General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Bolt Beranek and Newman Inc.

Report No. 4308

Physiological Correlates of Mental Workload

Final Report

Greg L. Zacharias

February 1980

Prepared for:
National Aeronautics and Space Administration
Langley Research Center

Report No. 4308

Physiological Correlates of Mental Workload

Contract No. NAS1-15192

Final Report

Greg L. Zacharias

February 1980

Prepared for:

National Aeronautics and Space Administration
Langley Research Center

Table of Contents

ABSTRACT

A literature review was conducted to assess the basis of and techniques for physiological assessment of mental workload. An assumption common to the studies reviewed is that a well-motivated "worker" undergoes a measurable change in physiological state with a change in the imposed workload. Most of the study findings support this notion, but all those reviewed had shortcomings involving one or more of the following basic problems: a) physiologic arousal can be easily driven by non-workload factors, confounding any proposed metric; b) the profound absence of underlying physiologic models has promulgated a multiplicity of seemingly arbitrary signal processing techniques; c) the unspecified multidimensional nature of physiological "state" has given rise to a broad spectrum of competing non-commensurate metrics; and d), the lack of an adequate definition of workload compels physiologic correlations to suffer either from the vagueness of implicit workload measures or from the variance of explicit subjective assessments.

Using specific studies as examples, this review discusses tw. basic signal-processing/data-reduction techniques in current use: time- and ensemble-averaging. Covered are methods of implementation, and their relative advantages and shortcomings. Methods for analyzing the reduced data are also discussed, with an emphasis on feature extraction, its current common use in statistical validation of measurement sensitivity, and its potential use in developing functional models of physiologic sensitivity to imposed workload. The current trend toward multi-channel physiologic monitoring and feature extraction is also evaluated, in terms of both potential sensitivity enhancement and degraded statistical reliability.

PREFACE

This report is submitted in partial fulfillment of the requirements of Contract No. NAS1-15192 entitled "Performance Measurement for Simulation Design, Evaluation and Planning". The contract was performed for the NASA Langley Research Center. Mr. George Steinmetz served as technical monitor.

1.0 INTRODUCTION

Considerable effort over the past thirty years has been devoted to understanding the relationship between a human operator's mental workload and his physiological "state". Although much of the work is aimed at a basic understanding of psychophysiological processes and interactions, a portion of the research has been devoted to the development of a reliable estimator of workload, one based solely on physiological measurements. The basic premise underlying this effort is that a well-motivated operator should undergo a change in physiological state with a change in the imposed workload, and that this state change should be reflected in one or more physiological measurements made on the operator. Of course, this is a highly simplified statement of the relationship between workload and physiological state, and one could append a number of definitions and qualifiers. In fact, part of this chapter will be devoted to just such an expansion of this simplified statement; the essence of the argument, however, is that the close connection between workload, physiological state, and physiological measurements should make it possible to develop a workload indicator which is a function solely of physiological measurements.

At the outset, it is appropriate to note that no such workload indicator has found widespread application to the

practical problems arising in studies of the human operator. This could be due to a number of reasons. First, one might suppose that the human operator research community is unaware of the existence of such indicators. This is not the case, however, since many of the recent workload studies have been conducted by workers active in human operator research, as evidenced by reports by Benson et al (1965), Jex and Allen (1970), Spyker et al (1971), and Wickens et al (1976). A second reason might be that the workload indicators which have been proposed from time to time may be of only limited utility, applicable only to well-controlled and unambiguous laboratory situations, and thus of little use in evaluation, for example, of cockpit workload. Whether this is true or not is difficult to tell from the literature, since "unsuccessful" workload indicators are rarely reported on, while critics of "successful" indicators can point to literally an infinite number of workload situations which have not been used to validate an indicator's accuracy. A third reason for the apparent lack of proven physiological workload indicators might be due to the relative newness of the field: the development of such indicators may be in its infancy, and the rush to apply proposed indicators to realistic workload situations may be premature, at best. As a final point, one might take issue with the basic premise which connects mental workload with physiological state. As Ursin and Ursin (1977)

argue, physiological measurements correlate more properly with "activation", a fairly ill-defined "state" of the human, but one highly dependent on emotional factors, past experience, and task expectancy (Lindsley (1951)). By this argument, any proposed physiological workload measurement is certain to fail because of the inability to control for individual personality differences. One can counter this argument, however, by noting that, for simple laboratory workload situations at least, some physiological indicators have been demonstrated to have some correlation with the imposed workload. Whether or not these indicators are highly sensitive to "irrelevant" personality factors is, of course, an open issue, and this report will proceed on the assumption that the workload indicators being proposed are more than mere indicators of emotional arousal.

This report is organized into four chapters. Chapter 2 builds on some of the introductory comments just made, and attempts to provide a simplified framework for relating workload and performance to the operator's physiological state. Some fundamental issues are briefly discussed here so as to provide a background for the descriptions of specific physiological measures given in Chapter 3. Finally, Chapter 4 concludes the chapter with a brief summary and evaluation.

2. TASK DIFFICULTY, PERFORMANCE AND ACTIVATION

This chapter will attempt to formulate some general relationships between task difficulty, performance, and physiological "activation." First, however, some working definitions are in order.

2.1 Definitions

Since the term workload is a fairly ill-defined term, its use will be temporarily suspended in favor of a (hopefully) more explicit term: task difficulty. The major motivation here is to divorce operator characteristics from task characteristics, and so provide a conceptual means of task classification independent of operator performance, motivation, activation, etc. In this sense, task difficulty can be viewed as equivalent to externally applied workload.

At first glance, the use of the term task difficulty would appear to provide no major obstacles when attempting to compare two different tasks. For example, we have an intuitive notion of when one task is more difficult than another: we recognize that memorizing a seven-digit number is more difficult than a two-digit number, and we recognize that sub-critical tracking (Jex et al (1966)) is more difficult with a 4 radian/sec instability than with a 2 radian/sec instability. However, when

more quantitative comparisons are attempted, one runs into difficulties: presumably a five-digit number is harder to memorize than a four-digit number, but is it 25% harder (one additional digit in four), or twice as hard? Is a 4 radian/sec instability 20% easier to control than one at 5 rad/sec, or twice as easy? In short, we have no obvious quantitative means of assessing task difficulty based on the task parameters themselves, and at best, we can only hope to order tasks in terms of relative difficulty. The problem of assigning them absolute difficulty levels would seem to be beyond our reach at present.

Even an ability to order tasks in terms of difficulty breaks down, however, when dissimilar tasks are compared, because no common task difficulty metric exists. As an example, it is futile to attempt to compare the difficulty of a five-digit memorization task with the difficulty of tracking against a 2 rad/sec instability, if only the characteristics intrinsic to each particular task are compared. One might thus argue that task difficulty is not a particularly useful concept, since it appears difficult or impossible to associate it with a numerical scale based on intrinsic characteristics of the task itself. However, the fact that task difficulty is not directly measurable need not argue for its non-existence, and researchers have recognized that there may be indirect methods of assessing task

difficulty. Thus, subjective rating scales have been used to estimate task difficulty (e.g., Harper and Cooper (1966), Spyker et al (1971), Nicholson et al (1970)). Secondary tasks have also been used to make similar inferences (e.g. Michon (1966), Benson et al (1965), Rolfe(1971)).

Presumably, the search for effective physiological indicators also reflects this belief in the existence of some essential element of the task which we label task difficulty. Although there are clearly problems involved with the assessment and measurement of task difficulty, it would appear that it is a useful concept and an important factor in any attempt to understand the relationship between workload, performance, and physiological activation. The arguments to follow will accept the premise of the existence of an inherent task difficulty level associated with a given task, even though that level may not be a directly quantifiable factor.

Task performance would appear to be a much more straight-forward concept, since the criterion is usually well-defined and actual performance is directly measurable. For example, one might refer to percent correct recall of a string of five-digit numbers, or RMS tracking error in a sub-critical tracking task. Of course, problems again arise when we attempt

comparisons between two fundamentally different tasks, because of the different performance metrics involved (e.g., percent recall vs. RMS error). To resolve this problem, one might propose a normalization procedure for performance scores, so that, for example, a normalized score of 100 corresponded to average performance on the task, when measured across some reference population of test subjects. In this manner, a score of 100 on one task would be, in some sense, comparable to a score of 100 on a distinctly different task. Presumably, deviations from average performance could be handled by a similar normalization procedure, based on the variance of the reference population. Other alternatives are clearly possible, but the main point is that task performance would seem to be a reasonably well-defined concept, and subject to fewer of the vagaries of definition associated with workload or task difficulty.

A more difficult situation arises when attempting to discuss "activation", "arousal", or "awareness", all often interchangeably used to define some central state of the human organism. In very simple terms, activation theory (Lindsley, 1951) proposes that sensory inputs are processed in two basic fashions: via the classical sensory pathways which subserve the sensory-motor reflexes and the higher cognitive functions of sensory integration; and via the reticular formation (RF), which

serves as a central system regulating the organism's "activation" and subsequent responsiveness to sensory cues. In essence, the RF acts as a gate for higher processing of sensory cues, with the gating mechanism and activation level in turn dependent on the cues themselves.

As just described, activation theory fails to take into account other factors which clearly influence cue integration and activation levels: emotional state, past experience, and current expectation, to name a few. However, this apparent shortcoming can be remedied by recognizing that the RF also receives inputs from higher centers, so that activation level can also be influenced by CNS components not directly related to sensory input. As Ursin and Ursin (1977) pointed out, the demonstrated pathways from higher centers to the RF are completely in line with Hebb's theory (Hebb (1955)) of a central drive contribution to activation and arousal. Thus, a more encompassing activation theory recognizes not only the effect of sensory inputs on the RF, but also the effect of psychological state determinants having origins in higher CNS centers.

A discussion of possible CNS activation pathways may serve as an aid in identifying possible determinants of overall activation state, but it is of little help in defining what exactly is meant by the term activation. Studies of activation

may be couched in either psychological or physiological terms, but few explicit definitions appear to exist. A possible remedy may be to avoid the vagaries associated with a central or internal "state", and simply define activation in terms of reasonably objective and measurable variables associated with the organism. In particular, one might choose to view activation as a physiological state of the organism, and consider activation to be a vector quantity having as many dimensions as there are measurable physiological quantities.

One might argue that this definition confuses physiological correlates of activation with the mental state of activation, but it is felt that a more concrete and measurable definition is in order, if activation theory is to make any direct contribution to the problems of workload and performance assessment. One might also argue against the multi-dimensional nature of the definition, and propose a more conventional one-dimensional metric for activation (e.g., low and high drive individuals as described by Burgess and Hokanson (1968, 1969)). However, this scalar approach ignores the fundamental multi-dimensional character of activation, and can lead to inconsistent or possibly erroneous conclusions, when one is limited to a single component of activation. Finally, one might argue that the definition fails to give any indication as to which physiological measures

should be incorporated into the activation "vector"; clearly, the only response to this objection is that the determination of an appropriate physiological set is the subject of current and future research, and, of course, is the subject of this review.

2.2 Relationships

The discussion so far has been concerned primarily with the problems of defining the terms task difficulty, performance, and activation. When one begins to look more closely at the relationships existing between these variables, additional problems arise. This section will briefly discuss some of the basic relationships felt to be involved, and their relevance to the development of physiological indicators of workload.

For the purpose of illustration, task difficulty, performance, and activation will be treated as if they were three independent scalar quantities. Based on the discussion of the previous section, scalar representations of task difficulty and performance are not unreasonable, assuming comparable tasks and performance metrics. A scalar representation of activation is clearly an oversimplification, however, but can be justified on either of two lines of reasoning. One can interpret the discussion to follow as one which simply illustrates trends more

appropriately represented in a higher dimensional space, an approach often used when attempting to understand the behavior of multi-dimensional systems when limited to two-dimensional illustrations. Alternatively, one can conceive of a scalar activation level, which is an explicit scalar function of the activation vector (for instance, the length of the vector). In either case, a scalar treatment allows for a two- dimensional representation of the relationships between all three factors, and seems to illustrate some of the basic ideas in workload measurement.

Figure 1a shows a not unreasonable relationship between performance and task difficulty: for a given activation level, performance drops with increasing task difficulty. The parametric dependence on activation level illustrates how, for a fixed level of task difficulty, performance improves with increasing levels of activation. If there exists some optimum activation level, above which performance degrades with increasing activation for all levels of task difficulty, then one might expect a "folding over" of the performance curves. Thus, a performance curve associated with a sub-optimal activation level may be nearly coincident with a curve associated with a supra-optimal activation level.

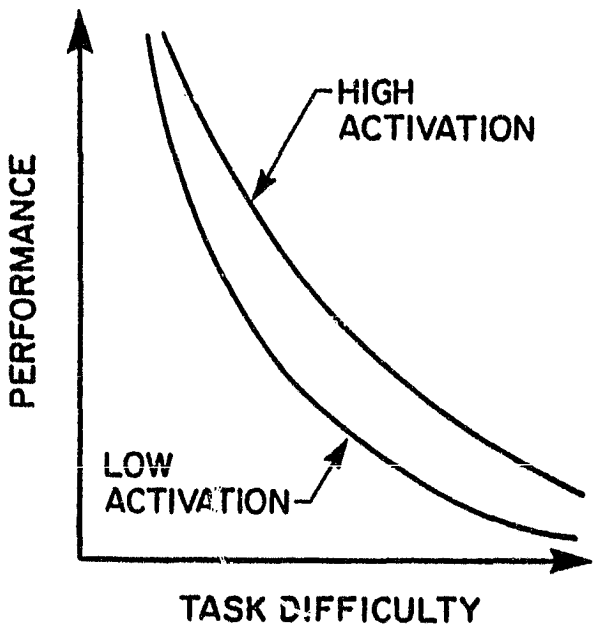

Figure 1a: Performance vs. Task Difficulty

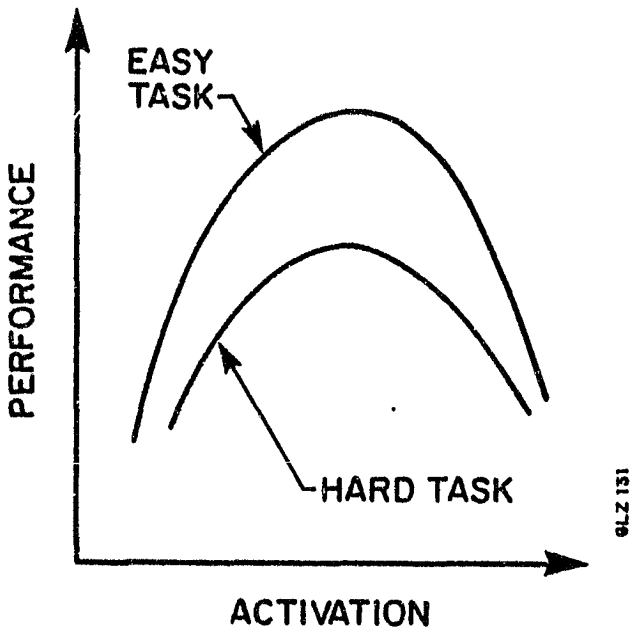

Figure 1b: Performance vs. Activation

Since this trend with activation is difficult to illustrate in the performance/task difficulty plane of figure 1a, one can make use of the performance/activation plane as shown in figure 1b. Here, task difficulty is the parametric variable, and the performance curves associated with each difficulty level illustrate the "inverted - U " function of activation, discussed by Malmo (1959). Optimum performance is associated with an

optimum activation level, and activation above or below this level results in degraded performance. The parametric dependence on task difficulty is chosen to illustrate how, for a fixed activation level, performance improves as task difficulty drops. The fact that the maxima of the curves are coincident implies an optimum activation level independent of task difficulty, a situation which is probably not realistic. However, this is consistent with the parametric curve dependence of figure 1a, in which no curve "crossover" is illustrated.

These curves can now be used to expand upon the basic premise underlying the development of a physiologically-based workload indicator. Figure 2 repeats the performance verses task difficulty curves of figure 1a, and illustrates possible "trajectories" in this plane, associated with a transition from one operating point to another. Consider the situation in which a subject is working on an easy task (T_1), at a performance level which places him at point A on the low activation curve. If he now performs a more difficult task* (T_2), while maintaining the same level of performance, then presumably he has transitioned to operating point B on the high activation curve. The change in activation level in going from point A to point B is presumably

* Comparable tasks are assumed, to allow for ordering of task difficulty levels.

