J. R. Tole, A. T. Stephens, R. L. Harris, & A. Eprath, Quantification of workload via instrument scan. *Proceedings of Workshop on Pilot Workload and Pilot Dynamics* (AFFTC-TR-82-5) Edwards AFB, 1982.

occur only at higher levels of load. Region B of the workload continuum can be assessed by primary task measures, which provide information regarding existing rather than potential information-processing overloads. Workload differences between systems that produce such differences in primary task performance can usually be evaluated easily, especially if one monitors operator strategies and procedures to detect compensatory behaviors. Subjective, physiological, or secondary task measures can also be applied in this region. However, they are actually required only when the primary task assessment is not sensitive enough to indicate small differences in workload.

In Region C of Figure 42.1 workload is extremely high and performance is catastrophically low. Again, primary task and other categories of measures would clearly indicate high levels of load in this region. Since primary performance is very low and variable, it would be extremely difficult to differentiate levels of workload within this region.

In many practical situations the workload questions asked of the human factors engineer involve levels of loading in the A or, less frequently, the B region of Figure 42.1. For this reason, assessment techniques emphasizing the sensitivity provided by secondary task, physiological, and subjective measures have recently received considerable attention (e.g., Wierwille & Williges, 1978). In Region A, these techniques may provide the only sensitive techniques for discriminating levels of workload and identifying potential overload situations. Therefore secondary task, subjective, or physiological measures should be considered in preference to primary task measures for workload assessments in this region. Primary task measures, on the other hand, can provide information regarding the presence of existing overloads and degraded performance in Region B and represent an important assessment technique in this region.

Although the use of secondary task, subjective, or physiological procedures can be recommended in Region A and primary task measures in Region B of Figure 42.1, no more specific recommendations for choice of a technique can be made on the basis of the sensitivity criterion alone. More extensive and specific recommendations would require data on the differences in sensitivity among secondary task, subjective, and physiological metrics. Unfortunately, there is a lack of such data (but see Acton, Crabtree, & Shingledecker, 1983; Casali & Wierwille, 1983; Hicks & Wierwille, 1979; Shingledecker, Acton, & Crabtree, 1983; Shingledecker, Crabtree, & Acton, 1982; Wierwille, & Casali, 1983a; Wierwille & Connor, 1983, for initial efforts in this direction). Until more complete data on the relative sensitivity of these measures are available, no specific conclusions concerning relative sensitivity can be made. Where available, specific data concerning sensitivity are presented under the appropriate technique description.

Choice of a metric on the basis of sensitivity is, therefore, related to the objective that is to be satisfied by the workload measure. If the objective is to determine if processing overloads leading to degraded operator performance actually exist within a task or system design option, primary task measures should prove satisfactory. On the other hand, if the objective is to identify the potential for overload and degraded performance, more sensitive techniques (physiological, secondary task, subjective) should be considered for application. It is important to note that in many instances both objectives should be addressed in a comprehensive assessment of operator load. For example, if primary task measures indicate that no overload currently exists and that performance is adequate on two tasks or design options, it is probable that an investigator would want to evaluate more specifically the potential for overload among the design options through use of a more sensitive procedure. This type of approach, of course, suggests the complementary use of various assessment procedures to address the different workload questions arising during the course of system or task evaluation.

1.2. Diagnosticity

The characteristic of diagnosticity (Shingledecker, 1983; Wickens, 1984b; Wickens & Derrick, 1981a) is based upon the multiple-resources approach to capacity limitations within the human information-processing system (e.g., Navon & Gopher, 1979; Wickens, 1984a; see also Sperling & Dosher, Chapter 2). This theory proposes that the overall processing system can be described as a series of independent capacities or resources that are not interchangeable (see Gopher & Donchin, Chapter 41, for a more complete treatment and discussion of the data supporting this theory). For example, as suggested by Wickens (1984a), the perceptual and central processing stages may draw upon one type of resource, while the response or motor input stage may draw from a separate resource. According to this position, a manual control or tracking task might place minimal demands on the perceptual/central processing resources, even though the motor resources may be exhausted. A monitoring or vigilance task might have the opposite resource-demand composition.

It has been proposed (e.g., Wickens & Derrick, 1981a) that workload measures vary in their degree of diagnosticity. For example, pupil diameter (Beatty, 1982; Beatty & Kahneman, 1966) and some subjective ratings scales (Eggemeier, Crabtree, Zingg, Reid, & Shingledecker, 1982; Notestine, 1983; Reid, Shingledecker, & Eggemeier, 1981; Wierwille & Casali, 1983b) appear to index workload across the entire processing system. With such measures, it would not be possible to diagnose which type of resource or capacity (e.g., perceptual versus motor output) had been affected, although an overall assessment of workload would still be possible. On the other hand, the event-related brain potential (e.g., Isreal, Chesney, Wickens, & Donchin, 1980; Isreal, Wickens, Chesney, & Donchin, 1980) and some secondary tasks (e.g., North, 1977; Shingledecker et al., 1983; Wickens & Kessel, 1980) show a greater degree of diagnosticity in that they appear to be maximally sensitive to particular types of resource/capacity expenditure. Use of such measures would permit more precise localization of the source of an overload, although they could be insensitive to loading in unmeasured resources.

Although the diagnosticity of specific measures is considered separately in the following sections, some general characterizations of the diagnosticity associated with major assessment categories can be supplied here. Subjective measures generally are considered to exhibit low diagnosticity owing to the inability of the operator to discriminate individual resources. Primary task measures similarly show low diagnosticity in most cases, since it is not usually possible to identify the specific source of a performance breakdown. Secondary task measures are typically considered highly diagnostic and therefore provide an index of the load imposed on specific resources. Physiological measures, as noted, can be either global (e.g., pupil diameter) or highly diagnostic (e.g., event-related potentials).

The choice of a global versus a specific diagnostic measure is directly related to the objective to be met by the workload assessment. If the goal of an evaluation is to determine if a workload problem exists at all, a technique with low diagnosticity

(e.g., subjective, primary task) should be used. On the other hand, when information about the specific locus of an identified problem is required to suggest an appropriate design modification, more diagnostic (e.g., secondary task, selected physiological) techniques should be chosen.

1.3. Primary Task Intrusion

Sensitivity (Section 1.1) and diagnosticity (Section 1.2) are derived from the theoretical bases of workload measures. The remaining three criteria from Table 42.1 refer to pragmatic considerations in implementing the measures in a variety of environments (e.g., operational, simulation, laboratory). Foremost among these is the degree to which the workload measure degrades primary task performance. In some applications, higher degrees of degradation or intrusion might be more acceptable than in others (e.g., a laboratory environment as opposed to a simulation or field test situation). Frequently, safety considerations preclude use of a technique that would contribute to degradations in primary task performance.

Intrusion also produces significant problems in interpreting the data that result from use of an assessment technique. The results of a procedure associated with significant degradations in primary task performance cannot accurately represent the degree of load required for unimpaired performance of the primary task.

There is no systematic data base that addresses the degree of intrusion that can be expected with all techniques. Although some efforts (e.g., Casasli & Wierwille, 1982, 1983; Rahimi & Wierwille, 1982; Wierwille & Connor, 1983) that permit some comparisons of intrusion among a number of measures have been conducted, the present data are not sufficient to draw general conclusions. However, current data and knowledge of the implementation procedures generally associated with each class of technique can be used to suggest a set of initial guidelines regarding the degree of intrusion to be expected with each category. Subjective measures, typically taken after completion of primary performance, and those physiological techniques not requiring additional operator processing or response will generally be the least intrusive types of measures. Secondary task methodology, on the other hand, has been frequently associated with significant levels of intrusion (Gartner & Murphy, 1976; Ogden, Levine, & Eisner, 1979, Rolfe, 1971; Williges & Wierwille, 1979). Although a variety of different procedures have been proposed to overcome their intrusiveness (see Section 4), it is desirable to consider seriously the utility of secondary task techniques when intrusion will present serious practical problems (e.g., in an operational environment).

1.4. Implementation Requirements

A variety of practical constraints dealing with the complexity of the measurement procedures and apparatus must be considered in the choice of a workload technique. These include such things as the instrumentation and software necessary for data collection and analysis and the amount of operator training required before valid results can be obtained.

Subjective techniques generally present the fewest implementation problems since typically they involve only paper and pencil or some other simple response apparatus. Primary task techniques also can frequently be used with minimal implementation difficulty. Apparatus and data analysis requirements are typically more extensive with physiological and secondary task procedures, and their value must be weighed against these requirements.

Many secondary tasks require some training to stabilize operator performance prior to their use. Generally, training requirements for most secondary tasks, and even for subjective techniques (see Section 3), can constitute an important consideration in selection of an assessment procedure. Such training requirements are not generally associated with physiological or primary task metrics. When operator training time is limited, these latter procedures should be considered.

1.5. Operator Acceptance

Workload measurement techniques used to answer operational questions must also be evaluated with respect to the subject's perception of the validity and utility of the procedures. This is particularly important where the subject population includes experienced operators of the system being evaluated. Workload assessment procedures perceived as intrusive or artificial incur the risk of being ignored or performed at substandard levels, thus compromising the potential effectiveness of the technique(s).

Unfortunately, with few exceptions (e.g., Hallsten & Borg, 1975; Katz, 1980), few data exist addressing the degree of acceptance associated with a particular technique. Acceptance of a technique will vary among subject populations, and it can generally be assumed that rejection of techniques that lack face validity will increase as experimental situations more closely approximate operational environments familiar to operators. For example, secondary tasks for use by experienced pilots in a high fidelity simulator should be chosen to reflect activities that might occur in the normal course of the pilot's performance (e.g., radio communications).

Care should be taken to explain the purposes and procedures of assessment techniques to the subject. This is particularly true for physiological measures, which can sometimes lead subjects to be apprehensive or suspicious of their use. Detailed explanations can increase operator acceptance of these measures and maximize their potential validity.

Operator acceptance and face validity do not, of course, ensure that a technique will reflect the actual levels of workload or the degree of capacity expenditure associated with task performance. Choice of an assessment technique should, therefore, not be based primarily on these considerations. However, the capability to meet these criteria can be an important factor in ensuring that the full potential of an assessment technique is realized, particularly in applications involving operational systems.

1.6. Summary of Guidelines for Choice of a Measurement Technique

From the preceding discussion, it is clear that the most important starting point in the choice of a workload assessment technique is determining the objective to be satisfied by the measure. Paradoxically, this is not always easy. Workload questions are frequently asked in generic, undefined ways: "Is workload too high?" or "What is the workload of this system?" Asked in this way, the questions are of course meaningless. The criterion against which "too high" is to be judged, or the characteristics and demands of the systems to be evaluated must be specified.

Once the objective is clearly specified, the sensitivity and diagnosticity required to answer the question can be determined. If several categories of techniques meet the sensitivity and di-

Table 42.2. Summary of Workload Assessment Technique Capabilities

	Sensitivity	Diagnosticity	Intrusiveness	Implementation Requirements	Operator Acceptance
Primary task measures	Discriminate overload from nonoverload situations. Capable of reflecting levels of capacity expenditure in overload conditions. Used to determine if operator performance will be acceptable with a particular design option, task, or operating condition.	Not considered diagnostic. Represents a global measure of workload that is sensitive to overloads anywhere within the operator's processing system.	Nonintrusive since no additional operator performance or report required.	Instrumentation for data collection can restrict use in operational environments. Use requires mock-ups, simulators, or operational equipment. No operator training required.	No systematic data. No reason to expect negative operator opinion.
Secondary task methods	Capable of discriminating levels of capacity expenditure in nonoverload situations. Used to assess reserve capacity afforded by a primary task. Can be used to assess the relative potential for overload among design options, tasks, or operating conditions.	Capable of discriminating some differences in resource expenditure (e.g., central processing versus motor). Diagnosticity suggests complementary use with more generally sensitive measures, with the latter initially identifying overloads and secondary tasks being used subsequently to pinpoint the locus of overload.	Primary task intrusion has represented a problem in many applications, particularly in the laboratory. Data are not extensive in operational environments. Several techniques (e.g., embedded secondary task, adaptive procedures, Section 4.4.2) have been designed to control intrusion. Potential for intrusion could limit use in operational environments.	Instrumentation for data collection can restrict use in operational environments, but some tasks have been instrumented for in-flight use. Use requires mock-ups, simulators, or operational equipment. Some operator training usually required to stabilize secondary task performance.	No systematic data. Requirement to perform secondary task could distract operator. Technique such as embedded secondary task (Section 4.4.2) should minimize any acceptance problems.

The capability of the major classes of workload assessment techniques to meet the listed criteria can be used to identify the class or classes of technique(s) that satisfy the objectives and constraints of a particular application. The criteria of sensitivity and diagnosticity relate to the objective to be satisfied by the workload measure; intrusiveness, implementation requirements, and operator acceptance deal with practical constraints to be be satisfied. This table summarizes the current status of each class of technique as it relates to each of the criteria. When the capabilities required in an assessment have been specified with respect to each of the criteria, the summary information can be used to guide initial choice of the most appropriate technique(s). When the initial choice has been made, consult appropriate sections of this chapter (Subjective, 2.0; Primary Task, 3.0; Secondary Task, 4.0; Physiological, 5.0) for details on specific techniques within each category. (From F. T. Eggemeier, Workload metrics for system evaluation, Proceedings of the Defense Research Group Panel VIII Workshop "Application of System Ergonomics to Weapon System Development," *Shrivenham, England, 1984. Reprinted with permission.)*

agnosticity criteria, practical constraints can serve as additional screening devices, with intrusion and implementation requirements being more heavily weighted than operator acceptance. Table 42.2 (Eggemeier, 1984) summarizes the current general status of each major category of technique as it relates to the five criteria proposed. The information in the table can be used to determine initially which categories of techniques might be considered for a particular application.

When a preliminary determination has been made regarding the category or categories of techniques under consideration for use, appropriate sections of this chapter and original references can be consulted for detailed information concerning individual techniques within each category.

1.7. Key References

Several comprehensive reviews of workload assessment techniques have been published in recent years. Reviews by Jahns (1973), Gartner and Murphy (1976), Wierwille and Williges (1978), and Chiles (1982) discuss the major categories of assessment techniques as outlined, and also identify criteria that should be met by the techniques. Because of their comprehensive nature, these reviews are particularly valuable as overviews of workload assessment techniques and should be among the first references consulted for an introduction to the area. Moray (1979) edited the proceedings of a conference on mental workload. Johannsen, Moray, Pew, Rasmussen, Sanders, and Wickens (1979) and Sanders (1979) treat the major classes of measures and provide excellent discussions of the relationship between theoretical positions regarding operator capacity limitations and workload assessment techniques.

For more detailed treatments of individual classes of techniques, consult Ogden et al. (1979) and Rolfe (1971) for secondary task methodology, Moray (1982) for subjective assessment procedures, and O'Donnell (1979) and Wierwille (1979) for psychophysiological techniques.

Table 42.2. (*continued*)

	Sensitivity	Diagnosticity	Intrusiveness	Implementation Requirements	Operator Acceptance
Physiological techniques	Capable of discriminating levels of capacity expenditure in nonoverload situations. Can be used to assess the relative potential for overload among design options, tasks, or operating conditions.	Some techniques (e.g., event-related brain potential) appear diagnostic of some resources, whereas other measures (e.g., pupil diameter) appear more generally sensitive. Choice of technique dependent on purpose of measurement (screening for any overload versus identifying locus of overload).	Intrusion does not appear to represent a major problem, although there are data to indicate that some interference can occur.	Instrumentation for data collection can restrict use in operational environments. Use requires mock-ups, simulators, or operational equipment. No operator training required.	No systematic data. Instrumentation and recording equipment could represent potential problems, but no significant problems reported in literature.
Subjective techniques	Capable of discriminating levels of capacity expenditure in nonoverload situations. Can be used to assess the relative potential for overload among design options, tasks, or operating conditions.	Not considered diagnostic. Available evidence indicates that rating scales represent a global measure of load. Lack of diagnosticity suggests use as a general screening device to determine if overload exists anywhere within task performance.	Intrusion does not appear to represent a significant problem. Most applications require rating scale completion subsequent to task performance and, therefore, present no intrusion problem.	Instrumentation required is usually minimal, permitting use in a number of environments. Traditional applications require mock-ups, simulators, or operational equipment. Imposes limits on use during early system development. Some familiarization with procedures can be required.	No systematic data. Informal evidence suggests that several rating scales enjoy a high degree of operator acceptance.

2. SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUES

2.1. Background

Subjective measures have been used extensively to assess operator workload (Moray, 1982; Williges & Wierwille, 1979). The reasons for the frequent use of subjective procedures include their practical advantages (e.g., ease of implementation, nonintrusiveness) and current data which support their capability to provide sensitive measures of operator load (see Section 2.2). The theoretical basis for the sensitivity of subjective measures is the assumption that increased capacity expenditure in Regions A and B of Figure 42.1 will be associated with subjective feelings of effort or exertion that can be reported accurately by the subject. Acceptance of this assumption has led a number of investigators (e.g., Gartner & Murphy, 1976; Johannsen et al., 1979; Sheridan, 1980) to suggest that subjective measures can provide valid and sensitive indicators of workload.

It is important to note, however, that the data base dealing with the various factors (e.g., task characteristics, operator experience levels) that can influence the degree of workload experienced and reported by an operator is not extensive (e.g., Johannsen et al., 1979; Moray, 1982). Also, although it can be assumed that capacity expenditure and experienced effort are related, subjective measures have not generally been validated regarding their specific capability to reflect the levels of capacity expenditure associated with task performance. Therefore, validation with respect to the present definition of workload is not complete (see Gopher & Donchin, Chapter 41, for a discussion of the problem of relating practical measures such as subjective techniques to workload theories and constructs). In addition, portions of the available data have suggested restrictions in interpreting the results of subjective techniques. Wickens and Yeh (1982, 1983), for example, have reported that some subjective measures can be heavily influenced by the number of tasks or task elements to be concurrently performed by a subject and can be relatively less sensitive than performance-based indices to some manipulations of single-task difficulty. These types of restrictions and the lack of comprehensive data related to other factors have led a number of investigators (e.g., Eggemeier et al., 1983; Gartner & Murphy, 1976; Sanders, 1979; Wickens & Yeh, 1983; Williges & Wierwille, 1979) to note limitations of subjective techniques or to suggest guidelines for their usage. These limitations and guidelines should be considered in any application of subjective techniques and are discussed more extensively in Section 2.4.

Although subjective techniques have been frequently used to assess operator load, the workload rating scale literature is characterized by limited standardization and application of individual scales, as well as by limited evidence of scales that were rigorously developed on the basis of psychometric theory (Williges & Wierwille, 1979). This section describes several

scales and psychometric techniques that have been used with some consistency to assess workload. Available data bearing on sensitivity, diagnosticity, and implementation requirements are also presented. Although no systematic data are available on the degree of intrusiveness or operator acceptance that can be expected with most of these techniques, there are sound reasons to expect that intrusiveness will typically be low (see Section 1.3).

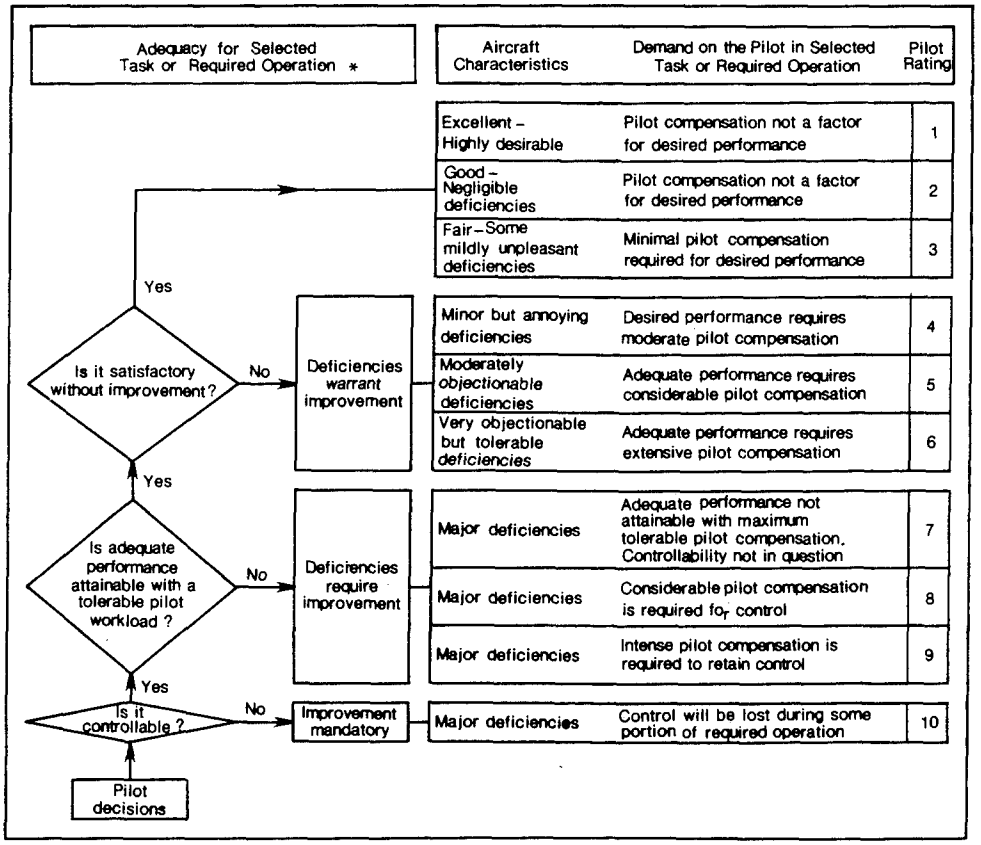
2.2. Rating Scales

2.2.1. Cooper-Harper and Related Scales. The most widely used rating procedure that can be related to workload is the Cooper-Harper Aircraft-Handling Characteristics Scale (Table 42.3) designed for use by test pilots (Cooper & Harper, 1969). Although it deals primarily with aircraft ease of control, numerous references to task demand and pilot compensation are included in the scale. Use of the Cooper-Harper scale as a workload index, therefore, involves the assumption that handling qualities and operator workload are directly related (Moray, 1982; Williges & Wierwille, 1979).

The limited available data do support a relationship between Cooper-Harper ratings and workload. R. A. Hess (1977), for example (see Wickens, Chapter 39, Figure 39.41), demonstrated a monotonic relationship between Cooper (1957) ratings and the optimal control model (e.g., Kleinman, Baron, & Levison, 1970) parameter of *fraction of attention* (see Wickens for definition and discussion) as applied to an aircraft hover control task. Since this attention parameter has been proposed as an index of workload (Levison, Elkind, & Ward, 1971), the Hess results can be interpreted as supporting a monotonic relationship between Cooper-Harper ratings and operator load. This type of conclusion has also been drawn by Moray (1982) on the basis of data (Wewerinke, 1974) that demonstrated a very high correlation ($r = +.99$) between load in a compensatory tracking task (as measured by observation noise) and ratings on a ten-point effort scale, which was highly correlated with a Cooper (1957) scale.

Additional evidence of this relationship was provided by McDonnell (1968). Subjects performed a cross-coupled secondary task (Jex, McDonnell, & Phatak, 1966) (see Section 4.4.2) in conjunction with a primary tracking task in a fixed-base flight simulator. Compensatory tracking in pitch was used for the primary task, while roll-axis tracking served as the secondary task. Basically, the instability of the secondary tracking task was increased until the concurrent primary task could no longer be maintained at criterion levels. High instability levels on the secondary task are therefore indicative of low primary task

Table 42.3. The Cooper-Harper Aircraft Handling Characteristics Scale



The Cooper-Harper aircraft handling qualities rating scale follows a decision-tree format in which a pilot initially considers the adequacy of the aircraft for some specified task or operation. Based on the initial judgment of adequacy, more detailed decisions regarding aircraft characteristics and the demands placed on the pilot are made, resulting in an eventual rating on the ten-point scale illustrated. (Rating scale adapted from Cooper & Harper, 1969.)

* Definition of required operation involves designation of flight phase and/or subphases with accompanying conditions.

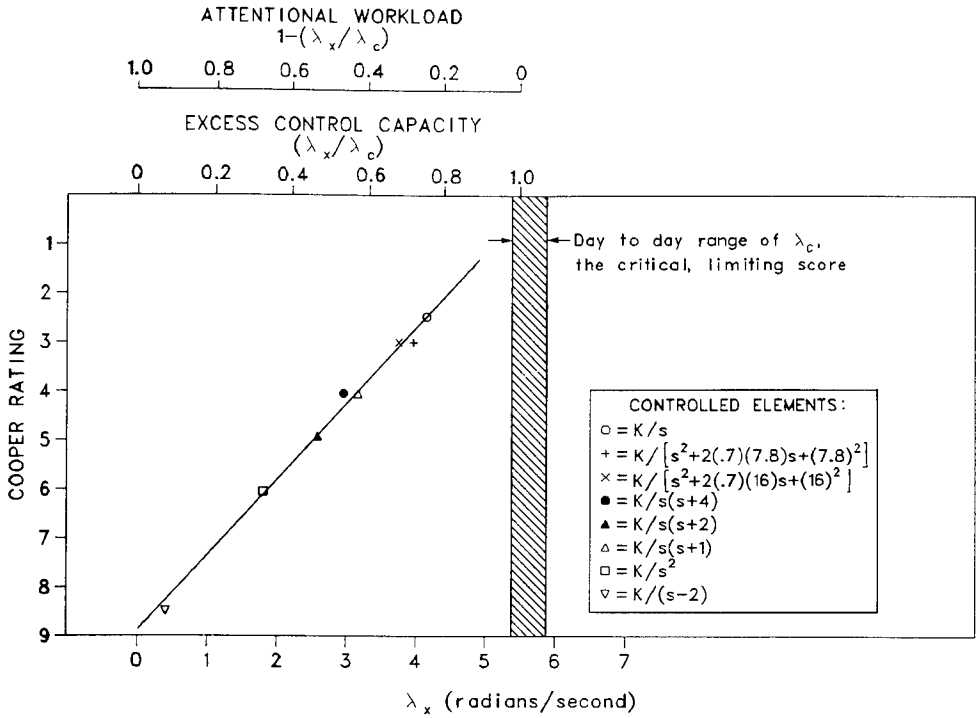


Figure 42.2. Cooper handling characteristics scale ratings as a function of excess control capacity afforded by several primary controlled elements. A cross-coupled secondary task was performed in conjunction with several primary controlled elements in a fixed-base flight simulator. Secondary task instability was increased until the concurrent primary task could no longer be maintained at criterion levels. High instability levels in the secondary task, therefore, indicate a relatively low degree of primary task workload, whereas low secondary task instability levels are indicative of high primary task loading. Cooper ratings were clearly related to the λ_x measure of excess control capacity afforded by the secondary task. This excess control capacity measure can be normalized by dividing λ_x by λ_c, the subject's instabilityscore on the secondary task without the concurrent primary task. The resulting measure of excess control capacity [*EC*] is not biased by individual differences in tracking skill. This measure is also illustrated in the figure, as is a measure of attentional workload given by $1 - EC$. Cooper ratings are highly correlated in the expected direction with each measure, thus supporting the usefulness of the Cooper scale as an index of pilot workload in the conditions tested. (Redrawn from McDonnell, 1968.) (For a more detailed treatment of the critical instability tracking task, see Chapter 39 by Wickens.)

loading, whereas low secondary task instability is associated with high primary task load.

Figure 42.2 illustrates the relationship between the Cooper ratings and the level of secondary task instability for several control plants. These data provide clear evidence of a very high correlation between Cooper ratings and the level of secondary task instability. It is important to note that the data confirm only that the Cooper-Harper scale indexes certain task demands external to the operator. It would, therefore, be premature to conclude that these ratings always reflect either the actual capacity expenditure or subjective effort incurred by task performance. However, in the absence of more complete data, it does appear reasonable to hypothesize that the Cooper-Harper handling characteristic ratings and subjective workload are related in some manner, although the relationship might not always be direct.

Since Cooper-Harper ratings have been shown to index many task variables such as control type and complexity, display sophistication, vehicle stability, and atmospheric turbulence (e.g., Crabtree, 1975; Krebs & Wingert, 1976; Lebacqz & Aiken, 1975; Schultz, Newell, & Whitbeck, 1970), the scale would appear to be a sensitive measure of a variety of handling qualities that have the potential to affect subjective workload. However, because of this wide range of sensitivity, the scale does not appear to be particularly diagnostic concerning the type of variable [e.g., display sophistication (perceptual) versus vehicle stability (motor output)] affecting the handling characteristics.

