A Review of the Mental Workload Literature

Brad Cain
Defence Research and Development Canada Toronto
Human System Integration Section
1133 Sheppard Avenue West
Toronto, Ontario M3M 3B9
CANADA
Telephone: 1-416-635-2195

E-mail: brad.cain@drdc-rddc.gc.ca

1.0 INTRODUCTION

The intent of this paper is to provide the reader with an overview of the mental workload literature. It will focus on other state of the art surveys with reference to some specific reports of the practical application of mental workload measurement. The surveys will be limited to those in English. Manzey[1] reportedly provides a review of psychophysiological methods in German; a comparable, recent review in English was not found although a NATO RTO report (Wilson 2004, pp. 64-65 and Chapter 8) provides some guidance in this respect. Appendix 1 lists a search for references to workload measurement techniques using the GOOGLE[2] search engine. The intent is to give the reader an appreciation of where work has been focused, or at least as reported on the Internet.

1.1 Definitions of Workload

Despite interest in the topic for the past 40 years, there is no clearly defined, universally accepted definition of workload. Huey and Wickens (1993, p. 54) note that the term "workload" was not common before the 1970's and that the operational definitions of workload from various fields continue to disagree about its sources, mechanisms, consequences, and measurement." Aspects of workload seem to fall within three broad categories: the amount of work and number of things to do; time and the particular aspect of time one is concerned with; and, the subjective psychological experiences of the human operator (Lysaght, Hill et al. 1989).

Workload is thought of as a mental construct, a latent variable, or perhaps an "intervening variable" (Gopher and Donchin 1986, p. 41-4), reflecting the interaction of mental demands imposed on operators by tasks they attend to. The capabilities and effort of the operators in the context of specific situations all moderate the workload experienced by the operator. Workload is thought to be multidimensional and multifaceted. Workload results from the aggregation of many different demands and so is difficult to define uniquely. Casali and Wierwille (1984) note that as workload cannot be directly observed, it must be inferred from observation of overt behaviour or measurement of psychological and physiological processes. Gopher and Donchin (1986, p. 41-2) feel that no single, representative measure of workload exists or is likely to be of general use, although they do not provide guidance on how many workload measures they feel are necessary or sufficient.

There are few formal definitions of workload. Most definitions are of the form:

[1] Manzey, D. (1998) Psychophysiologie mentaler beanspruchun. In F.Rösler (Ed.) Ergebnisse und Anwendungen der Psychopsychologie, Serie 1. Biologische Psychologie. Enzyklopãdie der Psychologie. Göttingen, Germany: Hogrefe.

[2] http://www.google.com

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 JUL 2007	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Review of the Mental Workload Literature	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defence Research and Development Canada Toronto Human System Integration Section 1133 Sheppard Avenue West Toronto, Ontario M3M 3B9 CANADA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002028.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**34**	

1) "Mental workload refers to the portion of operator information processing capacity or resources that is actually required to meet system demands." (Eggemeier, Wilson et al. 1991, p. 207)
2) "... mental workload may be viewed as the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time." (Gopher and Donchin 1986, p. 41-3)
3) "… the mental effort that the human operator devotes to control or supervision relative to his capacity to expend mental effort … workload is never greater than unity." (Curry, Jex et al. 1979)
4) "... the cost of performing a task in terms of a reduction in the capacity to perform additional tasks that use the same processing resource." (Kramer, Sirevaag et al. 1987)
5) "... the relative capacity to respond, the emphasis is on predicting what the operator will be able to accomplish in the future." (Lysaght, Hill et al. 1989, p. 27)

Gopher and Braune (1984) suggest that the workload construct was conceived to explain the inability of human operators to cope with the requirements of a task, and that workload measures are an attempt to characterize performance of a task relative to the operator's capability. They note that there is little knowledge to link the measurement of workload by any one paradigm to others and the lack of a formal theory of workload has lead to a proliferation of disparate methods with little chance of reconciliation. Gopher and Braune's early findings seemed to argue that workload reflects demands on a single, undifferentiated pool of resources (Gopher and Braune 1984, p. 530), where all tasks interact similarly and concurrent task demands were principally additive with a constant "overhead". This strict perspective is no longer held. It is now thought that the human information processor is appropriately represented as comprising multiple resources that are engaged differently according to the characteristics of the task demands (Jex 1988; Wickens and Hollands 1999). Although task demands and operator capabilities may be multidimensional, it is unclear whether the **conscious perception of workload** should be represented this way or as a single, scalar quantity.

Mental workload can be influenced by numerous factors that make a definitive measurement difficult. Jex (1988) implies that mental workload derives from the operator's meta-controller activities: the cognitive "device" that directs attention, copes with interacting goals, selects strategies, adjusts to task complexity, sets performance tolerances, etc. This supports the intuitive notion that workload can be represented as a function, and the utility of univariate workload measures as globally sensitive estimates of workload, while acknowledging that tasks of differing characteristics interfere differently. Alternatively, Wierwille (1988, p. 318) suggests that an operator faced with a task is fully engaged until the task is done, then is idle or engages in another task. It is not clear how this can be reconciled with multitask performance demonstrating interference effects without resorting to some manner of time sharing among concurrent tasks. Wierwille's position seems to preclude interleaving idle and active intervals during task execution.