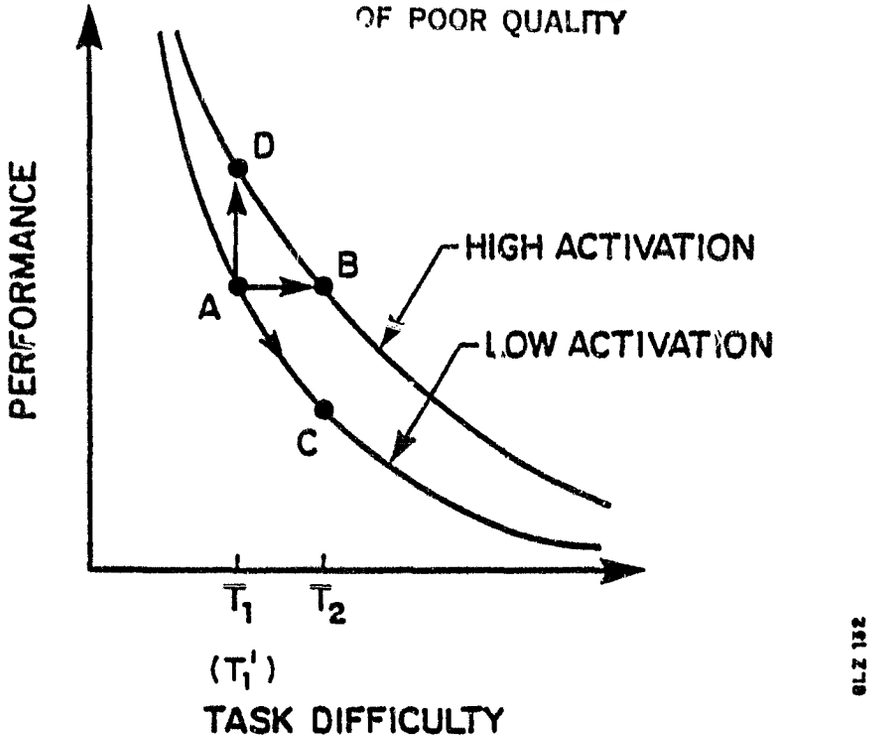

Figure 2: Transition Trajectories

reflected by changes in one or more physiological measures, and thus one should observe an increase in physiological "activity". In brief, the premise behind a physiological workload indicator is that an increase in physiological activity implies an increase in task difficulty, if performance is maintained at a constant level. The natural extension of this is that the observed change in activity is somehow proportional to the change in workload.

The requirement of a constant performance level is central to the workload definition. If performance is not maintained at a constant level, then other state transitions are possible. To illustrate, two alternatives to the constant performance transition are shown in figure 2. The A-to-C transition illustrates the situation in which a constant activity level is maintained, so that an increase in task difficulty results in a performance decrement. If performance were not monitored in this situation, one would conclude on the basis of physiological measures that task T_1 (associated with point A), was no easier to perform than task T_2 (associated with point B). The failure to monitor performance would thus result in an erroneous conclusion concerning task difficulty. Conversely, if performance were monitored, and a decrement observed, one would be led to the correct conclusion regarding the greater task difficulty level of task T_2 over task T_1.

One might argue that this A-to-C transition indicates that insufficient experimental control was provided to properly "motivate" the subject to maintain equal performance on the two tasks. This argument simply emphasizes the need for performance monitoring, either for the purpose of recognizing a decrement when it occurs (A-to-C), or for feeding back to the subject as an aid to ensuring a constant performance transition (A-to-B).

Figure 2 also illustrates a transition (A-to-D) involving a performance increment at constant task difficulty, due to an increase of activation level. If task T_1 is associated with operating point A and task T_1' with point D, then the activation level increase must be due to a change in motivation, rather than due to a need for maintaining constant performance (as in the A-to-B transition). If performance were not monitored in this situation, then, based solely on physiological measures, one would conclude that task T_1' was more difficult than task T_1, clearly an erroneous conclusion since the tasks are of equal difficulty.

Since this A-to-D transition occurs at a constant task difficulty level, the trajectory in the performance/activation plane of figure 1b must be along one of the inverted-U curves. Since performance improves in the transition from task T_1 to T_1', the activation level for task T_1' must be more nearly optimum than that for T_1. Thus, this type of transition implies a non-optimum activation level associated with task T_1, presumably due to insufficient subject motivation. Again, this simply emphasizes the need for performance monitoring, since activation level optimization is impossible in the absence of performance information being fed back to the subject.

Other transitions between different levels of performance, task difficulty, and activation could be discussed, but it is felt that the three examples just given serve to illustrate the basic relationships involved among the three metrics. These examples also illustrate the need for concomitant measurements, for both research and engineering applications. In a research situation, one is attempting to define the curves themselves, and thus all three coordinates are needed. For an application involving workload measurement one assumes the shape of the curves and infers task difficulty from the measured performance and the physiological correlates of activation. Clearly, performance measurements must be made concurrently with physiological measurements, if reasonable conclusions are to be made with respect to task difficulty and workload.

2.3 Physiological Measurements and Activation

To this point, the discussion has dealt with activation as if it were a well-understood and well-defined quantity, which, of course, it is not. This section addresses some of the issues associated with a better definition of activation.

Earlier, it was noted that activation could be viewed as a vector quantity, with each component corresponding to a particular measurable physiological variable. This viewpoint,

while more concrete than the psychological concept of an internal activation "state", lacks definition for at least three reasons. First, one is confronted with the problem of defining what is meant by a physiological variable, and, in particular, what defines a physiological "feature". Second, one must begin to answer questions regarding the type of signal processing most appropriate for transforming the measured physiological signal to one or more specific components of the activation vector. Here, basic questions arise concerning time verses ensemble averaging, and the use of physiological models to effect a data reduction. Finally, one must attempt to specify the dimension of the activation vector, with the particular goal of determining which physiological variables should be used to completely and efficiently specify the vector, so as to be useful in workload estimation. These interrelated questions are discussed in some detail in the remainder of this section.

2.3.1 Time Averaging

One of the basic problems confronting the researcher in physiological measures concerns the definition of a physiological variable. To illustrate the difficulties involved, consider the case of an EKG measurement made during a high workload situation. One might consider the voltage time history of the signal to be

the basic physiological variable being measured. However, when one attempts to correlate this variable with, say, degree of task difficulty, it immediately becomes apparent that some quantifiable characteristic of the signal must first be extracted from the signal itself. One of the more common choices has been heart rate, or its inverse, the interpulse interval. Of course, both of these derived signals are time varying, like the EKG signal itself, and are themselves subject to further processing; in particular, one might derive one or more time-varying signals from the signal originally derived from the EKG history. For example, one might first derive heart rate from the EKG, and, from this, derive a time varying signal specifying beat irregularity, or sinus arrythmia. Clearly, other options and further signal processing along these lines are possible, and one can envision a multitude of time-varying signals derived from the primary physiological measurement. The ingenuity of the physiological researcher would appear to be the only factor limiting the number of possible derived signals.

Regardless of the processing used to derive a time-varying signal from the basic physiological measurement, one is still faced with the problem of attempting to correlate a physiological variable with some facet of the workload environment. One approach common in the literature is to use time-domain averaging

to reduce to a scalar value the information contained in the theoretically infinite number of data points present in a continuous time-varying signal. Thus, in the example of attempting to correlate workload with the information in an EKG signal, one might first derive instantaneous heart rate from the signal, and then time average the heart rate over the workload interval. Alternatively one might simply count the number of beats within the workload interval and then divide by the interval length. In either case, this time-averaging approach provides the researcher with a single number, average heart rate over the workload interval, which can then be correlated with different workload levels.

Examples of time-averaging abound in the literature. The direct averaging approach was used by Burgess and Hokanson (1964, 1968), who used average heart rate as an indicator of arousal, and investigated the correlation between heart rate, psychological stress, and workload. Indirect time-averaging can be accomplished by taking cumulative measures over a fixed interval length: McCleary (1953) described an apparatus for measuring cumulative palmar sweat over a fifteen-minute period. Integration over time is another method of averaging, and Riehl (1961) proposed an index of "cerebral activity" derived from an EEG signal by dividing short-term average frequency by short-term

average amplitude. This time-varying index was then integrated over the measurement interval to yield a cumulative indicator of average mental activity.

Studies involving the simultaneous recording of a number of different physiological measurements often make use of time-averaging, because of the requirement for data compression in the face of voluminous amounts of data. In a study of physiological changes occurring during a 48-minute vigilance task, Poock et al (1969) measured average heart rate, blood pressure, skin temperature, and skin resistance. Averages were computed over four 12-minute segments of the task, and correlations were made with monitoring performance over corresponding segments. An earlier study by Eason et al (1965) also correlated average physiological measures with performance in a vigilance task. Here, heart rate, skin resistance, neck EMG activity, and eye activity were measured and averaged over the six 10-minute intervals comprising the one-hour vigilance task. By plotting these average values verses time, with one point every 10 minutes, Eason et al were able to illustrate some correspondence between physiological trends and trends in time-averaged performance.

A similar approach was used by Benson et al (1965) in a study of compensatory tracking performance with different

displays. Since the tracking task lasted for only four minutes, relatively short 30-second measurement intervals were used for determining average heart rate, skin resistance, respiration, and integrated EMG activity. These interval averages were in turn averaged, to obtain average values for an entire tracking run, so that comparisons could be made with average values obtained during non-tracking intervals. Differences in the average physiological values, obtained during tracking and non-tracking intervals, were then correlated with the type of display and imposed workload, to assess the reliability of each physiological measure for display/workload evaluation.

This time-averaging approach is also utilized to advantage in studies which make use of spectral analysis of the measured physiological data. Since spectral measurements are made over some finite length measurement "window", the window length becomes an effective averaging interval for the processed data: longer windows imply more averaging, and provide a better estimate of the long-term signal spectrum; shorter windows imply less averaging and provide a closer look at time variations in the spectrum. Typical examples of the use of spectral analysis have been described by Nicholson et al (1970) in an approach-to-landing workload study, and by Wisner (1971) in a discussion of an automobile driving study conducted by Pin et al

(1969). In the first study, a finger tremor index was derived based on the peak power spectral density occurring in the neighborhood of 10 Hz; in the second, an EEG alpha-wave index was used and based on the percentage of EEG power contained in the alpha frequency band (9 to 11 Hz). Clearly, both measures represent time-averaged behavior since they are based on fixed interval spectral estimates.

2.3.2 Ensemble Averaging

One of the basic problems of physiological measurements concerns the reduction of "noise" associated with the signal of interest. There are several possible sources in any measurement situation, but probably the major contribution stems from moment-to-moment variations associated with the individual. Occurring over a time period of seconds or minutes, these variations would appear to have little correlation with a steady workload imposed over the same time interval, and thus, might be regarded as noise. Time averaging of the data over this interval serves to average out this source of noise and allows the experimenter a means of determining relatively long term (on the order of a minute or more) changes in baseline physiological values.

One might argue, however, that this moment-to-moment variation should be closely monitored, and, rather than being considered "noise", be considered an indicator of short term changes in workload. If such short term changes exist, then time averaging over relatively long intervals effectively washes out potentially important physiological correlates of workload. Thus, one might argue that time averaging not only washes out the noise, but also the signal.

The idea of concentrating on short term changes in physiological variables has led many researchers to study "evoked responses" (ER) of one form or another, in hopes of correlating response measures with the imposed experimental condition. In brief, the method consists of presenting a subject with a task whose features are well-defined functions of time, and of making a physiological measurement over the same task interval at relatively high sample rates. By using the common time base between the task and the measurement, direct correlations can be made between task "features" and physiological response "features". Examples will be given shortly, but it is of interest to note that since this type of measurement precludes time averaging, some method is needed for noise reduction. Typically, researchers use ensemble averaging of multiple identical stimulus-response sequences, and average across the set

of responses at corresponding points in time. Relatively long term changes in the physiological measurement are thus averaged out of the data, and the researcher obtains an "average evoked response" (AER) which can then be correlated with the stimulus or imposed workload.

Evoked response measurements have typically been used to process short-term (approximately 500 milliseconds) segments of EEG signals, to evaluate stimulus specific responses. Spyker et al (1971) described a workload measurement technique utilizing the visual evoked response (VER), obtained by measuring a differential midline EEG in response to a 10 microsecond strobe light pulse. The EEG measurement interval was 500 milliseconds long and time locked to both the strobe pulse and the onset of a secondary task cue; ensemble averaging was performed over 50 consecutive evoked responses obtained during a test session, with each associated strobe pulse serving as the time marker for proper time alignment of all the responses. Wickens et al (1977) also described a workload measurement based on the evoked EEG response, but which utilized auditory tones as the stimuli. Again, ensemble averaging was used to reduce background noise presumably unrelated to the experimental conditions. Studies involving other sensory modalities (e.g., vestibular) for evoked EEG response studies have utilized the same time-locking and ensemble averaging for noise reduction and signal enhancement.

Evoked response measurements need not be limited to EEG signal processing, however, since other physiological indicators can readily be used for short-term stimulus-response measurements. Perhaps the classic example is pupil dilation measured over a short-time interval (approximately 10 seconds) and sampled at a relatively high rates (greater than one sample/second). Kahneman and Beatty (1966) made use of this response in evaluating the difficulty of a digit recall task, by tightly synchronizing to a common time base all three components of the experiment: the task, the pupil measurements, and the subject's verbal response. AER curves were then obtained by averaging across 20 pupil measurements at each point in time, with the recorded time base used to ensure proper time axis alignment. Kahneman et al (1969) used this same approach with a digit transformation task, and, in addition to obtaining pupil measurements once each second, simultaneously recorded heart rate, skin resistance, and respiration. Although the respiration records were not discussed in the paper, the AER's for pupil diameter, heart rate, and skin resistance were obtained by averaging 50 measurements at each point in time, to yield AER's of approximately 20 seconds duration, with a response point every second.

2.3.3 Frequency Response Considerations

These and other studies demonstrate how ensemble averaging over relatively short time intervals can be used to reduce physiological "noise", and how such averaging can be applied to the "classical" set of physiological measurements (e.g. skin resistance, heart rate, etc.) A basic issue underlying the use of ensemble averaging centers on the frequency content of the imposed workload one is trying to measure, and the frequency response of the physiological variable being monitored. If the workload level is changing rapidly with time (high frequency content), then one should: a) choose a physiological variable which is sufficiently responsive to rapid changes in workload (high frequency response); and b), use short-term ensemble averaging to generate ER curves which capture the rapid fluctuations present in the measured physiological variable. This line of reasoning is illustrated by the above-mentioned pupil response study conducted by Kahneman and Beatty (1966): they assume that pupil diameter is proportional to instantaneous workload, and from the changing time histories of the response curves, infer that the workload level changes continuously throughout the duration of their 20-second digit transformation task. Thus, short-term ensemble averaging is presumably most appropriate here.

The converse argument can be made for time-averaging of the data. If the workload level is relatively constant with time (low frequency content), then one should: a) choose a physiological variable which is relatively stable over the long term, but which need not be particularly responsive to rapid changes in workload (low frequency response); and b), use long-term time averaging to generate an average physiological level which can be correlated directly with the average imposed workload level. This approach is illustrated by the above-mentioned tracking task study conducted by Spkyer et al (1971): they assumed that workload level during their four-minute tracking task was relatively constant, and thus could be correlated against a long term average of one or more physiological "features" extracted from the measurements. Thus, long-term time-averaging is presumably most appropriate in this measurement situation.

2.3.4 Feature Extraction

Although the time- vs. ensemble-averaging question depends strongly on the time span of interest to the researcher, other factors influence the choice of data processing. With time-averaging, the average level of the monitored physiological variable can be directly correlated with some scalar attribute of

the imposed workload; with ensemble averaging, no such obvious choice is apparent. The researcher is thus forced to define one or more "features" associated with the evoked response history (e.g., total signal power, magnitude of the first positive peak, etc.) which, singly or in combination, can then be correlated with a scalar workload metric.