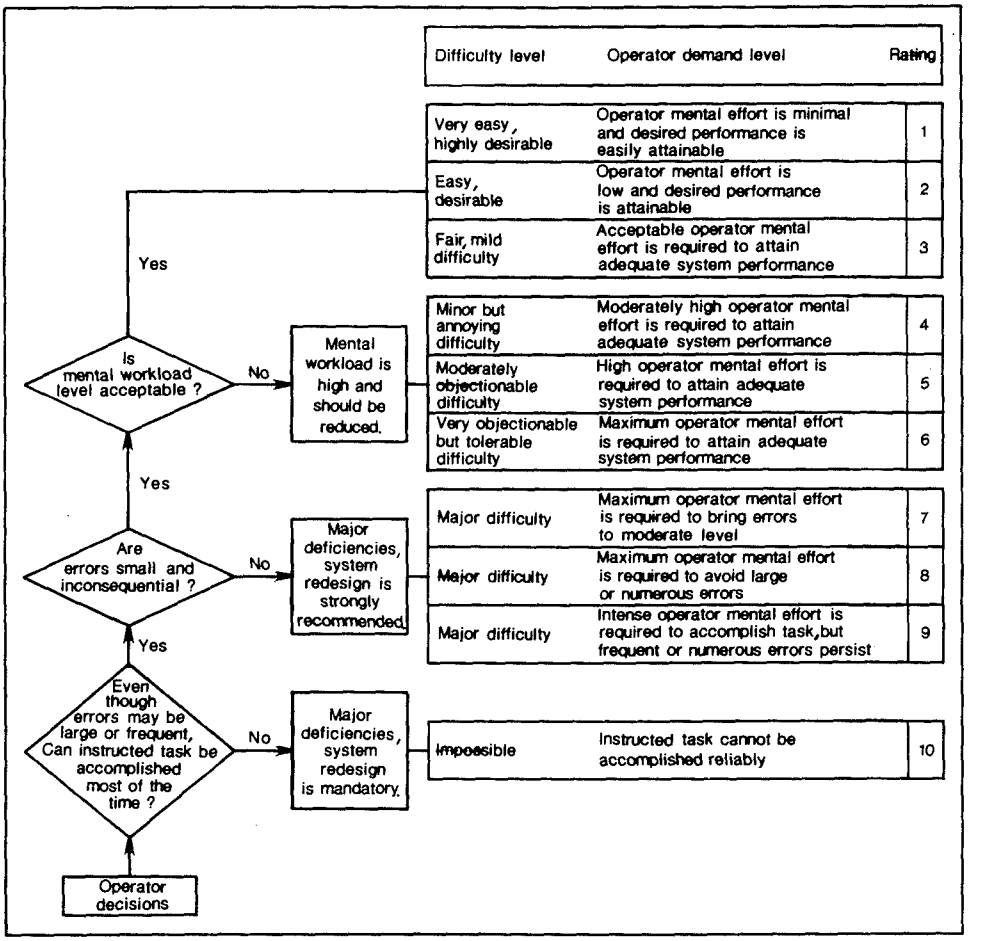
Several investigators (e.g., North et al., 1979; Wierwille & Casali, 1983b; Wolfe, 1978) have proposed mental workload rating scales modeled after the Cooper-Harper handling characteristics scale. Wolfe (1978) and North et al. (1979) reported use of a scale very similar to Cooper-Harper in both wording and format. However, references to aircraft-handling characteristics in the original scale were replaced by descriptors of pilot workload and effort. Available data support the sensitivity of this scale. Wolfe (1978) employed the scale in conjunction with instrument landings in a flight simulator, using flight control system degradation and wind gust levels to manipulate workload. The scale was sensitive to variations in difficulty and was also highly correlated ($r = +.80$) with a discriminant function that included primary task, secondary task, and opinion variables. North et al. (1979) also successfully used the scale during several types of landing approaches in a simulated vertical takeoff and landing aircraft. Mean subjective ratings varied as a function of several flight director display options and with the presence or absence of simulator motion. These variations were consistent with several flight performance measures.

As with the original Cooper-Harper scale, it is not likely that this scale would be diagnostic of the sources of workload variation. Neither Wolfe (1978) nor North et al. (1979) reported

extended training with the scale, although Wolfe (1978) did use the scale during a series of practice landings in the simulator prior to actual data collection. Use of the scale requires only paper and pencil, so overall implementation requirements are not extensive. Both this scale and the Cooper-Harper scale are tailored principally to the piloting and vehicular control environments and would require some modification if they were to be considered for more general application.

A more generally applicable modification of the Cooper-Harper scale (Table 42.4) has been proposed by Wierwille and Casali (1983b). In this modification references to aircraft handling, controllability, and pilot compensation have been replaced with terms more appropriate to workload and effort in the range of information-processing functions performed by a variety of systems operators. This scale has been evaluated in three flight simulator experiments (Casali & Wierwille, 1982, 1983; Rahimi & Wierwille, 1982). Each of the evaluation experiments involved manipulation of different types (i.e., perceptual, central processing, communications) of loading. Perceptual loading (C & Wierwille, 1982) was varied by manipulating the present rate and number of danger conditions to be detected by on the simulator instrument panel. Levels of central proce mediational (e.g., decision making, problem solving) load manipulated in the Rahimi and Wierwille (1982) stuc varying the number and complexity of the arithmetic an metric operations required to solve a series of navigat problems presented to pilots while flying the simulator. munications load (Casali & Wierwille, 1983) was vari changing the presentation rate of aircraft call signs an similarity of extraneous call signs to the targets that we be detected by pilots during segments of the simulated fl In every case, modified Cooper-Harper ratings demonstra monotonic relationship with loading levels. Representativ sults are illustrated in Figure 42.3 which shows mean stan modified Cooper-Harper ratings at three levels of central cessing load (Rahimi & Wierwille, 1982). Additional inform

Table 42.4. A Modified Version of the Cooper-Harper Handling Characteristics Scale



This version of the Cooper and Harper (1969) handling characteristics has been modified by replacing the references to aircraft handling, controllability, and pilot compensation in the original scale with terms that specifically deal with operator workload, mental effort, and performance. The decision-tree format of the original scale has been preserved so that the operator makes initial judgments regarding the adequacy of mental load and task performance and subsequently makes more refined estimates leading to a rating on the ten-point scale. In addition to dealing more directly with operator workload and effort, the wording on this scale should be applicable to a wide range of information processing and motor control tasks, thereby generalizing applications beyond the vehicular control environment treated in the original scale. (From W. W. Wierwille & J. G. Casali. A validated rating scale for global mental workload measurement applications. Proceedings of the Human Factors Society 27th Annual Meeting. *Copyright 1983 by Human Factors Society. Reprinted with permission.)*

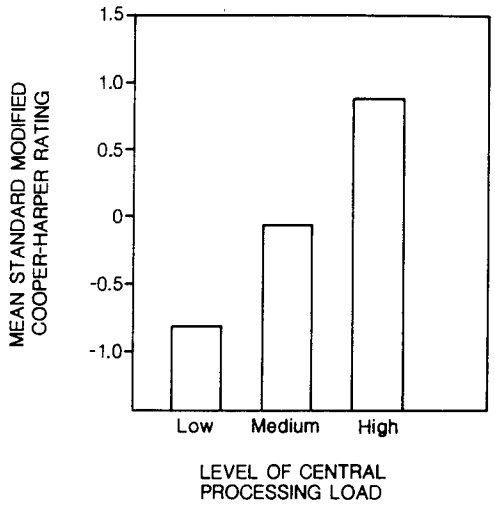


Figure 42.3. Mean standard scores on the modified Cooper-Harper scale at three levels of central processing load in a flight simulation experiment. Three levels of central processing load were imposed in a series of navigational problems that were solved by pilots in a moving-base flight simulator. Navigational load was varied by manipulating the number and complexity of the arithmetic and geometric operations required to solve a series of problems. Standardized modified Cooper-Harper ratings were significantly affected ($p < .001$) by the manipulations of central processing load. A post hoc multiple comparisons test indicated that ratings in the high loading condition differed significantly from both other conditions, but that medium load ratings were not significantly different from those in the low load condition. The modified Cooper-Harper scale therefore proved sensitive to different levels of central processing load. (From M. Rahimi & W. W. Wierwille, Evaluation of the sensitivity and intrusion of workload estimation techniques in piloting tasks emphasizing mediational activity. *Proceedings of the IEEE International Conference on Cybernetics and Society*. Copyright 1982 by IEEE. Reprinted with permission.)

bearing on the sensitivity of this modification of the Cooper-Harper scale is presented in Table 42.5, which shows the sensitivity of the scale (compared to two secondary behavioral tasks, time estimation and interval production; see Section 4.4.3) in discriminating between loading levels in perceptual, central processing, and communications tasks. The modified Cooper-Harper scale indicated significant differences in seven of the nine conditions, whereas both the time estimation and interval production tasks demonstrated differences in only four of the nine conditions. The scale also discriminated all but one of the differences demonstrated by the two secondary tasks. Thus, based on the noted monotonic relationships with load manipulations and the favorable comparison with the secondary task results, it can be concluded that the modified Cooper-Harper scale has demonstrated a high degree of sensitivity in work conducted to date. The results in Table 42.5 also indicate that the scale is sensitive to a variety of different loads employed and, therefore, suggest that the scale might provide a global rather than a highly diagnostic measure of workload.

As with other rating scales, the modified Cooper-Harper has minimal instrumentation requirements and apparently does not require extensive practice for successful application. Casali and Wierwille (1982, 1983), for example, report the use of one practice trial in the flight simulator prior to actual data collection. Some guidance on use of the scale is provided in Wierwille and Casali (1983b), and more complete instructions are included in Casali (1982).

2.2.2. University of Stockholm Scales. Few other rating scales have been used consistently to measure workload or associated factors, such as perceived difficulty. Two such scales have been evaluated at the University of Stockholm, one dealing with perceived difficulty and the other with perceived effort.

A nine-point category scale of perceived difficulty (Bratfisch, 1972; Bratfisch, Borg, & Dornic, 1972; Hallsten & Borg, 1975) has been used to assess perceived difficulty of intelligence test items involving reasoning, spatial ability, and verbal compre-

Table 42.5. Differences between Workload Levels as Discriminated by the Modified Cooper-Harper Scale and Two Secondary Tasks

	Load Level		
	Low Versus Medium		
Workload Measure	Perceptual[a]	Central Processing[b]	Communications[c]
Modified Cooper-Harper	X		X
Time estimation		X	
Interval production task			X

	Load Level		
	Medium Versus High		
Workload Measure	Perceptual[a]	Central Processing[b]	Communications[c]
Modified Cooper-Harper	X	X	
Time estimation	X		
Interval production task	X		

	Load Level		
	Low Versus High		
Workload Measure	Perceptual[a]	Central Processing[b]	Communications[c]
Modified Cooper-Harper	X	X	X
Time estimation	X	X	
Interval production task	X		X

The sensitivity of a modified version (Wierwille & Casali, 1983b) of the Cooper-Harper scale was evaluated in a series of three flight simulator experiments. A different type of loading (perceptual, central processing, and communications) was manipulated in each experiment. Three levels of loading (low, medium, high) were also employed in each experiment. Several measures of workload were used, including the modified Cooper-Harper scale and the secondary tasks (see Section 4) of time estimation (Hart, 1975) and interval production (Michon, 1966). This table shows the results obtained with each of these three metrics and indicates significant loading differences ($p < .05$) that were discriminated for each type of difficulty manipulation. As is clear from the table, the modified Cooper-Harper scale discriminated a larger number of differences than did either of the secondary task measures and failed to discriminate only one of the differences demonstrated by secondary task performance. The sensitivity of the modified Cooper-Harper scale to the types and levels of difficulty manipulations employed across the three experiments is, therefore, supported by the data. (Table based on the data of Casali & Wierwille, 1982, 1983; and Rahimi & Wierwille, 1982.)

X Denotes significant difference ($p < .05$) between loading levels as demonstrated with either the Newman-Keuls or Duncan post-hoc multiple comparison test.

[a] Casali and Wierwille, 1982.

[b] Rahimi and Wierwille, 1982.

[c] Casali and Wierwille, 1983.

hension. The scale used was symmetrical, with verbal labels associated with each of the nine categories of difficulty. Figure 42.4 (Bratfisch et al., 1972) illustrates a set of relationships between the difficulty estimates and an objective measure of difficulty provided by the sequence of items in the standardized intelligence test.

Spearman rank-order coefficients of correlation between perceived difficulty estimates and item sequence are shown in the figure. The scale produced difficulty ratings that were highly correlated with the objective index of difficulty in each task. Additional data bearing on the sensitivity of this scale are provided by Hallsten and Borg (1975), who obtained difficulty ratings on a similar nine-point scale for a number of spatial ability intelligence test items of known difficulty. Difficulty ratings were highly correlated ($r = -.81$) with the frequency with which items were solved, which served as the objective index of difficulty.

Available sensitivity data, therefore, support the scale, although it is important to note that the sensitivity results were obtained with judgments of perceived task difficulty. There are data (Dornic & Andersson, 1980) to indicate that, in some instances, perceived difficulty ratings differ from ratings of perceived effort expenditure in information-processing tasks. Therefore, some caution must be exercised in interpreting perceived difficulty ratings as direct indicators of operator effort or workload.

Diagnosticity of the nine-point scale cannot be addressed on the basis of current data, and most work to date has been with intelligence test items. Instrumentation for the scale is minimal, and there is no indication that extensive practice was required to familiarize subjects with the scale.

Another University of Stockholm scale (Dornic, 1980a, 1980b; Dornic & Andersson, 1980) derives judgments of the effort expenditure associated with task performance. It is a graphic scale anchored at the extremes by 0 and 10 and also by verbal labels.

Although the scale has not yet been used extensively, available sensitivity data have been favorable. Ratings have shown a monotonic relationship with difficulty levels in a visual discrimination task (perceptual demand) and in a letter-transformation task (central-processing demand) (Dornic, 1980a). An inverse relationship has also been demonstrated between effort ratings on a primary digit transformation task and performance levels on both visual and auditory variants of a secondary target detection task (see Figure 42.5). Thus the scale appears capable of providing an index of the spare capacity afforded by primary task performance (Region A of Figure 42.1).

As with the perceived difficulty scale, the available evidence supports the sensitivity of the effort scale, but a more substantial data base is required before definitive conclusions can be drawn. In addition, since the work conducted to date has been with tasks that emphasized perceptual or central-processing components, the diagnosticity of the scale cannot be evaluated adequately. Implementation requirements associated with the scale are not extensive, and it is less time-consuming than paired-comparison methods (Dornic & Andersson, 1980).

2.3. Psychometric Techniques

Psychometric techniques employed in workload scale development include magnitude estimation (e.g., Borg, 1978; Helm & Heimstra, 1981); paired comparisons (e.g., Daryanian, 1980; Wolfe, 1978); the method of equal-appearing intervals (e.g., Hicks & Wierwille, 1979); and conjoint measurement and scaling (e.g., Donnell, Adelman, & Patterson, 1981; Reid, Shingledecker, & Eggemeier, 1981). Given that certain assumptions are met, each method can produce interval-scaled data. The interval information provided by such scales can represent an advantage

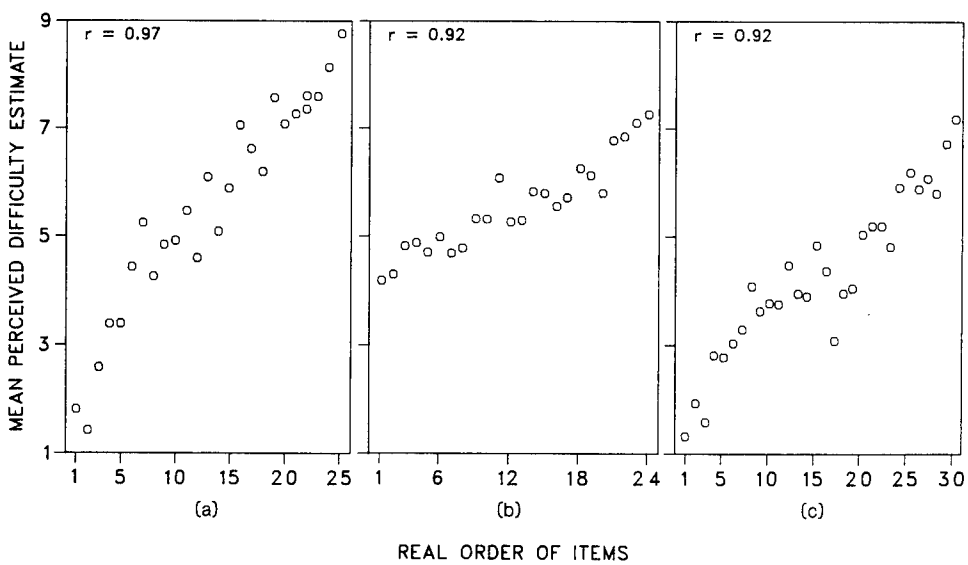


Figure 42.4. Means of perceived difficulty estimates plotted against the real order of items in a standardized intelligence test. Graphs (a–c) Data from tests of reasoning ability, spatial ability, and verbal comprehension, respectively. Item sequence served as an objective measure of task difficulty, with items arranged in increasing order of difficulty. It is clear that a monotonic relationship was obtained in each instance between estimates of difficulty and the position of the item in the test sequence. Rank-order correlation coefficients shown in each part also support a strong relationship between perceived and objective difficulty in each type of test. (Redrawn from Bratfisch, Borg, & Dornic, 1972.)

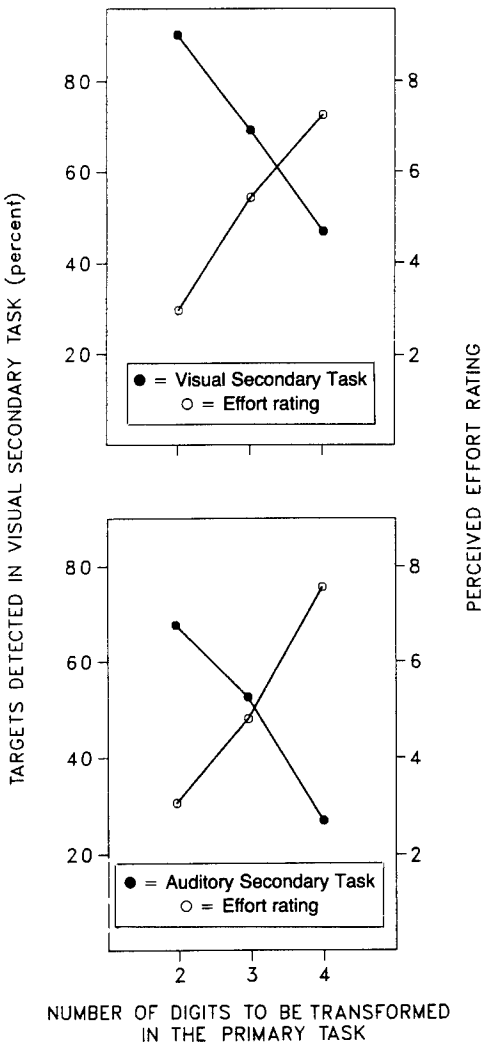


Figure 42.5. Performance on a secondary target detection task and perceived effort ratings as a function of primary task difficulty in two experiments. The primary task in both instances was a digit transformation task requiring subjects to make continuous mental transformations of a group of digits by adding one to each digit (e.g., 63 → 74) and then adding one to each of the resulting digits (e.g., 74 → 85), and so on. Difficulty was manipulated by varying the number of digits to be transformed (two versus three versus four). The secondary task in each instance required the subject to detect certain pairs of letters in a series of letter pairs. In one instance the pairs were presented visually and required a manual response; in the second instance the pairs were presented auditorily and required a verbal response. Secondary task performance declined with increases in primary task difficulty in both experiments, while effort ratings increased. The inverse relationship between the measure of spare capacity represented by the secondary task and perceived effort ratings supports the sensitivity of the ratings to manipulations of primary task workload. (Redrawn from Dornic, 1980b.)

over the ordinal information afforded by other scales and can facilitate interpretation of results by providing some information regarding the magnitude of workload differences between design options or tasks. Interval data also permit use of parametric data analysis procedures which afford such advantages as examining the interactive effects of variables on subjective ratings of load. Representative applications of psychometric techniques to workload scale development are reviewed in the following sections.

2.3.1. Magnitude Estimation. There is a sizable literature (e.g., Stevens, 1975) on the successful application of magnitude estimation to sensory scaling, and the technique is the most commonly used method of direct ratio scaling in psychophysical investigations (e.g., D'Amato, 1970; Falmagne, Chapter 1). In magnitude estimation, a subject makes direct numerical estimates of the magnitude of the experience produced by a particular stimulus situation. Application of the technique to task difficulty judgments requires that subjects perform a task and make a numerical estimate of its difficulty. There are two major magnitude estimation procedures (Stevens, 1958). In the first a standard stimulus is presented to a subject who is told that the sensation produced by the stimulus has a certain numerical value (e.g., 10) termed the modulus. Subsequent stimuli are assigned numerical values by the subject relative to the modulus. For example, if a stimulus has one-half the apparent magnitude of the modulus, a 5 would be assigned, whereas a 20 would be the response if the apparent magnitude were twice that of the modulus. In the second version of the method, no experimenter-defined modulus is presented. Subjects choose their own modulus and assign numbers to other stimuli in relation to it.

The most extensive application of magnitude estimation to task difficulty scaling has been conducted by the University of Stockholm group (e.g., Borg, 1978; Bratfisch, 1972). This work has generally employed the experimenter-defined modulus version of magnitude estimation. In these instances a task is given and the subject instructed to consider its difficulty level as 10. The remaining tasks are then rated by subjects in relation to the standard task. In some applications (e.g., Borg, Bratfisch, & Dornic, 1971a, 1971b), the standard task has been presented for comparison purposes on every trial, whereas in the remaining examples the standard task appears to have been presented only at the initiation of the sequence.

Current data indicate that the magnitude estimation technique provides sensitive estimates of perceived difficulty. Applications of the technique have generally resulted in monotonic and, in some instances, linear relationships between objectively defined task difficulty levels and perceived difficulty estimates. Figure 42.6 illustrates several examples of relationships between measures of objective task difficulty and perceived difficulty estimates derived through magnitude estimation.

Correlations reported between perceived and objective difficulty (see Table 42.6) have typically been high and in the expected direction. This was true over a large number of task characteristics manipulated to change the objective difficulty (e.g., number of stimuli and time limits). In each case the magnitude estimation technique has provided a monotonic index that has proven sensitive to some aspect of objective difficulty.

Intrarater reliability of the magnitude estimation technique using the experimenter-defined modulus has also been examined by Hallsten and Borg (1975) with standardized intelligence test items. Reliability coefficients were quite high and ranged from .62 to .98 for nine subjects, with a median of .93.

Helm and Heimstra (1981) also used a variant of magnitude estimation to assess differences in task difficulty in a series of information-processing tasks (e.g., visual discrimination task, Sternberg memory scanning task). In addition to magnitude estimation, a category scale was also used to gather task difficulty ratings, and the two procedures were compared with respect to their capability to reflect task performance levels. Results of both the magnitude and category scales were highly and significantly correlated with performance error on all tasks employed in the experiment. Correlation coefficients ranged from .85 to .96 for the category scale, and from .93 to .97 for the ratio estimation scale. The results of both rating techniques and the

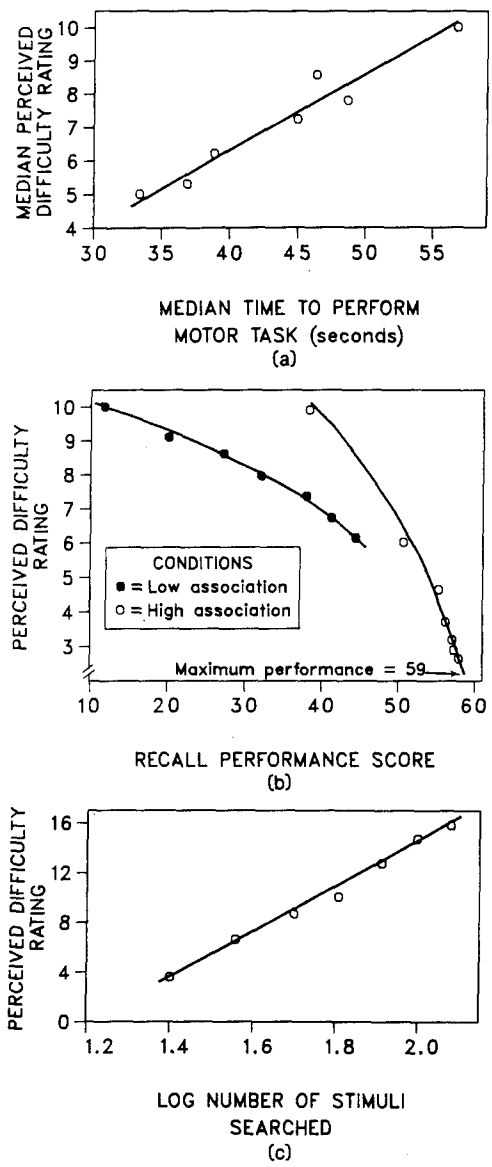


Figure 42.6. Perceived difficulty estimates obtained through magnitude estimation as a function of several variables in three experiments. Graph (a) The relationship between median perceived difficulty and median time to perform a motor task requiring the transfer of metal objects through a wire labyrinth (Bratfisch, Dornic, & Borg, 1970). The data points represent seven repetitions of the task and associated decreases in time to perform and estimate difficulty. The high correlation (r = .96) between difficulty estimates and the objective measure of performance time suggests that the latter represented a possible factor which influenced perceived difficulty. Graph (b) The relationship between perceived difficulty ratings and a combined recall performance score (based on omissions and displacement of order in recall of lists of words) in two difficult conditions in a word list acquisition task (Dornic, Bratfisch, & Larsson, 1973). The function at the left in the figure represents three difficulty conditions in which the between-word association was low; the function at the right represents three less difficult conditions with high between-word associations. Each point in each function represents one of seven repetitions of the difficult or easy conditions. The easy conditions were generally rated as less difficult than the more difficult conditions. Graph (c) The relationship between perceived difficulty estimates and the log number of stimuli presented in a visual target search task. (Borg, Bratfisch, & Dornic, 1971b). As is clear from the figure, perceived difficulty estimates increased systematically with increases in the log number of stimuli. The magnitude estimation technique has, therefore, provided indices of perceived difficulty that have proven to be at least montonically related to some aspect of objective difficulty in each case. (Redrawn from Borg, Bratfisch, & Dornic, 1971b; Bratfisch, Dornic, & Borg, 1970; Dornic, Bratfisch, & Larsson, 1973.)

percent error in task performance were monotonically related to task difficulty. The magnitude scale, however, showed a closer correspondence with performance at low to moderate levels of difficulty. Jenny, Older, and Cameron (1972) have also successfully applied magnitude estimation to workload assessment.

The data, therefore, support the sensitivity of estimates that result from application of magnitude estimation techniques. Complete data are not available on the diagnosticity that can be expected with the technique, but magnitude estimation has proven sensitivity to a range of perceptual, information-processing, motor, and communications functions. In view of this range of sensitivity, it appears that this technique can provide estimates of loading throughout the human processing system.

There are, however, implementation requirements that could potentially limit the practical use of magnitude estimation techniques. One difficulty is that tasks whose workload must be estimated do not always occur in close temporal proximity. Problems could arise with subjects retaining and effectively using a modulus over extended time periods. If the modulus is presented with every condition, this difficulty could be avoided. However, repeated presentation and performance of the same task in an operational environment might not be possible. The requirements for counterbalanced stimulus presentation orders normally implemented in magnitude estimation experimentation would also pose potential difficulties for application to operational environments. Before recommending the magnitude estimation techniques for extensive use beyond the laboratory environment, it is desirable that the impact of these methodological problems on the practicality of conducting workload assessments in actual systems contexts be better defined.

Table 42.6. Correlations between Perceived Difficulty Estimates Derived through Magnitude Estimation and Measures of Objective Task Difficulty

References	Measure(s) of Objective Task Difficulty	Correlation between Perceived and Objective Task Difficulty
Bratfisch, Dornic, and Borg (1970)	Time to complete a motor skill task	.96
Bratfisch, Borg, and Dornic (1972)	Sequence of items in a standardized intelligence task	.90
Dornic, Bratfisch, and Larsson (1973)	Number of correctly recalled words in learning task	−.98
Dornic, Sarnecki, and Svensson (1973)	Mean number of correct responses in a perceptual identification task	−.59
Hallsten and Borg (1975)	Solution frequencies of items in a standardized intelligence task	−.78

The relationship of perceived difficulty estimates based on magnitude estimation and objective measures of task difficulty has been investigated in a series of experiments utilizing a variety of tasks. In all instances, application of magnitude estimation resulted in at least monotonic relationships between objectively defined levels of task difficulty (e.g., number of correct responses) and perceived difficulty estimates. This table illustrates correlations obtained in several of the experiments between the measure of task difficulty and the perceived difficulty estimate. As is clear from the table, the correlations have been consistently high and in the expected direction, thereby supporting the capability of the difficulty estimates to reflect objective levels of task difficulty.

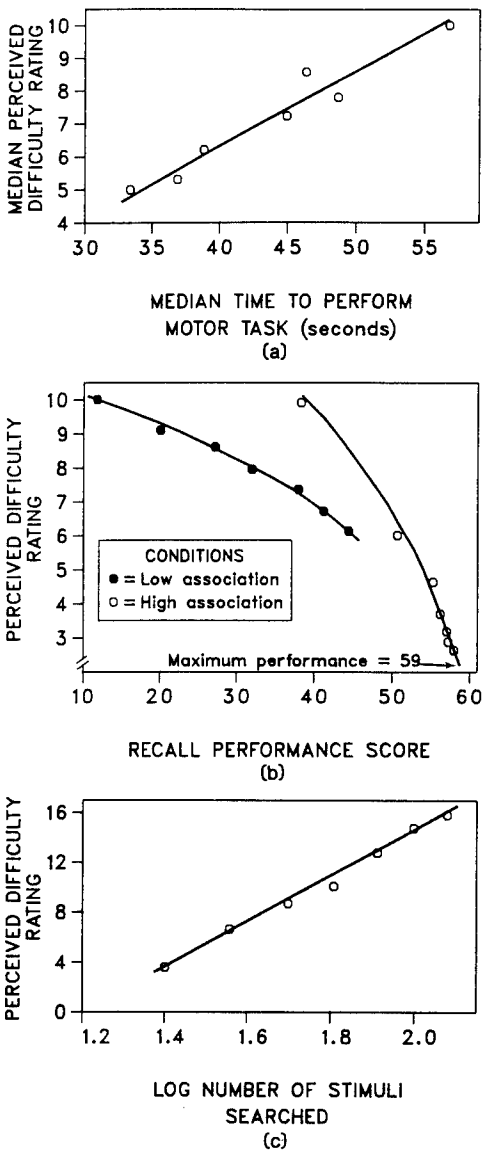


Figure 42.6. Perceived difficulty estimates obtained through magnitude estimation as a function of several variables in three experiments. Graph (a) The relationship between median perceived difficulty and median time to perform a motor task requiring the transfer of metal objects through a wire labyrinth (Bratfisch, Dornic, & Borg, 1970). The data points represent seven repetitions of the task and associated decreases in time to perform and estimate difficulty. The high correlation (r = .96) between difficulty estimates and the objective measure of performance time suggests that the latter represented a possible factor which influenced perceived difficulty. Graph (b) The relationship between perceived difficulty ratings and a combined recall performance score (based on omissions and displacement of order in recall of lists of words) in two difficult conditions in a word list acquisition task (Dornic, Bratfisch, & Larsson, 1973). The function at the left in the figure represents three difficulty conditions in which the between-word association was low; the function at the right represents three less difficult conditions with high between-word associations. Each point in each function represents one of seven repetitions of the difficult or easy conditions. The easy conditions were generally rated as less difficult than the more difficult conditions. Graph (c) The relationship between perceived difficulty estimates and the log number of stimuli presented in a visual target search task. (Borg, Bratfisch, & Dornic, 1971b). As is clear from the figure, perceived difficulty estimates increased systematically with increases in the log number of stimuli. The magnitude estimation technique has, therefore, provided indices of perceived difficulty that have proven to be at least montonically related to some aspect of objective difficulty in each case. (Redrawn from Borg, Bratfisch, & Dornic, 1971b; Bratfisch, Dornic, & Borg, 1970; Dornic, Bratfisch, & Larsson, 1973.)

percent error in task performance were monotonically related to task difficulty. The magnitude scale, however, showed a closer correspondence with performance at low to moderate levels of difficulty. Jenny, Older, and Cameron (1972) have also successfully applied magnitude estimation to workload assessment.