Workload is frequently described by terms such as *mental strain* ("... the concept of mental effort ...") and *emotional strain* ("... the excess mental effort that comes form anxiety evoking cognitive aspects of the task ..."). Boucsein and Backs (1999, p. 8) outline what is perhaps an alternative formulation for representing workload or strain, as a Three Arousal Model, more tightly coupling emotions and stress to workload. Gaillard (1993) maintains that workload and stress, while related, lack proper and distinct definitions. Both stress and workload involve environmental demands and the ability of the operator to cope with those demands, but these two concepts come from different theoretical backgrounds. Gaillard separates workload from emotion, with both under the control of a higher, mental mechanism, similar to a metacontroller. If this workload is a manifestation of the investment of effort by the metacontroller, then affective factors play a complementary role to information processing in the perception of workload, best represented as a two

dimensional model of cognitive energy mobilization. Information-processing models would be incomplete according to this perspective.

Colle and Reid (1999) state that "... the concept of mental workload is an applied construct and ... does not have a one-to-one relationship with attentional capacity or resources in information processing theories." Colle and Reid focus on the amount of mental work that can be accomplished within a period of time; "... mental workload is considered to be the average rate of mental work ...". They describe a procedure for defining workload or demand equivalence of tasks using double trade-off evaluations, but they do not present a philosophy to identify the appropriate time interval for a task in such an assessment. They present the results of three experiments as support for their proposal to develop a globally sensitive secondary task measure battery. Huey and Wickens (1993, pp. 57-68) provide a good overview of many of the external, task factors contributing to workload.

In summary, a commonly accepted, formal definition of workload does not exist. Workload can be characterized as a mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, coupled with the capability of the operator to respond to those demands. Operational definitions will likely continue to be proposed and tested, but unless an imperative need arises for a universal definition, each field and perhaps each investigator will continue with their "culturally preferred" definition of workload.

1.2 Reasons for Measuring Workload

The principal reason for measuring workload is to quantify the mental cost of performing tasks in order to predict operator and system performance. As such, it is an interim measure and one that should provide insight into where increased task demands may lead to unacceptable performance. Wickens (1992, p. 390) asserts "... performance is not all that matters in the design of a good system. It is just as important to consider what demand a task imposes on the operator's limited resources. Demand may or may not correspond with performance." Mulder, Mulder et al. (1999, p. 140) note "the main reason to measure physiological activity during and after mental work is to assess the costs involved in performing mental tasks and to measure the duration of the imposed effects upon the task performer." Presumably, these purposes are only interim objectives in applied or laboratory settings; the ultimate objective is assumed to be improved working conditions, intuitive workstation design, or more effective procedures.

There may also be legal reasons to measure workload. Workload measurement during the assessment of new user interfaces may be a requirement in order to attain certification for use; for example, certification of new aircraft cockpit designs[3]. The certification process may specify the method of workload measurement selected and, hopefully, there are some rational, validated criteria justifying its use as a surrogate for in-service performance.

In the comparison of system designs, procedures, or manning requirements, workload measurement can be used to assess the desirability of a system if performance measures fail to differentiate among the choices. Implicit in this approach is the belief that as task difficulty (workload) increases: performance usually decreases; response times and errors increase; control variability increases; fewer tasks are completed per unit time; task performance strategies change (Huey and Wickens 1993); and, there is less residual capacity to deal with other issues. There is, however, strong evidence to show this is not necessarily the case for monitoring or vigilance applications. In monitoring applications, workload may be considered low despite the difficulty of

[3] FAA (1993) Airworthiness Standards. FAR Parts 23 and 25, Appendix D. http://ecfr.gpoaccess.gov

maintaining attention. The dissociation between workload and performance is generally acknowledged although not well understood. Nevertheless, part of the system designer's objective is to optimize system performance and workload is considered one of the factors to be considered in the optimization process (Mitchell 2000).

1.3 Criteria for Workload Measurement Methods

If workload is being measured in an experimental setting, the measurement options are generally wider than those for an operational setting. Many of the workload measurement techniques can be used successfully to differentiate among empirical conditions, and perhaps even produce interval or ratio measures.

Some concerns about the practical application of workload measures based on laboratory studies are: the lack of ecological validity or context complexity; the lack of subject acceptance, commitment or expertise; the lack of assessment of the effect of strategy shifts both on performance, scheduling, and on the workload measurements themselves. To address these and other concerns, O'Donnell and Eggemeier (1986) proposed several criteria to guide the selection or development of mental workload measurement techniques:

1) The method must be reliably sensitive to changes in task difficulty or resource demand and discriminate between significant variations in workload.
2) The method should be diagnostic, indicating the source of workload variation and quantify contribution by the type or resource demand.
3) The method should not be intrusive or interfere with performance of the operator's tasks, becoming a significant source of workload itself.
4) The method should be acceptable to the subjects, having face validity without being onerous.
5) The method should require minimal equipment that might impair the subject's performance.