Hess and Polt (1964) studied the dependence of pupil dilation on multiplication task difficulty, and defined their evoked response feature as the percentage increase in pupil diameter, measured over the interval used by the subject to solve the problem presented him. Beatty and Wagoner (1977) also reported on evoked pupil response as a function of task difficulty. Although they illustrated their results with evoked response curves which demonstrated increasing dilation with task difficulty level, their formal conclusions were based on a statistical test based on a single response feature: average pupil dilation during the last second of the measurement interval. Effectively, they reduced the entire response curve to a single scalar measurement. Squires et al (1977) used auditory cues to elicit an evoked EEG response, and also used a single feature to infer response dependence on task workload. They used a scalar "discriminant score", described by Squires and Donchin (1976), which served as an estimate of the amplitude of the "P300

complex", a transient feature characteristically present in the evoked response. This approach to "scoring" each response provided a means of transforming the complete response curve to a single number, which could then be correlated with the experimental conditions being tested.

Feature extraction from the evoked response curve need not be limited to the extraction of a single feature, nor need each feature be based directly on the transient amplitudes characterizing the response curve. Spyker et al (1971) measured visually evoked EEG responses in their workload study. After averaging over 50 responses occurring over the workload interval, they extracted 28 features from the AER curve, effectively generating a 28-dimensional feature vector characterizing the average response. Although most of the 28 features consisted of the amplitudes and latencies of the AER maxima and minima, two features were based on "second-order" characteristics of the response: the overall response power, and the number of maxima in the signal. The inclusion of such features in a feature-vector demonstrates that response features need not be limited to simple attributes of the AER; in fact, any well-defined transformation of the time history can serve as a means of "feature extraction".

This flexibility in AER processing can be viewed as advantageous to the researcher, since it provides him with unlimited options for feature extraction. On the other hand, this flexibility can be viewed as a distinct disadvantage, since no obvious choices exist for the definition of one or more "ideal" features which completely and efficiently specify the AER. If feature choice is viewed in the context of representational efficiency verses accuracy, one can begin to appreciate the problem of feature extraction. If efficiency is paramount, then a single measure, say, peak amplitude, is appropriate. Of course, accuracy of representation is extremely low, since the single feature provides no information on the other important signal characteristics (e.g. total power, latency of peak, number of zero crossings, etc.). Conversely, if accuracy is most important, then many features are necessary; for instance, one might simply sample the response curve, and define every single time and amplitude pair to be a response "feature". Clearly, the efficiency of representation is low in this situation.

2.3.5 Functional Modelling

Many methods exist for both increasing the accuracy with which a signal is specified and for reducing the complexity of the specification. Typically, these methods all involve functional modelling of one sort or another, and presume the existence of some sort of structure associated with the signal's evolution in time. Examples abound in the signal processing literature, but few studies in the area of physiological measurements have taken advantage of the economics afforded by functional modelling. One exception has been described by John et al (1977), in a report on the use of evoked responses to help in the clinical classification of brain dysfunction. The AER's obtained from a test protocol were decomposed into three factor waveshapes and three associated coefficients. By multiplying each waveshape by its coefficient, and summing the three together, a fair approximation to the original AER was obtained. The details of specifying the set of factor waveshapes, and the approach for computing the coefficients, are discussed in the reference and need not be of concern here. The relation of this type of signal analysis to the notion of functional modelling deserves further comment, however.

Shown in figure 3 is a block diagram model of the AER "generating system", based on the analytic approach used by John et al (1977). Three signal generators respond synchronously to a stimulus trigger, and are responsible for generating the three factor waveshapes. Each signal is multiplied by a gain (the factor coefficient), and then summed with the other two to yield the signal which approximates the desired AER. A final noise summer accounts for modelling discrepancy between the actual AER and the model response.

Although not presented in this format, John et al (1977) proposed this as a model of AER generation, and thus effected a considerable reduction in the number of features required to characterize the AER. Since all of the AER's decomposed by this modelling approach utilize the same three factor waveshapes, the model has only four adjustable parameters: the three gains and the noise power. The gains can be taken as three features which economically specify the AER, and the noise power can be used to provide an indication of the modelling accuracy.

It is worth noting that no claims are made regarding the correspondence between elements of this functional model and elements of the physiological system being modelled. Thus, for example, the model does not purport to represent the contributions of three different regions in the brain to the

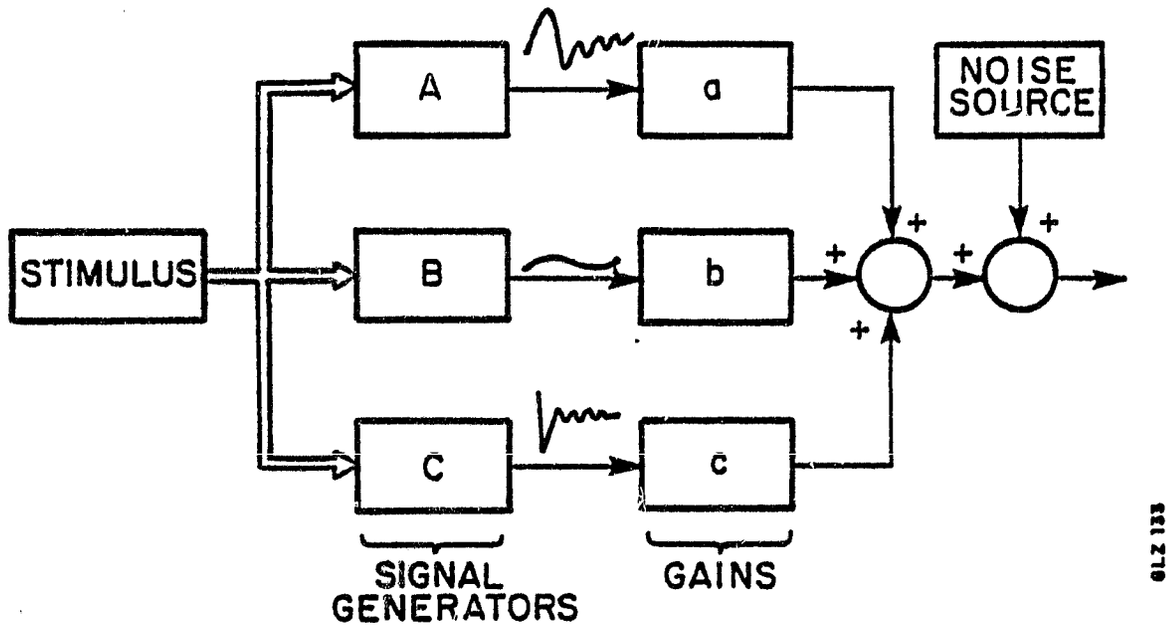

Figure 3: AER Functional Model

overall measured AER. Instead, it is a functional description using a modelling approach which simply attempts to accurately characterize the AER with a minimum number of model parameters. In short, it is a descriptive model.

Functional modelling of physiological responses need not be restricted to AER processing, since many measurement situations can benefit from the economics provided by modelling. An example was provided by Spyker et al (1971), who used a method of processing skin impedance measurements originally proposed by Khalafalla et al (1970). Skin resistance and reactance were measured at nine different frequencies, and the results plotted to provide a graphical representation of the frequency response of the cutaneous system. These data points were then fit with a smooth response curve by appropriately choosing the parameters of a four-element electrical circuit model, which, in turn, was chosen on the basis of the general frequency trends seen in the data. This circuit analog thus provided a means of reducing the nine data triplets (frequency, resistance, reactance) to four model parameters. Spyker et al were unsuccessful in their attempts at correlating these model parameters with the imposed workload of their task, but their approach illustrates how feature extraction can take advantage of functional modelling, and, in particular, how a multi-dimensional feature vector can be constructed from a number of model-based parameter values.

2.3.6 Dimensionality Considerations

This discussion of functional modelling and parameter identification also serves to emphasize one of the major problems associated with physiological monitoring: the multiplicity of derived measures. Thus, for any one physiological time history, a number of possible features can be extracted. If time averaging is used, a simple average value can be extracted from the data. If ensemble averaging is used, a feature vector can be comprised of all important peaks and latencies characterizing the signal. Alternatively, one can use a functional model of the physiological process, and derive a feature vector composed of model parameter values. In any case, several possibilities exist, even for a single time history of a single physiological variable.

This problem of multiple features becomes compounded when one attempts to derive a workload correlate by monitoring more than one physiological system. The philosophy behind multiple-response monitoring is primarily a practical one: the greater the number of responses monitored, the more likely is the experimenter to discover a physiological variable which changes in a consistent manner with the workload of the imposed

experimental task.* A similar but more sophisticated argument runs as follows: if no single physiological variable demonstrates adequate sensitivity to workload changes, then perhaps a number of these low-sensitivity measures can be combined in some manner to provide a single high-sensitivity physiological correlate of workload. In essence, workload might be more properly correlated with a number of small and barely significant changes in several monitored variables, rather than with a single large and highly significant change in only one variable.

Many workers have taken a multiple response approach in their attempts to isolate one or more reliable physiological correlates of workload. Workload studies which monitored more than one physiological variable have been described by Benson et al (1965), Eason et al (1965), Corkindale et al (1969), Poock et al (1969), Kahneman et al (1969), Jex and Allen (1970), and Spyker et al (1971). All of these studies attempted to find a simple correlation between each measured variable and the imposed

* One might argue that such an expansion in scope is bought at the expense of a less intensive analysis of each individual physiological measurement, and thus one might actually reduce one's chance of identifying a significant correlation by monitoring many variables. However, this presumes that depth of analysis must be traded off with breadth of monitoring, an experimental constraint which begins to have less and less relevance with today's computerized laboratory facilities.

workload, and some found one or more measures significantly more sensitive than the others. Thus, this "shotgun" approach served, for some of the studies, as a means of isolating one or more sensitive indicators.

Some of these same studies also incorporated a joint correlation of multiple measures, in an attempt to improve the sensitivity and accuracy of workload estimation based on physiological indicators. In their study of display effects on workload, Benson et al (1965) tested for changes in several physiological indicators, to see if a correlation could be made with display type. By restricting themselves to a separate consideration of each physiological variable on its own, they were unable to demonstrate any significant display-associated differences. However, by combining the measures in a non-parametic test of overall activation, they were able to demonstrate a significant difference between displays, at least for one of the two workload situations investigated in their experiment. Although their analysis approach made no formal use of multiple regression techniques, their results with combined measurements suggest that greater indicator sensitivity can be had by combining the physiological measures in some meaningful manner.

Multiple correlation techniques were used by Poock et al (1969) in their study of the physiological changes associated with a visual monitoring task. With measurements of heart rate, skin temperature, skin resistance, and blood pressure available for analysis, it was found that the best simple linear correlation with performance was obtained with systolic blood pressure. However, the correlation was fairly low ($r = 0.38$) and barely significant ($P < 0.08$). A pairwise multiple correlation, using systolic pressure and skin temperature, increased both the correlation and significance ($r = 0.50$; $P < 0.06$). The subsequent addition of pulse pressure resulted in a slight increase in the correlation coefficient ($r = 0.54$), but reduced the significance level to the previous low level associated with the simple linear correlation ($P < 0.08$). These findings suggest three things. First, correlation coefficients can be increased if several physiological measures are included in the analysis. Second, the rate of increase probably decreases as more measures are included, so that the marginal return becomes less with each additional measurement. Finally, there is probably some optimum measurement set which yields the highest level of significance; increasing the size of the measurement set drops the significance level because of the additional data spread introduced by less reliable measurements. Clearly, the problem of deciding upon a "best" measurement set involves a non-trivial trade between the

magnitude of the correlation coefficient and the level of significance associated with the correlation.

This problem becomes evident in the correlation analysis performed by Spyker et al (1971), in which an attempt was made to define a workload index based on physiological measurements. Two "feature vectors" were defined: the first had components associated with task performance, subjective assessment of task difficulty, and performance on a secondary task; the second was comprised of a subset of the 84 physiological features extracted from their physiological measurements made during task performance. Details of how the first "performance" vector was correlated with the second "physiological" vector are given in the next chapter; what is of interest here is how Spyker et al attacked the problem of defining the two feature sets comprising each vector. They noted that high correlations were easily achieved with a large number of physiological features and that "adding more features always increased [the correlation coefficient], even though the significance may be decreased" (Spyker et al (1971)). This drop in significance level apparently motivated the use of small subsets of both the performance and physiology features:

> Through a combination of classification ordering and multiple correlation ranking, a "best" subset of 10 [physiological] features was chosen to predict the [secondary task performance features of] miss rate and response time. (Spyker et al (1971))

No explanation of the feature selection protocol is given, nor is any indication given as to how rapidly significance levels degraded with ever increasing feature vector dimensions. This lack of an explanation suggests that some artistry is required in the selection of an appropriate feature set, and that careful attention must be given to the balance between correlation and significance.

2.3.7 Summary

This section has attempted to address the major issues involved in the definition and use of a physiologically-based "activation vector." Problems immediately arise when one attempts to define what is meant by a physiological feature, and the discussion here has centered on various approaches which have been used in the past. Most researchers have chosen to use either time or ensemble averaging to extract a meaningful physiological indicator from the noisy measurement environment. Time averaging provides the researcher with a simple scalar value which simplifies the data analysis, but the averaging procedure

tends to wash-out any high-frequency changes in physiological activation, changes which may be quite relevant to the workload one is trying to measure. The use of evoked responses and subsequent ensemble averaging circumvents this problem, but introduces the new problem of feature definition and extraction.

A review of the literature shows a multitude of approaches to the feature extraction problem, ranging from simple peak detection and quantification to more elaborate strategies involving functional modelling and parameter identification. The increasing sophistication in signal processing and the tendency to monitor several physiological variables simultaneously are two major factors contributing to a potentionally highly-dimensioned activation vector, composed of a large number of features representing several different aspects of the subject's physiological state. This dimensionality problem is not only a concern in terms of practical implementation, but also in terms of indicator accuracy and reliability: correlations between physiological indicators and task workload can apparently be improved with increasing feature vector dimensions, but an "excessive" number of features results in poorer statistical reliability. The definition of an appropriate physiological activation vector is clearly a goal of current research.

3. PHYSIOLOGICAL CORRELATES

This chapter discusses individual studies of specific physiological correlates, and reviews the findings in terms of the general concepts introduced in the previous chapter.

Each chapter section concentrates on a specific "physiological system", and the specific methods used for monitoring the activity of that system. The first three sections, devoted to correlates of eye, skin, and muscle activity, discuss many of the relevant studies in some detail, and provide a critical assessment of techniques and results. The next three sections, devoted to correlates of circulatory, respiratory, and brain electrical activity, provide listings of relevant references which were evaluated during the course of this review, but which, due to time limitations, are not accompanied by the discussion and evaluation comprising the first three sections of this chapter. The concluding section is devoted to multi-dimensional measures of physiological activity; again, due to time limitations, this section is restricted to a listing of those references evaluated during the course of the review.

3.1 Eye Measures

Physiological measures associated with the eye have classically fallen into three categories: eyelid blink activity, pupillary constriction and dilation, and visual scanning activity. Although visual scanning patterns have been proposed as a basis for inferring task workload (see, for example, Barnes (1977)), they will not be considered here because of the fact that scanning patterns, which are highly display- and task-dependent, are (normally) the result of a direct exercise of voluntary oculomotor control. Instead, the discussion will focus on the (almost) involuntary physiological metrics of blinking and pupillary constriction and dilation, in the following two subsections.

3.1.1 Eye Blinks

Eye blink rate might be considered a candidate measure for workload assessment, because of the generally held belief that blink rate levels increase with long-term visual fatigue. Few researchers have reported on the use of this measurement, however, and it is unclear from the literature how useful such a measure might be in either the laboratory or in a realistic working environment.

Eason et al (1965) investigated visual detection performance during a one-hour vigilance task, and monitored several physiological variables throughout the course of the vigil. One of these variables was "a composite recording of eye blinks, squints, and vertical eye movements," obtained from electrodes placed above and below one eye. Rectification and integration yielded an "eye region response" measurement every 20 seconds, 30 of which were averaged together to yield a single measurement every ten minutes during the one-hour vigil. The researchers found no significant change in this measurement over the course of the vigil, even though significant trends were observed in detection performance, and in two other monitored physiological variables (skin conductance and neck EMG activity). Furthermore, no significant correlations were found between eye activity and any of the other measured experimental variables.

One would be led to conclude from this study that eye blink rate (or generalized eye activity) is a considerably less sensitive physiological indicator than other possible choices, at least in the type of long-term vigilance situation described above. Whether or not shorter term measures (such as instantaneous blink rate) might be more highly correlated with short-term changes in workload would appear to be an open issue at present.