The data, therefore, support the sensitivity of estimates that result from application of magnitude estimation techniques. Complete data are not available on the diagnosticity that can be expected with the technique, but magnitude estimation has proven sensitivity to a range of perceptual, information-processing, motor, and communications functions. In view of this range of sensitivity, it appears that this technique can provide estimates of loading throughout the human processing system.

There are, however, implementation requirements that could potentially limit the practical use of magnitude estimation techniques. One difficulty is that tasks whose workload must be estimated do not always occur in close temporal proximity. Problems could arise with subjects retaining and effectively using a modulus over extended time periods. If the modulus is presented with every condition, this difficulty could be avoided. However, repeated presentation and performance of the same task in an operational environment might not be possible. The requirements for counterbalanced stimulus presentation orders normally implemented in magnitude estimation experimentation would also pose potential difficulties for application to operational environments. Before recommending the magnitude estimation techniques for extensive use beyond the laboratory environment, it is desirable that the impact of these methodological problems on the practicality of conducting workload assessments in actual systems contexts be better defined.

Table 42.6. Correlations between Perceived Difficulty Estimates Derived through Magnitude Estimation and Measures of Objective Task Difficulty

References	Measure(s) of Objective Task Difficulty	Correlation between Perceived and Objective Task Difficulty
Bratfisch, Dornic, and Borg (1970)	Time to complete a motor skill task	.96
Bratfisch, Borg, and Dornic (1972)	Sequence of items in a standardized intelligence task	.90
Dornic, Bratfisch, and Larsson (1973)	Number of correctly recalled words in learning task	−.98
Dornic, Sarnecki, and Svensson (1973)	Mean number of correct responses in a perceptual identification task	−.59
Hallsten and Borg (1975)	Solution frequencies of items in a standardized intelligence task	−.78

The relationship of perceived difficulty estimates based on magnitude estimation and objective measures of task difficulty has been investigated in a series of experiments utilizing a variety of tasks. In all instances, application of magnitude estimation resulted in at least monotonic relationships between objectively defined levels of task difficulty (e.g., number of correct responses) and perceived difficulty estimates. This table illustrates correlations obtained in several of the experiments between the measure of task difficulty and the perceived difficulty estimate. As is clear from the table, the correlations have been consistently high and in the expected direction, thereby supporting the capability of the difficulty estimates to reflect objective levels of task difficulty.

2.3.2. Methods of Paired Comparisons and Equal-Appearing Intervals. In addition to magnitude estimation techniques, paired comparisons and equal-appearing intervals are psychometric procedures that can be applied to workload scale development. Application of the method of paired comparisons (e.g., Edwards, 1957) involves presentation of pairs of stimuli that vary in some defined criterion attribute. Subjects must indicate which member of the pair under consideration possesses the greater degree of the criterion attribute (e.g., mental workload). All possible pairings of the stimuli to be scaled must be presented to subjects. The number of pairs is equal to $n(n - 1)/2$, where n is the number of stimuli to be judged. Therefore, with ten stimuli to be judged, the number of comparison pairs would equal 45. The number of comparison pairs rises substantially with the addition of stimuli (e.g., 30 stimuli would require 435 pairs). The scale value of the stimulus is derived from the proportion of times that a stimulus is judged to possess more of the criterion attribute than the other stimuli with which it has been paired. When several subjects have completed the paired comparison procedure, an $n \times n$ matrix can be derived to show the proportion of times that each stimulus was judged higher than every other stimulus on the criterion attribute.

There are a number of instances in which paired comparisons have been applied to workload scale development. Wolfe (1978), for example, used the procedure to develop a workload measure for instrument landing approaches in a flight simulator. Six different difficulty levels were achieved through manipulation of wind gust levels and the use of a nominal versus a degraded flight control system. Wolfe reported correlations of the paired-comparisons results with a modified Cooper-Harper scale (see Section 2.2.1) and several performance metrics. Results of the paired-comparison procedure were highly correlated with several performance indices ($r = .82$ with a performance discriminant function including primary and secondary task data), and with ratings from the modified Cooper-Harper scale ($r = .71$). A second successful application of the paired-comparison technique was reported by Daryanian (1980), who used the procedure to scale mental workload in a multielement decision-making laboratory task. Subjects performed and rated 27 different decision-making conditions that varied in difficulty. Stimulus presentation rate proved to be the most potent variable in determining subjective workload scale values.

Although the limited applications of the paired-comparison procedure have produced successful results, more data are needed before definitive conclusions regarding the sensitivity and diagnosticity of the technique can be drawn. The major drawbacks to the technique are the requirement that pairs of tasks be presented for comparison and the dramatic increase in the number of comparisons that must be completed as the number of tasks to be scaled rises. Use of the technique might, therefore, be eventually limited to laboratory environments or simulation experiments where the number of tasks is not large, the situation permits a high degree of control over task sequencing, and where individual tasks are not time-consuming.

An alternative to the paired-comparisons method which is useful when a large number of stimuli or tasks are involved is the method of equal-appearing intervals (e.g., Edwards, 1957). Typically, a group of stimuli or statements is presented to subjects, who must assign each stimulus to one of several categories according to the degree of a criterion attribute (e.g., workload) that it possesses. Labels are usually included for the extreme categories, with a center "neutral" point, and 11 points or categories are frequently used. Other than these anchors, points are left unlabeled so that the intervals between them are free to represent equal-appearing intervals or degrees of the criterion attribute for each subject. It is important to note that subjects are instructed to keep the distance between any two categories equal to that between any other two. To the extent this can be done, an 11-point scale can be created with interval-scale qualities.

Hicks and Wierwille (1979) applied the method of equal-appearing intervals to develop subjective measures of load in an automobile simulator. Difficulty of the automobile driving task was manipulated through application of crosswind gusts to the simulated vehicle. The rating scale consisted of 11 categories with a normal density function drawn above the categories. The subjective scale generated by the equal-appearing interval data proved quite sensitive. It yielded significant differences among all task difficulty levels and rated third among seven assessment techniques used on a relative sensitivity index that was constructed.

2.3.3. Conjoint Measurement and Scaling. All the psychometric techniques discussed thus far are considered unidimensional in that subjective workload or perceived difficulty is treated as a unitary construct by subjects completing their ratings. Some of the scales (e.g., Cooper-Harper) include references to several factors such as task difficulty and pilot compensation, but the subject must assign a single rating to characterize the combined effects of these factors. In addition to such procedures, there are multidimensional techniques with the advantage of reflecting several factors that can contribute to the experience of subjective mental load. The multidimensional procedure which has been applied specifically to the development of workload scales is the technique of conjoint measurement (e.g., Coombs, Dawes, & Tversky, 1970; Krantz & Tversky, 1971; Nygren, 1982; Tversky & Krantz, 1969).

Basically, the use of conjoint measurement involves taking separate ordinal ratings on a set of two or more dimensions and combining them into a one-dimensional scale with interval properties. In the case of workload measurement, the separate ordinal scales reflect dimensions (e.g., time stress, mental effort) that can be assumed to contribute to subjective mental load. Application of the conjoint measurement procedure involves two major phases: (1) scale development, and (2) event scoring.

During the scale development phase, the information necessary to combine the individual ordinal scales into one overall interval scale is generated. Typically this information is developed in several steps. First, levels of each dimension are described to the subject. Any number of levels may be used, but three to five are common. Then all possible groupings of levels and dimensions are combined into separate descriptions of the criterion factor. For instance, the highest level of time stress load is combined with each level of mental effort load, and so on. Thus if there are three levels of three dimensions, there are 27 possible composite descriptions of the criterion. The second step involves having a subject rank order these composites, from the description that the subject considers to possess the "most" of the criterion factor (e.g., workload) to that possessing the "least."

The third step involves submitting the subject's rankings on these composites to a series of axiom tests specified by the conjoint measurement procedure (e.g., Krantz &Tversky, 1971; Nygren, 1982). These axiom tests establish that certain logical consistencies expected in the data actually exist. If these consistencies are verified, the conjoint procedures identify the com-

bination rule or model (e.g., additive, distributive, dual distributive) that fits the ordered data. This model is then used iteratively to assign numerical values to each level of the separate scales for each dimension and to generate the single integrated scale. The criterion for an optimal scale is the assignment of values that best preserves the original ordering of the subject's ranks. The resulting best fit values yield a single score for each combination which, if all axiom tests are satisfied, has interval scale properties. In this way subjects are permitted to generate individualized scales reflecting their subjective combinations of dimensions making up the criterion factor.

During the subsequent or event-scoring phase, actual tasks or systems functions are performed and then rated by the subjects on each of the individual dimensions. The set of ordinal ratings on each of the dimensions can then be used to determine a corresponding value on the single interval scale developed during the scale development phase.

2.3.3.1. Mission Operability Assessment Technique. Conjoint measurement procedures have been applied to the measurement of workload and overall system operability in several aircraft environments (Donnell, 1979; Donnell, Adelman, & Patterson, 1981; Donnell & O'Connor, 1978). The specific procedure employed in these studies was the Mission Operability Assessment Technique (Donnell et al., 1981; Helm & Donnell, 1979). In this technique a number of factors, including pilot workload and system technical effectiveness, that is, the degree to which the system aids the operator in task accomplishment, are combined into an overall concept called "systems operability." Separate four-point ordinal rating scales for pilot workload and technical effectiveness have been developed (Table 42.7), resulting in a 16-element system operability matrix. A task analysis of the system under study is generated, and subjects estimate the pilot workload and technical effectiveness for elements of the system. The rank data generated during the scale development phase are subsequently used in a scaling program to develop an overall interval scale of systems operability.

Interrater ranking reliabilities during scale development have been high and statistically significant (Donnell, 1979; Donnell & O'Connor, 1978). On the other hand, ratings of particular tasks have produced interrater reliabilities that were statistically different from zero, but very low (Donnell et al., 1981). For each scale there were tasks on which pilot disagreement was quite substantial. This result can be expected if subjective workload and technical effectiveness estimates are affected by individual differences between subjects. The results did, however, lead Donnell et al. (1981) to strongly recommend the use of as many subjects as possible when implementing the Mission Operability Assessment Technique.

Some sensitivity data on the systems operability scale resulting from the combination of pilot workload/technical effectiveness ratings are reported in a study that manipulated psychomotor load by changes in aircraft pitch stability and random wind gust disturbance levels in a moving-base flight simulator (Wierwille & Connor, 1983). The results (Figure 42.7) indicated that the operability ratings demonstrated a monotonic relationship with manipulations of psychomotor load and also significantly discriminated each of the three workload levels employed in the experiment. In fact, the system operability ratings proved to be among the most sensitive of 20 different workload measures used.

The present evidence (Donnell, 1979; Donnell et al., 1981; Donnell & O'Connor, 1978; Wierwille & Connor, 1983), although limited, therefore supports the application of the conjoint measurement technique to the development of system operability estimates. As currently designed, the workload rating scale of the mission operability assessment technique is specifically dedicated to piloting tasks. However, only minor modifications to the scale would be necessary to extend its applicability. It should also be noted that this scale was not intended as a direct measure of workload. Although workload represents a prime factor in system operability as defined in the Mission Operability Assessment Technique, the specific relationship between workload and operability has not been specified. Therefore an investigator primarily interested in operator workload might wish to consider a more direct measure than that represented by this technique.

Table 42.7. Ordinal Rating Scales for Pilot Workload and Subsystem Technical Effectiveness That Are Included in the Systems Operability Measure of the Mission Operability Assessment Technique

Pilot Workload/Compensation/Interference: A measure of the degree of pilot workload/compensation/interference (mental and/or physical) required to perform a designated task.

Scale Values:

1. The pilot workload (PW)/compensation (C)/interference (I) required to perform the designated task is *extreme*. This is a *poor* rating on the PW/C/I dimension.
2. The pilot workload/compensation/interference required to perform the designated task is *high*. This is a *fair* rating on the PW/C/I dimension.
3. The pilot workload/compensation/interference required to perform the designated task is *moderate*. This is a *good* rating on the PW/C/I dimension.
4. The pilot workload/compensation/interference required to perform the designated task is *low*. This is an *excellent* rating on the PW/C/I dimension.

Subsystem Technical Effectiveness: A measure of the technical effectiveness of the subsystem(s) utilized in performing a designated task.

Scale Values:

1. The technical effectiveness of the required subsystem is *inadequate* for performing the designated task. Considerable redesign is necessary to attain task requirements. This is a *poor* rating on the subsystem technical effectiveness scale.
2. The technical effectiveness of the required subsystem is *adequate* for performing the designated task. Some redesign is necessary to attain task requirements. This is a *fair* rating on the subsystem technical effectiveness scale.
3. The technical effectiveness of the required subsystem *enhances individual task performance*. No redesign is necessary to attain task requirements. This is a *good* rating on the subsystem technical effectiveness scale.
4. The technical effectiveness of the required subsystem *allows for the integration of multiple tasks*. No redesign is necessary to attain task requirements. This is an *excellent* rating on the subsystem effectiveness scale.

Systems operability in the mission operability assessment technique (Donnell, Adelman, & Patterson, 1981; Helm & Donnell, 1979) includes the factors of pilot workload and technical effectiveness, both of which are represented by the four-point ordinal scales illustrated in the table. The mission operability assessment technique involves application of conjoint measurement and appropriate scaling techniques permitting the ordinal ratings on the pilot workload and technical effectiveness scales to be combined into one overall interval scale of systems operability. See text for further details.

Source: (From Donnell, M. L., Adelman, L., & Patterson, J. F. A systems operability measurement algorithm (SOMA): Application, validation, and extensions *(Report No. TR-81-11-156)*. McLean, Va., 1981. Reprinted with permission.)

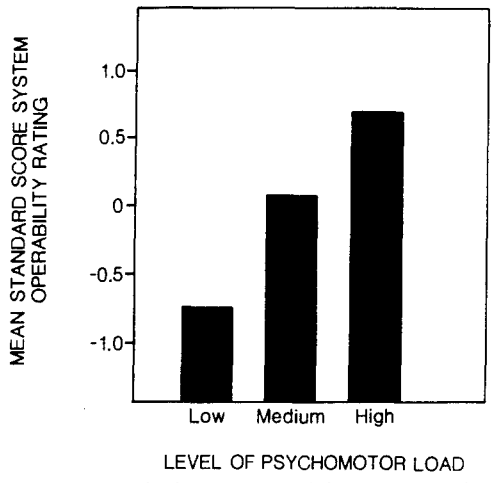


Figure 42.7. Mean standard system operability scores resulting from the combination of pilot workload (PW) and technical effectiveness (TE) scales (Donnell, 1979) as a function of psychomotor load. Three levels of psychomotor load were achieved through manipulation of wind gust disturbance level and aircraft pitch stability in a flight simulator. Standardized system operability scores were significantly affected ($p < .0001$) by the load manipulation. Post hoc multiple comparisons tests showed that all standardized ratings differed from one another, thereby supporting the sensitivity of the systems operability scale to differences in psychomotor loading. (From W. W. Wierwille & S. A. Connor, Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator, *Human Factors, 25*. Copyright 1983 by Human Factors Society. Reprinted with permission.)

2.3.3.2. Subjective Workload Assessment Technique. Conjoint measurement techniques have also been applied in the development of a rating scale specifically designed for workload assessment, the Subjective Workload Assessment Technique (SWAT) (e.g., Reid, Eggemeier, & Shingledecker, 1982; Reid, Shingledecker, & Eggemeier, 1981; Reid, Shingledecker, Nygren, & Eggemeier, 1981).

In SWAT, subjective workload is defined as being primarily composed of three dimensions: time load, mental effort load, and stress load. These dimensions are adaptations of factors proposed as major contributors to subjective workload by Sheridan and Simpson (1979) and other theorists (e.g., Jahns, 1973; Johannsen et al., 1979; Kahneman, 1973; Moray, 1982). Each dimension is represented in SWAT by an individual three-point rating scale with verbal descriptors (Table 42.8).

Interrater reliabilities determined by the Kendall coefficient of concordance for scale development rank orderings have ranged from $W = 0.68$, $p < .01$ (Eggemeier et al., 1983) to $W = 0.87$, $p < .01$ (Reid, Shingledecker, & Eggemeier, 1981). Therefore in all studies reported to date, levels of agreement regarding the amount of workload imposed by the various combinations of time, effort, and stress load have been reasonably high and statistically significant, permitting use of a single overall group scale instead of individual scales for each subject.

The sensitivity of ratings gathered during the event-scoring phase (Section 2.3.3) of SWAT has been demonstrated in a variety of different tasks, including central processing, motor output, and communications (see Figure 42.8). Graph (a) (Reid, Shingledecker, & Eggemeier, 1981) shows the results of an experiment that employed two levels of difficulty in a primary critical tracking (motor output) task (e.g., Jex & Clement, 1979; see Section 4.4.2.1) and a secondary radio communications task (Shingledecker, Crabtree, Simons, Courtright, & O'Donnell, 1980; see Section 4.4.2.2). SWAT ratings successfully discriminated a condition in which the communications task was performed alone versus a more difficult dual task condition, and also discriminated levels of difficulty in the tracking task. Graph (b) (Eggemeier et al., 1982) illustrates the effects of variations in interstimulus interval and number of information categories to be recalled in a short-term memory (central-processing) task. The memory task required that subjects monitor a display for the occurrence of each category of information, and mentally tabulate the number of presentations per category. Both interstimulus interval and number of memory categories significantly ($p < .01$) affected SWAT ratings. Notestine (1983) also demonstrated the sensitivity of SWAT ratings to variations in the difficulty of a visual display monitoring (perceptual) task. Ratings in a condition involving easy signal detection (indicator biased to the left or right of a center line 85% of the time) were significantly lower than in a more difficult detection (indicator biased 75% of the time) condition. It has also been demonstrated that SWAT is sensitive to workload variations (e.g., the presence or absence of threats to an aircraft) in high fidelity flight simulations (Skelly, Reid, & Wilson, 1983).

Additional data on SWAT sensitivity are provided by comparisons between SWAT ratings and other measures of operator load. For example, SWAT ratings in the Reid, Shingledecker,

Table 42.8. Three-Point Rating Scales for the Time, Mental Effort, and Stress Load Dimensions of the Subjective Workload Assessment Technique (SWAT)

Time Load:
1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.
2. Occasionally have spare time. Interruptions or overlap among activities occur frequently.
3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.

Mental Effort Load:
1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.
2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.
3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.

Stress Load:
1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.
2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.
3. High to very intense stress due to confusion, frustration, or anxiety. High to extreme determination and self-control required.

In SWAT (Reid, Shingledecker, & Eggemeier, 1981; Reid, Shingledecker, Nygren, & Eggemeier, 1981), subjective workload is defined as being primarily composed of three dimensions: time load, mental effort load, and stress load. The three dimensions are represented by the three-point ordinal rating scales illustrated in the table. In applications of SWAT, an operator performs the task(s) of interest and provides separate three-point ratings on each dimension. SWAT is based on conjoint measurement and scaling which permit operator ratings on the three separate dimensions to be converted into one overall interval scale of workload. See text for additional details. (From Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting. *Copyright 1981 by Human Factors Society.)*

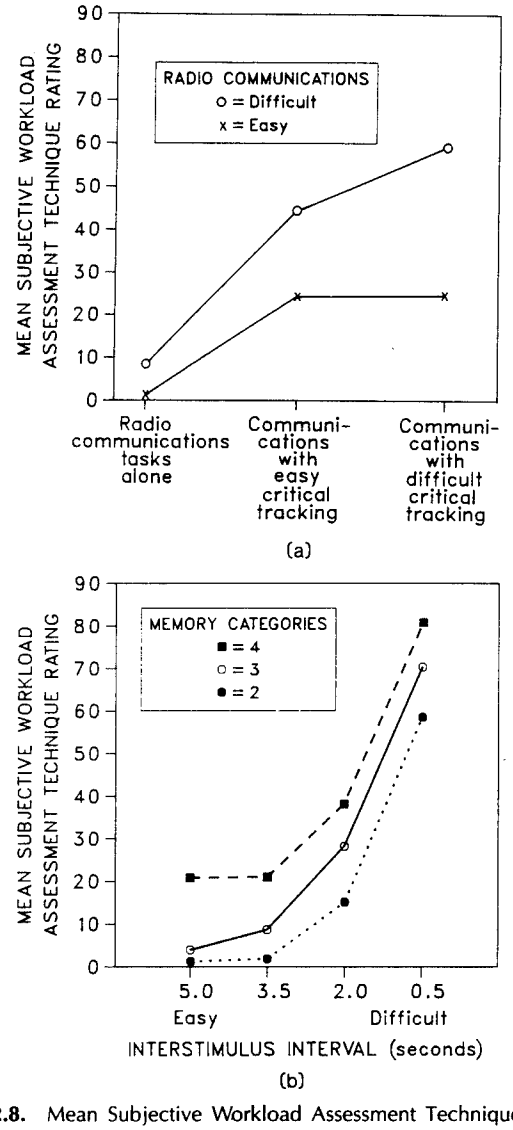


Figure 42.8. Mean Subjective Workload Assessment Technique (SWAT) ratings as a function of task difficulty in several different types of tasks. (a) The effects of two levels of primary task tracking difficulty with a simple and difficult version of a secondary aircrew radio communications task (Reid, Shingledecker, & Eggemeier, 1981). Both radio communications condition and tracking-task difficulty significantly affected ($p < .01$) mean SWAT ratings. Post hoc multiple comparisons tests indicated that low-difficulty tracking ratings were significantly different from those associated with high-difficulty tracking ($p < .01$), and that ratings from the single-task conditions were lower than ratings from the dual-task conditions. SWAT ratings, therefore, distinguished levels of difficulty in the tracking task and reflected the additional demands imposed by the more difficult dual task condition. (b) The effects of variations in memory task difficulty on mean SWAT ratings (Eggemeier, Crabtree, Zingg, Reid, & Shingledecker, 1982). Difficulty of the memory task was manipulated by varying the number of information categories (two, three, or four) that were to be retained in memory and the interstimulus interval (0.5, 2.0, 3.5, and 5.0 seconds). Both manipulations produced significant effects ($p < .01$) on SWAT ratings, thereby supporting the sensitivity of the procedure to difficulty manipulations in this task. Taken together, the results of Graphs (a) and (b) support the capability of the SWAT technique to discriminate workload differences in both central-processing and motor output tasks. (Redrawn from G. B. Reid, C. A. Shingledecker, & F. T. Eggemeier. Application of conjoint measurement to workload scale development. *Proceedings of the Human Factors Society 25th Annual Meeting*. Copyright 1981 by Human Factors Society. And redrawn from F. T. Eggemeier, M. S. Crabtree, J. J. Zingg, G. B. Reid, & C. A. Shingledecker, Subjective workload assessment in a memory update task. *Proceedings of the Human Factors Society 26th Annual Meeting*. Copyright 1982 by Human Factors Society. Reprinted with permission.)

and Eggemeier (1981) experiment were significantly related to performance scores obtained on the secondary communications task ($r = .76$, $p < .01$). Eggemeier et al. (1982) also compared the sensitivity of SWAT ratings with memory task errors in the short-term memory task described previously. Figure 42.9 (Eggemeier et al., 1982) depicts normalized error scores versus normalized SWAT ratings as a function of variations in interstimulus interval within three memory category sizes. It is clear that SWAT ratings varied more substantially than memory performance, and this is confirmed by slopes of the least squares regressions lines computed as descriptive indices of SWAT and memory error sensitivity. These slopes indicate that SWAT varied from six times more sensitive in the least difficult (two-category) condition to twice as sensitive in the most difficult (four-category) case. Similar results were obtained with variations of memory categories within interstimulus intervals, although memory error proved more sensitive than SWAT in the most difficult condition. This pattern of sensitivity is, of course, consistent with the rationale discussed earlier for use of subjective workload measures (Section 1.1). At lower levels of load (Region A of Figure 42.1), the subjective measure is capable of reflecting increased effort expenditure not demonstrated by the primary task measure of memory errors, and SWAT, therefore, demonstrates greater sensitivity. As load increases and degradations in primary task performance increase (Region B of Figure 42.1), the relative sensitivity of the primary task measure increases and, in one instance, actually exceeds that of SWAT.

Current results (e.g., Eggemeier et al., 1982; Reid et al., 1981) which support the sensitivity of SWAT to difficulty manipulations in a variety of task types suggest that the procedure is not diagnostic in the sense of distinguishing perceptual, central-processing load, and motor loading. Implementation requirements with SWAT are not extensive, and consist of the paper and pencil materials required with the other rating scales. The scale development phase does require approximately one hour of subject time to complete the required rank orderings, and analysis of the scale development data requires access to proper axiom testing and scaling programs. Descriptions of these programs and some initial work on development of a means of evaluating the fit of an additive conjoint measurement model to a three-factor design can be found in Nygren (1982, 1983).

2.4. Limitations of Subjective Techniques and Guidelines for Usage

Despite the advantages of subjective techniques, there are a number of important restrictions in their interpretation and several guidelines for their usage that should be considered in any application (e.g., Gartner & Murphy, 1976; Sanders, 1979; Sheridan & Simpson, 1979; Williges & Wierwille, 1979). These limitations generally refer to the potential influence on subjective estimates of (1) factors (e.g., confounding of mental and physical load) that can influence the degree of load actually experienced by the operator or (2) methodological constraints (e.g., delay in reporting workload ratings) that can influence the reported levels of load.

One potential limitation with interpretation of subjective measures is the possibility of confounding mental and physical load by the operator. Several theorists (e.g., Johannsen et al., 1979; Moray, 1982) have suggested that subjective feelings of load might be related to physiological activation. Since both mental and physical load an be related to activation, it appears very feasible that some confounding could occur. If an inves-

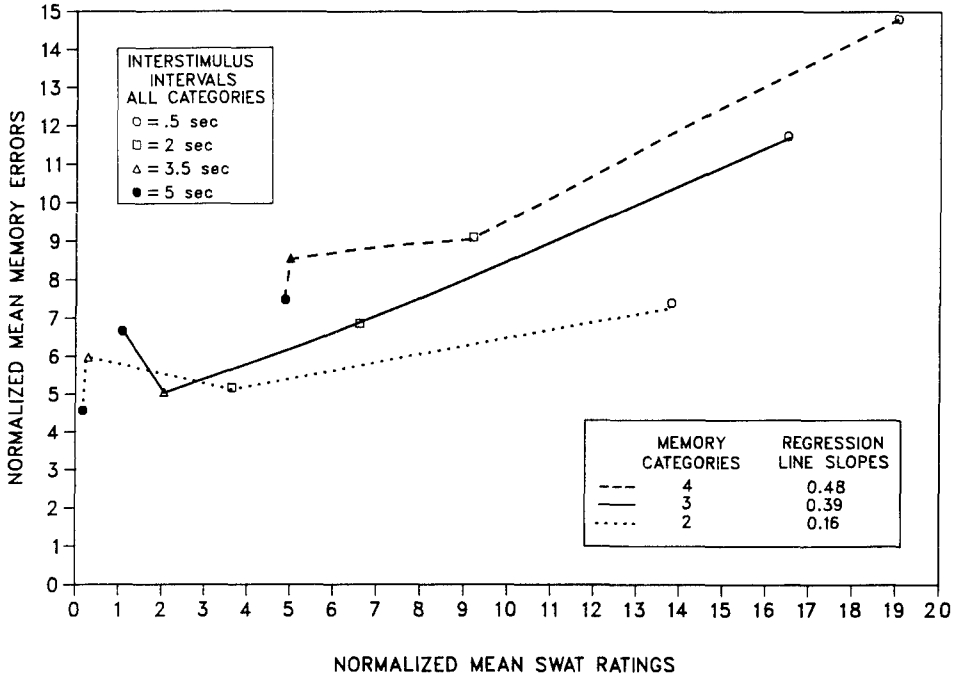


Figure 42.9. Normalized mean Subjective Workload Assessment Technique (SWAT) ratings and memory error as a function of number of memory categories and interstimulus interval. The functions represented depict the effect of variations in stimulus presentation rate (interstimulus intervals of 0.5, 2.0, 3.5, 5.0 sec) within each of three memory load levels (two, three, or four categories of information to be retained in memory) on both normalized memory errors (a primary task measure) and SWAT ratings. As is clear from the figure, SWAT ratings varied more substantially than errors as a function of interstimulus interval within each level of memory load. This sensitivity difference is confirmed by the slopes of the least-squares regression lines that were computed as descriptive indices of the relative sensitivity of memory errors and SWAT ratings to variations in interstimulus interval. The slopes depicted in the figure indicate that SWAT was approximately six times more sensitive in the two-category conditions and approximately twice as sensitive as primary task errors in the three- and four-category conditions. The results support the relative sensitivity of the SWAT ratings versus the primary task workload measure in the conditions studied. (Redrawn from F. T. Eggemeier, M. S. Crabtree, J. J. Zingg, G. B. Reid, & C. A. Shingledecker, Subjective workload assessment in a memory update task. *Proceedings of the Human Factors Society 26th Annual Meeting.* Copyright 1982 by Human Factors Society. Reprinted with permission.)

tigator is interested in an overall assessment of load, this is not a serious problem. However, if one wishes to make inferences about mental load or physical load, per se, the potential for confounding should be considered in interpreting results.

A second possible limitation associated with subjective measures is an inability of an operator to distinguish external demand or task difficulty (characteristics of the physical and mental tasks that must be performed) from the actual effort or workload experienced in dealing with these demands (Gartner & Murphy, 1976). Such confounding could result in biased estimates in which actual workload is over- or underestimated because the operator feels the task "should" require more or less work than may be actually experienced. There are data (Dornic & Andersson, 1980) that suggest that task demand or perceived difficulty does not always determine rated estimates of effort expenditure or workload. Dornic and Andersson reported results in which subjects rated both perceived difficulty and perceived effort expenditure in a series of six information-processing tasks. The rank orderings of the tasks on perceived effort differed appreciably from the orderings derived from the perceived difficulty ratings.

Although current evidence regarding the dissociation of task difficulty and effort is not extensive, the data do suggest that care should be exercised in choice of the wording on scales used to obtain subjective judgments. The researcher interested in subjective workload or perceived effort expenditure should be careful to clearly request that type of information from subjects (see, for example, the modified Cooper-Harper scale, Table 42.4 in Section 2.2.1) as opposed to system demand factors such as perceived task difficulty or vehicular-handling characteristics (see, for example, the Cooper-Harper scale, Table 42.3 in Section 2.2.1).

A third factor that could limit the general usefulness of subjective techniques deals with the nature of the relationship between actual capacity expenditure and the effort experienced by the operator. The assumption that increased capacity expenditure will be associated with subjective feelings of effort forms the theoretical basis for the sensitivity of subjective measures (Section 2.1). However, as pointed out by Gopher and Donchin (Chapter 41), it is probable that not all of the processing done by an individual is available to conscious introspection. When all processing is not open to introspection, the capability of the subjective technique to reflect capacity expenditure would be affected, thereby limiting the sensitivity of the measure.

A fourth restriction in interpretation of subjective estimates of load has been suggested by several investigators (Wickens & Derrick, 1981a; Wickens & Yeh, 1982, 1983) on the basis of some dissociations between ratings of mental workload and primary task performance in a series of information-processing and motor control tasks (e.g., tracking, memory search). Subjects

in the inferenced experiments performed a number of tasks under several conditions of difficulty (e.g., varied bandwidths in single-task tracking, the addition of a concurrent memory search task with tracking) and rated the task difficulty or mental load associated with each task configuration. The pattern of dissociation between the subjective ratings and task performance (Section 3) suggested that subjective measures were heavily influenced by such factors as the number of tasks or task elements to be performed, with relatively little regard for whether concurrent tasks required common processing resources (e.g., perceptual resources demanded in both tasks; see also Sections 1.2 and 4.4.3) or separate resources (e.g., perceptual resources demanded by one task, motor resources by the other task). Performance, on the other hand, was more substantially influenced by the requirement to perform concurrent tasks which shared common resources versus those which did not. This dissociation can be related to the criterion of diagnosticity (Section 1.2) and suggests that dual-task performance can provide a highly diagnostic measure of load, whereas subjective measures may be more globally sensitive to processing load anywhere within the human system. These dissociations led Wickens and Yeh (1983) to note that minimizing subjective ratings during system design could discourage the use of systems with multiple tasks competing for separate resources and encourage the use of those involving single tasks, even though the latter could produce relatively poorer performance in some instances. Such disociation patterns should be considered when interpreting the results of subjective assessments of load, and the results of subjective and primary task measures compared when such dissociations appear probable.

All of the restrictions outlined deal with factors that can potentially influence the degree of workload experienced by an operator. A second class of limitations related to subjective techniques deals with methodological factors than can influence the degree of load actually reported by an operator. One very important methodological guideline in the use of subjective techniques is to specifically determine what tasks or elements of system operation are to be rated by the operator. A primary consideration in the use of subjective techniques is the type of question that a subject can answer with reasonable confidence (Sanders, 1979). Sanders maintains that subjects should not be asked to make judgments about the loading associated with global levels of activities since workload is specific to particular task elements or conditions. Therefore, the use of subjective techniques should be limited to situations that involve clearly defined questions regarding the influence of specific variables on subjective load.

A related limitation of subjective techniques is their dependence on the short-term memory of the operator who completes the rating scales. By their very nature, subjective techniques require the operator to recall the level of subjective loading experienced during task performance. If the operator is required to remember several ratings, or if delays are introduced in the completion of task ratings, distortions may occur during relatively short-term intervals that intervene between performance of the task and completion of workload estimates. Unfortunately, requests for delayed ratings are quite common in simulation or operational environments where it is often maintained that the operator is too busy with subsequent task activity to complete a rating scale. Few data are available on the possible loss and distortion of rating scale information over retention intervals or on the effects of intervening task performance on subjective ratings. Although investigations to date (Eggemeier et al., 1983; Notestine, 1983) have not demonstrated differences in mean ratings as a result of delays, the current data are limited to the restricted number of laboratory situations that have been tested. Until more complete data are developed, the best guidance that can be provided is that workload ratings should be completed as soon after task performance as possible.

In summary, a number of limitations and associated guidelines for use should be considered when employing subjective assessment techniques. It appears possible that physical and mental task loading might be confused by an operator, so mental workload ratings for tasks that involve extreme degrees of physical activity should be interpreted with caution. In requesting subjective ratings, it is desirable to have reports made on specific tasks or aspects of system performance to avoid global assessments of workload where these would be of little practical value. On the basis of currently available evidence, it appears that ratings of perceived effort expenditure or some similar construct (e.g., stress, experienced mental load) would be the rating of choice in most instances related to workload questions. Current evidence also suggests potentially important dissociations between subjective and primary task measures of load, and these should be considered in interpreting the results of subjective estimates. In implementing subjective techniques, it also appears desirable to obtain workload ratings as soon as possible after task performance has been completed to minimize loss of rating information from short-term memory.

2.5. Key References

The review of subjective mental workload by Moray (1982) provides an excellent discussion of factors related to the subjective workload experienced by the operator. The comprehensive reviews of workload assessment methodology by Gartner and Murphy (1976) and Williges and Wierwille (1979) include sections dealing with subjective measures of mental load and address some methodological considerations associated with use and interpretation of such techniques. The paper by Ellis (1978) on subjective assessment of pilot workload also discusses several techniques and treats practical considerations related to implementation. On the general background level, Johannsen et al. (1979) provide an excellent review of the theoretical bases for expecting that subjective measures should prove sensitive to variations in workload, including the concept of effort (Jahns, 1973; Kahneman, 1973).

3. PRIMARY TASK MEASURES

3.1. Background

Primary task measures to assess workload by measuring actual performance on the task or design option of interest. It is assumed that, as workload increases, the additional processing resources/capacity utilized will necessarily result in some change (usually degradation) in the quality of operator performance (Sanders, 1979; Williges & Wierwille, 1979). It is argued that measurement of such changes should provide an index of the workload of the task. Importantly, such measures of the overall effectiveness of the person/machine interaction should be a very meaningful index of workload since it directly reflects the outcome of the operator's efforts. As a consequence, primary task measures are frequently used as workload assessment techniques (Williges & Wierwille, 1979).

As noted previously (Section 1.1), there are some potential problems related to the sensitivity of primary task measures. This is particularly true of Region A (Figure 42.1), where it is hypothesized that the operator has sufficient spare processing capacity to deal with increases in load and maintain primary task performance. Primary task measures would, therefore, be insensitive to workload changes within this region. A similar situation exists for Region C of Figure 42.1. Here, the operator's capacity has been reached and exceeded. Further changes in workload will show no primary task performance changes after performance has reached some asymptotic level. As a consequence, primary task measures are expected to demonstrate their greatest sensitivity in Region B of Figure 42.1, where a monotonic relationship is hypothesized to exist between performance and workload.

One clear difficulty in deciding whether to rely exclusively on primary task measures lies in determining what region of the workload-performance relationship is represented by a particular situation. There is no adequate technique that can be applied on an a priori basis to determine if the levels of load will fall in Region A or in Region B of Figure 42.1. Also, there are no adequate data dealing with the relative range of the various regions. Changes in information-processing strategies (e.g., Sperandio, 1971, 1978; Welford, 1978) or training and experience could extend the range of Region A by enabling the operator to cope with increased workload without associated decreases in performance errors.

In addition to a restricted range of sensitivity, some practical problems with primary task measures should be considered in their application. In general, unique measures must be developed for each experimental situation (e.g., Hicks & Wierwille, 1979; Williges & Wierwille, 1979). The desirability of developing workload measures that are not task specific is clear and has been noted by other investigators (e.g., Johanssen et al., 1979; Welford, 1978).

Despite these possible limitations in their use, primary task measures have been used quite frequently in workload assessment research (e.g., Chiles & Alluisi, 1979; Gartner & Murphy, 1976; Wierwille & Williges, 1980; Williges & Wierwille, 1979). One major use of such measures has been to address the adequacy of operator performance under particular experimental conditions or with certain design options, thereby distinguishing levels of load in Region B (Figure 42.1), or discriminating overload (Region B) from nonoverload (Region A) conditions. In some instances the primary task measure was used as the only metric of operator performance and load, although more typically the primary task metric was used in conjunction with some other assessment technique. Examples of both types of use are presented in the following sections, which deal with single and multiple primary task measures of load. Another purpose of primary task measurement, to provide baselines for evaluation of secondary task effects, is discussed more extensively in Section 4.

3.2. Single Primary Task Measures

In this approach a single aspect of primary task performance (number of errors, speed of performance) is used as an indicant of workload. To maximize the utility of this technique, the primary task measure should be chosen to reflect a parameter of performance that is expected to be influenced by the manipulation of load. For example, if the workload question involves increasing the number of displays or indicators to be monitored, a reaction-time measure to the display signals should precisely tap the monitoring behavior being loaded. Selection of a measure is critical to the success of the workload evaluation and in many cases constitutes a difficult task for the experimenter.

A number of successful applications of single primary task measures have been carried out. Error and latency scores have shown sensitivity to workload manipulations (e.g., Dorfman & Goldstein, 1971; Helm, 1981; Isreal, Chesney, Wickens, & Donchin, 1980; Kraus & Roscoe, 1972; McKenzie, Buckley, & Sarlanis, 1979; Percival, 1981). Dorfman and Goldstein (1971), for example, investigated the effect of increases in rate of signal presentation on performance of a display-monitoring task. Increases in speed of presentation led to systematic decrements in the number of correct responses. Kraus and Roscoe (1972) examined the effects of two types of aircraft control systems on procedural errors by pilots in a flight simulator. Pilot errors were approximately ten times greater for a normal controller versus one that permitted direct control over aircraft-maneuvering performance. A number-cancellation secondary task also showed significant differences as a function of the controller. More recently, Percival (1981) used a reaction-time measure to examine the effects of two different target types, increases in the number of background characters, and time on watch in a visual search task. Analyses revealed that both the number of background characters and type of target significantly affected mean search time whereas time on watch did not. In many of these successful applications, the primary task parameter has been selected carefully to tap a direct, meaningful consequence of the type of workload to be expected (e.g., measurement of pilot errors as a result of different kinds of air traffic control procedures). Consequently in these cases single primary task measures were sensitive to the workload manipulation.

However, there are also instances in which appropriate single primary task measures failed to reflect manipulations of task load. Several of the studies that provide examples of primary task insensitivity also used an additional measure of workload (e.g., secondary task, subjective rating) which indicated that a significant manipulation of load did, in fact, occur. Schultz, Newell, and Whitbeck (1970), for example, examined the effect of increases in the amount of turbulence on the glide-slope error in a fixed-base aircraft simulator. The glide-slope measure failed to reflect significant performance differences as a function of handling difficulty, even though ratings on a Cooper-Harper scale (Section 2.2.1) were significantly different for some of the conditions tested. Eggemeier et al. (1983) obtained similar results with a short-term memory task in which subjects recalled the frequency of occurrence of several information categories in visually presented sequences. A primary task measure of errors in recall failed to reflect variations in stimulus presentation rate, but subjective ratings of load increased substantially ($p < .01$) as a function of increases in presentation rate. Similar sensitivity differences have been reported in a number of experiments (e.g., Bahrick et al., 1954; Bell, 1978; Boggs & Simon, 1968; Burke, Gilson, & Jagacinski, 1980; Finkelman & Glass, 1970) using secondary task methodology and single primary task measures of load. In each instance, secondary task performance reflected differences in task performance conditions (e.g., degree of training; type of primary task display; presence of an environmental stressor such as noise) that were not revealed in primary task measures. (See Section 4 for an explanation of secondary task methodology and for a more detailed treatment of differences in primary and secondary task sensitivity.)

Clearly, the use of single primary task measures of workload has produced some instances in which levels of load were discriminated and others in which they were not. All of the unsuccessful applications of primary task measures noted included some other measure of workload that did discriminate the conditions employed. Therefore the results can be interpreted within the framework provided in Figure 42.1, where primary task performance measures will distinguish variations in loading in Region B, but are expected to be relatively insensitive when compared with some other measures under loading conditions in Region A. As a consequence, although single primary task measures can provide important information about expected levels of operator performance in Region B, their failure to demonstrate differences across experimental conditions or design options should not be interpreted as an indication that workload is equivalent.

3.3. Multiple Primary Task Measures

In simulated or real work environments, and in some complex laboratory task situations, it is possible to collect performance data on multiple aspects of the primary task. In this case, error or latency data are gathered on several dependent variables. The intent is to provide greater sensitivity to changes in workload by (1) permitting combined analysis of the multiple measures to decrease measurement error or, more frequently, (2) to provide assessment of a number of resources or skills so that the precision of measurement will be increased.

Obviously, the experimenter's task in selecting measures is not as demanding if one is not limited to a single measure. However, this potential for increased sensitivity is frequently purchased at considerable practical expense. Although it may be possible to collect data on literally hundreds of aircraft control parameters in simulators or flight test, decisions concerning which parameters to analyze, and the analyses themselves, can be extremely difficult. Too often, data are collected because it is possible to do so, and this can lead to nonproductive effort.

Practically, one should select multiple parameters of the primary task based on some theoretical framework. However, there are usually few data on how a given primary task measure relates to a particular theory because these metrics are unique to each application and, therefore, have not been systematically validated or parametrically studied. The investigator will, therefore, be in a position of hypothesizing which primary task measures best relate to the theoretical position chosen. This limitation means, in practice, that multiple primary task measures of workload will always be most useful as an overall screening device but will be somewhat limited with respect to diagnostic capability.

As with single measures, multiple primary task measures have produced mixed results with respect to capability in distinguishing different levels of load. In some instances the measures failed to discriminate variations in load that were detected by other assessment techniques. In other applications at least some aspect(s) of primary task performance reflected manipulations in load.

Both speed and accuracy measures have been used successfully in multiple primary task measures of load. Dorfman and Goldstein (1975), for example, used response latency, percentage of correct responses, and response failures to assess the effects of different rates of stimulus presentation in a display-monitoring task. All three measures were affected by the workload factor at all levels tested. In a subsequent experiment Goldstein and Dorfman (1978) varied both rate of presentation and the number of elements to be searched in the display-monitoring task. Two measures of response latency were computed, and both showed sensitivity to both stresses, although there were some differences between the two measures in their capability to discriminate individual levels of the two independent variables. Hicks and Wierwille (1979) used a driving simulator and compared the sensitivity of steering reversals, yaw deviation, and lateral vehicle deviation to various workload levels. All three measures of primary task performance proved sensitive to the different levels of workload achieved through varying the application of crosswind gusts to the simulated vehicle.

Not all applications of multiple primary tasks measures found equivalent sensitivity to workload for all the indices used (e.g., Brecht, 1977; Finkelman, Zeitlin, Filippi, & Friend, 1977; Huddleston & Wilson, 1971; Whitaker, 1979). For example, Whitaker (1979) examined the effects of two levels of stimulus-response compatibility as well as the number of stimulus alternatives in a choice reaction time task. Stimulus-response compatibility significantly affected choice reaction time, but errors in primary task performance were not significantly affected. Finkelman et al. (1977) also examined both time and error measures in a primary task to assess the affects of noise on driver performance. Subjects drove an automobile course and were periodically subjected to bursts of white noise either while performing the driving task alone or while performing it in conjunction with a secondary delayed digit recall task. Driving task performance was scored on the basis of the time to complete the course and by the number of course pylons that were struck. Both the presence of the noise and the requirement to perform the secondary task significantly degraded time to complete the course. However, significant increases in errors occurred only in the highest workload condition which included both the presence of the noise and the requirement to perform the secondary task.

Experiments using multivariate analyses have suggested that different primary task measures may be sensitive to different types of loading, as well as to different levels of load. Kreifeldt, Parkin, Rothschild, and Wempe (1976) investigated the effects of three air traffic control management schemes on pilot simulator performance. They recorded 16 objective flight performance measures such as aileron, elevator, and throttle activity. A series of multivariate analyses of variance (MANOVAs) and discriminant analyses revealed that eight of the 16 measures were useful in discriminating the type of load being manipulated. Similarly, North et al. (1979) used a series of MANOVAs to evaluate the capability of 14 flight performance variables in discriminating the effects of crosswind, motion, and displays on pilot performance and workload. Results indicated that three of the 14 primary flight performance variables varied significantly across all flight segments as a function of display, nine as a function of crosswind amplitude, and one as a function of motion condition. Pitch acceleration was the only primary task variable significantly affected by all three independent variables across all flight segments. See studies by Reising, Bateman, Herron, and Calhoun (1977) and Wolfe (1978) for additional examples demonstrating the need for caution in assuming that all primary task measures are equally valid as measures of any type of workload.

Although most applications of multiple primary task measures have been successful in demonstrating that at least some aspect of performance changed as a function of levels or types of load, there are instances in which this has not been the case.

Schori (1973), for example, found no significant differences in tracking performance as a function of the use of visual, auditory, or cutaneous information displays. A secondary visual monitoring task did, however, discriminate among the displays. Rolfe, Chappelow, Evans, Lindsay, and Browning (1974) used five primary task performance measures to evaluate their capabilities in assessing three different types of load (physical, perceptual, and mental) in an aircraft simulator. Although observational measures and subjective ratings revealed significant workload differences, none of five primary task measures (e.g., glide-scope deviation, airspeed variability) did so.

It is clear from the preceding discussion that although multiple primary task measures have considerable face validity, they are no easier to interpret than other measures of load. Any noted performance differences do not provide the basis for clear diagnostic statements concerning the resources being overloaded. As a consequence, multiple primary task measures, like their single-task counterparts, must be considered global rather than diagnostic measures of load. Although, with the use of appropriate multivariate statistical treatment, they appear to offer somewhat greater sensitivity than single-task measures, multiple primary task measures must also be assumed to represent a restricted range of sensitivity. Failure to demonstrate performance differences with such measures should, therefore, not be interpreted as reflecting equivalent levels of load between conditions or design options being compared.

The principal reason for use of multiple primary task measures would therefore appear to be the need to meet the objective of determining whether the workload of a system will compromise operator performance. In such applications, diagnostic capability is not needed, and it is only necessary to determine which aspects of the primary task the user considers critical to the criterion of good performance. Either multiple or single primary task measures may then be perfectly adequate to provide the needed answers.

4. SECONDARY TASK MEASURES

4.1. Background

A very frequently used workload assessment procedure is secondary task methodology, which requires concurrent performance of two tasks by the operator. The task of central interest is termed the primary task, and an estimate of primary task workload is derived from performance of an additional or secondary task. In most applications, the procedure is used to measure the presumed spare or reserve processing capacity afforded by a primary task. This measure is derived from levels of performance on the secondary task, which serves as an indicant of the spare capacity available while the operator performs the primary task. Secondary task measures are generally considered more sensitive (see Section 1) to differences in capacity expenditure than are primary task procedures (Section 3). As noted previously (Section 1), the technique is also considered to be highly diagnostic of primary task demand.

Application of the procedure requires individual and concurrent performance of both the primary and secondary tasks. Individual performance levels are used as baselines for assessing the effects of concurrent task performance. Without such baselines, suitable interpretation of concurrent performance is impossible. (See Sperling & Dosher, Chapter 2, for an extensive discussion of general issues in measuring concurrent task performance.)

In the usual application, the subject or operator is instructed to maintain error-free performance on one task at the expense of the other. Depending on the experimenter's choice, either primary or secondary task performance may be emphasized. Two major categories of secondary task methodology (Knowles, 1963) can be distinguished by this differing emphasis on either primary or secondary task performance: (1) the loading task and (2) the subsidiary task paradigms.

4.2. Categories of Secondary Task Measures

4.2.1. Loading Task Paradigm. In the loading task paradigm the subject is instructed to maintain secondary task performance, even if decrements in primary task performance result. It is assumed that the additional load imposed by the secondary task will shift total workload from Region A to Region B of Figure 42.1, thereby inducing breakdowns in primary task performance. Under equal levels of secondary task loading, performance on more difficult primary tasks will deteriorate more than will performance on less difficult tasks. In this paradigm, secondary task performance is measured to ensure that specified criterion levels are maintained and that the loading imposed by the task is, in fact, equated across the various experimental conditions. Degradations in primary task performance that occur at specific levels of secondary task loading can then be used as an index of primary task workload.

Secondary loading tasks are used principally to simulate the effects of information-processing requirements or demands that are absent from the laboratory or simulation environment but are expected in an operational environment. For example, an investigator conducting an evaluation of two cockpit display options might be concerned that pilot performance under single-task laboratory conditions (e.g., Region A of Figure 42.1) would not reveal differences in display loading that would be apparent with the addition of demands imposed by concurrent activities in the flight environment (e.g., Region B of Figure 42.1). In this situation, a secondary loading task could be employed to increase total workload in the laboratory environment, thereby making it more representative of the operational environment and increasing the sensitivity of primary task performance. In addition to this function, loading tasks have been used to simulate stressors or to aid in evaluating the effects of other stressors (e.g., noise, heat) by permitting evaluation of workload in a more sensitive task environment (Ogden et al., 1979). As in the example just noted, primary task performance is expected to be more sensitive under the loading condition to the distracting nature of any additional processing requirements associated with the stressor being evaluated.

Loading tasks have been used in a variety of applications to evaluate the adequacy of displays, configurations, methods of task performance, and the effects of various types of stressors on primary task performance (Ogden et al., 1979; Rolfe, 1971). An experiment by Dougherty, Emery, and Curtin (1964) provides an example of increasing the sensitivity of workload assessment through use of this paradigm. A conventional aircraft instrument panel was compared with a pictorial display that portrayed comparable information. Pilots flew standard profiles in a flight simulator while performing a secondary task requiring that they read a series of digits presented at varying rates on a cockpit display. Figure 42.10 shows combined mean absolute error in the primary flight task as a function of display type

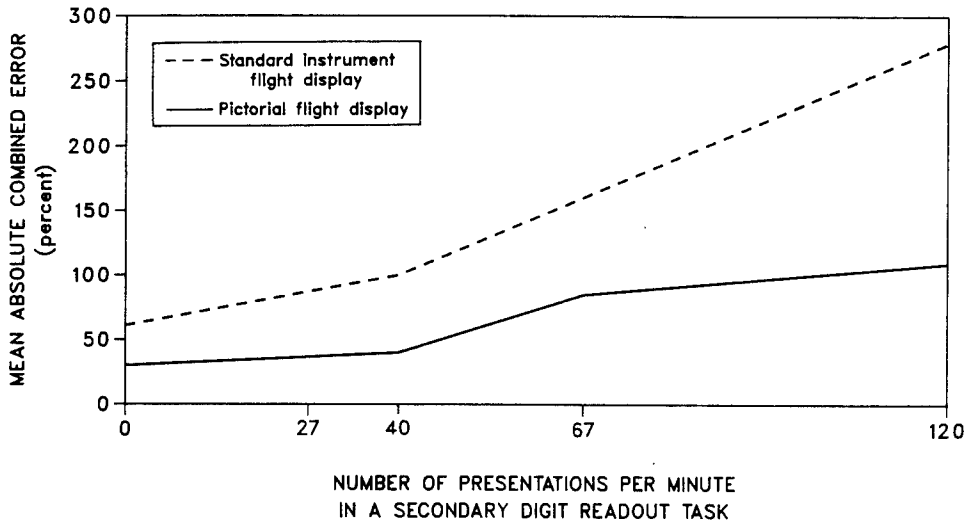


Figure 42.10. Mean absolute combined error of primary flight performance as a function of display type and stimulus presentation rate in a secondary digit readout task. The secondary task required that pilots read a series of digits presented at varying rates on a cockpit display. The flight performance measure illustrated is a combined score representing four measures of error: altitude, heading, airspeed, and track deviations. The combined error score represents the sum of a proportional error score in each condition for altitude, airspeed, error, and track. The proportional error of each parameter was derived by dividing the individual error scores in each condition by the overall mean for that parameter. A standard baseline was thus provided for weighting the errors contributed by each of the four dependent variables. The combined error score, therefore, provided an overall index of performance in each condition. The effects of displays, secondary task presentation rate, and their interaction on combined errors were significant ($p < .01$). Post hoc multiple comparison tests indicated no significant differences between displays under single-task baselines and at the slowest two presentation rates under concurrent-task performance. However, there were significant differences between displays at the fastest two presentation rates. Use of the secondary loading task at the fastest two presentation rates, therefore, permitted differences in workload between displays to be detected. These differences were not distinguishable under primary task baselines or under the lowest two secondary task presentation rates. (Redrawn from D. J. Dougherty, H. H. Emery, & J. G. Curtin, *Comparison of perceptual workload in flying standard instrumentation and the contact analog vertical display* (Rept. JANAIR D228-421-019). Copyright 1964 by Bell Helicopter Co. Reprinted with permission.)

and digit presentation rate. There were no significant differences in flight performance under single-task baseline conditions. Simple primary task measures, therefore, indicated no difference in workload between the two displays. However, primary flight performance did vary significantly as a function of the display at the two fastest digit presentation rates, with the standard instrument display showing significant decrements relative to baseline. The more difficult versions of the secondary loading task, therefore, shifted total workload with the conventional display into Region B of Figure 42.1, thereby causing decrements in flight performance. Since equivalent levels of secondary processing load did not lead to significant flight performance decrements with the pictorial display, it can be concluded that the latter imposed less load on the pilot than did the conventional display. This example illustrates the proper use of the loading task paradigm to enhance the measurement sensitivity of a display workload assessment: the highest levels of secondary task demand were sufficient to shift overlall load into the sensitive portion of the performance curve, and interpretation was carried out relative to single primary task baseline performance.

4.2.2. Subsidiary Task Paradigm. The second and more frequent application of the secondary task technique is the subsidiary or reserve capacity task paradigm. In this paradigm the subject is instructed to avoid degraded primary task performance at the expense of the secondary task. The secondary task in this paradigm is not used to load the primary task, but rather is used to determine how much additional work can be undertaken while the primary task is performed at single-task baseline levels (Knowles, 1963). The subsidiary task paradigm is based on the assumption that the addition of the secondary task will shift total workload from Region A to Region B (Figure 42.1) and that decrements in secondary task performance will result. Such decrements should reflect the spare or reserve capacity that remains when the primary task is being performed. The theoretical basis of this method is depicted graphically in Figure 42.11, adapted from Brown (1964).

A number of studies (e.g., Bahrick et al., 1954; Bell, 1978; Burke et al., 1980; Dornic, 1980a; Schifflet, Linton, & Spicuzza, 1982) illustrate the use of the subsidiary task paradigm to measure reserve capacity differences not revealed by primary task metrics. Bell (1978), for example, employed a primary pursuit rotor tracking task and a subsidiary number-processing task to investigate the effects of high ambient temperatures and noise stress on operator performance. Neither the heat nor the noise stress reliably affected primary tracking performance, but both stressors significantly affected performance in the subsidiary number-processing task (Figure 42.12).

Evidently the primary task performance measures failed to reflect differences in workload attributable to the stress con-

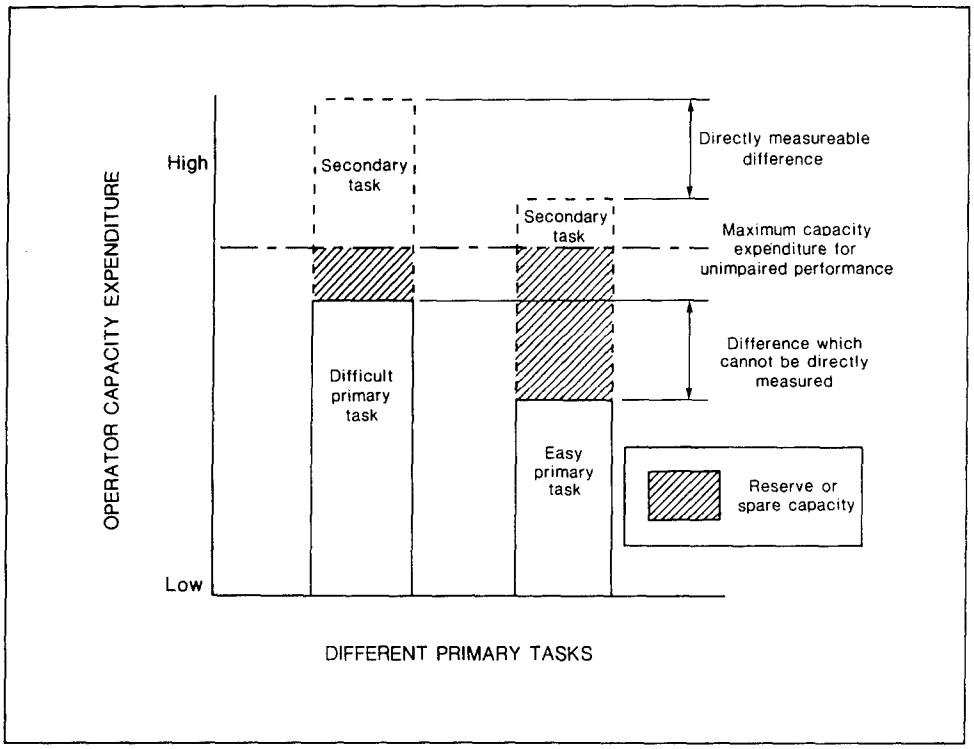


Figure 42.11. Representation for use of the secondary task to measure operator reserve processing capacity. The two tasks represented differ in the capacity expenditure required for their performance, but this difference cannot be directly measured since neither task exceeds operator processing capacity for unimpaired performance. Addition of the secondary task exceeds processing capacity in both instances, leading to decrements in secondary task performance. These directly measurable differences can be assumed to reflect the differences in primary task capacity expenditure that cannot be measured through use of primary task procedures. A number of assumptions associated with the subsidiary task paradigm are also depicted. The straight line depicting maximum capacity expenditure for unimpaired performance reflects the assumption that overall processing capacity remains fixed across all levels of primary task difficulty. A second assumption of the paradigm is that the constituents of workload are linearly additive, regardless of the source of the load. This assumption is depicted by the simple addition of the secondary task capacity expenditure to that of the primary task, with no interaction or intrusion of the secondary task on primary task capacity requirements. A third assumption originally associated with the paradigm is that operator processing capacity is unitary or undifferentiated, as reflected in the unitary capacity expenditure index depicted on the ordinate. Each of these assumptions has been questioned, and the text should be consulted for further discussion of the data related to violations of each. (Adapted from I. D. Brown, 1964.)

ditions because the subjects were able to compensate for the increased load through additional capacity expenditure. Use of the secondary task measure, however, permitted a more sensitive analysis of the capacity expenditure than that afforded by the primary task. This type of approach has been used to derive estimates of reserve capacity for a variety of purposes, including evaluation of instrumentation and displays, evaluation of the workload imposed by different operating conditions and procedures, assessment of practice effects on performance, and ordering the difficulty of various primary tasks. (Consult reviews by Ogden et al., 1979; Rolfe, 1971; and Williges & Wierwille, 1979, for an extensive list of examples.)