3.1.2 Pupillary Constriction/Dilation

Pupil constriction and dilation has been classically associated with the regulation of light flux falling on the retina, but it has long been known that pupil diameter changes can occur completely independently of illumination changes. Goldwater's review article (1972) noted that

> Lowenstein (1920) reported that pupillary dilation accompanied such suggestion-induced states as 'excitement', 'comfort', 'pleasure', and 'displeasure', as well as suggestions of impending pain and threat.

and that

> Hess (1965) has interpreted pupil dilation ...as an index of 'interest', 'emotion', and 'motivation'.

Although many of the early studies were concerned with such "psychological" factors, more recent efforts have been directed at correlating pupillary changes with mental workload. A typical protocol has a subject solving a "problem" while the experimenter measures the concurrent changes in pupil diameter. The "problem" portion of the protocol is usually a well-defined function of time, consisting of a presentation phase, a solution phase, and a response phase. Pupil measurements made over the same time intervals can then be directly associated with these different phases.

Problems presented to the subject may be simple auditory or visual discrimination tasks (Simpson and Hale (1969), Beatty and Wagoner (1978)), arithmetic or digit transformation tasks (Hess and Polt (1964), Kahneman and Beatty (1966)), or tasks requiring verbal skills (Bradshaw (1968)). By using the same class of problem throughout the experiment, but varying the level of difficulty, researchers have been able to demonstrate significant differences in pupil dilation with task difficulty level, with harder tasks associated with larger dilations. The findings suggest that pupil diameter provides a direct measure of the imposed workload.

Some of the relevant studies in this area are briefly summarized in the following several paragraphs.

Hess and Polt (1964) measured pupil size during the solution of an aurally presented multiplication problem. Pupil diameter was measured photographically every half second, but no time histories were presented in the paper. The authors noted, however, that

> Typically, the pupils of each subject showed a gradual increase in diameter, reached a maximum dimension immediately before an answer was given, and then reverted to the previous control size.

The physiological measurement extracted from each response history consisted of the mean pupil size at maximum dilation, which was obtained by averaging over the five photographic frames immediately preceding the subject's verbal response. This maximum diameter value was then used to compute a percentage increase figure for each of the five subjects in the study, and for each of the four multiplication problems presented each subject. Averaging across subjects, Hess and Polt found that as problem difficulty increased from the "simple" problem of 7 x 8 to the "difficult" problem of 16 x 23, the dilation percentage also increased. Their conclusion was that the more difficult problem imposed a greater mental workload on the subject, and was reflected in larger associated pupil dilations. However, no attempt was made to independently evaluate the workload associated with each multiplication problem, and thus no direct correlation could be made between workload and pupil dilation. Their conclusion is based on the observed progression of dilation magnitude with apparent problem difficulty, but even this lacks rigorous support, because of their failure to perform statistical tests on the significance of the observed differences due to different problems.

Kahneman and Beatty (1966) overcame some of these obstacles by presenting subjects with a class of problems which,

intuitively at least, are more accurately characterized in terms of their associated workload. In one set of experiments, subjects were aurally presented a string of digits for later verbal recall; by increasing the string length from three to seven digits, the authors were able to identify a progression in pupil dilation, due presumably to the incremental workload associated with each additional digit requiring recall. This finding was made possible by the incorporation of several features in the experimental design and data analysis. First, aural string presentation and verbal recall were highly structured functions of time, so that a common time base could be used for ensemble averaging of pupil responses across subjects and replications; in effect, an evoked pupil response was obtained for each problem difficulty level. Second, maximum pupil diameter was measured for each subject, and ordered as a function of string length; a steady progression of maxima was found in four of the five subjects tested, with the fifth subject contributing a single inversion. The authors noted that this was highly significant ($P<10^{-4}$), and thus provided a statistical basis for their conclusions.

One final feature of this study by Kahneman and Beatty (1966) deserves comment. In addition to the digit recall task, they imposed two additional problems on their subjects: the

recall of four words, and the transformation of a four-digit string to another four-digit string. By measuring pupil response during these two tasks and comparing with the response evoked by the digit recall task, they found a significant ordering of dilation maxima: digit recall resulted in the smallest peaks, while digit transformation resulted in the largest; word recall was intermediate. If pupil response is directly related to task workload, then one would infer that the digit recall task to be easiest, and transformation to be the most difficult. In fact, this is exactly what the authors found, based on recall performance:

> The two tasks [word recall and digit transformation] are clearly more difficult than the recall of digits; for our subjects the mean span for recall of digits was 7.8, while the mean span was 5.7 for words and 4.5 for the transformation task.

Of course, the assumption here is that task difficulty is inversely related to maximum attainable performance, but at the least, this performance-based assessment of task difficulty provides some independent measure of workload with which to correlate the measured pupil dilations. In this sense the study is exceptional, since most such studies fail to provide any independent measure of task workload.

In a follow-up study, Kahneman et al (1967) took advantage of the time dependence of the evoked pupil response, and demonstrated its utility as a measure of short-term workload. As in the previous study, subjects were presented with a four-digit string which was to be transformed to a second string, by adding 1 to each digit. With a time-locked sequence of aural string presentation and verbal response, the authors were able to obtain an average evoked pupil response similar to that obtained in the previous study. In a second series of tests, subject workload was increased by the imposition of a visual detection task, with the visual "target" appearing at different time points in the interval allotted to the concurrently imposed digit transformation task. In this case, one might have expected to see an overall increase in pupil dilation, because of the increased workload associated with the detection task. This was not the case however: the authors found the evoked response curves for both the single- and double-task situations to be almost indistinguishable.

The resolution of this apparent contradiction was made possible by the measurement of visual detection performance, as a function of time of appearance of the visual target. The authors found a clear parallel in time between the percent of targets missed and the pupil dilation time history, and they noted that

their subjects were, to some degree, "functionally blind" when pupil dilation was at a maximum. One is led to the conclusion that this perceptual deficit was due to the ongoing mental workload associated with the digit transformation task, and that the pupil response history adequately reflected the changing workload level throughout the task. Thus, by measuring detection performance concurrently with pupil dilation, the authors were able to provide a reasonable demonstration of pupillary sensitivity to short-term variations in mental workload.

Kahneman et al (1969) utilized a digit transformation task for loading the subject, and concurrently measured short term changes in pupil diameter, heart rate, and skin conductance. With the recordings time-locked to the stimulus and verbal response, the authors were able to demonstrate evoked responses in all three measures, with the response magnitudes roughly proportional to problem difficulty. They found pupil response to be the most sensitive indicator of problem difficulty, followed by skin conductance, and then heart rate. No attempt was made to combine all three measures into a single workload indicator.

Bradshaw (1968) used anagrams in his study of pupil response and workload, and found no significant response differences between "easy" and "hard" problems. Whether this null result is due to a poor choice of problems is unclear from the paper, since

no examples are given. The author also claimed to have investigated the effects of verbalizing verses withholding the problem answer, but the results are unclear as presented.

Simpson (1969) and Simpson and Hale (1969) considered the effects of overt subject response on overall pupil response, using a tone discrimination task. In the first study, Simpson found that pupil response was enhanced when a subject was required to overtly indicate his decision on the discrimination task, and concluded that pupil response was not only a function of the mental workload imposed on the subject, but was also dependent on the overt response requirements specified at task completion. In the second study, Simpson and Hale found that pupil response was also affected by the complexity of the overt response: when subjects were given the freedom to choose the type of overt response used to signal completion of the tone discrimination task, pupil response was greater than that obtained with a fixed overt response pattern. Both studies suggest that pupil dilation is not entirely determined by problem workload, and that overt response patterns must be carefully controlled in any experimental study of pupil sensitivity to mental workload.

Beatty and Wagoner (1978) looked at pupil response in a visual discrimination task, and suggested that the response

patterns could be used to differentiate between different levels of cognitive function. They used letter pairs as the stimuli and required subjects to determine if the two letters were the same on the basis of: a) their physical characteristics (e.g., AA or aa); b) their name (e.g., Aa); or c) their membership in the class of vowels or consonants (e.g., Ae or BR). The evoked pupil response was found to be smallest with the physical matching problem, and largest with the membership matching problem. The authors concluded that this finding was consistent with the notion of a "hierarchically organized cognitive system" which exhibits greater degrees of activation (pupil dilation) with increasing levels of problem abstraction. A simpler interpretation may be that the pupil response merely reflects the mental workload associated with tasks of different difficulty levels, an interpretation which is entirely consistent with the results of the previously described studies.

At this point it is appropriate to make some general comments regarding the use of pupillary dilation as a workload indicator.

All of the studies described here utilized a discrete problem "event" to trigger the pupil response, and thus the measurements belong in the general category of evoked response measurements. As described in the previous chapter, noise

reduction is accomplished through the use of ensemble averaging, a technique which requires multiple stimulus-response sequences. However, there is little information in the literature regarding the number of such sequences needed to adequately extract the response from the noise, and thus there is little guidance for the researcher attempting to design a workload monitor based on pupil response. One can envision a pupillometric monitor with a "probe" stimulus (such as suggested by Wickens et al (1976), for EEG monitorinG activity) to be used over a relatively long time interval for workload monitoring. If the stimulus were repeated at, say, ten-second intervals, and six pupil responses were required for each ensemble average evoked response, then a pupillometric indicator of workload would be available once per minute. Whether or not this would be an adequate sample rate would clearly depend on overall task length and the anticipated "frequency content" of the workload. Of course, higher or lower sample rates could be obtained by varying the stimulus repetition rate and/or the number of responses used in ensemble averaging, but little attention has been paid to these factors in the literature.

The notion of using a "probe" stimulus with ensemble averaging presumes that pupil measurements can only be reliable indicators in an evoked response or discrete event context. None

of the cited studies considered the continuous monitoring of pupil diameter during a "continuous" workload situation (e.g., tracking a visual target). Whether this is due to a bias of the research community primarily engaged in pupil response research, or due to the inappropriateness of such monitoring, is unclear from the literature. It would appear that some effort should be directed toward determining the usefulness of continuous monitoring of pupil diameter.

A comment should also be made on the visual environment in which pupil measurements are made. Most of the studies cited here used auditory stimuli to generate the evoked responses, and this presumably was done to avoid the confounding effects of a visual stimulus. As Goldwater (1972) discussed, factors such as intensity, color, and depth of visual target can all affect the pupil response, and any sensitive study must provide for careful control of the visual environment. This is clearly difficult even in a highly-structured experimental research situation; it may prove to be impossible in a rich visual environment (such as an aircraft cockpit), especially if the experimenter has control over neither the subject's direction of gaze nor his accommodative behavior. Thus, pupil response measurements may prove to be of quite limited utility in the relatively unstructured workload situations associated with realistic tasks.

A final comment should be made regarding the validation of pupillometric workload indicators. Most of the studies cited here simply demonstrated significant responses differences due to tasks having apparently different levels of difficulty, and little work has been done to provide an independent measure of the actual workload imposed on the subjects. As discussed earlier in section 2, this is not a fault which is unique to pupillometric research, but it clearly suggests that further validation work must be conducted, if pupil response is ever to become a reliable workload indicator.

3.2 Skin Measures

Physiological measures associated with the skin fall into two categories: temperature measurements and conduction measurements. These are discussed in the following two sections.

3.2.1 Skin Temperature

Skin temperature might be considered a candidate measure for workload assessment, because of its apparent dependence on several factors which are directly influenced by autonomic system arousal: metabolic rate, sweat rate, and blood flow volume. Few researchers have reported on the use of this measurement, however, and, at present, it is unclear as to how potentially useful such a measurement might be.

Poock et al (1969) studied changes in skin temperature occurring over the course of a 48-minute long vigilance task. By partitioning the vigil into four 12-minute segments, and computing an average value for each segment, they found a barely significant ($P < 0.08$) correlation between skin temperature and detection performance ($r = 0.38$). The suggestion is that a boredom-induced decrease in arousal lead to corresponding decreases in both temperature and performance. This notion that skin temperature may be an effective indicator of arousal (and thus possibly workload) is supported by the study's finding that temperatures averaged approximately 2° F. higher ($P < 0.01$) when target signals were detected than when missed. Presumably, arousal levels are higher with detected events, and these higher levels are reflected in corresponding changes in skin temperature.

If these inferences are valid, then it would appear that skin temperature measurements may prove of some use in workload assessment. Considerable research needs to be done however, to demonstrate reliable sensitivity.

3.2.2 Skin Conductance

Increases in skin conductance have long been associated with autonomic arousal, and, over the years, this type of response has

come to be labelled the "galvanic skin reflex", or GSR. The physical basis for the apparent resistance drop appears to be still an open issue: Venables and Martin (1967) note that although the eccrine sweat glands are most certainly implicated in the response, it is unsettled as to whether the response is due solely to the presence of increased sweat on the skin surface, or whether additional response contributions can be ascribed to presecretory activity of the sweat gland cell membranes. If the former, then a conductance measurement is in some sense equivalent to the type of measurement one would obtain from a sweat rate monitor; if the latter, then a conductance measurement must be viewed as a combined measurement, including the effects of both sweat rate and membrane conductivity. In either case, an electrically-based measurement of skin "response" has long been recognized as a physiological correlate of arousal, and has served as an objective indicator in many studies of arousal, awareness, and anxiety.

Since the greatest concentration of eccrine sweat glands occurs in the palms of the hands and the soles of the feet, most researchers have chosen to measure palmar skin resistance, because of both the response sensitivity and monitoring convenience afforded by this location. The general nature of the response is succinctly summarized by Venables and Martin (1967):

> The palmar skin resistance levels (SRLs) of normal human subjects range from a few Kohm to several hundred Kohm, depending on such factors as current density, type of electrode, etc. These SRLs often show slow and gradual changes as a function of the changing state of the individual, and in the case of sleeping or drugged subjects, resistance levels may rise to a Mohm or thereabouts.
>
> [Galvanic skin responses (GSRs)] occur as a sudden drop in resistance with a latency range of approximately 1.5 to 3.5 sec; these responses follow specific stimuli and may range from a few hundred ohm to several Kohm. In addition, spontaneous changes frequently occur which resemble stimulus induced responses in waveform. Measures commonly made of [GSRs] are latency, amplitude, duration and number.

Skin conductance measurements have classically been associated with studies of general arousal or anxiety, but more recent research has attempted to utilize this type of measurement in the direct evaluation of mental workload. A typical experimental protocol involves a subject performing a "continuous" task, such as tracking a visual target (e.g., Benson et al (1965)) or detecting an event during an imposed vigil (e.g., Eason et al (1965)). Skin conductance is monitored throughout these relatively long-term tasks which last from a few minutes to over an hour, and some form of interval time-averaging is typically used to process the raw measurements. Thus, the task interval might be divided into ten sub-intervals, with an

average measurement associated with each, and computed by time-averaging the conductance signal during the corresponding sub-interval. In effect, the signal is passed through a low-pass filter and a slow sampler, a processing method which contrasts sharply with the transient response analysis typically used in the early studies of autonomic arousal (e.g., measurement of transient waveform amplitudes and latencies). This dichotomy in signal processing approach will be considered again later in this section.

Some of the relevant studies investigating skin conductance changes as a function of mental workload are summarized in the following paragraphs.

Eason et al (1965) investigated visual detection performance during a one-hour vigilance task, and included skin conductance measurements in their chosen set of monitored physiological variables. A conductance estimate was obtained for each ten-minute segment of the task, in a rather unique manner:

> ...the [conductance] score derived for a given 10-min. segment was an average of the measures obtained during the third and eighth minutes of that particular segment. Within these 2-min. samples, measures of skin conductance were obtained for each 20-sec interval by drawing a visually "best fit" line through that segment of the record and measuring the distance from the baseline, which had a known ohm value, to the midpoint

> of the best fit line in millimeters. The values obtained were converted to conductance units...

Although this type of signal processing is difficult to characterize analytically, it suggests a considerable amount of low-pass filtering and interval sampling.

The across-subject averages obtained by Eason et al showed a significant ($P < 0.01$) decrease in skin conductance through the course of the one-hour vigil, a trend which the authors attributed to a decrease of arousal, induced by the boredom associated with the detection task. The percentage change in skin conductance was fairly small however, with an average initial value of approximately 17 micromhos* and an average final value of approximately 15 micromhos. This low sensitivity may have been a contributing factor in the study's subsequent failure to differentiate between two different experimental conditions considered in the study: a high and low presentation rate for the signal detection task. No significant skin conductivity differences were found for the two presentation rates, and one might conclude from this finding that conductivity is an insensitive workload indicator, in this type of vigilance

* A micromho is a conductance measure, and corresponds (inversely) to a resistance of one Mohm. Thus, 17 micromhos correspond to 1/17 Mohm or 58.8 Kohm.

situation. However, the three other monitored variables (heart rate, neck EMG activity, and eye blink activity) also failed to serve as differentiating indicators between the two presentation rates, and thus, it seems more reasonable to conclude that there was, in actuality, little significant workload difference between the two imposed presentation rates. The lack of significant performance differences between the two rates also supports this notion, and thus the null results suggest that task workload was not a variable in this study.