4.3. Assumptions of the Subsidiary Task Paradigm

Although the subsidiary task paradigm has proven useful in many situations, a number of assumptions made in its use must be evaluated before the technique is chosen (refer to Figure 42.11). First, it is clear that overall processing capacity is assumed to remain fixed across levels of task demand (Hawkins & Ketchum, 1980; Senders, 1970). If human processing capacity is not fixed, but can vary as a function of task demand, reserve capacity indices are reduced to ordinal measures of workload (Hawkins & Ketchum, 1980). Although there are theoretical positions (e.g., Kahneman, 1973; see Gopher & Donchin, Chapter 41, for a more extensive discussion) that suggest that capacity is capable of expansion, current evidence to support such a notion is equivocal (Hawkins & Ketchum, 1980) and does not now pose a serious problem for secondary task methodology.

A second assumption originally associated with the subsidiary task paradigm was that the information-processing capacity of the human system is unitary or undifferentiated (Hawkins & Ketchum, 1980; Senders, 1970). (See Gopher & Donchin, Chapter 41, for a more extensive discussion of this theoretical position.) This assumption was based on theories that attributed processing restrictions in the human system to limits of a single processing channel (e.g., Broadbent, 1958) or of a single pool of processing resources (e.g., Moray, 1967). If capacity is unitary, there should be no substantial difference in sensitivity between secondary tasks, and comparative eval-

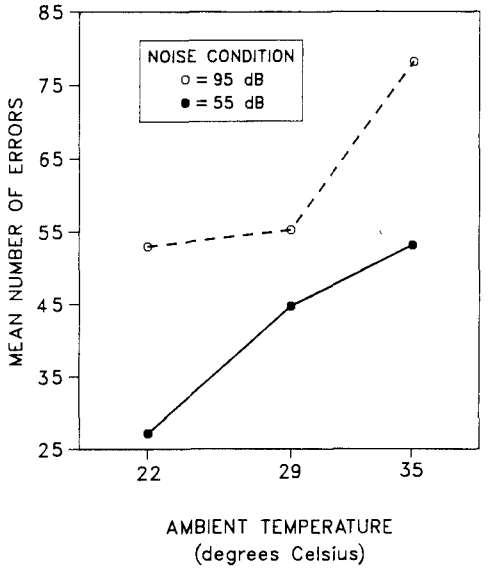


Figure 42.12. Errors in a secondary digit-processing task as a function of ambient noise and temperature conditions. A subsidiary auditory digit-processing task was performed concurrently with a primary pursuit rotor tracking task to evaluate the effects of three ambient temperature conditions (22, 29, and 35°C) and two noise levels [55 dB (A) background noise versus 95 dB (A) white noise bursts of 1–9 seconds duration] on operator performance. The digit-processing task required that subjects monitor a sequence of two-digit numbers and give an appropriate key press response to indicate if a number was numerically higher or lower than the preceding number. Primary tracking task performance as indexed by time on target did not vary significantly as a function of either noise or temperature conditions. Errors in the subsidiary processing task revealed significant effects of both temperature ($p < .01$) and noise ($p < .005$). Post hoc multiple comparisons tests on the temperature factor indicated a significant difference ($p < .05$) between the 22 and 35°C conditions. Use of the subsidiary digit-processing task, therefore, permitted differences in capacity expenditure to be distinguished that were not apparent in primary task performance. (Drawn from the data of P. A. Bell, Effects of noise and heat stress on primary and subsidiary task performance. *Human Factors, 20*. Copyright 1978 by Human Factors Society. Reprinted with permission.)

uation of the workload imposed by different primary tasks should be possible, regardless of the secondary tasks employed. Recent evidence (e.g., Navon & Gopher, 1979, 1980; Wickens, 1980, 1984a) favoring a multiple-resources approach to human capacity limitations (see also Section 1.2), however, indicates that sensitivity is a function of overlap in processing resources between the primary and secondary tasks and suggests that the notion of a universal secondary task is unworkable. Several investigators (e.g., Gopher, 1978; Wickens, 1979) have suggested the alternative of establishing a battery of secondary tasks, each tapping a different resource, that would be applied to different primary tasks. This approach would enable construction of a resource/load profile for each primary task. Rather than providing a single metric of overall workload, such a battery would assess the degree of load in each of a number of resources, thereby providing a more diagnostic analysis of workload (see Section 1.2). Such an approach does not appear to be particularly restrictive for practical applications, if the number of secondary tasks included in a battery is not excessive. Some guidelines developed from current theory for matching primary and secondary task demands are discussed in Section 4.4.3.

A third major assumption of the subsidiary task paradigm is that the constituents of workload are linearly additive, regardless of the source of the load (Senders, 1970). The importance of primary task intrusion has been extensively discussed (Gartner & Murphy, 1976; Knowles, 1963; Ogden et al., 1979; Rolfe, 1971; Senders, 1970) and relates not only to the practical considerations discussed in Section 1.3, but also to the theoretical basis of the paradigm as a measure of reserve capacity. From a theoretical perspective, changes in both primary and secondary task performance under dual task conditions suggest that resource expenditure associated with the primary task has been altered by the addition of the secondary task. Secondary task performance measures, therefore, would not represent a pure index of the reserve capacity associated with the primary task. When this occurs, clear interpretation of the results is extremely difficult. Obviously, in view of the importance of this assumption, it is critical that the primary task be measured alone in every experiment, and that the degree of intrusion, if any, be specified for every secondary task. Failure to do this is a common methodological flaw in many reported studies.

A number of methodological guidelines that should be followed when applying secondary task techniques are suggested by several of the assumptions discussed. These and other considerations related to use of secondary task methodology are discussed in the following section.

4.4. Methodological Guidelines

4.4.1. General Methodological Considerations. Several general guidelines in applications of secondary task methodology are shown in Table 42.9. These guidelines are based on the assumptions discussed here, current theory regarding the nature of processing limitations in the human, and considerations resulting from previous applications of secondary task methodology. More detailed guidelines dealing with the issues of primary task intrusion and secondary task sensitivity are treated in Sections 4.4.2 and 4.4.3.

4.4.2. Techniques to Minimize Primary Task Intrusion. As noted previously (Section 1.3), primary task intrusion has represented a major problem in applications of secondary task methodology (e.g., Gartner & Murphy, 1976; Ogden et al., 1979; Williges & Wierwille, 1979). There are a number of potential sources of such intrusion, including peripheral interference (Wickens, 1984a) and failure to adhere to the resource allocation policy (e.g., maintain primary task performance at the expense of the secondary task) stipulated by the experimenter (Pew, 1979).

Peripheral interference (Wickens, 1984a) results from physical (e.g., the inability of the eye to focus simultaneously at two locations) rather than resource or capacity constraints within the processing system. In attempting to reduce or eliminate peripheral interference, a variety of secondary tasks have been proposed that are designed to minimize input and output constraints imposed by the necessity to perform two tasks concurrently. A typical pattern in these attempts has been to use sensory and motor modalities in the secondary task that differ from those required in the primary task, as suggested by Knowles (1963). Another approach (Ogden et al., 1979) has been to use tasks that reduce stimulus input or immediate response requirements. Examples of such tasks include random digit generation (e.g., Zeitlin & Finkelman, 1975), silent addition (e.g., McLeod, 1973), subjective time estimation (e.g., Casali & Wierwille, 1982, 1983; Hart, 1978), and a time interval production task (e.g., Casali & Wierwille, 1982, 1983; Johannsen, Pfendler, & Stein, 1976; Michon, 1964, 1966; Shingledecker, 1980). See

Table 42.9. Methodological Guidelines for Applications of Secondary Task Methodology

1. In the loading task paradigm, subjects should be instructed to maintain secondary task performance at single-task baselines under concurrent task conditions.
2. In the subsidiary task paradigm, subjects should be instructed that primary task performance should be maintained at single-task baseline levels under concurent task conditions.
3. In both paradigms baseline measures of single-task performance on both the primary and secondary tasks should be taken. In the loading task paradigm, primary task baselines are required to assess differences in primary task performance that might occur under concurrent task conditions. Secondary task baselines are required to ensure that the secondary task is performed to the criterion set by the experimenter. In the subsidiary task paradigm, primary task baseline performance is required to evaluate any intrusion effects that might occur. Baseline secondary task measures are required to evaluate properly the degree of single to dual task decrements which might occur.
4. In both paradigms employ several levels of secondary task difficulty. As illustrated in Figure 42.10, higher levels of secondary task difficulty may distinguish differences in workload between design options or tasks that are not distinguished by lower levels of secondary task difficulty. The theoretical basis for such difficulty effects is that lower levels of secondary task difficulty may not be sufficient to shift total workload from Region A to B (Figure 42.1), whereas more difficult levels may do so.
5. In the subsidiary task paradigm, consider the use of various techniques that have been proposed to reduce or eliminate primary task intrusion. Two major classes of these techniques include adaptive secondary methodology and embedded secondary tasks. Both of these techniques are treated in more detail in Section 4.4.2.
6. In both paradigms attempt to ensure maximum secondary task sensitivity through choice of an appropriate task and through use of sufficient practice to achieve stable performance on the secondary task prior to its use. Guidelines related to secondary task sensitivity are discussed more extensively in Section 4.4.3.

Several general guidelines that should be followed in applications of secondary task methodology are suggested by assumptions associated with the technique, current theory regarding the nature of human information-processing limitations, and considerations resulting from previous applications of the technique. Based on the purpose of the study/evaluation to be conducted, choose either the loading task paradigm (Section 4.2.1) or the subsidiary task paradigm (Section 4.2.2). After the paradigm has been chosen, implement the appropriate guidelines illustrated in the table.

Section 4.4.3 for representative studies that have used several of these tasks.

In addition to choice of particular tasks to minimize input and output interference, other techniques that are potentially applicable to a variety of secondary tasks have been proposed to deal with the problem of primary task intrusion. These techniques have the potential to deal with other probable sources of intrusion, such as failure to adhere to resource allocation policies stipulated by an experimenter. Two such techniques are adaptive and embedded secondary tasks.

4.4.2.1. Adaptive Task Techniques. In the usual form of the adaptive technique, primary task performance is maintained at specified levels by manipulating secondary task loading. The degree of secondary task loading that can be achieved without intrusion then constitutes one measure of primary task workload. By manipulating secondary task difficulty, primary task performance can be stabilized to permit clearer interpretation of any secondary task decrements that occur.

The cross-adaptive technique is one such procedure that has been applied to secondary task methodology. In this technique primary task criterion levels are maintained under concurrent task conditions by varying secondary task loading as a function of primary task performance. The cross-adaptive technique does not necessarily eliminate intrusion, but rather standardizes primary task performance levels in all conditions according to an experimenter-defined criterion. Kelly and Wargo (1967) demonstrated the feasibility of one version of this paradigm with a primary tracking task and a discrete secondary monitoring task. In the cross-adaptive condition, the secondary task was turned on or off depending on whether tracking scores were above or below a specified criterion. Primary task performance levels were stabilized in this condition, but not in a condition that employed the same secondary task with a fixed difficulty. Thus use of the cross-adaptive technique facilitated interpretation of the results.

Cross-adaptive secondary tasks similar to that employed by Kelly and Wargo have been used to assess reserve capacity with somewhat mixed results (Brecht, 1977; Schori, 1973; Schori & Jones, 1975). Brecht (1977), for example, compared the degree of intrusion that resulted in a primary arithmetic task from both self-paced and cross-adaptive versions of a secondary task that required subjects to respond to visual signals on a display panel. In the cross-adaptive condition, the secondary task was turned on or off as a function of the errors in the primary arithmetic task. Primary task errors were nearly the same in both secondary task conditions. In this case, therefore, primary task accuracy was not significantly affected by the cross-adaptive versus self-paced secondary task.

In addition to discrete monitoring tasks, a version of the critical tracking task (e.g., Jex & Clement, 1979) has been used in the cross-adaptive paradigm. (See Wickens, Chapter 39, for a more detailed discussion of the critical tracking task.) The critical tracking task measures the limits of the operator's capability to control an unstable target in a single axis (Jex & Clement, 1979). The principal measure of operator performance in this task is a "critical" level of instability at which tracking performance breaks down.

The cross-coupled critical tracking task (Jex & Clement, 1979; Jex, Jewell, & Allen, 1972) is a cross-adaptive version of the task that involves two-axis tracking. In the cross-coupled task, instability on the secondary axis is varied adaptively as a function of the operator's performance on the primary axis. Workload in this task is defined as the level of difficulty that can be controlled on the secondary axis while maintaining primary task performance at its specified level.

Several investigators have used cross-coupled tracking tasks to assess operator workload. Burke et al. (1980), for example, used the task to demonstrate differences in workload between a kinesthetic-tactile and several visual displays. The cross-coupled instability task has been successfully used (Jex & Clement, 1979) to evaluate workload associated with a variety of other factors, including different display types, such as a cockpit moving-map display versus a horizontal situation indicator (Clement, 1976; Clement, McRuer, & Klein, 1972); and control devices involving different levels of kinesthetic information (Merhav & Ya'acov, 1976).

Several observations concerning the applicability of adaptive secondary tasks can be made on the basis of current evidence. First, the cross-adaptive loading task appears capable of eliminating the joint variation in primary/secondary task performance that results from primary task intrusion. However, a

number of factors related to implementation requirements can limit application of the technique. For example, effective use of cross-adaptive techniques requires a stable, sensitive, and continuous measure of primary task performance (Kelly & Wargo, 1967). If discrete or instantaneous performance measures are used, the technique can be much less effective in stabilizing primary task performance. The cross-coupled critical tracking task can also be difficult to use with discrete primary tasks, either because they do not permit a continuous measure of performance or because they demand short-term attention to the extent that a secondary continuous control task is impossible (Jex & Clement, 1979). Therefore, if cross-adaptive tasks are to be considered as a general solution to primary task intrusion, more work must be performed to extend their usage to discrete tasks. Aside from the constraints imposed by the requirement for a primary task with continuous output, the instrumentation necessary to measure primary task performance and adapt the secondary task can impose potential limits on the applicability of cross-adaptive techniques, particularly in operational environments. Use of the technique, therefore, appears most feasible in laboratory or simulation environments.

In addition to the constraints noted, Brown (1978) has raised some objections to the use of adaptive secondary tasks. Brown maintained that to provide a reliable scale of measurement, only secondary tasks that impose a constant load should be employed. Secondary tasks with variable loads were regarded as providing a variable index of primary task workload. Therefore the use of cross-adaptive loading tasks was criticized because their load varies in some inverse relationship with primary task performance. Brown, therefore, recommended that only forced-paced secondary tasks be used.

4.4.2.2. Embedded Secondary Tasks. Another technique that has been proposed to minimize primary task intrusion and that is applicable to operational, simulation, and laboratory environments is the embedded secondary task (Shingledecker et al., 1980). An embedded secondary task is a calibrated task that already exists as a part of the operator's role in the system environment. Although it represents a component of operator activity in the system, it can be treated as distinct from primary task performance. The rationale underlying the embedded secondary task is that selection of a component of operator behavior with a secondary priority in the system will ensure that the task will be relegated to a secondary role by the operator. Therefore it is anticipated that intrusions on primary system performance can be minimized without employing artificial, experimenter-imposed task priorities.

Shingledecker et al. (1980) demonstrated the possibility of using radio communications activities as an embedded secondary task. A number of fighter aircraft communications activities were identified and their workload scaled. Activities chosen for scaling required a sequence of verbal responses and manual radio switching activities by the pilot to meet the demands of a communicated request. Pilot responses therefore provided the means to evaluate secondary task performance.

Shingledecker and Crabtree (1982) evaluated the sensitivity of the communications tasks to variations in load imposed by a critical tracking task (e.g., Jex et al., 1966) in a low-fidelity flight simulator. The total time to complete a communications secondary task served as the performance measure. Four of eight communications tasks reliably discriminated two levels of tracking difficulty, thereby supporting the feasibility of communications activities as secondary tasks. Subjects in the experiment were not pilots, so the population was not appropriate for a rigorous evaluation of the primary task intrusion that can be expected in operational or high-fidelity simulation environments.

Although the results of current work support the embedded secondary task technique, additional data are required to evaluate primary task intrusiveness in high-fidelity environments and sensitivity to a wider variety of primary tasks. If successful, the embedded secondary task technique would have considerable practical utility because of its potentially nonintrusive applicability to a variety of complex environments. Since the technique involves activities normally performed during system operation, the tasks should not appear artificial, and operator acceptance should be high. Also, instrumentation requirements and learning/practice effects associated with the tasks should be minimal.

4.4.3. Secondary Task Sensitivity. Recommendations regarding the choice of a secondary task to ensure greatest sensitivity should, ideally, be based on systematic comparative data relating performance on various secondary tasks to workload variations in a standard set of primary tasks. Although not extensive, some work toward developing such a data base has been conducted (e.g., Shingledecker et al., 1983; Wierwille & Casali, 1983a). Wierwille and Casali, for example, summarized the results of four experiments (Casali & Wierwille, 1982, 1983; Rahimi & Wierwille, 1982; Wierwille & Connor, 1983) that examined the sensitivity of a large number of workload measures (secondary task, physiological, primary task, subjective) to several types of load variations (perceptual, central processing, psychomotor, communication) in a general aviation flight simulator. The secondary task of time estimation (Hart, 1975, 1978) was used in all four experiments and successfully discriminated two of three levels of psychomotor, central-processing, and communication load, and one of three levels of perceptual load. A secondary interval production task (Michon, 1966) used in three of the experiments demonstrated a different pattern of sensitivity. This task successfully discriminated two of three levels of perceptual load, but failed to demonstrate any significant differences in the central-processing or communications loading employed in the experiments. Results from these experiments and from other comparative evaluations of secondary tasks (e.g., Brown, 1965; Huddleston & Wilson, 1971; Wetherell, 1981; Wickens & Kessel, 1979, 1980; Zeitlin & Finkelman, 1975) confirm that, in many cases, varying sensitivity can be expected from individual secondary tasks when they are used to evaluate different levels and types of primary task demand.

Additional comparative research of the type summarized by Wierwille and Casali (1983a) is required before extensive empirical recommendations regarding choice of secondary tasks for particular applications can be made. However, in the absence of such comparative data, some guidelines can be suggested. These guidelines are based on current theory and data and, when followed, should increase the sensitivity of secondary task assessments of primary task loading. These guidelines include: (1) the desirability of choosing a secondary task that imposes some continuous demand on the operator's information-processing system; (2) the desirability of providing practice on the secondary task prior to its use in the dual task situation; and (3) the need for the secondary task to be representative of the processing resources expended by the primary task. Each of these guidelines is discussed in more detail in the sections that follow.

The guideline of choosing a secondary task that imposes some continuous demand on the operator's processing system is supported by several comparative evaluations of secondary tasks (Brown, 1965; Huddleston & Wilson, 1971; Zeitlin & Finkelman, 1975). These evaluations can be interpreted as supporting the position that tasks such as monitoring and continuous short-term memory which require sustained operator attention or processing can be more sensitive to variations in loading than tasks that require only momentary allocations of attention. This guideline would appear to be particularly critical when the primary task is one that involves a high degree of temporal variability in its loading. In such instances, a secondary task that does not impose some continuous demand (e.g., memory load, monitoring requirements) might be effectively interleaved with the varying requirements of the primary task, thus obscuring "peaks" in primary task demand and reducing the capability of the secondary task to reflect transient or short-term loads.

The guideline of providing practice on the secondary task (e.g., Knowles, 1963; Pew, 1979) prior to its use in a dual-task paradigm reflects the necessity of some stability in secondary task performance as a prerequisite to reliable and sensitive estimation of the workload imposed by the primary task. It is particularly critical when a secondary task is to be used repeatedly to assess the workload associated with different levels of primary task difficulty, different design options, and so forth.

One potential problem related to this guideline is that practice on the primary-secondary task combination also may improve an operator's capability to perform the two tasks concurrently (e.g., Pew, 1979). Several investigators (e.g., Damos, 1977; Gopher & North, 1977) have reported the development of specific time-sharing capabilities during concurrent task performance subsequent to practice on the individual tasks themselves. An important implication of this finding is that when such time-sharing strategies significantly influence either primary or secondary task performance, workload estimates and conclusions will be specific to that type of primary-secondary task combination, since pure estimates of reserve capacity cannot be derived. Another potential problem is related to the fact that with extended practice, the secondary task may become automatized (e.g., Fisk, Derrick & Schneider, 1982; Schneider & Fisk, 1982a, 1982b), thereby minimizing the capacity demands of the task and making it less senitive to variations in primary task load. The notion of an automatized secondary task is based on a distinction between controlled and automatic processing (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). (See Gopher & Donchin, Chapter 41, for a more extensive discussion of this distinction.) Controlled processing is said to occur when subjects respond to novel or inconsistent stimuli or have received minimal training. Automatic processing develops as subjects are trained to respond consistently to stimuli, and it is characterized by fast, effortless processing when compared to controlled processing. As noted by Fisk et al. (1982), some problems in interpretation of secondary task experiments could arise if one secondary task that could become automatic (e.g., simple reaction time to an easily discriminable stimulus) is used throughout the course of an experiment. On the basis of current data (e.g., Schneider & Fisk, 1982a, 1982b), concern about such automatic processing should be most pronounced in situations involving stimuli that require a consistent response (e.g., always respond to the presence of a particular stimulus) over a relatively large number of repetitions (e.g., in excess of 1000). Where such concern does exist, one solution is to vary stimuli and responses so that no stimulus requires a consistent response. Fisk et al. (1982), for example, used a visual choice reaction time secondary task in which stimulus-response mappings were changed several times to preclude the possible development of automaticity and to ensure relatively constant secondary task load thoughout the course of the experiment. Analyses of the reaction time data confirmed the effectiveness of the procedure in that there were no significant practice effects during the experiment.

The guideline that the secondary task be representative of the processing resources expended by the primary task is based on the multiple resources theory of capacity limitations in the human system (Section 4.3). Basically, the multiple resources approach predicts relative insensitivity if a mismatch exists between the processing resources demanded by the primary and secondary tasks (Wickens, 1984b). Some data from Shingledecker et al. (1983) illustrate sensitivity differences that can be attributed to the degree of overlap in secondary and primary task resources. Shingledecker et al. (1983) used a Michon (1966) interval production secondary task in a series of experiments that included several levels of difficulty in three different primary tasks: (1) a critical-tracking (Jex & Clement, 1979) task; (2) a memory-search (Sternberg, 1966) task; and (3) a display-monitoring (Chiles, Alluisi, & Adams, 1968) task. These three primary tasks can be regarded as having placed heaviest demands on motor, central-processing, and perceptual resources, respectively. The interval production secondary task required that subjects produce a series of regularly timed finger-tapping responses. Primary task difficulty was varied through manipulation of instability (lambda) in the tracking task, the size of the memory set in the Sternberg task, the number of displays to be monitored, and the ease of signal detection in the monitoring task. Levels of primary task difficulty were chosen on the basis of previous parametric analyses that indicated that each demand level used produced significant variations in primary task performance. The interval production workload measure was based on differences between the variability in the intervals produced under single-task baseline and concurrent task conditions. The results (Figure 42.13) clearly demonstrate a high degree of differential sensitivity on the part of the interval production measure and indicate that it was sensitive to the difficulty variations employed in the psychomotor task—but was relatively insensitive to demand manipulations used in the central-processing and perceptual tasks. With the assumption that the interval production-tapping task places its heaviest demands on motor output, the sensitivity differences are consistent with the predictions of current multiple resources frameworks (e.g., Gopher, Brickner, & Navon, 1982; Wickens, 1984a) which hold that motor output functions draw on a resource pool that is separate from that used for perceptual and central-processing functions. Confirmation of the predicted effects of dimension overlap on concurrent task performance has also been provided in a number of other recent studies (e.g., North, 1977; Wickens, 1980; Wickens & Kessel, 1980; Wickens, Mountford, & Schreiner, 1981). Taken together, these results suggest that to ensure greatest sensitivity, a secondary task should be chosen to demonstrate maximum possible overlap with the demand of the primary task.

A central issue in implementing this guideline involves identification of those dimensions that define separate resources. Although several theorists (e.g., Friedman, Polson, Dafoe, & Gaskill, 1982; North, 1977; Sanders, 1979) have addressed this issue, the most extensive theory has been advanced by Wickens (1984a), who proposed that stages of information pro-

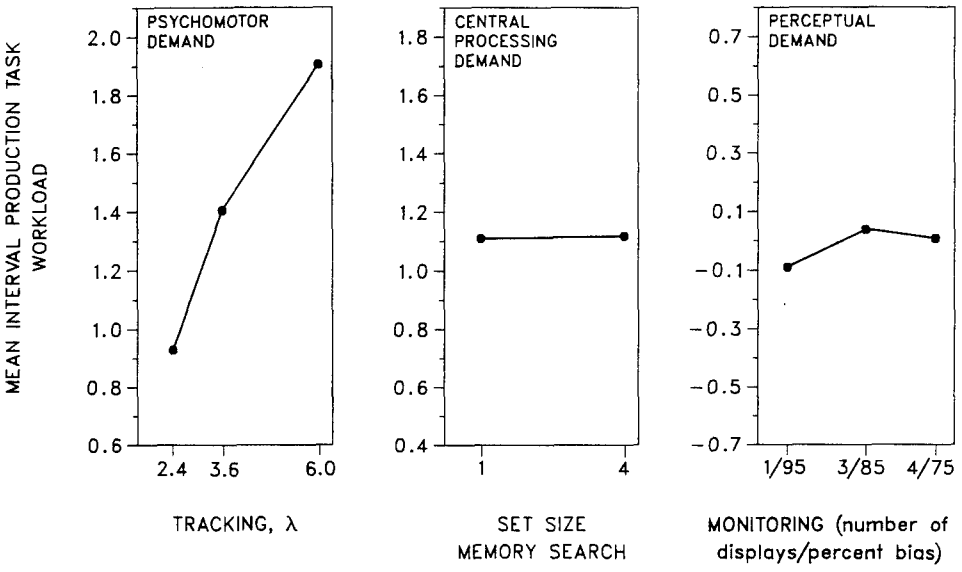


Figure 42.13. Mean interval production task (IPT) workload as a function of tracking, memory, and monitoring task difficulty in three experiments. The IPT served as the secondary task in three experiments which involved either a primary critical tracking (Jex, McDonnell, & Phatak, 1966) task (psychomotor demand), a primary Sternberg (1966) memory search task (central-processing demand), or a primary probability monitoring (Chiles, Alluisi, & Adams, 1968) task (perceptual demand). Primary task difficulty was varied through manipulations of tracking task instability (lambda), the size of the memory set, or the number of displays to be monitored and the ease of signal detection (percentage of time that a signal bias occurred) in the monitoring task. The IPT workload measure was based on differences between the variability in duration of the intervals produced under single-task baseline and concurrent task conditions. As illustrated in the figure, IPT performance was significantly affected by manipulations of tracking task demand ($p < .01$), but not by manipulations of central-processing load in the memory task ($p > .05$) or by demand manipulations in the monitoring task ($p > .10$). Each of the tracking task loading levels differed reliably from every other ($p < .01$). The results, therefore, illustrate the differential sensitivity of the IPT measure to different types of task demand and indicate that the IPT demonstrates its greatest sensitivity to psychomotor/response output demands. (Redrawn from C. A. Shinlgedecker, W. H. Acton, & M. S. Crabtree, *Development and application of a criterion task set for workload metric evaluation*. Copyright 1983 by Society of Automotive Engineers, Inc. Reprinted with permission.)

cessing (perceptual/central-processing operations versus response selection and execution), modalities of perception (auditory versus visual), and codes of information processing and response (spatial-manual versus verbal-vocal) represent dimensions that appear to define separate resources (Figure 42.14). Detailed discussions of the data that support this view can be found in Wickens (1980, 1984a) and in Gopher and Donchin (Chapter 41).

Although current data provide some support for the dimensions outlined by Wickens (1984a), more extensive data are required before definitive conclusions can be drawn regarding the number and types of dimensions required by multiple resources theory. At present, however, the most acceptable guideline that can be followed to ensure secondary task sensitivity is to choose a task that is representative of primary task-processing demands as outlined by the current theory.

As an aid in identifying particular secondary tasks that can be considered for use, Section 4.4.5 briefly discusses major classes of tasks that have been used previously and provides representative examples of successful uses of each class.

4.4.4. Interpretations of Single-to-Dual Task Performance Decrements. An additional important consideration in applications of secondary task methodology is the proper interpretation of differences between single- and dual-task performance levels. It has been general practice in the subsidiary task paradigm to interpret decrements in concurrent secondary task performance relative to single-task baselines as an indicant of primary task capacity/resource expenditure. Single-to-dual primary task decrements have been similarly interpreted in the loading task paradigm.

However, as noted by several investigators (e.g., Kantowitz & Knight, 1976; Navon & Gopher, 1979; Roediger, Knight, & Kantowitz, 1977; Wickens, 1984b), there are a number of sources of single-to-dual task decrements that are not directly related to the capacity or resource expenditure associated with either the primary or secondary task. These sources have been variously referred to as qualitative changes in single-to-dual performance (Roediger et al., 1977) or concurrence costs (Navon & Gopher, 1979). Nonresource interference can be related to such factors as: (1) interference between primary and secondary tasks occasioned by competition for structures or mechanisms within the processing system (e.g., a single sensory or motor system); and (2) capacity or resource expenditure unrelated to either task individually, but necessary to coordinate or schedule the concurrent performance of both tasks. Single-to-dual task decrements that are attributable to such concurrence costs clearly confound interpretation of the resulting data, which are meant to represent a pure measure of the capacity or resource expenditure associated with the primary task.

To aid in distinguishing resource/capacity related performance decrements from those associated with concurrence

costs or nonresource factors, several investigators (e.g., Kantowitz & Knight, 1976; Roediger et al., 1977) have indicated that it is necessary to demonstrate in dual-task situations that difficulty manipulations in one task produce performance changes in the other. In the subsidiary task paradigm, this could be achieved through manipulating the difficulty of the primary task and observing concomitant variations in secondary task performance. Alternately, the same objective could be achieved in the loading task paradigm by varying secondary task difficulty and observing changes in primary task performance. The basic argument here is that the observation of systematic performance changes in one task with difficulty variations in the other suggests that such changes are not completely attributable to those nonresource factors that operate in an all-or-none manner. For example, certain forms of competition for processing structures (e.g., a single sensory system) that could cause dual-task decrements can be assumed to operate in an all-or-none fashion. Therefore, graded or systematic secondary task decrements with increases in primary task difficulty would indicate that such all-or-none structural interference could not constitute the entire explanation of the noted decrements. It is also reasonable to assume that, in some instances, capacity or resource expenditure associated with the requirement to coordinate dual-task performance would operate in such an all-or-none manner.

Of course, any form of structural interference or capacity expenditure attributable to dual-task coordination that does not operate in an all-or-none manner could not be discriminated with the proposed method. It is possible, for example, that a graded form of structural interference could occur if the time required for use of a common processing structure was systematically related to the difficulty of each task. In these cases, performance levels of one task could be significantly affected by variations in the difficulty of the second task, and the resulting dual-task decrement patterns would not discriminate resource from nonresource competition. Therefore, systematic performance variations in one task that result from difficulty manipulations in the other can provide a strong but not absolute basis to infer that dual-task decrements are attributable to resource/capacity competition. Failure to find such performance variations, however, suggests that some form of nonresource competition could have contributed to any noted single-to-dual task decrements. In this case, straightforward interpretation of dual-task decrements as representing the resource/capacity expenditure associated with task performance is not possible.

A second major factor in secondary task decrements that is unrelated to resource/capacity expenditure is the possibility that subjects will vary their allocation of processing resources to tasks as a function of experimental conditions. Of course, the subsidiary task paradigm should incorporate clear instruc-

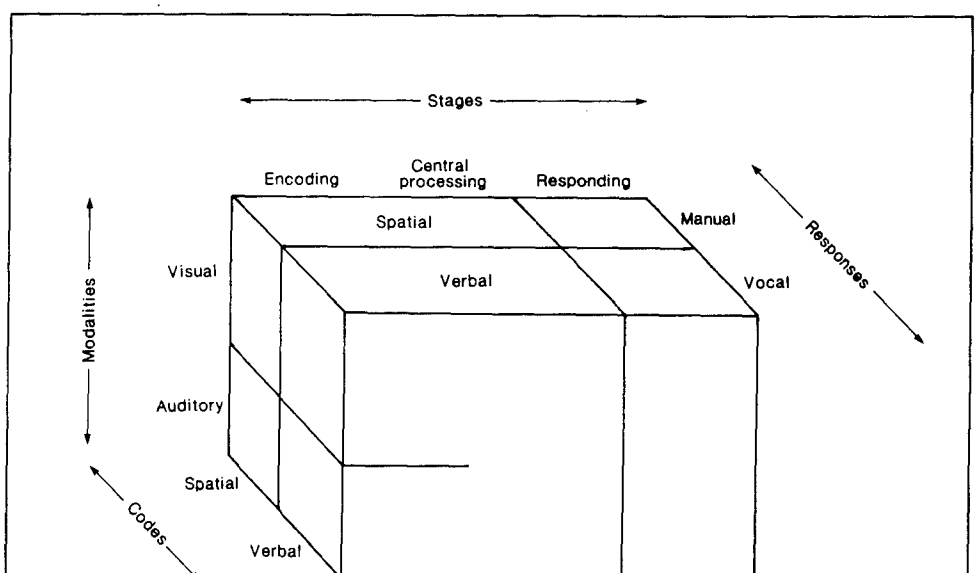


Figure 42.14. A proposed structure of processing resources. The multiple resources approach to processing limitations within the human information-processing system maintains that it can be best described as a series of independent pools or processing structures, each with its own limited supply of resources which are not interchangeable. One critical element in this theory is identification of dimensions that define the various resources that make up the processing system. Based on evidence from the dual-task literature, Wickens (1980, 1984a) has proposed that processing resources may be defined by three dichotomous dimensions represented in the figure. These dimensions and their components include: (1) stages of processing, which include a resource pool dedicated to perceptual/central-processing functions and a pool for response selection and execution; (2) codes of information processing and response, which include a pool for processing verbal information and vocal response and a separate pool for processing spatial materials and manual responses; and (3) modalities of perception, which include separate pools for auditory materials and for visual materials. The proposed structure of the noted resources has been depicted by Wickens (1984a) in the heuristic representation outlined in the figure. (From C. D. Wickens, Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention*. Academic Press, Inc., 1984. Reprinted with permission.)

tions to subjects to maintain primary task performance at single-task levels when it is performed in conjunction with the secondary task. Similar instructions to maintain unimpaired secondary task performance should be part of the loading task paradigm. However, as noted previously (Section 4.4.2), subjects frequently fail to maintain single-task performance levels on the designated task. One potential reason for such failure is a shift in the amount of processing resources allocated to each task under different conditions. Such shifts can be particularly troublesome when they occur between primary tasks or conditions whose workload is to be directly compared on the basis of secondary task performance. Consider the case where easy and difficult variants of a primary task are to be compared. During dual performance in the easy primary task condition, subjects could allocate resources to favor the primary task, but shift resource allocation to favor the secondary task during the difficult primary task condition. In this instance, secondary task metrics expressed only as single-to-dual decrements could lead to the erroneous conclusion that capacity expenditure typically associated with the easier primary task was actually greater than that of the more difficult version.

Because of the potential effects of both nonresource competition and allocation policy shifts on dual-task performance, some investigators (e.g., Gopher et al., 1982; Navon & Gopher, 1979, 1980) have argued that the nature of interactions between concurrently performed tasks can best be investigated if both task difficulty and task emphasis are jointly manipulated in a dual-task situation. The effect of varying allocation policy between two concurrently performed tasks can be graphically depicted in the form of a performance operating characteristics (POC) curve, which plots joint levels of performance in a single graph. (See Sperling & Dosher, Chapter 2, and Gopher & Donchin, Chapter 41, for additional discussion of this type of analysis.) Figure 42.15 (Wickens, 1984b) depicts a hypothetical representation of concurrent performance of tasks A and B within a POC space, and serves to illustrate several important aspects of the POC. Single-task performance levels are indicated on the appropriate axes, as is a hypothetical intersection at point P, which represents perfect time-sharing or no single-to-dual decrement for either task. Hypothetical performance levels for tasks A and B are depicted for three allocation policies, and the curve joining the points represents the POC. As illustrated, the single-task baselines on the axes may not be continuous with the extension of the POC to the single-task axes. If actual single-task baseline performance is better, the difference between these points represents concurrence costs incurred in the dual-task performance condition. Time-sharing efficiency between tasks can be assessed by the average distance of the curve from the origin O. As the distance from the origin increases and dual-task performance levels approach P, more efficient time-sharing performance is indicated. Of course, as time-sharing efficiency increases, the sensitivity of a secondary task to changes in primary task loading levels decreases. Finally, the allocation bias of a given dual-task performance situation is indicated by the proximity of a given point on the POC to one axis over the other. Points along the positive diagonal represent an equal allocation of resources between tasks. Provided that measures on the two axes have been converted to common units (e.g., standardized scores, see, for example, Wickens et al., 1981), spatial relations along the axes may be interpreted in the manner described.

The shape of POC curve can provide some indication of the degree of resource or capacity interference that is present under concurrent task conditions. A smooth or linear POC of the type depicted indicates that a trading relationship exists between the tasks, since a number of resource units removed from one task can be utilized to improve performance on the other. On the other hand, a discontinuous or rectangular POC suggests that resources are not exchangeable between tasks and that resources withdrawn from one task cannot be used to improve performance on the other. One reason for a discontinuous POC could be that the two tasks draw on separate resources and are, therefore, not interchangeable. Another possible reason is that performance on either task cannot be improved by allocation of additional resources (i.e., the task is data limited; see Norman & Bobrow, 1975).

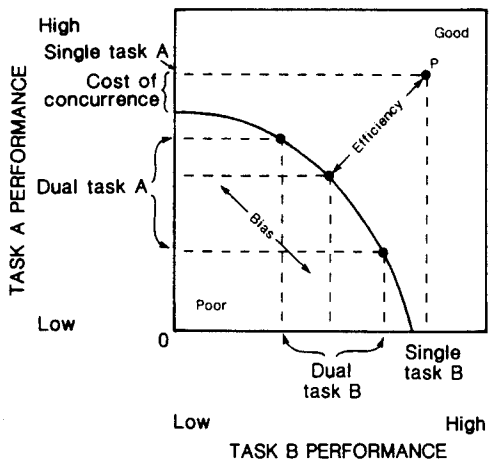


Figure 42.15. Hypothetical representation of dual-task performance within a performance operating characteristic space. The joint performance of two hypothetical tasks A and B are depicted in a performance operating characteristic (POC) space. Single-task performance levels are indicated on each axis. Point P is a hypothetical intersection point which is based on single-task performance levels and represents perfect time-sharing, or no single-to-dual task performance decrement for either task. Individual points within the POC represent joint performance levels for each task under a particular allocation policy that requires that some specified proportion of resources be dedicated to performance of each task. Three such points are depicted in the figure, and these have been joined to form the POC curve. As illustrated, single-task baseline levels as depicted on the axes may not be continuous with the extension of the POC curve to the axes. Under the condition illustrated, actual baseline performance exceeds the projection, and the noted difference represents concurrence costs associated with the dual-task performance condition. Time-sharing efficiency between tasks is indicated by the average distance of the curve from the origin (0), and as the curve approaches P, more efficient performance is indicated. Efficient time-sharing is associated with low sensitivity in a secondary task. The allocation policy adopted for a particular dual-task performance condition can be evaluated by the proximity of the joint performance point to one axis versus the other. Points falling along the positive diagonal represent an equal allocation of resources. See text for additional information regarding interpretations of POCs. (From C. D. Wickens, *Engineering psychology and human performance*. Charles Merrill Publishing Co., 1984. Reprinted with permission.)

Plotting dual-task data in a POC space can, therefore, afford some advantages to an investigator in terms of identifying the possible existence of concurrence costs; depicting shifts in allocation policy that could affect levels of secondary task performance; and, through examination of the shape of the POC, permitting some inferences about the existence of resource-related interference between tasks.

However, from a practical perspective, neither observation of systematic changes in task performance with difficulty manipulations in the concurrent task nor representation of dual-

task data in a POC space provides a ready index of primary task workload. Neither approach affords a single measure of load, and each has other potentially serious practical constraints associated with it. For example, it may be impossible in some applications to systematically manipulate primary task difficulty. It is also very likely that many applications (e.g., piloting an aircraft) would preclude the variations in allocation of resources between tasks required to develop a POC. Therefore it appears that, although both methods can provide stronger evidence of capacity/resource interference than that afforded by single-to-dual task decrement, the latter still provides the most practically viable means of indexing primary task workload in the secondary task paradigm. When they can be applied, the former techniques can provide useful information in properly interpreting any noted single-to-dual task decrements. As such, both techniques appear to represent useful adjuncts to the single-to-dual task decrement metric, particularly in application-oriented situations.

4.4.5. Major Classes of Secondary Tasks. Ogden et al. (1979) reported that the four general classes of secondary tasks most commonly used since 1964 included choice reaction time, monitoring, tracking, and memory tasks. Other classes of tasks frequently used included mental mathematics, time estimation paradigms, shadowing, and simple reaction time. Table 42.10

Table 42.10. Major Classes of Frequently Used Secondary Tasks

1. *Choice Reaction Time*

   Numerous studies have employed choice reaction time as a secondary task. Choice reaction time typically involves presentation of more than one relatively simple stimulus, with the requirement that a subject generate a different response for each stimulus. Reaction time stimuli have been presented both visually and auditorily, and the predominant response mode has been manual. Choice reaction time tasks can be generally assumed to impose greater central-processing and response selection demands than simple reaction time tasks. Representative examples of the use of choice reaction time can be found in Fisk, Derrick, and Schneider (1982); Isreal, Chesney, Wickens, and Donchin (1980); and McLeod (1977).

2. *Tracking*

   Tracking tasks have frequently been used as secondary tasks in workload evaluations. Both pursuit and compensatory tracking tasks have been commonly used in workload assessment. These tasks employ visual stimulation and continuous manual response. Depending on the order of control dynamics, various degrees of central-processing and motor demands are involved in tracking performance. The critical tracking task (Jex, McDonnell, & Phatak, 1966) imposes heavy loads on motor resources. Representative examples of the use of this task can be found in Jex and Clement (1979); Martin (1970); Whitaker (1979); and Wickens and Kessel (1980).

3. *Monitoring*

   Monitoring tasks have been frequently used as secondary tasks. Monitoring is typically characterized by the requirement to detect the occurrence of a stimulus from among several alternatives, and it is generally considered to place a relatively heavy emphasis on perceptual processes. Representative examples of the use of this technique can be found in Brecht (1977) and Schori (1973).

4. *Memory*

   A very large number of memory tasks have been utilized as workload assessment techniques. Most have been short-term memory tasks, and a number of different types of materials and specific memory requirements have been employed. Memory tasks are generally considered to impose their heaviest demands on central-processing resources. One of the most commonly used memory tasks is the Sternberg (1966) memory search paradigm. The Sternberg task has the potential to permit central-processing effects to be discriminated from stimulus encoding/response execution effects, and so it has been frequently employed in studies of multiple resources theory (e.g., Wickens & Derrick, 1981b) and also in evaluations of pilot workload (e.g., Crawford, Pearson, & Hoffman, 1978; O'Donnell, 1976; Schifflet, Linton, & Spicuzza, 1982). In addition to the references cited, representative examples of the use of memory tasks can be found in Allport, Antonis, and Reynolds (1972); Huddleston and Wilson (1971); and Wickens and Kessel (1980).

5. *Mental Mathematics*

   Various forms of mental mathematics have been effectively used as secondary tasks. Different forms of addition tasks have been most frequently employed, but subtraction and multiplication tasks have also been used. Mental mathematics is typically considered to draw most heavily on central-processing resources. Representative examples of the use of this task can be found in Green and Flux (1976); Huddleston and Wilson (1971); and McLeod (1973, 1977).

6. *Shadowing*

   Shadowing tasks typically require that a subject repeat sequences of verbal or numerical material as they are being presented. No transformations of the material are usually required, so that such tasks are typically considered to exert their heaviest demands on perceptual resources. Representative examples of the use of this task can be found in Anderson and Toivanen (1970); Fournier and Stager (1976); McLeod (1973); and Price (1975).

7. *Simple Reaction Time*

   In addition to choice reaction time, simple reaction time tasks which employ one discrete stimulus and response have also been used as secondary tasks. The use of such tasks typically would be suggested when an investigator wishes to minimize central-processing and response selection requirements associated with secondary task performance. Representative examples of use of this task can be found in Eysenck and Eysenck (1970); Lansman and Hunt (1982); Schwartz (1976); and Tyler, Hertel, McCallum, and Ellis (1979).

8. *Time Estimation Paradigms*

   Two major paradigms related to the production of time intervals have been used for workload assessment. These paradigms include the Michon (1966) interval production task and the time estimation technique of Hart (1975, 1978).

   The interval production task requires that a subject generate a series of regular time intervals by performing a motor response at a specific rate. Performance of the task requires no sensory input, and the output modality can be chosen to reduce conflicts with the output modality of the primary task if the experimenter chooses to do so. The Shingledecker, Acton, and Crabtree (1983) data indicate that this task places heaviest demands on motor output/response resources.

   Hart (1978) has reviewed the rationale for use of estimates of time interval duration as a measure of the workload imposed by concurrent task demands. Time estimation was chosen for evaluation because of its acceptability to pilots, ease of implementation and scoring, and minimal learning effects. A basic distinction is drawn between two modes of time estimation, active and retrospective. Choice of a particular mode by a subject is theoretically related to the level of demand imposed by concurrent task performance. Active time estimation involves actively keeping track of time during a specific interval. In retrospective time estimation subjects may also make

(*Table continues on p. 34.*)

Table 42.10. (*continued*)

8. *Time Estimation Paradigms* (*continued*)
time estimates without attending to time as it passes by estimating the duration on an interval at its conclusion. Representative examples of the use of the interval production task can be found in Casali and Wierwille (1982, 1983); Michon (1966); and Shingledecker, Acton, and Crabtree (1983). Examples of the use of time estimation can be found in Casali and Wierwille (1982, 1983); Gunning (1978); and Hart (1975).

This table is intended to serve as an initial guide in choosing a class (or classes) of secondary tasks for a particular application. Eight of the most frequently used classes of secondary tasks are described, and references to several successful applications of each task are provided so that additional methodology and implementation details can be accessed.

provides a brief description of each of the eight classes of tasks and also includes listings of representative studies that successfully employed each technique. It should be noted that many of the secondary tasks described in Table 42.10 are derived from the experimental paradigms of cognitive psychology and an extensive literature is associated with each of the methods. This literature is ably described in the chapters comprising Section V of this handbook, which deal with information processing and human performance.

When a class of task has been selected for possible application, it is recommended that a number of the representative studies listed be consulted for additional details regarding methodological considerations and implementation requirements. More extensive listings of tasks within each category and additional secondary tasks can be found in Ogden et al. (1979), Williges and Wierwille (1979), Chiles and Alluisi (1979), and Wierwille and Williges (1980).

4.5. Key References

Reviews of secondary task methodology in Knowles (1963), Rolfe (1971), Ogden et al. (1979), and Williges and Wierwille (1979) provide comprehensive overviews of the technique, including methodological issues associated with its usage and examples of tasks that have been used in the paradigm. The review by Rolfe (1971) and papers by Kantowitz and Knight (1976), Roediger et al. (1977), Navon and Gopher (1979), Fisk et al. (1982), and Gopher and Sanders (1982) should be consulted for more detailed treatment of methodological considerations involved in applications of the paradigm. The review of multiple resources theory by Wickens (1984a) provides an excellent summary of the evidence bearing on the theory and its implications for secondary task assessment of workload.

5. PHYSIOLOGICAL MEASURES

5.1. Background

The concept of measuring workload through physiological processes such as heart rate, muscle tension, or eye movements is disarmingly simple. It would appear that effort, proposed as a major determinant of workload (Johannsen et al., 1979), should be quantifiable through direct measurement of physiological arousal or activation level. Unfortunately, success in achieving this diagnostic simplicity has been less than spectacular (see O'Donnell, 1979; Wierwille & Williges, 1978, for reviews). Many studies failed to find consistent patterns of physiological change with known changes in workload. This initially discouraging trend was reversed when it was realized that many measures must be viewed as potential indices of specific psychological processes rather than as global measures of effort, arousal, or activation (Hassett, 1978). When the hope that physiological measures could be used interchangeably was abandoned, unexpected degrees of diagnosticity and sensitivity were revealed for measures such as the cortical evoked response, whereas others (e.g., pupil diameter) were found to index combinations of resource utilization.

This section discusses those physiological techniques used most often to assess workload, with special emphasis on their diagnosticity and applicability to specific test environments. Representative data supporting these applications are presented, but the reader is referred to key references presented throughout the section for more specific coverage. Since the potential value of these measures as workload assessment procedures lies in applications falling in the A or B (low or medium workload) regions of Figure 42.1, methods applicable to lower workload, short-term tasks are emphasized. Thus several procedures such as biochemical analysis and long-term monitoring of operator's state are excluded (see Hockey, Chapter 44, and Moray, Chapter 40).

5.2. Measures of Brain Function

The electroencephalogram (EEG) recorded from surface electrodes placed directly on the scalp offers an attractive procedure for directly tapping the brain's activity during performance of a task (see Gopher & Donchin, Chapter 41). Although some attempts to quantify the amount of EEG power in specific bands (alpha, beta, theta, delta, etc.) of the EEG spectrum have been carried out, these have been generally disappointing as indices of workload, except where overall activation clearly changes as a function of the load imposed. Crude measures of alpha abundance during performance of a task have been taken, and it has been postulated that increased alpha correlates with low involvement of particular cortical areas in the task. These procedures, however, have generally shown low reliability and considerable variability except where grossly different functional tasks were used (spatial versus verbal tasks). For this reason, the gross analysis of power in the EEG over long epochs has not developed as a common workload assessment procedure.

On the other hand, the development of the cortical evoked response in its various forms has shown considerable promise in assessing specific workload variables. Several useful and apparently sensitive procedures for assessing workload have evolved, and the following sections focus on these.

5.2.1. Methods of Signal Analysis. Progress in utilizing the electroencephalogram in workload assessment parallels the evolution of signal analysis techniques and the development of more sophisticated procedures for isolating the brain's response to a specific discrete stimulus. The most common procedure for accomplishing this involves ensemble averaging of EEG records which are time-locked to the presentation of the stimulus (Callaway, Teuting, & Koslow, 1978). The effect of this time-lock averaging is to enhance the signal-to-noise ratio, and to isolate the brain's response to the stimulus from ongoing EEG activity generated by other sources.

Although ensemble averaging constitutes the most commonly used method for isolating the evoked response, several

alternative procedures have been developed. Among these, the linear stepwise discriminant analysis (LSDA) has been used most often (Donchin & Herning, 1975; Horst & Donchin, 1980; Squires & Donchin, 1976). This procedure uses a theoretical or empirically determined set of features postulated to differentiate between the brain's response to different classes of stimuli or situations. One-way analysis of variance is used to select the most discriminating feature, and this procedure is repeated until addition of a feature fails to improve discrimination between the samples. The selected features are then used with appropriate filtering techniques to scan EEG epochs suspected of containing the evoked response. Matching of the EEG signal to the selected features results in a single trial classification of the EEG as containing the evoked response in question.

More sophisticated analysis techniques can be used to improve detection of the evoked response in some cases. The quadratic discriminant function (QDF) has been described by Aunon, McGillem, and O'Donnell (1982). This classifier uses a decision rule that attempts to minimize error or risk in the classification. It has been shown to be somewhat more sensitive than the LSDA for detecting differences in the evoked response to stimuli occurring in different parts of the visual field. However, differences between this and less sophisticated techniques have been relatively small considering the increased complexity of the analysis required.

5.2.2. Transient Cortical Evoked Response. In the "transient" evoked response, stimuli are presented at a relatively slow rate (e.g., 1 second or longer between stimuli). In this mode the essential effects of the stimulus on the brain are allowed to dissipate before a second stimulus is presented, and the transient response of the brain is therefore isolated in the evoked response. An idealized picture of the typical transient visual evoked response is presented in Figure 42.16. Early components (less than 250 msec) have been related to sensory characteristics of the stimulus, such as image sharpness, color, and intensity (O'Donnell, 1979; Regan, 1972) and to some early cognitive events.

With respect to workload, attention has been focused on what is identified as the third major positive peak, which frequently occurs in the time period between 250 and 500 msec (depending on the task). This P300 wave (sometimes called the P3) was described by Sutton, Tueting, Zubin, and John (1967) as occurring to an unpredictable stimulus that reduced uncertainty. Subsequently, many studies confirmed that the P300 was elicited only when the subject was actively processing information, and that it was elicited only by stimuli that had some relevance to the task being performed by the subject (Beck, 1975). Numerous studies have confirmed that the P300 is not synonymous with the contingent negative variation, showing a different scalp distribution and different pattern of sensitivity to stimulus conditions (Donchin, 1976). In addition, the amplitude and latency of the P300 wave appears sensitive to different aspects of the stimulus situation. Amplitude has been shown to vary monotonically with stimulus probability (Duncan-Johnson & Donchin, 1977; Squires, Wickens, Squires, & Donchin, 1976). It has been proposed that P300 amplitude is directly proportional to the degree of subjective surprise at the appearance of a stimulus. Conversely, the greater the expectancy, the smaller the P300 amplitude. On the other hand, the latency of the P300 wave has been suggested as indexing the amount of time taken by the subject in evaluating a stimulus (Donchin, 1981). In this view, P300 latency is seen to be independent of response selection and execution time. Thus P300 may index a cognitive component that, while generally correlated with reaction time, can be independent of the overt measure. Together, the latency and amplitude of P300 may be used to assess differences in task-induced difficulty of processing and responding to information.

The implications of the foregoing levels of specificity in analysis of the P300 amplitude and latency for workload assessment are clear. Insofar as surprise, expectancy, and task

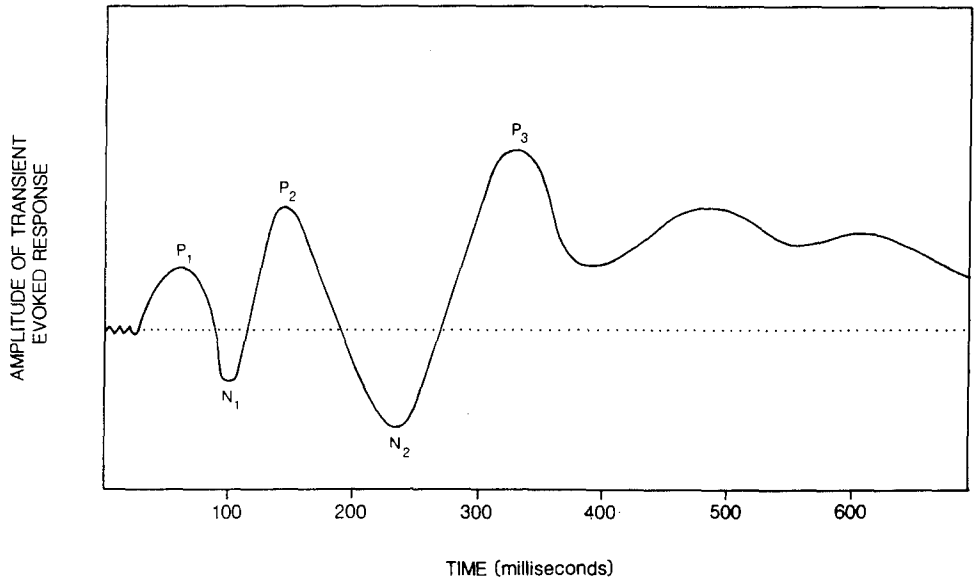


Figure 42.16. Idealized components of a typical transient evoked response (visual). The components occurring before about 250 msec (P_1, N_1, P_2, N_2) have been related to sensory characteristics such as image sharpness, color, and movement, and to some early cognitive events. The components between about 250 and 500–600 msec have generally been related to "cognitive" activities involved in processing information, as well as its meaning or information value. Peaks occurring after 500 or 600 msec in a simple response situation are usually due to motor factors such as movement or muscle contraction.

predictability determine the difficulty of a task, paradigms that emphasize analysis of amplitude differences should measure workload. If difficulties in stimulus evaluations are the major determinant of task workload, latency differences in P300 should be explored. Although the diagnostic limits of these observations are still being defined, two relatively standardized paradigms have evolved that can be recommended for use in applied situations. These are discussed next.

5.2.2.1. The "Oddball" Paradigm. A technique has been reported (see Gopher & Donchin, Chapter 41) that permits assessment of certain types of workload through analysis of the P300 amplitude generated to a relatively nonintrusive secondary task. A typical procedure involves presenting an auditory stimulus to a subject during performance of a visual-motor task. The auditory stimulus may be one of two clearly discriminable types (e.g., high tones versus low tones). One type of stimulus occurs more frequently than the second, and the subject is instructed to monitor (e.g., count) either class of stimuli. The P300 amplitude to the rarer of the two classes of stimuli is then obtained, and this amplitude has been shown to vary as a function of a number of conditions (Donchin, 1981).

This "oddball" paradigm, developed at the Cognitive Psychophysiology Laboratory of the University of Illinois, has been employed in a number of workload studies. Wickens, Isreal, and Donchin (1977) utilized a visual tracking task and manipulated the workload of the task by displacing the cursor in either the horizontal (one-dimensional) or both the horizontal and vertical (two-dimensional) directions. During tracking, the auditory oddball task was presented in a Bernoulli series, with two easily discriminable pitches presented at a 1.5-second interstimulus interval. A clear reduction in P300 amplitude was seen with the imposition of the tracking task, as compared to a control condition with no tracking task. However, no differences in P300 amplitude were found between the one- and two-dimensional tracking. Thus it would appear that amplitude measures of the P300 are sensitive to the imposition of the workload but do not show good sensitivity between levels of tracking workload.