In their study of display effects on compensatory tracking, Benson et al (1965) monitored a number of physiological variables, including skin conductance (actually, resistance). A continuous record was not maintained; instead, a measurement was made every 15 seconds, and two such measurements were averaged together to yield a sample value every 30 seconds, which was then plotted against time for graphic presentation. Figure 4 is a sketch based on these time histories, illustrating the corresponding conductance trends during adjacent tracking and resting segments of the experiment. Two points are worth noting. First, one sees a gradual conductance decrement with time during both intervals, a trend which is consistent with the findings of Eason et al (1965)) just described: a gradual reduction of activation due to boredom and/or habitation. The second point

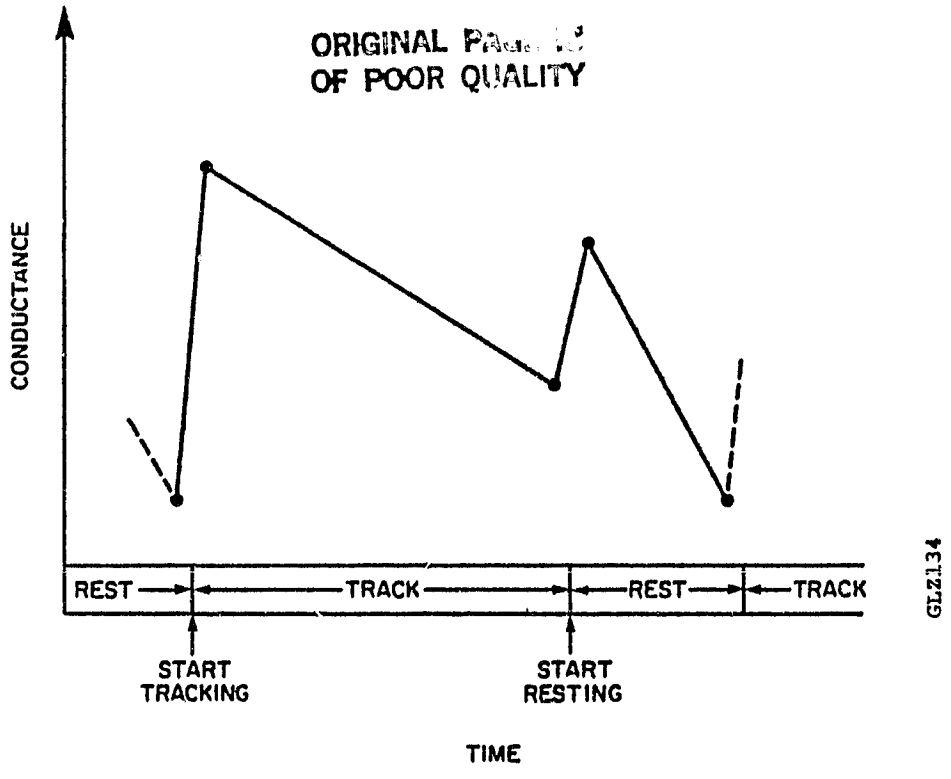

Figure 4: Conductance Trends During Tracking and Resting (after Benson et al (1965))

concerns the dramatic conductance increases associated with a track/rest transition: these occur within the time span of the intersample interval (30 seconds), and imply that relatively rapid increases in arousal are associated with the initiation of both tracking and resting. The arousal increase (i.e. conductance increase) associated with tracking initiation might be explained in terms of a step increase in task workload, but

this explanation fails to account for the arousal increase associated with the start of the rest interval. However, both can be consistently explained by attributing the arousal increase to simply a change in the subject's environment, from one of tracking to resting, and vice versa. This argument is also consistent with the converse behavior seen within an interval (tracking or resting): arousal (i.e., conductance) decreases with time, presumably due to a lack of environmental change. The suggestion then, is that the conductance change observed in the course of this experiment is a strong function of environmental stimulation and subject boredom, and may not necessarily reflect only the workload level imposed on the subject.

This study by Benson et al (1965) also examined conductance changes due to the type of tracking display used, and due to the imposition of a concurrent secondary task. The authors found no significant response difference with display type, and thus were unable to rank the displays in terms of associated arousal and/or workload, based on the skin conductivity measurements. They did, however, demonstrate a significant response difference with the imposition of a secondary task, and showed that conductance levels during tracking were enhanced when the secondary task was imposed. One might presume that the two-task situation was a higher workload situation, and that the observed conductivity

increase provided a reliable indication of the workload increment. However, it is questionable as to whether the workload did increase significantly since the authors found a significant ($P < 0.001$) decrement in primary task performance when the secondary task was imposed. Thus, one could argue that the subjects "traded-off" task demands to maintain a relatively constant workload level across the one- and two-task situations, in effect cancelling out any workload variations which may have been planned for in the experimental design.

It would appear, then, that a simpler explanation of the observed conductivity changes could be built around the notion that the two-task condition was simply a more interesting task environment, and motivated higher levels of arousal. The basic appeal of this argument, of course, is that it also provides a consistent explanation of the task-rest conductance differences described and illustrated above. Thus, from the results of this study, it would appear difficult to argue that skin conductivity measurements provided a sensitive or reliable indication of the task workload.

Poock et al (1969) studied changed in physiological activation occurring over the course of a 48-minute long visual monitoring task. Variables were recorded continuously and then analyzed by first dividing their time histories into four

12-minute segments, and then computing a (time averaged) mean value for each segment. Significant within-vigil trends were demonstrated for heart rate, blood pressure, and skin temperature, but no significant changes were seen in the skin conductance measurements. This contrasts with the decreasing trends observed by Eason et al (1965) during their vigilance experiment, but it should be recalled that their trends, though significant, were quite small. The suggestion is that skin conductance remains relatively constant over intervals lasting on the order of an hour, and that if one wishes to discern reliable conductance changes, one should limit the length of the measurement interval.

Jex and Allen (1970) monitored several physiological variables while subjects performed compensatory tracking. Based on 100-second time-averages obtained from one subject's data, they found palmar skin conductance to be consistently higher while tracking than while resting. The results confirm the findings of Benson et al (1965) discussed earlier. Although Jex and Allen also show conductance to increase with the difficulty of the tracking task, it would appear that their results are confounded with task order. Aside from the track/rest differences, it appears that their conductance measurements fail to provide any discrimination of workload differences.

Spyker et al (1971) also monitored several physiological variables while subjects performed compensatory tracking at varying levels of difficulty. As described previously in section 2.3.5, skin resistance and reactance were measured at nine different frequencies, and then fit by an adjustable four-parameter electrical model. With 18 resistance/reactance measurements, four model parameters, and five more derived parameters, the authors were able to construct a 27-dimensional conductance "feature vector" for their workload study. No significant correlations were found between any of these components and any of the objective or subjective measures of tracking task difficulty.

This null result could be explained in a number of ways. First, there may have simply been no significant workload differences among the various tracking tasks. This seems unlikely, however, since the different tasks received subjective difficulty ratings which differed significantly from one another, indicating that a real workload difference between tasks probably did exist. Since some of the other physiological measures did show significant correlations with the performance variables, the null result may have been due to the insensitivity of skin conductivity measurements; that is, conductivity may simply have been a poor workload indicator. Finally, one could ascribe the

null result to the particular method used in this study for monitoring conductivity. Instead of continuously recording conductance while tracking, and then time averaging the data over the run, Spyker et al chose a rather unusual protocol: at three minutes and ten seconds into the four-minute run, five seconds of data were recorded and saved for later frequency analysis. In effect, this five-second segment was presumed to represent conductivity over an interval almost 50 times as long. Whether or not this was the major cause of their null finding is difficult to tell, but it does suggest that it is inappropriate to conclude, on the basis of this particular study, that skin conductivity measurements fail to correlate with workload.

This study by Spyker et al used, in effect, a single point sample of skin conductivity; the other studies cited above tended towards long-term time-averaging. An intermediate approach was taken by Kahneman et al (1969), who looked at conductance changes occurring within a 20-second interval, with a measurement made every second. Using a digit-string transformation task for loading the subject, and carefully time-locking the conductance measurements within the problem presentation and verbal response, the authors were able to demonstrate an evoked skin conductance response similar to the pupil response histories described in section 3.1.2. Averaging across subjects and replications, they

obtained an average evoked response curve for each of three digit transformation tasks, and showed that the response peaks were ordered with (presumed) task difficulty. Specifically, the smallest response was associated with the easiest transformation task (adding 0 to each digit in the string) and the largest response was associated with the hardest transformation (adding 3 to each digit).

A non-parametric ordering test conducted by Kahneman et al showed their results to be fairly significant ($P < 0.02$), and supported their conclusions that, in this type of experimental situation at least, skin conductance can provide a reliable and sensitive indication of the imposed workload. It should be recognized that this study avoided the possible confounding effects of arousal and interest generated by an imposed secondary task (i.e., recall the study by Benson et al (1965)); presumably, adding 3 to a digit string is as boring a task as is adding 1 to the same string.

Perhaps more important to the success of this study, however, is the particular method used to process the conductivity measurements. Ensemble averaging over a short time span revealed a characteristic evoked response whose features could be used directly in a correlational study of imposed workload; time-averaging would have "washed-out" any significant

features, and probably would have resulted in considerably poorer correlations. In light of the success of this study, and the essentially null results of the earlier studies, it would appear that ensemble averaging techniques offer considerably greater promise for extracting meaningful conductivity "features" from monitored data. The suggestion is that conductivity studies should be patterned after the pupillometric studies discussed in section 3.1.2, with the standard complement of features associated with any evoked response study: relatively short measurement intervals and high sample rates, a workload stimulus/response protocol which is closely time-locked to the conductivity measurements, and ensemble averaging across replications (and possibly subjects).

If skin conductivity measurements were obtained and processed in this manner, then all of the earlier comments made regarding pupillometric measurements are relevant, with the obvious exception of visual field effects. The reader is referred to the latter part of section 3.1.2 for further discussion.

3.3 Muscle Measures

Physiological measures associated with muscular activity can be broken down into three categories: those associated with the

electrical activity of the muscle (electromyographic), those associated with the mechanical measurement of "involuntary" muscular activity (such as tremor), and those associated with the direct exercise of voluntary motor control. As in the case of eye measurements, the discussion here will exclude from consideration those measurements of voluntary motor control, and instead concentrate on electromyographic and mechanical measures of involuntary activity, in the following two subsections.

3.3.1 Electromyographic Measures (EMG)

Introspective accounts of fatigue often include references to a generalized muscle "tenseness," and it has been recognized for some time that both physical and mental effort can contribute to elevating the overall level of muscular activity (Golla (1921)). Since fatigue-inducing tasks often involve some sort of overt motor response, researchers have been careful to avoid confounding measurements of generalized "tenseness" with active, voluntary muscular activity. This has been accomplished by recording from muscle groups which are deemed "passive" or "irrelevant" to the overt motor response associated with the task: for example, if a subject is required to press a button with his right hand, then the "tenseness" recording might be made from his left leg. This separation of irrelevant from relevant muscle groups may be difficult or impossible in very complex

motor tasks (e.g., swinging a golf club), but appears to be well-suited to tasks which involve primarily mental effort and a minimum of motor activity.

Measurements are typically made with electrodes attached to the skin surface in the neighborhood of the muscle group(s) to be monitored. This type of recording is termed electromyographic, and in the discussion to follow, the abbreviation EMG will be used to denote not only this method of recording, but also its restricted application to the monitoring of "irrelevant" or "passive" muscle activity.

The notion that generalized muscle activity is centrally-mediated is discussed briefly by Lippold (1967):

> Benson and Gedye (1961, have investigated the mechanism of irrelevant muscle action and conclude that it is supraspinal in origin because it affects all the musculature and not only those groups supplied by the primary or nearby motorneurone pools.
>
> They have also found that the level of integrated EMG from an irrelevant muscle varies from one subject to another although the same task is being carried out by the primary muscle. Emotional factors play a large part in this phenomenon; patients with anxiety have greatly increased irrelevant responses...

This suggests that central factors such as anxiety and arousal may be the significant determinants of generalized muscle activity.

Since increases in EMG activity are associated with the onset of fatigue, and since fatigue is a function of the imposed workload (among other factors), it is not unreasonable to search for some correlation between workload and EMG activity.

Several studies have attempted to demonstrate such a correlation, by appropriate EMG signal processing. Typically, workload is imposed via a tracking task (e.g., Benson et al (1965)) or via a vigilance task (e.g., Eason et al (1965)). Since the time span for the former is on the order of minutes, while the latter may last for more than an hour, most researchers use some form of data compression, because of the fairly large bandwidth characterizing EMG signals. Signal processing might proceed as follows. First, the signal is passed through a high-pass filter to remove any drift which might be associated with electrode movement or electrochemical artifacts. Next, low-pass filtering is performed to limit instrumentation noise. Following rectification, the signal may then be passed through yet another low-pass filter, to provide a running estimate of the RMS signal power (alternatively, the rectified signal may be passed through a resetting integrator, to yield integrated EMG values on a periodic basis). The EMG "values" reported in the literature are thus often complex transformations of the original signal time history.

Some of the studies investigating EMG dependence on workload are summarized in the following several paragraphs.

In their study of display effects on compensatory tracking, Benson et al (1965)) monitored EMG activity from the left forearm flexor muscle group and the left calf extensor muscle group; neither group was involved in the overt control stick movements required by the primary tracking task, nor in the button pushes required by the associated secondary task. Results were presented only for the forearm recording, since the authors noted that the calf recordings showed similar trends.

The results of this study are difficult to interpret because it is unclear how the raw EEG signal was processed. The authors state that the signal was first bandpassed (for drift compensation and noise reduction), rectified, integrated, and finally, sampled every 30 seconds throughout the course of the four-minute run. The authors pointed out that sampling was not accompanied by resetting of the integrator, so that the sampled signal should have represented cumulative EMG activity throughout the course of a run. Since the integration was performed on a rectified signal, one would expect the sampled time history to have been a monotonic increasing function of time. The data, however, show no such behavior, and thus one is at a loss in attempting to interpret the results.

Corkindale et al (1969) used the physiological test battery developed by Benson et al (1965) to monitor pilot activity during final approach, and studied physiological dependence on three factors: time along the approach, visibility, and instrument display type. Two of the test battery measures were integrated EMG activity, one obtained from a set of leg electrodes, and the other from a set of arm electrodes. Average measurements were obtained for each 30-second interval comprising the final two minutes of the approach, but no time histories were presented. Instead, the authors conducted an analysis of variance and found the following: arm EMG activity increased significantly with time along the approach and with degraded visibility conditions, but showed no dependence on display type; leg EMG activity showed no dependence on any of the three factors.

Since the EMG activity increases are consistent with subjective accounts of increased workload near touchdown and/or under degraded visibility conditions, one might be led to conclude that EMG monitoring could provide useful workload information in realistic working environments. Unfortunately, there are some problems with this conclusion. First, the EMG measurement technique was presumably patterned after that used by Benson et al (1965), and the discussion above suggests that it is unclear what this measure of EMG "activity" actually consisted

of. Second, Corkindale et al provided neither correlation coefficients nor parameterized time histories, making it impossible to ascertain, even qualitatively, the sensitivity of their EMG index. Finally, and perhaps most significantly, the authors probably failed to discriminate between relevant and irrelevant muscle activity: they recorded from the arm and leg of a pilot actively engaged in flying the aircraft, and most likely confounded their results with voluntary motor activity associated with vehicle control. Since no data is provided regarding control activity (e.g., RMS rudder pedal deflection), it is impossible to tell whether their EMG measurements reflected generalized muscular "tension" or simply a change in the level of the pilot's overt control activity.

Eason et al (1965) monitored neck EMG activity while subjects were engaged in a one-hour vigilance task. Recordings were periodically integrated every 20 seconds, and 30 sequential integrator readings were averaged together to yield an integrated EMG measurement once every 10 minutes. The across-subject averages showed a significant ($P<0.05$) increase in integrated EMG activity through the course of the one-hour vigil, with values at the end of the vigil approximately 20% higher than those at the start of the vigil. Although detection performance decreased significantly over the course of the vigil, the authors found no significant correlation between performance and EMG trends.