The failure of the P300 to differentiate between different levels of tracking difficulty has been interpreted to mean that the resources tapped by the evoked response peak may be specific to the perceptual demands of a task and may not be sensitive to the response load (Isreal, Wickens, & Donchin, 1979). This was confirmed in an experiment in which the workload of a display-monitoring task was manipulated independently of the response. Subjects monitored four or eight targets that moved about on a television screen. Half of the targets were squares and half were triangles, and periodically either a square or a triangle increased in brightness or changed its direction of movement. Subjects were required to monitor one class of targets and to detect their intensification or their change in course. Workload was manipulated by varying the number of targets to be monitored. Auditory-evoked responses were obtained to an oddball paradigm as described. Figure 42.17 reveals that the P300 amplitude was monotonically related to the number of display elements to be monitored for the course change detection condition (Isreal, Wickens, Chesney, & Donchin, 1980). This sensitivity to a task requiring more perceptual than motor activity indicates that the P300 may be specific to perceptual resource loading.

The use of the P300 amplitude in response to the oddball paradigm as a workload assessment technique clearly requires further validation. However, the technique has shown some ability to detect workload differences in a simulator environment.

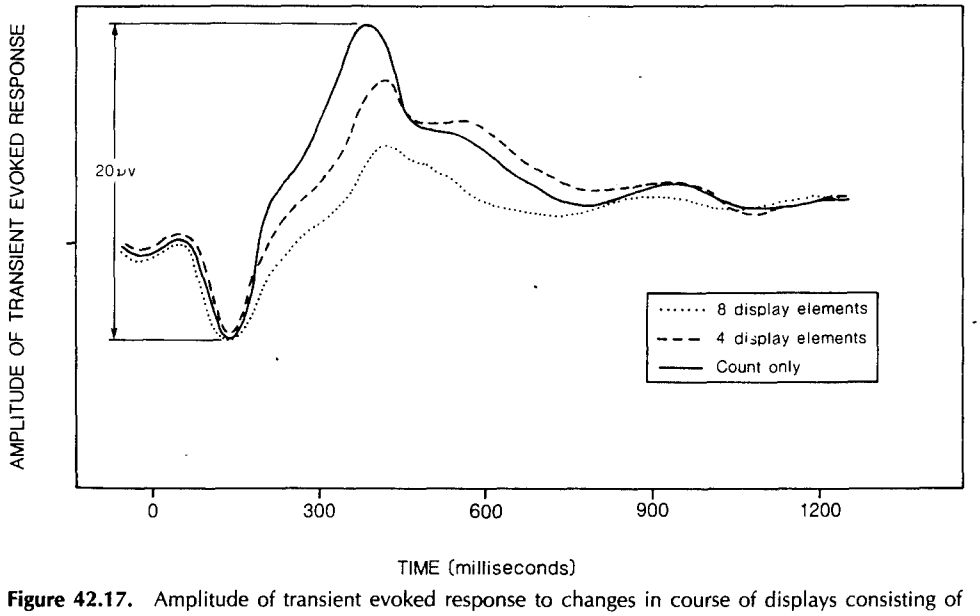


Figure 42.17. Amplitude of transient evoked response to changes in course of displays consisting of differing numbers of elements. Workload was highest when eight elements had to be monitored, and the P300 peak showed reliable differences between the control condition (count only) and the four-element condition ($p < .01$) and between the four-element and the eight-element condition ($p < .01$). This measure, therefore, showed sensitivity to perceptual loading in a task requiring little motor activity. Parietal evoked responses to auditory stimuli presented in a Bernoulli series of low- and high-pitched tones presented at 2-second intervals are shown. Probability of the high-pitched tone was .33 on any trial. (Redrawn from J. B. Isreal, C. D. Wickens, G. L. Chesney, & E. Donchin, The event-related brain potential as an index of display-monitoring workload. *Human Factors, 22*. Copyright 1980 by Human Factors Society. Reprinted with permission.)

Natani and Gomer (1981) used a low-fidelity flight simulation and controlled workload by manipulating task difficulty. They found significant differences, attributable to workload, which paralleled performance scores. Although this was a relatively austere demonstration in the sense that the simulator and procedures only approximated actual operations, it suggests the type of methodology that could utilize P300 amplitude as obtained in the oddball paradigm in field settings.

The overall pattern of results obtained with this technique suggests that it is effective in assessing workload because the subject's normal ability to establish a pattern or expectancy based on past experience is disrupted with the imposition of additional central-processing loads. The major requirement in adapting the paradigm to any particular situation is to ensure that the secondary task remains relevant in a consistent way to the individual. This permits the expectancy determinant of P300 amplitude to contribute predominantly to the measure. This task relevance can be introduced into a real-life situation without the necessity of having artificial tones or counting requirements. For instance, the ordinary auditory stimuli occurring to a pilot or driver which require attention (e.g., threat warning tones in a combat aircraft) could be used to generate the P300. In this way the oddball paradigm can be used as a highly face valid, nonobtrusive measure in an operational situation.

The major limitation of the oddball paradigm is the lack of evidence concerning its sensitivity. Wickens et al. (1977) have pointed out that both the ensemble average evoked response and the slope measure based on the sequential probability requires sampling the EEG over a considerable period of time. Thus moment-to-moment fluctuations in the workload of a given task may be averaged into the measure, reducing its overall sensitivity. These authors suggest techniques for providing such moment-to-moment assessement, including single-stimulus isolation of the P300, but these have yet to be systematically tested in controlled experiments manipulating workload over a broad range.

5.2.2.2. Transient Response to the Primary Task. The transient response can be evoked by the presentation of any number of primary task stimuli, either visual or auditory. Such a measure should directly assess the evaluation time and expectancy associated with the task. The amplitude and latency of the P300 could then be used to quantify the workload associated with the primary task. Only a limited number of applications of this measure have been carried out, principally using laboratory paradigms in which processing load was varied.

In one such demonstration the Sternberg memory-scanning paradigm (Sternberg, 1969; also see Chase, Chapter 28) was used to manipulate the memory-scanning workload. Subjects were presented with probe letters of the alphabet and were required to respond differently depending on whether the probes were members of a previously memorized "positive" set of letters. Memory load was manipulated by changing the number of letters in the memorized set (Gomer, Spicuzza, & O'Donnell, 1976). Visual evoked responses were obtained to the presentation of the probe items, and the amplitude and latency of the P300 were analyzed. Figure 42.18 shows both reaction time and P300 latency for one to six items held in memory. Although both showed a generally linear increase with workload, the P300 latency revealed less deviation from linearity than reaction time. P300, therefore, appears able to index the cognitive workload involved in a memory-scanning task, at least insofar as this load is reflected in stimulus evaluation processes.

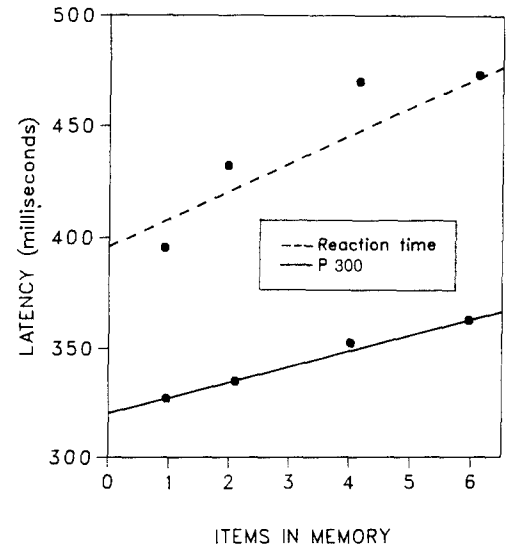


Figure 42.18. Reaction time and P300 latency in a memory-scanning task. The Sternberg (1969) paradigm was used to generate transient evoked responses to postive set probe items, and P300 latencies in subjects were averaged. Reaction time differences over memory sets were significant ($p < .01$), with 80% of the variance accounted for by the linear component. P300 differences were significant over memory sets ($p < .01$), with 99% of the variance accounted for by the linear component. Results indicate that P300 discriminates between small differences in memory-scanning workload at the levels tested and, compared to reaction time, is more linearly related to the number of items to be remembered. (From F. E. Gomer, R. D. Spicuzza, & R. D. O'Donnell, Evoked potential correlates of visual item recognition during memory-scanning tasks. *Physiological Psychology*, 1976, 4. Reprinted with permission.)

Although the sensitivity of the evoked response to a primary stimulus has not been defined for many tasks, it would appear that it is at least able to discriminate between memory load differences as low as one or two letters of the alphabet. This would suggest good sensitivity for the measure. However, it is unclear as yet whether such sensitivity will be reflected at higher loading levels. Until such data are provided, the use of transient evoked response and P300 analyses to primary task measures can be recommended only for assessing lower levels of memory and processing load, which can be expected to affect stimulus evaluation time.

5.2.3. Other EEG Analyses. Although the transient evoked response has shown actual utility in assessing workload, two other procedures are in early stages of development and are discussed because of this potential future utility as workload metrics. Both are attractive on theoretical and practical grounds and are being actively investigated.

5.2.3.1. The Steady-State Evoked Response. The "steady-state" evoked response is generated from a relatively rapid presentation of the stimulus (typically greater than four per second). In this case, before the brain's response to one stimulus can totally die out, a second stimulus occurs. The resulting evoked response, therefore, represents that portion of brain activity entrained to the evoking stimuli. If this procedure is carried on for a long enough period of time, a steady state evolves, and the evoked response describes this condition (Regan, 1977; Spekreijse, 1973).

The steady-state response of the brain has been demonstrated to assess sensory function with good sensitivity (Marg, Freeman, Peltzman, & Goldstein, 1976; Regan, 1977). It can be generated foveally or peripherally, with patterned or unpatterned stimuli, and can be isolated from the human visual

system at frequencies well above the critical flicker fusion point (Moise, 1978; Spekreijse, 1973). These characteristics make the steady-state evoked response extremely attractive as a potential measure of sensory loading.

Attempts to demonstrate the application of steady-state procedures to workload assessment (O'Donnell, 1983) use a technique first described by Regan (1977) for calculation of the apparent transmission speed of the visual system. If the phase lag between the input stimulus and the evoked response is calculated at several temporal frequencies, the resulting phase versus frequency plot can be used to provide an estimate of neuro/visual transmission time. Quite reliable estimates with good stability over time have been obtained by driving a single light with multiple sine waves, causing a complex flickering pattern (O'Donnell, 1983). The brain is still able to resolve the individual components of the flickering light, and the resulting plot agrees well with estimates obtained by separate flickering sine waves. Using such procedures, updated estimates of neural transmission time can be obtained every 20 seconds. This measure would be expected to change as a function of subject variables (fatigue, injury, drug ingestion, etc.), and it has been suggested as a possible on-line monitoring technique to assess the subjective effects of workload on an operator. It is as yet unclear whether the steady-state evoked response will ultimately provide a good on-line measure of workload factors. Its diagnosticity, sensitivity, and even its validity as a workload measure are untested. However, it is included here because of its ease of implementation and low intrusiveness, and because of its practical significance if indeed it is ultimately validated as a measure of factors influencing workload.

5.2.3.2. Multiple-Site Recording. Multiple resource theory is consistent with the postulation that various brain areas are differentially involved in specific resources. Multiple-site recordings of either the overall EEG or cortical evoked responses should, therefore, be capable of revealing those areas that are active in specific tasks and should permit identification of specific resources being utilized in the task. For example, hemispheric specialization of some degree has been well established, with performance on verbal tasks grossly related to the left hemisphere and spatial tasks activating the right hemisphere (Doyle, Ornstein, & Galin, 1974). Studies attempting to define this resource specificity have been carried out, but have been criticized on methodological grounds (Donchin, Kutas, & McCarthy, 1976). It has become clear that the complexity of cortical processing will require complex analysis of a large number of sites to make realistic inferences of the spatiotemporal activity of the brain during performance of specific tasks. In addition, experimental techniques to ensure that tasks used to develop such inferences are truly tapping single resources will be extremely difficult. Efforts are under way to develop complex algorithms to carry out such spatiotemporal analyses (Gevins, 1983; Gevins, Doyle, Cutillo, Schaffer, Lannehill, Ghannam, Gilcrease, & Yeager, 1981). These are extremely cumbersome and, at the moment, EEG techniques to probe specifically for the cortical manifestation of resource utilization must be considered experimental and highly tentative. Again, however, they provide a theoretically attractive approach and, for this reason, should continue to be developed.

5.3. Measure of Eye Function

Since the eye is an important source of information input to the individual, and since it is readily accessible to observation, a large number of visual functions have been studied as potential workload assessment techniques. Procedures that have been developed to measure eye movements and other parameters are discussed extensively in Young and Sheena (1975). Table 42.11 summarizes the parameters of significance for each of these techniques. Hallett (Chapter 10) also summarizes information on techniques for measuring eye movements.

In workload-related reserch, the corneal reflex, EOG, and the pupil-center-corneal reflection distance techniques have been most frequently employed. Of these procedures, it can be seen that the EOG technique can be expected to provide low intrusiveness, high operator acceptance, and minimal implementation requirements. However, vertical and horizontal accuracy tends to be somewhat lower than that seen with other techniques, and calibration can be cumbersome. The pupil-center-corneal-reflex camera is highly accurate but more expensive and elaborate system. However, it does have good operator acceptance, since head movements are permitted and since required subject training and cooperation are reasonably low.

5.3.1. Pupillary Response. The pupil of the eye shows small but highly consistent variations in size as a function of several variables (Hess, 1965). Kahneman and Beatty (1966) demonstrated the value of this technique as a workload assessment device when they showed that the averge pupil diameter changes by as much as 0.6 mm during presentation of seven digits for short-term recall. Pupil diameter reaches maximum size in the period between presentation of the stimulus and the report by the subject, when the memory load is presumably highest. It falls off monotonically to baseline levels as the report is given and the workload is decreased.

Subsequently the sensitivity of pupilometery to workload in other types of tasks was demonstrated (Beatty, 1982) (see Table 42.12, on page 42-42). Semantic difficulty in classifying letter pairs (the Posner paradigm) was reflected in small (less than 0.2 mm) but reliable differences in pupil dilation. Complexity of grammatical and arithmetic reasoning was similarly reflected in increased dilation, as was the difficulty of a perceptual task. Finally, during a sustained attention task, the amount of pupil response to nontarget stimuli decreased from 0.07 mm during the first third of the task to 0.04 mm during the last third. Since these changes paralleled decrements in performance, they offer an attractive index (and perhaps predictor) of performance effects of sustained attention or workload. Beatty (1982) finds these results "physiologically reasonable" and suggests that they are so consistent across different experiments that the pupil size might be used to assess the *relative* workload of very different tasks. In this way workload could be scaled in any task, relative to known levels of other tasks. Such scaling data are not available for any other physiological index of workload. The ability of pupil size to make such fine distinction between workload levels within a task as well as between tasks recommends it as one of the most sensitive workload measures available. However, this very sensitivity generates other problems that limit its use in applied settings.

The implementation requirements of the pupillary measurement technique can be quite severe. Commercial apparatus is available, but extreme care must be exercised in experimental design (Janisse, 1973a, 1973b). It is very difficult to use this measure in applied, nonlaboratory environments because eye movements, changes in ambient lighting, and even emotional effects can cause pupillary responses that are larger than those attributable to workload (Hasett, 1978). The measure, therefore,

appears particularly well suited to the laboratory environment, but has limited utility in any other setting.

In addition to implementation problems, the sensitivity of the measure to various kinds of workload and tasks clearly limits its diagnosticity. For this reason it should be viewed as a global screening device, with little ability to identify the resources utilized in a task. A possible explanation for this resource independence has been suggested by Beatty (1982), who argued that task-evoked pupillary dilation reflects an interaction between the cortex and the reticular activating system during cognitive processing. The reticular system discharge is seen as the principal determinant of pupil response; therefore, this response indexes the individual's general capacity (Kahneman, 1973).

It is not necessary that this particular psychometric interpretation and physiological basis be confirmed before the measure can be used productively in answering workload questions. In fact, the available evidence indicates that pupil size will remain one of the most valuable indices of cognitive workload when used properly in the laboratory. Care in implementation, design, and interpretation is amply rewarded by the sensitivity of measurement, and the technique can be readily recommended as the foundation of a physiological, laboratory-based mental-workload assessment screening.

5.3.2. Eye Point of Regard and Scan Patterns. The absolute position of the eye at any point in time can be used to infer the information required to carry out a task, and many studies have used this type of measure to determine the processing requirements of a task (see Moray, Chapter 40). However, to assess the dynamic moment-to-moment workload, the *pattern* of eye movements in carrying out the task is of greater interest (Krebs & Wingert, 1976; Stern & Bynum, 1970). The hypothesis underlying the study of scan patterns assumes that as the individual's workload increases, time pressure will force modification of the pattern of visual scan. Changes in such information-gathering strategies imply the operator is load-shedding, or otherwise attempting to reduce the overall cognitive load.

Results of scan pattern studies generally reveal that increased workload is reflected in longer dwell times in each position, and the use of a smaller number of display elements. In addition, the pattern of scans becomes much more variable between display elements. These changes appear to be related more to the subject's perception of workload than to the actual load imposed (Dick, 1980). It would appear therefore that changes in scan patterns reflect the subject's response to a *perceived* load and, with proper controls, can be used to differentiate between objective and subjective workload.

Although specific data on the diagnosticity of these measures are not available, it is reasonable to assume that they are relatively global indicators of both perceptual and central-processing load, at least in situations where there are no externally imposed visual-motor output differences. Clearly, to utilize scan patterns as a workload assessment technique, the situation must be structured so that (1) critical information must be gathered from multiple locations, (2) the relative importance of data obtained from each location is different, and (3) the subject can adjust or change the imposed load by a change in strategy. Under such conditions, scan pattern measures can be considered relatively sensitive but minimally diagnostic measures of workload.

5.3.3. Eye Blinks and Movement Speed. Most early studies relating eye blink to workload have been criticized because of problems in design, analysis, or experimental control (Hall & Cusack, 1972). Simple measures of blink frequency per unit time appear to show great variability and must be used in the most rigidly controlled experimental settings. (See Tursky, 1974, for a discussion of instrumentation techniques and problems.) Because of these problems, simple measures of eye blink frequency do not appear promising as workload assessment techniques.

Other types of eye blink analysis have generated considerable interest as useful metrics for assessing longer-term effects of workload (Oster & Stern, 1980). Measures of closure duration and blink pattern have been successful in indexing time-on-task effects that might indirectly reflect levels of workload. In a similar way, the speed of the eye in making a controlled-distance fixation has been suggested as a measure of the same effects. In addition, the frequency of large amplitude eye movements (greater than 9.5°) has been shown to decrease with time-on-task in automobile drivers (Ceder, 1977) and helicopter pilots (Troy, Chen, & Stern, 1972). In all of these measures, however, it is difficult to determine whether the observed effects were truly due to workload differences or simply resulted from changes in motivation or fatigue. Any attempt to utilize this measure as a workload assessment device must clearly consider these possible confounding factors. Lacking such data, eye blinks, closure duration, and speed of eye movement must be considered, at best, global indications of the long-term effects of workload on the individual, rather than as specific diagnostic techniques.

5.4. Measures of Cardiac Function

The electrocardiogram (EKG), blood pressure, blood volume, and oxygen concentration have all been used as physiological indices of performance, stress, or workload (Gunn, Wolf, Block, & Person, 1972). With respect to workload, emphasis has been placed on the cardiac rate itself, since this is obtainable with relatively noninvasive, nonintrusive techniques. Typically, surface electrodes permit identification of the pulse beat, recognizable in an EKG by a typical pattern (the QRS pattern). The number of QRS peaks per unit time constitutes the heart rate. Similarly, the time between successive QRS peaks gives the interbeat time. Absolute heart rate is affected by so many subtle psychological processes that it is probably not useful as a workload measure. However, several studies have suggested that the beat-to-beat variability seen in subjects at rest may measure mental workload (see *Ergonomics*, 1973, *16*, entire issue).

Kalsbeek and Ettema (1963) found decreases in the heart rate variability with increased mental work, and several studies confirmed this general relationship (Kalsbeek, 1971). More complex analysis of the heart beat intervals revealed similarly consistent changes with changes in mental work. Spectral analysis of the interbeat intervals revealed several peak frequencies: a 0.1-Hz (6 cpm) component, a peak between 0.2 and 0.35 Hz corresponding to the respiration frequency of the subject, and a task frequency corresponding to the number of signals to be processed (Mulder & Mulder-Hajonides van der Meulen, 1973). Studies have found the 0.1-Hz component to be highly correlated to the workload of a reaction time task and to fatigue in a driving task (Egelund, 1982; see also Erikson, 1977). Another complex measure, the vector cardiogram, used by Spyker, Stackhouse, Khalafalla, and McLane (1971), and a complex of several variability measures was used successfully to predict workload in a helicopter (Stackhouse, 1976). Results such as

Table 42.11. Comparison of Eye-Movement Measuring Techniques[a]

Method	Measurement Range (degree)		Accuracy		Speed or Frequency Response
	Vertical	Horizontal	Vertical	Horizontal	
Corneal reflex (Mackworth Camera)					
Polymetric Lab V1164	±9	±9	0.5°	0.5°	Photographic rate: 12–64 frames/sec Television: 60 fields/sec
Polymetric Mobile V0165	±10	±10	1°	1°	Same
NAC-REES	±10 20	±10 20	2°	2°	Same
Contact lens with lamp or radiant spot	Both ±10 30 Larger than others	±10 30	Precision 3 sec 15 sec	 3 sec 15 sec	High High
Coil mirror	±10	±10	2 sec	2 sec	High
EOG		±50 80	2°	1.5°	dc or 0.01–15 Hz limited by filtering
Limbus boundary					
Narco Eye Trac	±10		4°	2°	2 msec: 30 msec with recorder
Narco Model 200	−10 −20	±20	2°	1°	4 msec: 26 msec with filtering
Wide-angle Mackworth camera					
Polymetric V1166	40	40	2.5°	2.5°	Same as V1164
Pupil-center-corneal-reflection distance					
Honeywell oculometer	−30 −10	±30	1°	1°	0.1 sec time constant
Whittaker Eye View Monitor	±15 Higher possible	±22	1°	1°	30–60 samples/sec
U.S. Army Human Engineering Lab	30	40	2°	2°	60 samples/sec filtered
Double Purkinje					
Image eye tracker	25°	25°		Noise of 1 min	300 Hz

Interference with Normal Vision	Subject Cooperation Required	Subject Training Required	Calibration and Setup Time	Head Attachments Required
Medium	High	Low	High/Low	Chinrest or biteboard
			High: biteboard	Biteboard
High: Weight on head optics near eye	High	Low	Medium: fit headband, set light source	Head-mounted optics
High: Weight on head	High	Low		
High	High	High	High: lens must be filtered	Contact lens
None	Medium	Low	High: requires electrode stabilization and light adaptation	Yes, 2–6 electrodes
Medium	High	Low	Low	Head bracket and chinrest
Medium	High	Low	Low	Spectacles
Medium Subject looks through apertures; special lighted stimuli are required	High	Low	Low	Viewing through aperture
Low	Low	Low	Low: higher for maximum linearization	None
Low	Low	Low	Low	None
Low	Low	Low	Low	None
Low	Low	Low	Low	Chinrest or biteboard

Source: From Young, L., and Sheena, D. Survey of eye movement recording methods. *Behavior Research Methods and Instrumentation,* 1975, 7, 397–429.

[a]The major techniques used in most workload studies are compared with respect to several relevant factors affecting the practicality and applicability of each.

Table 42.12. Task-Evoked Pupillary Responses Obtained in Several Studies to a Variety of Cognitively Different Tasks

Peak Pupillary Response (mm)	Memory	Language	Reasoning	Perception
0.5			Multiply (hard)	
			Multiply (medium)	
	7 Digits			
		Grammatical		
0.4	6 Digits	reasoning		
	4 Words			
		Word match (hard)	Multiply (easy)	
0.3	5 Digits	Word match (easy)		
	4 Digits			
				Discrimination (hard)
	3 Digits			
0.2		Sentence encode-1		Auditory detection
			Store multi-plicand	
	2 Digits	Single word		
				Discrimination (easy)
		Sentence		
0.1		encode-2		Visual
	1 Digit	Letter match		detection

Maximum pupil dilation during task peformance is shown. In all cases reasonable ordering of the presumed mental workload is achieved with the pupillary measure, leading to the suggestion that pupil response may be used to index workload both within a task and between qualitatively different cognitive tasks. (From J. Beatty, Task-evoked pupillary responses, processing load and the structure of processing resources, Psychological Bulletin, 91. *Copyright 1982 by the American Psychological Association. Reprinted with permission.)*

these have encouraged several investigators, and techniques for obtaining heart rate variability measures in applied, nonlaboratory settings are currently being developed (O'Donnell, 1983).

However, not all attempts to relate cardiac variability to workload have been successful, and there continues to be skepticism concerning its value as a workload measure. A possible explanation for these apparently contradictory findings may lie in the method of calculating cardiac variability. It has been noted (Kalsbeek, 1973) that more than 30 techniques have been presented (see Table 42.13 for a partial list). These differ in the amount of emphasis that they give to volume, amplitude, and timing. Because of this, some analysis techniques are reasonably independent of others, and each analysis could tap different resources in the human. For the present, therefore, heart rate and heart rate variability must be considered an attractive and promising but unvalidated measure of workload. Until studies establish a clearer picture of how heart rate variability changes with different kinds of workload, its use must be considered experimental.

5.5. Measures of Muscle Function

The myoelectric signal generated by the motor units involved in contraction of a muscle can be measured either with needle electrodes placed directly into the muscle, or with surface electrodes placed over or near it (see Basmajian, 1978; Licht, 1971; for reviews of methodology). In most operational situations, needle electrodes are not feasible, whereas surface electromyography (EMG) is a relatively easy procedure.

The amplitude of the EMG signal has been shown to be related to the force exerted by a muscle, at least within certain limits, and to the tension level of the muscle. Because increased muscular tension has been associated with both physical and mental work, the EMG has been proposed as a measure of both kinds of workload.

5.5.1. Physical Work. To assess physical workload, one is interested in the activity of the muscle per se as it resists externally imposed forces. Typically electrodes are placed on the limbs or other major muscle groups. It is generally believed that there is an essentially linear relationship between muscle activity and recorded electrical activity under both isometric and isotonic contraction if certain conditions are met (Basmajian, 1978). Further, as fatigue increases in a muscle, synchronization of motor neurons results in a characteristic change in the EMG spectrum. Lower frequencies tend to become more dominant as motor units fire in more regular "volleys" (O'Donnell, Rapp, & Adey, 1973).

These observations permit assessment of physical workload in at least two ways. First, the absolute force required for an individual to operate a system can be quantified. Thus the motor/strength requirements of the system could be defined. Although it may often be easier to do this by direct behavioral measures (e.g., strain gauges), there are many situations where such intrusion is not possible. For these cases, the EMG provides an ideal alternative. A second procedure would use the fatigue-induced changes in the EMG spectrum as an indicator of the physical workload involved over time. Differences in physical

Table 42.13. Representative Formulae for Calculating Simple Estimates of Heart Rate Variability

Mean beat-to-beat interval

$$\bar{X}_{R\text{-}R} = \frac{\Sigma \bar{X}_i}{N}$$

Variance of beat-to-beat interval

$$S^2_{R\text{-}R} = \frac{\Sigma(X_i - \bar{X})^2}{N}$$

Mean difference in successive beat intervals

$$\bar{D}_{R\text{-}R} = \frac{\Sigma(X_i - X_{i-1})}{N - 1}$$

Variance of difference in successive beat intervals

$$D^2_{R\text{-}R} = \frac{\Sigma(X_i - X_{i-1})^2}{N - 1}$$

Source: From Mulder and Mulder-Hajonides van der Meulen (1973).
X_i = ith beat-to-beat ($R - R$) interval in milliseconds.

workload between systems are revealed by differences in the speed and degree of shift in the spectral characteristics of sequential EMG samples.

Although obvious, the EMG has not been used extensively to measure physical workload. This appears particularly unfortunate since it is a low-intrusion, direct measure that exhibits reasonable stability and sensitivity within a subject, although showing variability between subjects. Unlike some physiological measures, interpretation of the measure is straightforward with respect to the overall workload construct, and the physiological basis for the measure is unambiguous.

5.5.2. Mental Work. For assessing mental workload, the relatively static tension level of a muscle not directly involved in task performance is usually monitored. This may involve placing electrodes on a limb not being used in the task, or on another muscle such as the neck or forehead. General activation theory (Duffy, 1962; Malmo, 1969) predicts that an increase in mental work or stress will be accompanied by a corresponding increase in the EMG tension level. Indeed, a general muscle tension factor appears to exist for muscles in the upper body at least. In practice this means that muscles of the head, neck, shoulder, and forearm should all be sensitive to activation resulting from various types of mental work. There is, however, considerable variability in the EMG absolute values between subjects, limiting this measure to within-subject designs.

Generally, higher tonic EMG levels are found to correlate with higher workloads, at least up to a point on the activation curve (Stern, 1966; Wisner, 1973). However, as with the cardiac measures of workload, negative results have been frequent enough to cast doubt on the universal applicability of the EMG (Jex & Allen, 1970; Spyker et al., 1971). It has been proposed, for instance, that sympathetic nervous system activity may decrease with decreasing vigilance and arousal, but that this is counteracted by a somatic increase as part of the body's efforts to overcome any impeding decrement in performance. The EMG, then, might reflect these contradictory trends in ways that are not yet understood.

In view of this discussion, it is clear that the EMG cannot now be recommended as a simple, diagnostic measure of mental workload. More sophisticated analysis techniques, however, could rapidly change this picture.

6. SUMMARY

This chapter has attempted to bring together the laboratory and field-based techniques currently in use to assess workload. No doubt, many specific procedures of interest to particular applications have been left out of this survey. In no sense is this meant to summarily exclude these from any list of valid workload assessment techniques. In fact, several of these are acknowledged to show considerable promise (e.g., occlusion techniques and respiratory rhythms). They are not discussed here partly because of space limitations and partly because a judgment had to be made concerning the practicality and general applicability of each measure. It is hoped that the inclusion of general references will serve to point the interested reader to the individual techniques not included here.