Two different signal presentation rates were used in this study to see if rate-induced workload effects could be demonstrated. No such effects were found in the measured EMG activity, but this result is consistent with the fact that presentation rate also had no effect on performance nor on any of the other monitored physiological variables (blink rate, skin conductance, and heart rate). As suggested earlier in section 3.2.2, this failure to find a workload-related effect may have been due simply to an inappropriate choice of presentation rates, and not due to any inherent insensitivity of the physiological measurement set.

This study by Eason et al found a curious relationship among the observed physiological trends, one which may have a direct implication on the eventual utility of EMG monitoring. The study found that, over the course of the one-hour vigil, EMG activity increased, skin conductance decreased, and heart rate remained approximately constant. Since these trends cannot be explained on the basis of a simple change in arousal level, the authors put forth the following explanation:

> The vigilance task employed in the present experiment was of such a nature that one would expect the activity levels of the autonomic and somatic nervous systems to change in a differential manner during the course of the vigil.....

> The decrease in skin conductance, resulting from a decrease in the activity level of the sympathetic branch of the autonomic nervous system, may be attributed to the calming and drowsiness inducing effects of the environmental situation. The simultaneous increase in neck tension level [EMG activity], due to an increase in the activity level of the somatic nervous system primarily, may have resulted from an attempt to compensate for the detrimental effects of fatigue, boredom, drowsiness...
>
> The fact that heart rate remained relatively constant throughout the vigil may be interpreted as being due to concomitant changes in autonomic and somatic activity which tended to offset one another.

In effect, the authors suggest that the subjects attempted to maintain an adequate level of arousal during the vigil by gradually increasing their voluntary muscle activity.

If this is indeed the case, EMG activity must, at best, be regarded as an ambiguous indicator of arousal. Normally, an increase in EMG activity would be regarded as indicative of an <u>increase</u> in arousal. In this study by Eason et al it seems reasonable to associate the observed EMG activity increase with a <u>decrease</u> in arousal, since the observed decrement in performance and drop in skin conductance both point to an arousal decrease. Of course, these two apparently contradictory relationships are reconciled by recognizing that the human can exercise voluntary control over his own EMG activity, in attempting to control his

own arousal level. This implicit feedback structure within the human suggests that it is naive to consider EMG activity as an "open-loop" indicator of arousal; as an indicator of workload, EMG activity may even be more suspect.

Jex and Allen (1970) monitored passive forearm EMG activity while subjects performed a compensatory tracking task. After rectifying the amplified EMG signal, they did not integrate it in the conventional manner; instead, they chose to pass the signal through a very low-pass filter (5Hz), and then time-average this signal over the course of a 100-second tracking run. This long term time-averaged EMG activity then served as the basis for comparing EMG levels associated with tracking and with resting. Although the authors presented no statistical tests, it would appear from their individual subject data that the tracking-resting differences are statistically significant, suggesting that EMG level is a good indicator of the presence or absence of the imposed tracking task.

Three levels of task difficulty were investigated, with the hope of differentiating the three on the basis of the physiological measurements taken. However, no statement, pro or con, was made by the authors as to the applicability of the EMG measurements in this regard, although individual subject data indicate a slight decreasing trend in EMG levels with increasing

task difficulty. Unfortunately, as with the skin conductance measurements obtained during this study (recall the discussion of section 3.2.2), it would appear that the data are confounded by task order. Thus, fatigue effects may also be contributing to the apparent EMG trend. In short, no definitive statements can be made regarding the dependence of EMG levels on workload, based on the results of this study.

Spyker et al (1971) also monitored passive forearm EMG activity while subjects performed compensatory tracking at varying levels of difficulty. After bandpass filtering to eliminate baseline drift and instrumentation noise, a two-minute record of the signal was processed to provide an average RMS level for the middle two minutes of the four-minute tracking run. Although no comparisons were made with resting activity, the authors found significant correlations between EMG levels and both tracking error ($r=0.49$, $P<0.01$) and subjective estimates of task difficulty ($r=0.44$, $P<0.01$). Since tracking error correlated rather well with the subjective difficulty estimates ($r=0.80$), these finding suggest that EMG monitoring may indeed provide a reasonable basis for inferring workload; at the least, the finding suggest that RMS level may be a more appropriate metric for EMG quantification, when compared with some of the other schemes discussed above.

At this point it is appropriate to make some general comments regarding the use of EMG activity as a workload indicator.

All of the studies cited above used some form of long-term time-averaging to process the EMG signal (with averaging intervals ranging from 30 seconds (Benson et al (1965)) to 10 minutes (Eason et al (1965)). As argued earlier in Section 2.3, this type of long-term time-averaging provides the researcher with a convenient method of extracting a "feature" from the physiological record, but carries with it the potential of "washing-out" possibly relevant short-term trends. Thus, a 30-second or 10-minute time-averaged EMG signal may provide the researcher with a convenient method for compressing a relatively long record of high frequency data to a single numerical value; however, this procedure also misses any short-term trends in the EMG signal which may have been present and which may have proved to provide better correlations with the experimental conditions considered.

The EMG signal is characterized by a power spectrum encompassing frequencies which are orders of magnitude higher than those frequencies typically of interest to workload researchers. Thus, it is entirely reasonable to forgo detailed information on signal waveform by utilizing signal processing

techniques which extract waveform-independent features from the raw data. In classical signal processing applications, this is most often done by simply calculating the signal's average power, or its equivalent, RMS signal level. All but one of the studies cited here, however, failed to use this method of signal processing, and instead relied on various means of bandpassing, rectification, and low-pass filtering or integration. The one study which characterized the raw EMG data by its RMS signal level (Spyker et al (1971)) also reported the highest and most significant correlations between this EMG "feature" and the workload levels imposed by the experiment. The implication is that RMS level may be a more appropriate metric for EMG quantification, when compared with the other methods described earlier.

These points suggest that EMG workload correlations might be improved by avoiding long-term time-averaging and by calculating RMS signal levels. A promising approach might be based on extracting the RMS signal level for short-term intervals (say on the order of one second), thus providing a high sample rate RMS measurement. Coupled with ensemble averaging across replications and subjects, this approach might provide sufficient noise reduction to reveal meaningful short-term trends in EMG activity. Of course, evoked response techniques might also be implemented,

by using a workload stimulus/response protocol closely time-locked to the high sample rate EMG measurements. Whether or not this approach would provide improved workload correlations is of course speculative, but it would appear deserving of further consideration.

The comments so far have been concerned with signal processing techniques, but it is also appropriate to comment on a more basic aspect of measured EMG activity: its dependence on voluntary inputs. Although muscle "tenseness" might be regarded as a result of "involuntary" central factors such as anxiety or arousal, it must also be recognized that humans voluntarily regulate muscle activity levels to meet actual or perceived needs imposed by the environment. A good example is provided by the vigilance study of Eason et al (1965) cited above, which showed an apparent drop in subject arousal during the course of the vigil, but also showed a corresponding rise in EMG activity. As discussed earlier, this apparent contradiction can be resolved by assuming that each subject voluntarily increased his EMG activity level in an unsuccessful attempt to sustain his arousal level. The implication is that voluntary self-regulatory efforts may contribute to the trends seen in EMG activity levels, thus rendering them ambiguous indicators at best. Clearly, this is an open issue at present.

A final comment should be made regarding the utility of measurements in realistic workload situations. All but one of the studies cited here involved very simple motor tasks which confined voluntary muscular activity to a small set of well-defined muscle groups; thus, choosing a site for the recording of "irrelevant" muscle activity was fairly simple. In more complex tasks (such as piloting an aircraft) many more muscle groups are involved, and it may prove to be difficult or impossible to record from a muscle group which is truly "irrelevant" to the motor commands needed for the task. Even in relatively sedentary tasks (such as monitoring a control panel or display scope) one would expect to see voluntary movement which may be entirely unrelated to the task demands, but which, nonetheless, involves fairly complex motor behavior and a large number of supposedly "irrelevant" muscle groups. Unless some means is provided for identifying and isolating this type of recording artifact, it would appear that EMG monitoring may be of quite limited utility in practical workload assessment.

3.3.2 Tremor Measures

In addition to generalized muscular "tenseness", stress-induced muscular activity can also be manifested by a post-stress muscular tremor accompanied by high-frequency limb movement. A possible "stress" pathway may be hypothesized by

recognizing that emotional stress is known to induce a release of catecholamines into the bloodstream (Levi (1965)) and that limb tremor varies directly with blood concentration of adrenaline (Mardsen et al (1967)). Although this neurohumoral pathway may not be the sole route for stress-induced tremor, it does suggest that limb tremor measurement might provide a means for inferring internal stress levels. Whether or not such measurement could be extended to provide inferences regarding externally imposed workload levels is, however, unclear from the literature.

Nicholson et al (1970) conducted a study of commercial aircraft landings, and attempted to correlate subjective assessments of approach difficulty with two objective physiological measures: finger tremor and heart rate. A tremor record a few seconds long was obtained within a minute after landing, via a finger-mounted accelerometer; subsequent signal processing allowed the authors to compute the power spectral density (PSD) of the tremor, over a frequency range from 1 to 20 Hz. A comparison of the resulting spectral curves showed that "uneventful" landings seemed to be characterized by relatively flat PSD curves, whereas "stressful" landings seemed to result in PSD curves which were distinctly peaked in the neighborhood of 10 Hz. The authors thus chose as their tremor "feature" the peak power spectral density in the 10 Hz neighborhood.

Pilots were requested to provide an overall subjective assessment of landing difficulty, along with assessments of five individual factors which might influence workload (e.g., favorability of the landing aids, meteorological conditions, etc.). With a data base consisting of 34 landings, the authors then attempted to correlate the subjective assessments with the tremor measurement. No significant correlations were found.

In view of the fact that the authors also failed to find significant correlations between the subjective assessments and the other monitored physiological variable (peak heart rate near touchdown), one might be led to conclude that the approach conditions were sufficiently alike so as to preclude any differentiation based on physiological measures. This does not appear to be the case, however. Figure 5 is a schematized scatter-plot of the 34 tremor/heart-rate data pairs, based on a similar plot presented by the authors. Most of the landings were associated with low heart rate and low tremor (quadrant I); presumably, this subset comprises the majority of the "uneventful" landings. Eight of the landings were associated with a high heart rate, half of which were associated with low tremor (quadrant II) and half with high tremor (quadrant III). The authors analyzed each of the eight corresponding landings on an individual basis, and concluded that: a) the low-tremor

subset (II) was associated with high-workload approaches in which the pilot anticipated any approach/landing problems; and b) the high-tremor subset (III) was associated with high-workload approaches compounded by unfavorable conditions (e.g., poor visibility) which could give rise to unanticipated events. The authors concluded that tremor measurements could be used to differentiate between two classes of high-workload situations: those with and without possible "surprises." It should be noted, however, that this conclusions was based on anecdotal inference, and not on any firm statistical analysis of the data.

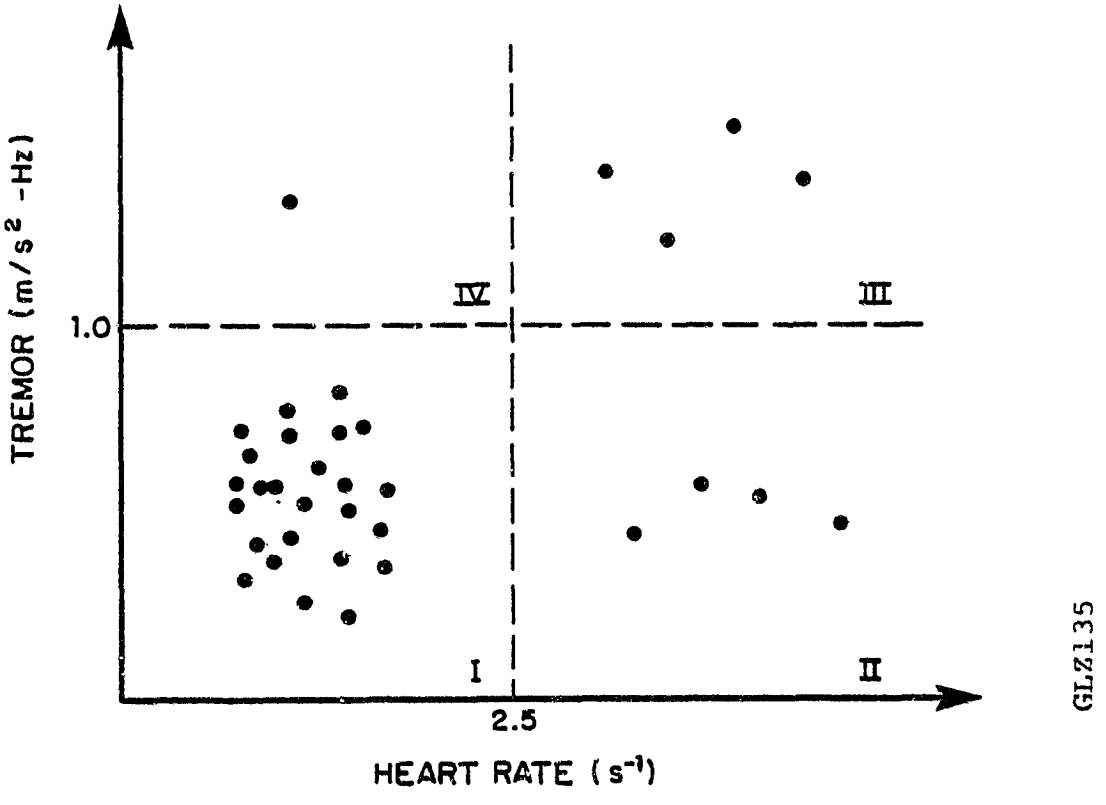

Figure 5: Finger Tremor vs. Heart Rate for 34 Landings (after Nicholson et al (1970))

Whether or not tremor may serve to differentiate among different "qualities" of workload is unclear from this study. It would seem appropriate, however, to continue research in the area of tremor measurements, to determine if there exist more direct correlations with imposed workload levels.

3.4 Circulatory Measures

Physiological measures associated with the circulatory system generally involve either a heart rate measurement, a blood pressure measurement, or a measure derived from one or both of these "primary" signals.

A review of the recent literature shows a concentration of research on one such "derived" signal: heart rate variability, or, in clinical terms, "sinus arrythmia". This measure of heart beat regularity is the subject of most of the reviewed references listed below.

Benson et al (1965)
Bergstroem and Arnberg (1971)
Boyce (1974)
Burgess and Hokanson (1964, 1968)
Connor and Lang (1969)
Corkindale et al (1969)
Kalsbeek and Ettema (1963)
Luczak and Laurig (1973)
Mulder (1967)
Mulder and Mulder-Hajonides (1973)
Nicholson et al (1970)
Nicholson et al (1973)

Deane and Zeaman (1958)
Eason et al (1965)
Ettema and Zielhuis (1971)
Firth (1973)
Haider (1972)
Jex and Allen (1970)
Jex and Clement (1977)
Kahneman et al (1969)
Kalsbeek (1971, 1973)
Opmeer (1973)
Poock et al (1969)
Rohmert et al (1973)
Sekiguichi et al (1978)
Smit and Wewerinke (1978)
Soliday and Schohan (1965)
Spyker et al (1971)
Strasser (1977)
Strasser et al (1973)
Westcott and Huttenlocher (1961)

3.5 Respiratory Measures

Physiological measures associated with the respiratory system can generally be grouped in one of two categories: those reflecting the mechanical aspects of respiratory activity (e.g., respiration rate and flow volume) and those reflecting the chemical (e.g., blood gas partial pressures). Since the latter generally require invasive measurement techniques and/or long-interval sample periods, this review has concentrated on the more simply instrumented measurements of the former.

Given below is a list of reviewed references reporting on "mechanical" measures of respiratory activity, and their capabilities for reflecting task workload.

Benson et al (1965)
Corkindale et al (1969)
Deane and Zeaman (1958)
Ettema and Zielhuis (1971)
Jex and Allen (1970)
Jex and Clement (1977)
Kalsbeek (1971)
Mulder and Mulder-Hajonides (1973)
Pettyjohn et al (1977)
Smit and Wewerinke (1978)
Soliday and Schohan (1965)
Spyker et al (1971)

3.6 Brain Electrical Measures

Electroencephalographic (EEG) recording of the brain's electrical activity can be characterized as measuring either the "free-running" electrical activity of the brain, or the "evoked" response to stimuli which elicit specific temporal patterns in the EEG. Early workload studies concentrated on how the temporal and frequency characteristics of the "free-running" signal(s) reflected imposed workload levels; more recent studies have focused on how the temporal patterns in the evoked response(s) change as a function of task workload.

Given below is a list of reviewed references reporting on EEG measures of task workload. Only a few of the studies concentrate on the free-running EEG, since the majority deal with evoked response measures.

Connor and Lang (1969)
Donchin (1978)
Donchin and Cohen (1967)
Eason et al (1969)
Haider (1972)
Low and McSherry (1968)
Moise (1978)
Murdoch (1977)
O'Donnell (1978)
O'Donnell and Hartman
O'Donnell and Spicuzza (1977)
Offenloch (1977)
Riehl (1961)
Smith et al (1970)
Spyker et al (1971)
Squires et al (1976)
Strasser et al (1973)
Wickens (1978)
Wickens et al (1976)
Wickens et al (1977)

3.7 Combined Measures

Although several of the reviewed studies monitored two or more distinct physiological signals, only a handful attempted to combine these measures into a single physiologically-based correlate of workload. Discussion of the basic approach has already been given in section 2.3; the relevant studies are listed below.

Benson et al (1965)
Corkindale et al (1969)
Poock et al (1969)
Smith and Wewerinke (1978)
Spyker et al (1971)

5.0 SUMMARY AND CONCLUSIONS

This review was conducted to assess the basis of and techniques for physiological assessment of mental workload. Although many of the studies reviewed support the notion that one or more physiological correlates may eventually provide a valid means for evaluating task workload, it would appear that particular approaches suffer not only from specific problems associated with the type of physiological monitoring involved, but also from a number of fundamental problems which underly the current development of reliable physiological workload indicators. These specific and fundamental problems are summarized here.

Correlates of eye activity fall into three categories: oculomotor visual scanning activity, eyelid blink activity, and pupillary constriction/dilation. Studies falling into the first category have not been reviewed here, because of the direct voluntary control involved in scanning. Little work has been done on blink measures, and the one study cited here showed blink rate to be a relatively insensitive measure. In contrast, a considerable amount of research has centered on pupillometric measures, and response sensitivity to changes in imposed workload has been convincingly demonstrated, at least for discrete cognitive tasks. Pupillometric measures do, however, require the

imposition of an "evoking" stimulus, and fairly sophisticated instrumentation for accurate and timely monitoring of pupil size. In addition, since non-workload related visual display factors (such as intensity, color, and depth-of-field) can significantly affect pupil size, it is questionable whether pupillometric correlates can be reliably applied in rich visual environments in which subjects are free to control their gaze and focal point (a situation which occurs with a pilot in an aircraft cockpit).

Physiological measurements associated with the skin are usually based on either temperature or conductance readings. Little effort has been devoted toward the development of a temperature-based workload metric, although the one study cited here showed some sensitivity to arousal level. Whether this effect can eventually serve as the basis for a workload indicator is unclear at present, however. Skin conductance measurements have a long association with studies of autonomic system arousal, but only a few studies have attempted to use such measures for assessing the level of imposed workload. Of the studies cited here, all but two utilized long-term time-averages of skin conductivity, with results which either were negative, or which could be attributed to gradual changes in subject arousal. Of the remaining two, only one showed significant sensitivity, due to the evoked response measurement techniques used. Clearly,

more research is needed in this area, if a sensitive and reliable skin-conductance metric is to be developed.

Physiological measurements of muscular activity fall into two categories: those based on mechanical measurements of high-frequency tremor, and those based on electromyographic (EMG) recordings. Few studies have concentrated on the former, although the one cited here did suggest at least a qualitative relationship between peak tremor power and high-workload situations in which there existed a potential for "surprise" events. The tremor metric did, however, fail to differentiate between high- and low-workload situations. In contrast, EMG-based metrics have demonstrated some sensitivity to workload variations, as shown from the studies cited here. However, there appear to be at least three specific drawbacks to the use of such metrics. First, the long-term time-averaging typically used in EMG processing would appear potentially capable of masking out significant short-term responses to changes in the imposed workload. In addition, voluntary tensing of the muscles, in response to low-workload task boredom, may give rise to totally false indications of high-workload situations. Finally, the requirement for monitoring the activity of "irrelevant" muscle groups (those not involved in task-related motor control) may prove difficult to satisfy in a complex psychomotor task (such as piloting an aircraft).

As with skin measurements, circulatory and respiratory measurements have a long association with studies of autonomic system arousal, and a correspondingly lower utilization in studies of mental workload. Circulatory measures typically involve either EKG-derived indicators of heart rate or heart rate variability (sinus arrythmia), or blood pressure measurements; non-invasive respiratory measures are usually based on measurements of respiration rate or flow volume. A few of the studies cited here have concentrated on measures involving other than heart rate variability, and have demonstrated mixed results, in terms of indicator sensitivity to imposed workload. The remaining studies have concentrated on the use of sinus arrythmia measures, and, although encouraging, it must be recognized that such measures are easily confounded both by physical workload and by arousal factors not directly associated with imposed mental workload levels.

Electroencephalographic (EEG) measures of brain activity fall into two categories: those derived from the record of a "free-running" EEG, or those derived from the time history of a stimulus-induced evoked response. Because of the great sensitivity of free-running EEG records to factors other than task workload, few studies have attempted to develop a physiologic metric based on such a measure; those that have

attempted have failed to demonstrate anything approaching a quantitative correlation between "free-running" parameters and imposed workload. In contrast, evoked response measures have shown a fair amount of promise, and this potential has motivated a number of current study efforts to concentrate in this area. At this time, however, it is unclear whether these efforts will eventually overcome some of the basic shortcomings associated with this type of physiologic correlate: the need for high signal-to-noise recording techniques, to allow for recordings in realistic (and electrically noisy) workload environments; the need for an unobtrusive evoking stimulus, which does not interfere with the basic task; and a requirement for even a rudimentary functional model of the stimulus-response relationship, to allow for the development of rational response metrics which can replace the arbitrarily-specified ones now in common use.

In addition to these specific problems associated with each type of physiological correlate, there appear to be a number of fundamental generic problems, one or more of which is associated with every study cited here. In brief, these are: the lack of an adequate definition of workload; poor or non-existent control of the workload/performance/activation triad; the lack of a rationale for dealing with the multi-dimensional nature of

physiologic state; and the failure to use commonly-accepted signal processing and functional modelling techniques. These problems are summarized below.

The search for effective physiological indicators of mental workload reflects a basic belief in the existence of some essential task-related element, which we label "task difficulty". However, although we have an intuitive notion of when one task is more difficult than another like task, we have no means of quantifying this difficulty difference. When comparing dissimilar tasks, we are even more at a loss in assessing the relative difference in imposed workload. In these situations we usually rely on either theoretical arguments concerning implicit task factors, or empirical studies of subjectively-reported task difficulty, neither of which are particularly satisfying in terms of rigorously defining and quantifying task workload.

Naturally, this is not a problem peculiar to studies directed at developing physiologically-based workload metrics. However, it should be recognized that such studies are attempting to correlate a noisy and not well-understood dependent variable, the physiologic measurement, with an unmeasurable and only poorly-understood independent variable, the task workload. Thus, the apparent failure of a physiologically-based metric need not necessarily be attributed to a lack of physiologic insensitivity;

it may simply be due to poor inferences regarding task workload. Likewise, the failure of such a metric to generalize to non-laboratory tasks may be due to the original and inappropriate quantification of the workload levels imposed by the laboratory tasks used in the metric's development.

Compounding this inherent problem is a procedural problem which characterizes many of the studies cited here: a failure to recognize the three basic dimensions involved in human task execution---physiologic activation level, inherent task difficulty, and resulting task performance. By concentrating on quantifying the first dimension, several studies have failed to control and/or identify the coordinates of the other two. The implications of such a protocol failure become evident when one considers the basic premise behind the development of a physiologic indicator: an increase in physiologic "activity" indicates an increase in task difficulty, if performance is maintained at a constant level. If performance is not monitored during or after task execution, then there is no way to ascertain if performance levels were maintained at constant levels. Thus, there is no way to validate the condition of the premise, which thus places the study results on uncertain ground.

There are several other ways by which study results may be invalidated, and which were discussed earlier in this review, but

they all stem from the failure to provide concomitant measures along all three dimensions, during the course of conducting a workload experiment. Any study which fails to meet this condition must therefore have its results viewed as tentative, if not invalid.

It should be recognized that the "dimension" of physiologic activity is, in reality, multi-dimensional, since responses have been demonstrated in several physiologic "channels". This has motivated some workers to take a multiple response approach, and, in effect define a physiologic activity "vector" which has the potential for successful correlation with imposed workload. However, it should be recognized that although correlation coefficients can be increased with the number of vector components, the rate of increase most likely decreases as more measures are included, thus reducing the marginal return. More relevant to the reliable application of such a vector metric, however, is the rate of reduction in correlation significance, as an increasing number of "noisy" physiologic indicators are included in the actuation vector. This suggests a need for developing a means of determining the "optimum" measurement set, which provides an appropriate compromise between significance and reliability. How this is to be accomplished is unclear at present.

Most of the studies cited here make use of one of two basic signal processing techniques for reducing noise in the recorded physiologic signals: time- and ensemble-averaging. The former is normally used to generate point estimates for a given measurement interval, while the latter is often used to generate signal-enhanced time histories. Time-averaging can be implemented quite simply, provides for a fair amount of data compression, and provides a means of reducing short-term fluctuations in the data (high frequency noise reduction). Ensemble-averaging requires more sophisticated processing techniques, and since it is typically used with evoked response studies, requires the use of an evoking stimulus which may interfere with the primary task. However, ensemble-averaging provides a means of detecting short-term fluctuations in the data, if the user has the "failure extraction" capability needed for processing and comparing data which has been ensemble-averaged.

Both approaches have a long and well-validated history in the signal processing research and applications conducted by researchers in other fields. Although a few of the studies cited here utilize rather unorthodox (and questionable) means of obtaining time- or ensemble-averaged data, there appears to be a more widespread problem which is of greater concern: the use of

seemingly arbitrary algorithms for transforming the averaged data into sensitive metrics of task workload.

It would appear that two factors contribute to this situation: a lack of signal processing expertise (and familiarity with orthodox and well-tested means of specifying signal characteristics), and a fundamental lack of understanding concerning the characteristics of the "black box" responsible for generating the physiologic signal of interest. The former leads to an often capricious selection of signal metrics (which are likely to be incommensurate between similar studies conducted by different researchers) while the latter promotes an unending search for a single "key" feature (of a specific physiologic variable) which can be correlated directly with task workload. It would seem that current efforts in this area are likely to continue to fail, if signal processing expertise does not begin to infuse the field, and if attempts are not made to understand the functional input/output characteristics of the physiologic system being monitored.

REFERENCES

Barnes, J.A., "Use of Eye-Movement Measures to Establish Design Parameters for Helicopter Instrument Panels," in Methods to Assess Workload, AGARD Conference Proceedings CP-216, April, 1977.

Bartoshuk, A.K., "Electromyographic Gradients as Indicants of Motivation," Canadian J. of Psychology, 9(1955) 215-230.

Beatty, J., "Pupillometric Measurement of Cognitive Workload," Proceedings of the Twelfth Annual Conference on Manual Control, NASA TMX-73,170 (May 1976).

Beatty, J., "Pupil Dilation as an Index of Workload," Proceedings of the Symposium on Man-Machine System Interface: Advances in Workload Study, Air Line Pilot's Association, Washington, D.C., 1978.

Beatty, J. and B.L. Wagoner, "Pupillometric Signs of Brain Activation Vary with Level of Cognitive Processing," Science, 199 (1978), 1216-1218.

Becker, W.L., et al., "Techniquees of Physiological Monitoring," AMRL-TDR-62-98, Aerospace Medical Research Laboratories, Wright-Patterson AFB, Sept. 1962.

Benson, A.J. and J.L. Gedye, "Some Supraspiral Factors Influencing Generalized Muscle Activity," in Proceedings of the Symposium on Skeletal Muscle Spasm, (Leicester, England: Franklin Ward and Wheeler, Ltd.) 1961.

Benson, A.J., H.F. Huddleston, and J.M. Rolfe, "A Psychophysiological Study of Compensatory Tracking on a Digital Display," Human Factors, 7 (1965), 457-472.

Bergstroem, B. and P. Arnberg, "Heart Rate and Performance in Manual Missile Guidance," Perceptual and Motor Skills, 32(1971), 352-354.

Boyce, P.R., "Sinus Arrhythmia as a Measure of Mental Load," Ergonomics, 16(1974), 177-183.

Bradshaw, J.L., "Load and Pupillary Changes in Continuous Processing Tasks," Br.J. of Psychology, 59(1968), 265-271.

Bradshaw, J.L., "Pupil Size as a Measure of Arousal During Information Processing," Nature, 216 (1967), 515-516.

Bradshaw, J.L., "Pupil Size and Problem Solving," Quarterly J. of Exp. Psychology, 20 (1968), 116-122.

Brener, J., "Heart Rate," in A Manual of Psychophysiological Methods, Eds. P.H. Venables and I. Martin, (Amsterdam: North-Holland Publishing Co.) 1967.

Burgess, M. and J. Hokanson, "Effects of Increased Heart Rate on Intellectual Performance," J. Of Abnormal and Social Psychology, 68 (1964), 85-91.

Burgess, M. and J. Hokanson, "Effects of Autonomic Arousal Level, Sex, and Frustration on Performance," Perceptual and Motor Skills, 26 (1968), 919-930.

Callaway, E., "Evoked Responses in Psychiatry," J. of Nervous and Mental Disorders, 143 (1966), 80-94.

Campos, J.J. and H.J. Johnson, "The Effects of Verbalization Instructions and Visual Attention on Heart Rate and Skin Conductance," Psychophysiology, 4(1966), 305-310.

Caspers, H., "Changes of Cortical D. C. Potentials in the Sleep-Wakefulness Cycle," in The Nature of Sleep, (London: Churchill) 1961.

Clynes, M., "Respiration Control of Heart Rate: Laws Derived From Analog Computer Simulation," IRE Transactions, ME-7 (1960), 2-14.

Connor, W.H. and P.J. Lang, "Cortical Slow-Wave and Cardiac Rate Responses in Stimulus Orientation and Reaction Time Conditions," J. of Experimental Psychology, 82(1969), 310-320.

Cooley, W.W. and P.R. Lohnes, Multivariate Procedures for the Behavioral Sciences, (New York: John Wiley and Sons) 1965.

Corkindale, K.G., F.G. Cumming and A. M. Hammerton-Fraser, "Physiological Assessment of Pilot Stress During Landing," in Measurement of Aircrew Performance, AGARD Conference Proceedings, CP #56, Brooks Air Force Base, Texas, 1969.

Deane, G.E. and D. Zeaman, "Human Heart Rate During Anxiety," Perceptual and Motor Skills, 8(1958), 103-106.

Defayolle, M., J.P. Dinand, and M.T. Gentille, "Averaged Evoked Potentials in Relation to Attitude, Mental Load, and Intelligence," in Measurement of Man at Work, Eds. W.T. Singleton, J.G. Fox, D. Whitfield. (London: Taylor and Francis Ltd.) 1971.

Donchin, E. "Brain Electrical Correlates of Pattern Recognition," in Signal Analysis and Pattern Recognition in Biological Engineering, Ed. G.F. Inbar. (New York: John Wiley and Sons) 1975.

Donchin, E., "Brain Electrical Activity as an Index of Mental Workload in Man-Machine Systems," in Proceeding of the Symposium on Man-Machine System Interface: Advances in Workload Study, Air Line Pilot's Association, Washington, D.C., 1978.

Donchin, E. and L. Cohen, "Average Evoked Potentials and Intramodality Selective Attention," EEG and Clinical Neurophysiology, 22(1967), 537-546.

Eason, R.G., A. Beardshall, and S. Jaffee, "Performance and Physiological Indicants of Activation in a Vigilance Situation," Perceptual and Motor Skills, 20(1965), 3-13.