Similarly, a class of techniques frequently used to assess workload was deliberately excluded from this chapter. Task analytic methods, particularly as they are used with computer models of whole missions or operations (see e.g., Lane, Strieb, Glenn, & Wherry, 1981) constitute an important tool for workload investigations during design and other stages of aircraft and systems development. These techniques, however, are primarily off-line analyses that utilize the kind of laboratory and field data gathered with the techniques such as those described in this chapter. They provide an overall systems answer to the workload question and as such deserve separate treatment from highly specific workload measures. The interested reader is referred to Chubb (1981), Geer (1981), Lane et al. (1981), Parks (1979), and Wherry (1984) for reviews and introductions to some of the modeling techniques used in these areas.

This chapter clearly shows that techniques currently exist for assessing workload sensitively, validly, and reliably. Although it would be naive to claim that these techniques have reached the level of theoretical or practical sophistication desired and necessary, a reasonable assessment indicates that they are capable of forming the basis for standardization in the area. It is most critical that existing techniques be given adequate laboratory and field tests and that investigators become aware of the need for attention to questions of sensitivity, diagnosticity, intrusiveness, implementation, and operator acceptance, as described in Table 42.1. Standardized evaluation and continual refinement of assessment techniques according to these criteria will certainly result in maximum progress toward a comprehensive theory and practical measures of workload.

REFERENCES

Acton, W. H., Crabtree, M. S., & Shingledecker, C. A. Development of a standardized workload metric evaluation methodology. *Proceedings of the IEEE National Aerospace and Electronics Conference*, 1983, 1086–1089.

Allport, D. A., Antonis, B., Reynolds, P. On the division of attention: A disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology*, 1972, *24*, 225–235.

Anderson, P. A., & Toivanen, M. L. *Effects of varying levels of autopilot assistance and workload on pilot performance in the helicopter formation flight mode* (Report No. JANAIR 690610). Minneapolis, Minn.: Honeywell, Inc., March 1970.

Aunon, J. I., McGillem, C. D., & O'Donnell, R. D. Comparison of linear and quadratic classification of event-related potentials on the basis of their exogenous or endogenous components. *Psychophysiology*, 1982, *19*, 531–537.

Bahrick, H. P., Noble, M., & Fitts, P. M. Extra-task performance as a measure of learning a primary task. *Journal of Experimental Psychology*, 1954, *48*, 298–302.

Basmajian, J. V. *Muscles alive. Their functions revealed by electromyography*. Baltimore, Md.: Williams & Wilkins, 1978.

Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 1982, *91*(2), 276–292.

Beatty, J., & Kahneman, D. Pupillary changes in two memory tasks. *Psychonomic Science*, 1966, *55*, 371–372.

Beck, E. C. Electrophysiology and behavior. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual Review of Psychology*, 1975, *26*, 233–262.

Bell, P. A. Effects of noise and heat stress on primary and subsidiary task performance. *Human Factors*, 1978, *20*, 749–752.

Boggs, D. H., & Simon, J. R. Differential effect of noise on tasks of varying complexity. *Journal of Applied Psychology*, 1968, *52*, 148–153.

Borg, G. Subjective aspects of physical and mental load. *Ergonomics*, 1978, *21*, 215–220.

Borg, G., Bratfisch, O., & Dornic, S. *Perceived difficulty of an immediate memory task* (Report No. 15). Stockholm, Sweden: University of Stockholm, Institute of Applied Psychology, 1971. (a)

Borg, G., Bratfisch, O., & Dornic, S. *Perceived difficulty of a visual search task* (Report No. 16). Stockholm, Sweden: University of Stockholm, Institute of Applied Psychology, 1971. (b)

Bratfisch, O. *Experienced intellectual activity and perceived difficulty of intelligence tests* (Report No. 30). Stockholm, Sweden: University of Stockholm, Institute of Applied Psychology, 1972.

Bratfisch, O., Borg, G., & Dornic, S. *Perceived item difficulty in three tests of intellectual performance capacity* (Report No. 29). Stockholm, Sweden: University of Stockholm, Institute of Applied Psychology, 1972.

Bratfisch, O., Dornic, S., & Borg, G. *Perceived difficulty of a motor skill task as a function of training* (Report No. 11). Stockholm, Sweden: University of Stockholm, Institute of Applied Psychology, 1970.

Brecht, M. *Cardiac arrhythmia and secondary tasks as measures of mental load.* Unpublished master's thesis, California State University at Northridge, 1977.

Broadbent, D. *Perception and communication.* Oxford: Pergamon, 1958.

Brown, I. D. The measurement of perceptual load and reserve capacity. *Transactions of the Association of Industrial Medical Officers*, 1964, *14*, 44–49.

Brown, I. D. A comparison of two subsidiary tasks used to measure fatigue in car drivers. *Ergonomics*, 1965, *8*, 467–473.

Brown, I. D. Dual task methods of assessing workload. *Ergonomics*, 1978, *21*, 221–224.

Burke, M. W., Gilson, R. D., & Jagacinski, R. J. Multimodal information processing for visual workload relief. *Ergonomics*, 1980, *23*, 961–975.

Callaway, E., Teuting, P., & Koslow, S. (Eds.), *Brain event-related potentials in man.* New York: Academic, 1978.

Casali, J. G. *A sensitivity/intrusion comparison of mental workload estimation techniques using a simulated flight task emphasizing perceptual pilot behaviors.* Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, 1982.

Casali, J. G., & Wierwille, W. W. A sensitivity/intrusion comparison of mental workload estimation techniques using a flight task emphasizing perceptual piloting activities. *Proceedings of the IEEE International Conference on Cybernetics and Society*, 1982, 598–602.

Casali, J. G., & Wierwille, W. W., Communications-imposed pilot workload: A comparison of sixteen estimation techniques. *Proceedings of Second Ohio State University Symposium on Aviation Psychology*, 1983, 223–235.

Ceder, A. Driver's eye movements as related to attention in simulated traffic flow conditions. *Human Factors*, 1977, *19*, 571–581.

Chiles, W. D. Workload, task, and situational factors as modifiers of complex human performance. In E. A. Alluisi & E. A. Fleishman (Eds.), *Human performance and productivity.* Hillsdale, N.J.: Erlbaum, 1982.

Chiles, W. D., & Alluisi, E. A. On the specification of operator or occupational workload with performance measurement methods. *Human Factors*, 1979, *21*, 515–528.

Chiles, W. D., Alluisi, E. A., & Adams, O. S. Work schedules and performance during confinement. *Human Factors*, 1968, *10*, 143–196.

Chubb, G. P. SAINT, A digital simulation language for the study of manned systems. In J. Morssel & K. F. Kraiss (Eds.), *Manned systems design methods, equipment, and applications.* New York: Plenum, 1981.

Clement, W. F. Investigating the use of a moving map display and a horizontal situation indicator in simulated powered-lift short-haul operations. *Proceedings of the Twelfth Annual NASA University Conference on Manual Control*, University of Illinois, May 1976, 201–224.

Clement, W. F., McRuer, D. R., & Klein, R. H. Systematic manual control display design. *Proceedings of the AGARD Conference on Guidance and Control Displays* (AGARD CP-96), February 1972, 6/1-6/10.

Coombs, C. H., Dawes, R. M., & Tversky, A. *Mathematical psychology: An elementary introduction.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Cooper, G. E. Understanding and interpreting pilot opinion. *Aeronautica Engineering Review*, 1957, *16*, 47–52.

Cooper, G. E., & Harper, R. P., Jr. *The use of pilot rating in the evaluatio of aircraft handling qualities* (Report No. NASA TN-D-5153). Mo fett Field, CA: Ames Research Center, National Aeronautics an Space Administration, 1969.

Crabtree, M. S. *Human factors evaluation of several control system con figurations, including workload sharing with force wheel steerin during approach and flare* (Report No. AFFDL-TR-75-43). Wright Patterson Air Force Base, Ohio: USAF Flight Dynamics Laborator April 1975.

Crawford, B. M., Pearson, W. H., & Hoffman, M. *Multipurpose digita switching and flight control workload* (Report No. AMRL-TR-78 43). Wright-Patterson Air Force Base, Ohio: USAF Aerospac Medical Research Laboratory, December 1978.

D'Amato, M. R. *Experimental psychology: Methodology psychophysic and learning.* New York: McGraw-Hill, 1970.

Damos, D. L. The development and transfer of time-sharing skills *Proceedings of the Human Factors Society Twenty-First Annua Meeting*, San Francisco, October 1977, 53–57.

Daryanian, B. *Subjective scaling of mental workload in a multitas environment.* Unpublished master's thesis, Massachusetts Institut of Technology, 1980.

Dick, A. O. *Instrument scanning and controlling: Using eye movemen data to understand pilot behavior and strategies* (Report No. NASA CR-3306). Langley Air Force Base, VA: National Aeronautics an Space Administration, 1980.

Donchin, E. The relationship between P300 and the CNV (a corre spondence). In W. C. McCallum & J. R. Knott (Eds.), *The responsiv brain.* Bristol, England: John Wright, 1976.

Donchin, E. Event-related brain potentials: A tool in the study of human information processing. In H. Begleiter (ed.), *Evoked potentials in psychiatry.* New York: Plenum, 1981.

Donchin, E., & Herning, R. I. A simulation study of the efficiency o stepwise discriminant analysis in the detection and comparison o event-related potentials. *Electroencephalography and Clinical Neurophysiology*, 1975, *38*, 51–68.

Donchin, E., Kutas, M., & McCarthy, G. Electrocortical indices of hemispheric utilization. In S. Harnad (Ed.), *Lateralization in the nervous system.* New York: Academic, 1976.

Donnell, M. L. *An application of decision-analytic techniques to the tes and evaluation phase of a major air system: Phase III* (Report No TR-PR-79-6-91). McLean, Va.: Decisions and Designs, Inc., May 1979.

Donnell, M. L., Adelman, L., & Patterson, J. F. *A systems operability measurement algorithm (SOMA): Application, validation, and ex tensions* (Report No. TR-81-11-156). McLean, Va.: Decisions and Designs, Inc., April 1981.

Donnell, M. L., & O'Connor, M. F. *The application of decision analyti techniques to the test and evaluation phase of the acquisition of a major air system: Phase II* (Report No. TR-78-3-25). McLean, Va. Decisions and Designs, Inc., April 1978.

Dorfman, P. W., & Goldstein, I. L. Spatial and temporal information cues in a time-sharing task. *Journal of Applied Psychology*, 1971 *55*, 554–558.

Dorfman, P. W., & Goldstein, I. L. The effects of task coherency, preview and speed-stress in timing and anticipation. *Journal of Motor Behavior*, 1975, *7*, 45–55.

Dornic, S. Language dominance, spare capacity, and perceived effort in bilinguals. *Ergonomics*, 1980, *23*, 366–377. (a)

Dornic, S. *Spare capacity and perceived effort in information processing* (Report No. 567). Stockholm, Sweden: University of Stockholm, Department of Psychology, December 1980. (b)

Dornic, S., & Andersson, O. *Difficulty and effort: A perceptual approach* (Report No. 566). Stockholm, Sweden: University of Stockholm, Department of Psychology, November 1980.

Dornic, S., Bratfisch, O., & Larsson, T. *Perceived difficulty in verbal learning* (Report No. 41). Stockholm, Sweden: University of Stockholm, 1973.

Dornic, S., Sarnecki, M., & Svensson, J. *Perceived difficulty, learning time, and subjective certainty in a perceptual task* (Report No. 43). Stockholm, Sweden: University of Stockholm, Institute of Applied Psychology, 1973.

Dougherty, D. J., Emery, J. H., & Curtin, J. G. *Comparison of perceptual workload in flying standard instrumentation and the contact analog vertical display* (Report No. JANAIR D228-421-019). Fort Worth, Tex.: Bell Helicopter Company, December 1964.

Doyle, J. C., Ornstein, R., & Galin, D. Lateral specialization of cognitive mode: II. EEG frequency analysis. *Psychophysiology*, 1974, *11*, 567–578.

Duffy, E. *Activation and behavior*. New York: Wiley, 1962.

Duncan-Johnson, C. C., & Donchin, E. On quantifying surprise. The variation in event-related potentials with subjective probability. *Psychophysiology*, 1977, *14*, 456–467.

Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.

Egelund, N. Spectral analysis of heart rate variability as an indicator of driver fatigue. *Ergonomics*, 1982, *25*, 663–672.

Eggemeier, F. T. Workload metrics for system evaluation. *Proceedings of the Defense Research Group Panel VIII Workshop "Application of System Ergonomics to Weapon System Development,"* Shrivenham, England, 1984, C/5–C/20.

Eggemeier, F. T., Crabtree, M. S., & LaPointe, P. A. The effect of delayed report on subjective ratings of mental workload. *Proceedings of the Human Factors Society Twenty-Seventh Annual Meeting*, 1983, 139–143.

Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., & Shingledecker, C. A. Subjective workload assessment in a memory update task. *Proceedings of the Human Factors Society Twenty-Sixth Annual Meeting*, 1982, 643-647.

Ellis, G. A. Subjective assessment pilot opinion measures. In A. H. Roscoe (Ed.), *Assessing Pilot Workload* (Report No. AGARD-AG-233). Neuilly-sur-Seine, France: Advisory Group on Aerospace Research and Development, February 1978.

Erikson, C. G. On the psychophysiology of heart rhythms. *Goteborg Psychological Reports*, 1977, *7* (3).

Eysenck, M. W., & Eysenck, M. C. Processing depth, elaboration of encoding, memory stores, and expended processing capacity. *Journal of Experimental Psychology: Human Learning and Memory*, 1979, *5*, 472–484.

Finkelman, J. M., & Glass, D. C. Reappraisal of the relationship between noise and human performance by means of a subsidiary task measure. *Journal of Applied Psychology*, 1970, *54*, 211–213.

Finkelman, J. M., Zeitlin, L. R., Filippi, J. A., & Friend, M.A. Noise and driver performance. *Journal of Applied Psychology*, 1977, *62*, 713–718.

Fisk, A. D., Derrick, W. L., & Schneider, W. *The use of dual task paradigms in memory research: A methodological assessment and evaluation of effort as a measure of levels of processing* (Report No. HARL-ONR-8105). Champaign, Ill.: University of Illinois, Human Attention Research Laboratory, Psychology Department, March 1982.

Fournier, B. A., & Stager, P. Concurrent validation of a dual-task selection test. *Journal of Applied Psychology*, 1976, *61*, 589–595.

Friedman, A., Polson, M. C., Dafoe, C. G., & Gaskill, S. J. Dividing attention within and between hemispheres: Testing a multiple resources approach to limited capacity information processing. *Journal of Experimental Psychology*, 1982, *8*, 625–650.

Gartner, W. B., & Murphy, M. R. *Pilot workload and fatigue: A critical survey of concepts and assessment techniques* (Report No. NASA-TN-D-8365). Washington, D.C.: National Aeronautics and Space Administration, November 1976.

Geer, C. W. *Human engineering procedures guide* (Report No. AFAMRL-TR-81-35). Wright-Patterson Air Force Base, Ohio: Air Force Aerospace Medical Research Laboratory, 1981.

Gevins, A. Brain potential evidence for lateralization of higher cognitive functions. In J. B. Healige (Ed.), *Asymmetry: Method, theory and application*. New York: Praeger, 1983.

Gevins, A., Doyle, J., Cutillo, B., Schaffer, R., Lannehill, R., Ghannam, J., Gilcrease, V., & Yeager, C. New method reveals dynamic patterns of correlation of human brain electrical potentials during cognition. *Science*, 1981, *213*, 918–922.

Goldstein, I. L., & Dorfman, P. W. Speed and load stress as determinants of performance in a time-sharing task. *Human Factors*, 1978, *20*, 603–609.

Gomer, F. E., Spicuzza, R. J., & O'Donnell, R. D. Evoked potential correlates of visual item recognition during memory-scanning tasks. *Physiological Psychology*, 1976, *4*, 61–65.

Gopher, D. Human performance and residual capacity. *Proceedings of the Airline Pilots Association Symposium on Man-System Interface: Advances in Workload Study*, Washington, D.C., July 1978, 6–20.

Gopher, D., Brickner, M., & Navon, D. Different difficulty manipulations interact differently with task emphasis: Evidence for multiple resources. *Journal of Experimental Psychology: Human Perception and Performance*, 1982, *8*, 146–157.

Gopher, D., & North, R. A. Manipulating the conditions of training in time-sharing performance. *Human Factors*, 1977, *19*, 583–593.

Gopher, D., & Sanders, A. F. *S-Oh-R: Oh stages! Oh resources!* (Report No. HEIS-82-8). Haifa, Israel: Technion, Israel Institute of Technology, Research Center for Work Safety and Human Engineering, 1982.

Green, R., & Flux, R. Auditory communication and workload. *Proceedings of the AGARD Conference on Methods to Assess Workload* (AGARD-CPP-216), April 1976, A4/1–A4/8.

Gunn, C. G., Wolf, S., Block, R. T., & Person, R. J. Psychophysiology of the cardiovascular system. In N. S. Greenfield & R. A. Sternback (Eds.), *Handbook of psychophysiology*. New York: Holt, Rinehart & Winston, 1972.

Gunning, D. Time estimation as a technique to measure workload. *Proceedings of the Twenty-Second Annual Meeting of the Human Factors Society*, 1978, 41–45.

Hall, R. J., & Cusack, B. L. *The measurement of eye behavior: Critical and selected reviews of voluntary eye movement and blinking* (U.S. Army Technical Memorandum 18-72). Aberdeen Proving Ground, Maryland: Human Engineering Laboratory, 1972.

Hallsten, L., & Borg, G. *Six rating scales for perceived difficulty* (Report No. 58). Stockholm, Sweden: University of Sweden, Institute of Applied Psychology, 1975.

Hart, S. G. Time estimation as a secondary task to measure workload. *Proceedings of the Eleventh Annual Conference on Manual Control* (Report No. NASA TMX-62). Moffett Field, CA, National Aeronautics and Space Administration, Ames Research Center, May 1975.

Hart, S. G. Subjective time estimation as an index of workload. *Proceedings of the Airline Pilots Association Symposium on Man-System Interface: Advances in Workload Study*, Washington, D.C., July 1978, 115–131.

Hassett, J. *A primer of psychophysiology*. San Francisco: Freeman, 1978.

Hawkins, H. L., & Ketchum, D. *The case against secondary task analyses of mental workload* (Report for Contract No. N0014-77-C-0643). Arlington, Va.: Office of Naval Research, January 1980.

Helm, W. R. Psychometric measures of task difficulty under varying levels of information load. *Proceedings of the Human Factors Society Twenty-Fifth Annual Meeting*, 1981, 518–521.

Helm, W. R., & Donnell, M. L. *Mission operability assessment technique: A system evaluation methodology*. (Technical Publication No. TP-79-31). Point Magu, Cal.: Pacific Missile Test Center, October 1979.

Helm, W. R., & Heimstra, N. W. *The relative efficiency of psychometric measures of task difficulty and task performance in predicting task performance* (Report No. HFL-81-5). Vermillion, S.D.: University of South Dakota, Human Factors Laboratory, Psychology Department, August, 1981.

Hess, E. H. Attitude and pupil size. *Scientific American*, 1965, *212*, 46–54.

Hess, R. A. Prediction of pilot opinion ratings using an optimal pilot model. *Human Factors*, 1977, *19*, 459–476.

Hicks, T. G., & Wierwille, W. W. Comparison of five mental workload

Schneider, W., & Shiffrin, R. M. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 1977, *84*, 1–66.

Schori, T. R. A comparison of visual, auditory, and cutaneous tracking displays when divided attention is required to a cross-adaptive loading task. *Ergonomics*, 1973, *16*, 153–158.

Schori, T. R. & Jones, B. W. Smoking and workload. *Journal of Motor Behavior*, 1975, *7*, 113–120.

Schultz, W. C., Newell, F. D., & Whitbeck, R. F. A study of relationships between aircraft system performance and pilot ratings. *Proceedings of the Sixth Annual NASA University Conference on Manual Control*, Wright-Patterson Air Force Base, Ohio, April 1970, 339–340.

Schwartz, S. P. Capacity limitations in human information processing. *Memory and Cognition*, 1976, *4*, 763–768.

Senders, J. W. The estimation of operator workload in complex systems. In K. B. DeGreene (Ed.), *Systems psychology*. New York: McGraw-Hill, 1970.

Sheridan, T. B. Mental workload: What is it? Why bother with it? *Human Factors Society Bulletin*, 1980, *23*, 1–2.

Sheridan, T. B. & Simpson, R. W. *Toward the definition and measurement of the mental workload of transport pilots*. (FTL Report No. R79-4). Cambridge, Mass.: Massachusetts Institute of Technology, Flight Transportation Laboratory, January 1979.

Shiffrin, R. M., & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 1977, *84*, 127–190.

Shingledecker, C. A. Enhancing operator acceptance and noninterference in secondary task measures of workload. *Proceedings of the Human Factors Society Twenty-Fourth Annual Meeting*, 1980, 674–677.

Shingledecker, C. A. Behavioral and subjective workload metrics for operational environments. *Proceeding of the AGARD (AMP) Symposium Sustained Intensive Air Operations: Physiological and Performance Aspects* (AGARD-CP-338). November 1983, 6/1–6/10.

Shingledecker, C. A., Acton, W. H., & Crabtree, M. S. *Development and application of a criterion task set for workload metric evaluation*. (Paper No. 831419). Warrendale, Pa.: Society of Automotive Engineers, SAE Technical Paper Series, October 1983.

Shingledecker, C. A., & Crabtree, M. S. *Subsidiary radio communications tasks for workload assessment in R&D simulations: II. Task sensitivity evaluation* (Report No. AFAMRL-TR-82-57). Wright-Patterson Air Force Base, Ohio: U.S. Air Force Aerospace Medical Research Laboratory, September 1982.

Shingledecker, C. A., Crabtree, M. S., & Acton, W. H. Standardized tests for the evaluation and classification of workload metrics. *Proceedings of the Human Factors Society Annual Meeting*, 1982, 648–651.

Shingledecker, C. A., Crabtree, M. S., Simons, J. C., Courtright, J. F., & O'Donnell, R. D. *Subsidiary radio communications tasks for workload assessment in R&D simulations: I. Task development and workload scaling* (Report No. AFAMRL-TR-80-126). Wright-Patterson Air Force Base, Ohio: U.S. Air Force Aerospace Medical Research Laboratory, December 1980.

Skelly, J. J., Reid, G. B., & Wilson, G. R. *B-52 full mission simulation: Subjective and physiological workload applications*. Paper presented at the Second Aerospace Behavioral Engineering Technology Conference, Long Beach Cal., 1983.

Spekreijse, H. Contrast evoked responses in man. *Vision Research*, 1973, *13*, 1577–1601.

Sperandio, J. C. Variation of operator's strategies and regulating effects on workload. *Ergonomics*, 1971, *14*, 571–577.

Sperandio, J. C. The regulation of working methods as a function of workload among air traffic controllers. *Ergonomics*, 1978, *21*, 193–202.

Spyker, D. A., Stackhouse, S. P., Khalafalla, A. S., & McLane, R. C. *Development of techniques for measuring pilot workload* (Report No. NASA CR-1888). Washington, D.C.: National Aeronautics and Space Administration, November 1971.

Squires, K. C., & Donchin, E. Beyond averaging: The use of discriminant functions to recognize event-related potentials elicited by single auditory stimuli. *Electroencephalography and Clinical Neurophysiology*, 1976, *1*, 449–459.

Squires, K. C., Wickens, C., Squires, N. K., & Donchin, E. The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*, 1976, *193*, 1142–1146.

Stackhouse, S. P. *The measurement of pilot workload in manual control systems* (Report No. F0398 FR1). Minneapolis, Minn.: Honeywell, January 1976.

Stern, J. A., & Bynum, J. A. Analysis of visual search activity in skilled and novice helicopter pilots. *Aerospace Medicine*, 1970, *41*, 330–335.

Stern, R. M. Performance and physiological arousal during two vigilance tasks varying in signal presentation rate. *Perceptual and Motor Skills*, 1966, *23*, 691–700.

Sternberg, S. High-speed scanning in human memory. *Science*, 1966, *153*, 652–654.

Sternberg, S. The discovery of processing stages: Extension of Donder's method. In W. G. Koster (Ed.), *Attention and performance II*. Amsterdam: North-Holland, 1969.

Stevens, S. S. Problems and methods of psychophysics. *Psychological Bulletin*, 1958, *55*, 177–196.

Stevens, S. S. *Psychophysics*. New York: Wiley, 1975.

Sutton, S., Tueting, P., Zubin, J., & John, E. R. Information delivery and the sensory evoked potential. *Science*, 1967, *155*, 1436–1439.

Tole, J. R., Stephens, A. T., Harris, R. L., & Eprath, A. Quantification of workload via instrument scan. *Proceedings of the Workshop on Flight Testing to Identify Pilot Workload and Pilot Dynamics* (Report No. AFFTC-TR-82-5). Edwards Air Force Base, California: Air Force Flight Test Center, May 1982, 234–250.

Troy, M. E., Chen, S. C., & Stern, J. A. Computer analysis of eye movement patterns during visual search. *Aerospace Medicine*, 1972, *43*, 390–394.

Tursky, B. Recording of human eye movements. In R. F. Thompson and M. M. Patterson (Eds.), *Bioelectric recording techniques* (Part C). New York: Academic, 1974.

Tversky, A., & Krantz, D. H. Similarity of schematic faces: A test of interdimensional additivity. *Perception and Psychophysics*, 1969, *5*, 124–128.

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. D. Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, 1979, *5*, 607–617.

Welford, A. T. Mental workload as a function of demand, capacity, strategy, and skill. *Ergonomics*, 1978, *21*, 151–167.

Wetherell, A. The efficacy of some auditory-vocal subsidiary tasks as measures of the mental load on male and female drivers. *Ergonomics*, 1981, *24*, 197–214.

Wewerinke, P. H. Human operator workload for various control situations. *Proceedings of the Tenth Annual Conference on Manual Control*, Wright-Patterson Air Force Base, Ohio, 1974, 167–192.

Wherry, R. J. Prediction of human and system performance and effectiveness. *Proceedings of the Defense Research Group Panel VIII Workshop: Applications of System Ergonomics to Weapon System Development*, Shrivenham, England, 1984, C/37–C/58.

Whitaker, L. A. Dual-task interference as a function of cognitive load processing. *Acta Psychologica*, 1979, *43*, 71–84.

Wickens, C. D. Measures of workload, stress, and secondary tasks. In N. Moray (Ed.), *Mental workload: Its theory and measurement*. New York: Plenum, 1979.

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

Wickens, C. D. Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention*. New York: Academic, 1984. (a)

Wickens, C. D. *Engineering psychology and human performance*. Columbus, Ohio: Merrill Publishing 1984. (b)

Wickens, C. D., & Derrick, W. Workload measurement and multiple resources. *Proceedings of the 1981 IEEE Conference on Cybernetics*

and Society, 1981, 600–603. (a)

Wickens, C. D., & Derrick, W. *The processing demands of second order manual control: Application of additive factors methodology* (Report No. EPL-81-1/ONR-81-1). Champaign, Ill.: University of Illinois, Engineering Psychology Laboratory, January 1981. (b)

Wickens, C. D., Isreal, J., & Donchin, E. The event-related cortical potential as an index of task workload. *Proceedings of the Twenty-First Annual Meeting of the Human Factors Society*, San Francisco, 1977.

Wickens, C. D., & Kessel, C. The effect of participatory mode and task workload on the detection of dynamic system failures. *IEEE Transactions on Systems, Man, & Cybernetics*, 1979, *13*, 21–31.

Wickens, C. D., & Kessel, C. The processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, *6*, 564–577.

Wickens, C. D., Mountford, S. J., & Schreiner, W. Multiple resources, task-hemispheric integrity, and individual differences in time sharing. *Human Factors*, 1981, *23*, 211–229.

Wickens, C. D., & Yeh, Y. Y. The dissociation of subjective ratings and performance. *Proceedings of the IEEE International Conference on Cybernetics and Society*, Seattle, Wash., October 1982, 584–587.

Wickens, C. D., & Yeh, Y. Y. The dissociation of subjective ratings and performance: A multiple resources approach. *Proceedings of the Human Factors Society Twenty-Seventh Annual Meeting*, October 1983, 244–248.

Wierwille, W. W. Physiological measures of aircrew mental workload. *Human Factors*, 1979, *21*, 575–593.

Wierwille, W. W., & Casali, J. G. *The sensitivity and intrusion of mental workload estimation techniques in piloting tasks* (Report No. 8309). Blacksburg, VA: Virginia Polytechnic Institute and State University, Vehicle Simulation Laboratory, Department of Industrial Engineering and Operations Research, September 1983. (a)

Wierwille, W. W., & Casali, J. G. A validated rating scale for global mental workload measurement applications. *Proceedings of the Human Factors Society Twenty-Seventh Annual Meeting*, 1983, 129–133. (b)

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

Wierwille, W. W., & Williges, R. C. *Survey and analysis of operator workload assessment techniques* (Report No. 2-78-101). Blacksburg, Va.: Systemetric Corporation, September 1978.

Wierwille, W. W., & Williges, R. C. *An annotated biliography on operator mental workload assessment* (Report No. SY-27R-80). Patuxent River, Md.: Naval Air Test Center, March 1980.

Williges, R. C., & Wierwille, W. W. Behavioral measures of aircrew mental workload. *Human Factors*, 1979, *21*, 549–574.

Wisner, A. Electrophysiological measures for tasks of low energy expenditure. In W. T. Singleton, J. G. Fox, & D. Whitfield (Eds.), *Measurement of man at work*. London: Taylor & Francis, 1973.

Wolfe, J. D. *Crew workload assessment: Development of a measure of operator workload* (Report No. AFDL-TR-78-165). Wright-Patterson Air Force Base, Ohio: Air Force Flight Dynamics Laboratory, December 1978.

Young, L., & Sheena, D. Survey of eye movement recording methods. *Behavior Research Methods and Instrumentation*, 1975, *7*, 397–429.

Zeitlin, L. R., & Finkelman, J. M. Research note: Subsidiary task techniques of digit generation and digit recall as indirect measures of operator loading. *Human Factors*, 1975, *17*, 218–220.