Eason, R.G. and M.R. Harter, and C.T. White, "Effects of Attention and Arousal on Visually Evoked Cortical Potentials and Reaction Time in Man," Physiology and Behavior, 4(1969), 283-289.

Eason, R.G. and C.T. White, "Relationship Between Muscular Tension and Performance During Rotary Pursuit," Perceptual and Motor Skills, 10(1960), 199-210.

Ettema, J.H. and R.L. Zielhius, "Physiological Parameters of Mental Load," Ergonomics, 14(1971), 137-144.

Firth, P., "Psychological Factors Influencing the Relationship Betwen Cardiac Arrhythmia and Mental Load," Ergonomics, 16(1973), 5-16.

Gardner, R.M., J.S. Beltramo, R. Krinsky, "Pupillary Changes During Encoding, Storage, and Retrieval of Information," Perceptual and Motor Skills, 41(1975), 951-955.

Gartner, W. and M. Murphy, "Pilot Workload and Fatigue: A Critical Survey of Concepts and Assessment Techniques," NASA TN D-8365, 1976.

Gerathewohl, S.J., E.L. Brown, J.E. Burke, K.A. Kimball, W.F. Lowe, and S.P. Stackhouse, "Inflight Measurement of Pilot Workload: A Panel Discussion," Aviation, Space, and Environmental Medicine, 49(1978), 810-822.

Goldwater, B.C., "Psychological Significance of Pupillary Movements," Psychological Bulletin, 77(1972), 340-355.

Groll-Knapp, E., "Evoked Potentials and Behavior" in Measurement of Man at Work, Eds. W.T. Singleton, J.G. Fox, D. Whitfield, (London: Taylor and Francis, Ltd.) 1971.

Haider, M., "Measurement of Mental Load During Complex Psychomotor Skills," in Displays and Control, Eds. R.K. Bernotat and K.P. Gartner, (Amsterdam: Swets & Zeitlinger) 1972.

Haider, M., "Comparison of Objective and Subjective Methods of the Measurement of Mental Workload," in Displays and Controls, Eds. R.K. Bernotat and K.P. Gartner, (Amsterdam: Swets & Zeitlinger) 1972.

Hakerem, G., "Pupillography," in A Manual for Psychophysiological Methods Eds. P.H. Venables and I. Martin (Amsterdam: North-Holland Publishing Co.) 1967.

Hammerton-Fraser, A.M. and G.F. Morgan, "An Index of Mental Activity from Digital Sampling of the Galvanic Skin Response," RAF Institute of Aviation Medicine Report No. 431.

Harper, B. and G. Cooper, "A Revised Pilot Rating Scale for the Evaluation of Handling Qualities," Cornell Aeronautical Laboratory Report, No. 153, Sept. 1966.

Hess, E.H. and J.M. Polt, "Pupil Size in Relation to Mental Activity During Simple Problem Solving," Science, 143(1964), 1190-1192.

Jex, H.R. and R.W. Allen, "Research on a New Human Dynamic Response Test Battery," Proceedings of the Sixth Annual Conference on Manual Control, Wright-Patterson Air Force Base, Ohio, April 1970.

Jex, H.R. and W.F. Clement, "On Defining and Measuring Perceptual-Motor Workload in Manual Control Tasks," STI Working Paper 1104-1, Systems Technology Inc., July 1977.

Jex, H.R. , J.D. McDonnell, and A.V. Phatak, "A Critical Tracking Task for Man-Machine Research Related to the Operator's Effective Delay Time," NASA CR-616, November 1966.

John, E.R., Neurometrics: Clinical Applications of Quantitative Electrophysiology, (New York: John Wiley and Sons) 1977.

John, E.R., et al, "Neurometrics," Science, 196(1977), 1393-1408.

Kahneman, D., Attention and Effort, (Englewood Cliffs, NJ: McGraw-Hill) 1973.

Kahneman, D. and J. Beatty, "Pupil Diameter and Load on Memory," Science, 154(1966), 1583-1585.

Kahneman, D. and J. Beatty, "Pupillary Responses in a Pitch Discrimination Task," Perception and Psychophysics, 2(1967), 101-105.

Kahneman, D., J. Beatty and I. Pollack, "Perceptual Deficit During a Mental Task," Science, 157(1967), 218-219.

Kahneman, D., B. Tursky, D. Shapiro and A. Crider, "Pupillary, Heart Rate, and Skin Resistance Changes During a Mental Task," J. of Experimental Psychology, 79(1969), 164-167.

Kalsbeek, J.W.H., "Sinus Arrythmia and the Dual Task Method in Measuring Mental Load," in Measurement of Man at Work, Eds. W.T. Singleton, J.G. Fox, D. Whitfield (London: Taylor and Francis Ltd.) 1971.

Kalsbeek, J.W.H, "Do You Believe in Sinus Arrhythmia," Ergonomics, 16(1973), 99-104.

Kalsbeek, J.W.H. and J.H. Ettema, "Continuous Recording of Heart Rate and Measurement of Perceptual Load," Ergonomics, (1963), 306-307.

Kalsbeek, J.W.H. and J.H. Ettema, "Sinus Arrhythmia and the Measurement of Mental Load," Communication at the London Conference of the British Psychological Society, December 1965.

Khalafalla, A.L., Turner, and D. Spyker, "An Electrical Model to Simulate Skin Dielectric Dispersion," Fifth Annual Meeting, Association for the Advancement of Medical Instrumentation, March 1970.

Lacey, J.I., "Individual Difference in Somatic Response Patterns," J. of Comparative and Physiological Psychology, 43 (1950), 338-350.

Levy, E.Z., G.E. Johnson, J. Serano, V.H. Thaler, G.E. Ruff, "The Use of Skin Resistance to Monitor States of Consciousness," Aerospace Medicine 31, (1961), 60-66.

Lindsley, D.B. "The Reticular Activating System and Perceptual Integration," in Electrical Stimulation of the Brain, Ed. D.E. Sheer, (Austin, Texas) 1961.

Lippold, O.C.J., "Electromyography," in A Manual of Psychophysiological Methods, Eds. P.H. Venables and I. Martin, (Amsterdam: North-Holland Publishing Co.) 1967.

Lorens, S.A. and C.W. Darrow, "Eye Movements, EEG, GSR, and EKG During Mental Multiplication," EEG and Clinical Neurophysiology, 14(1962), 739-746.

Low, M. and J. McSherry, "Further Observations of Psychological Factors Involved in CNV Genesis," EEG and Clinical Neurophysiology, 25(1968), 203-207.

Luczak, H. and W. Laurig, "An Analysis of Heart Rate Variability," Ergonomics, 16 (1973), 85-97.

Lywood, D.W., "Blood Pressure," in A Manual of Psychophysiological Methods, Eds. P.H. Venables and I. Martin, (Amsterdam: North-Holland Publishing Co.) 1967.

Malmo, R.B., "Activation: A Neurological Dimension," Psychological Review, 66(1959), 367-386.

Margerison, J.H., P. St. John-Loe, C.D. Binnie, "Electroencephalography," in A Manual for Psychophysiological Methods Eds. P.H. Venables and I. Martin, (Amsterdam: North-Holland Publishing Co.) 1967.

McCleary, R.A., "A Simple Method for the Physiological Measurement of Anxiety," Aviation Medicine, (1953), 508-513.

Michon, J.A., "Tapping Regularity as a Measure of Perceptual Motor Load," Ergonomics, 9(1966), 401-412.

Moise, S., "Brain Electrical Activity and Eye Movement," in Proceedings of the Symposium on Man-Machine System Interface: Advances in Workload Study, Air Line Pilot's Association, Washington, D.C., 1978.

Mulder, G., "The Heart of Mental Effort," Position Paper for NATO Symposium on Mental Workload, Mati, Greece, Sept. 1967.

Mulder, G. and W.R.E.H. Mulder-Hajonides, "Mental Load and the Measurement of Heart Rate Variability," Ergonomics, 16(1973), 69-83.

Murdoch, B.D., "The Electroencephalogram in Aircrew Selection and Aviation Medicine: A Survey of Literature," National Institute for Personnel Resesarch, Council for Scientific and Industrial Research, Johannesburg, Republic of South Africa, CSIR Special Report PERS 268, July 1977.

Nicholson, A.N., "Aircrew Workload During the Approach and Landing," Aeronautical Journal, 77(1973), 283-289.

Nicholson, A.N., L.E. Hill, R.G. Borland, and H.M. Ferres, "Activity of the Nervous System During the Let-down, Approach and Landing: A Study of Short Duration High Workload," Clinical Aviation and Aerospace Medicine, 41(1970), 436-446.

Nicholson, A.N., L.E. Hill, R.G. Borland, and W.J. Krzanowski, "Influence of Workload on the Neurological State of a Pilot During the Approach and Landing," Aerospace Medicine, 44(1973), 146-152.

O'Donnell, R., "Brain Electrical Activity in the Assessment of Workload," in Proceedings of the Symposium on Man-Machine System Interface: Advances in Workload Study, Air Line Pilot's Association, Washington, D.C., 1978.

O'Donnell, R.D. and B.O. Hartman, "Contributions of Psychophysiological Techniques to Aircraft Design and Other Operational Problems," Aerospace Medical Research Laboratories, Wright-Patterson AFB, Ohio (in press).

O'Donnell, R. and R. Spicuzza, "Visually Evoked Brain Potentials as Aids in Display Design," Frontiers in Medical Signal Processing, MIDCON 77, Nov. 1977.

Offenloch, K., "Neurophysiological Assessment of Functional States of the Brain," in Methods to Assess Workload, AGARD Conference Proceedings, CP-216, April 1977.

Opmeer, C.H.J.M., "The Information Content of Successive RR-Interval Times in the ECG," Ergonomics, 16(1973), 105-112.

Payne, D.T., M.E. Parry and S.J. Harasymiw, "Percentage of Pupillary Dilation as a Measure of Item Difficulty," Perception and Psychophysics, 4(1968), 139-143.

Pettyjohn, F.S., R.J. McNeil, L.A. Akers, and J.M. Faber, "Use of Inspiratory Minute Volume in Evaluation of Rotary and Fixed-Wing Pilot Workload," in Methods to Assess Workload, AGARD Conference Proceedings, CP-216, April 1977.

Pin, M.C., F. Lecret, et M. Pottier, "Les niveaux d'activation lors de differentes situations de conduite," Bulltein de l'Organisme National du de Securite Routiere, 19(1969), 1-11.

Poock, G.K., G.A. Tuck, and J.H. Tinsley, "Physiological Correlates of Visual Monitoring," Perceptual and Motor Skills, 29(1969), 334.

Pribram, K.H. and D. McGuiness, "Arousal, Activation, and Effort in the Control of Attention," Psychology Review, 82(1975) 116-149.

Regan, D., Evoked Potentials in Psychology, Sensory Physiology, and Clinical Medicine, (London: Chapman and Hall, Ltd.) 1972.

Reihl, J.H., "Analog Analysis of EEG Activity," Aerospace Medicine, 32(1961), 1101-1108.

Rohmert, W., W. Laurig, U. Philipp, H. Luczak, "Heart Rate Variability and Work-Load Measurement," Ergonomics, 16(1973), 33-44.

Rolfe, J.M., "The Secondary Task as a Measure of Mental Load," in Measurement of Man at Work, Eds. W.T. Singleton, J.G. Fox, and D. Whitfield, (London: Taylor and Francis Ltd.) 1971.

Ruff, G.E., "Psychological and Psychophysiological Indices of Stress," in Unusual Environments and Human Behavior, Ed. N.M. Burns, (The Free Press of Glencoe, NY) 1963.

Sekiguichi, C., Y. Handa, M. Gotoh, Y. Kurihara, A. Nagasawa, and I. Kuroda, "Evaluation Method of Mental Workload Under Flight Conditions," Aviation, Space and Environmental Medicine, 49(1978), 920-925.

Shackel, B., "Eye Movement Recording by Electro-Oculography," in A Manual of Psychophysiological Methods, Eds. P.H. Venables and I. Martin, (Amsterdam: North-Holland Publishing Co.) 1967.

Simpson, H.M., "Effects of a Task-Relevant Response on Pupil Size," Psychophysiology, 6(1969), 115-121.

Simpson, H., and S. Hale, "Pupillary Changes During a Decision-Making Task," Perceptual and Motor Skills, 29(1969), 495-498.

Smit, J. and P.H. Wewerinke, "An Analysis of Helicopter Pilot Control Behavior and Workload During Instrument Flying Tasks," National Aerospace Laboratory NLR, The Netherlands, Report NLR MP 78003 U, 1978.

Smith, D.B., E. Donchin, L. Cohen, and A. Starr, "Auditory Averaged Evoked Potentials in Man During Selective Binaural Listening," EEG and Clinical Neurophysiology, 28(1970), 146-152.

Soliday, M.S. and B. Schohan, "Performance and Physiological Responses of Pilots in Simulated Low-Altitude High-Speed Flight," Aerospace Medicine, 36(1965), 100-104.

Spong, P., M. Haider, and D.B. Lindsley, "Selective Attentiveness and Cortical Evoked Responses to Visual and Auditory Stimuli," Science, 148(1965), 395-397.

Spyker, D.A., S.P. Stackhouse, A.S. Khalafalla, and R.C. McLane, "Development of Techniques for Measuring Pilot Workload," NASA CR-1888, Nov. 1971.

Squires, K., and E. Donchin, "Beyond Averaging: The Use of Discriminant Functions to Recognize Event Related Potentials Elicited by Single Auditory Stimuli," Electroencephalography and Clinical Neurophysiology, 41(1976), 449-459.

Squires, K., C.D. Wickens, N. Squires, and E. Donchin, "The Effect of Stimulus Sequence on the Waveform of the Cortical Event Related Potential," Science, 193(1976), 1142-1146.

Stennett, R.G., "The Relationship of Performance Level to Level of Arousal," J. of Experimental Psychology, 54(1957), 54-61.

Strasser, H., "Physiological Measures of Workload: Correlations Between Physiological Parameters and Operational Performance," in Methods to Assess Workload, AGARD Conference Proceedings, CP-216, April 1977.

Strasser, H., G. Brilling, K.P. Klinger, and W. Muller-Limmoth, "Physiological and Operating State of a Group of Aeroplane Pilots Under the Conditions of Stressing Tracking Tests," Aerospace Medicine, 44(1973), 1040-1047.

Surwillow, W.W., "Psychological Factors in Muscle Action Potentials: EMG Gradients," J. of Experimental Psychology, 52(1956), 263-272.

Ursin, H. and R. Ursin, "Physiological Indicators of Mental Workload," presented at NATO Workshop on Mental Workload, Mati, Greece, 1977.

Venables, P.H. and I. Martin, "Skin Resistance and Skin Potential," in A Manual of Psychophysiological Methods, Eds. P. H. Venables and I. Martin, (Amsterdam: North-Holland Publishing Co.) 1967.

Walter, W.G., R. Cooper, V.L. Aldridge, W.C. McCallum and A.L. Winter, "Contingent Negative Variations: An Electric Sign of Sensorimotor Association and Expectancy in the Human Brain," Nature (London), 204(1964), 380-384.

Westcott, M. and J. Huttenlocher, "Cardiac Conditioning: The Effects and Implications of Controlled and Uncontrolled Respiration," J. Experimental Psychology, 61(1961), 353-359.

Wickens, C., "Brain Electrical Activity as an Index of Workload," in Proceeding of the Symposium on Man-System Interface: Advances in Workload Study, Air Line Pilot's Association, Washington, D.C., 1978.

Wickens, C.D., J. Isreal, and E. Donchin, "The Event Related Cortical Potential as an Index of Task Workload," Proceedings of the Human Factors Society, 21st Annual Meeting, 1977, 282-286.

Wickens, C.D., J. Israel, G. McCarthy, D. Gopher, and E. Donchin, "The Use of Event Related Potentials in the Enhancement of System Performance," Proceedings of the Twelfth Annual Conference on Manual Control, NASA TM X-73,170, 1976.

Wisner, A., "Electrophysiological Measures for Tasks of Low Energy Expenditure," in Measurement of Man at Work, Eds. W.T. Singleton, J.G. Fox, B. Whitfield. (London: Taylor and Francis Ltd.) 1971.

Young, L.R., and D. Sheena, "Survey of Eye Movement Recording Methods," Behavior Research Methods and Instrumentation, 7 (1975), 397-429.

Zwaga, H.J.G., "Psychophysiological Reactions to Mental Tasks: Effort or Stress?", Ergonomics, 16(1973), 61-67.