Other criteria have since been added to this list:

6) The method should be timely and sufficiently rapid to apply to capture transient workload changes.
7) The method should be reliable, showing repeatability with small variance compared with main effects.
8) The method should be selectively sensitive to differences in capacity demand and not to changes unrelated to mental workload (such as emotional stress; others such as Gaillard (1993) or Gaillard and Kramer (1999) might debate this restriction).
9) The method should be insensitive to other task demands, such as physical activity beyond the conduct of the tasks. (Casali and Wierwille 1984, p. 1034)

The measurement technique selected should also meet formal, axiomatic constraints. Colle and Reid (1997) note that only after workload measurement techniques adopt this approach, such that subjective and performance (and physiological) measures can be directly compared, can we fully understand the concept of workload. Others had related concerns that would feed into formal measurement methods, characterizing not only the tasks, but also the operator-abilities and costs incurred performing tasks (Derrick 1988). Objective measurement techniques, while attending to the technical requirements of measuring the physical quantity involved, do not address how these measurements can be transformed into a workload measure. The work of Colle and colleagues on subjective scale development has potential in objective performance and physiological measures as well the subjective scales.

Some measurement techniques attempt to address individual differences, either through developing weights to proportion scale ratings to an overall workload calculation that are operator specific (as in SWAT and NASA TLX) or through base-lining, such as in the physiological measures. Many researchers feel that the concept of individual differences is central to workload measurement. As such, measurement techniques should be designed to capture those differences and reflect them in the values obtained from a sound theoretical framework.

As mentioned, the choice of workload measurement technique for an operational application is more constrained than in a laboratory setting, or an empirical setting. Measurement of workload by several, unobtrusive techniques has been proposed to take some pre-emptive, mitigating action to maintain performance, such as automated aiding, so real time responses are necessary. The measurement technique has to have minimal interference with operator activity, either mental or physical.

2.0 NOTABLE REVIEWS OF WORKLOAD MEASUREMENT TECHNIQUES

The most organized discussion of workload assessment and its measurement was found in Chapters 41 and 42 of the "Handbook of Perception and Human Performance" (Boff, Kaufman et al. 1986). In Chapter 41, Gopher and Donchin (1986) present both a historical and a state of the art review of workload, its definition and early models up until the mid 1980s. Gopher and Donchin give an introduction to the various classes of workload measurement, noting advantages, disadvantages, and controversies about their use. In Chapter 42, O'Donnell and Eggemeier (1986) expand on Gopher and Donchin's overview, discussing numerous specific measurement techniques for each workload measurement class. These are useful references for learning the fundamentals of workload measurement as well as identifying potential measurement techniques.

Moray (1979), Roscoe, Ellis et al. (1978) provide collections of papers in separate monographs that also review the state of knowledge up to the 1980s. The papers cover a broad range of topics from the development of specific measurement techniques to more general reviews of the philosophy and problems associated with workload measurement. Moray's (1979) NATO workshop proceedings are among the early attempts to collate and organize "… the enormous amount of knowledge about workload and many models for it …" into a coherent model that would be theoretically sound and practicable. It involved psychologists, engineers, modellers, and ergonomists, resulting in almost 30 papers. While it did not create a functional workload model, it did serve to focus and inform the various communities on various aspects of workload measurement. Some of the work in Moray (1979) is still relevant and useful, although the science has advanced somewhat in the past 25 years; nevertheless, this monograph also makes a good starting point on various aspects of workload measurement. Roscoe, Ellis et al. (1978) provide an early review of the state-of-the-art, referencing works from the 1950s that predated current workload theories, associating workload both with mental effort (without defining effort formally) and with the extent to which an operator is engaged by his duties. They acknowledge task demands, operator capabilities, and contextual temporal demands as being components of workload.

An interesting proposal from the experimental psychology group of Moray's NATO symposium was that not only is workload multidimensional, it should be considered as a vector rather than the more typical scalar quantity. Further, this vector representation will be task specific (or perhaps specific classes of tasks), although they readily admit they do not know the dimensions of such a vector. An analogous problem is that of speed and velocity. Speed is the magnitude of the velocity vector, but in some situations, direction counts such as in the fuel requirements for aircraft flying in winds of different headings, covering similar distances in similar durations: to the airline the result is important to their profitability, but to the passenger, the result isn't

noticeable. Some measurement techniques may be sensitive to specific components of the workload vector only and insensitive to others. Alternatively, some measures may be sensitive to several dimensions, but they may not be able to differentiate the contributions, resulting in an apparent dissociation of causes and effects.

Wierwille and colleagues reported a focused series of experiments stressing different aspects of mental demand on workload methodology in aircraft simulators. Each study confirmed that measurement techniques are differentially sensitive to the load manipulations. In the study of psychomotor loading (Wierwille and Connor 1983), only 5 of the 20 workload measurement techniques were judged appropriate for the piloting task and of those 5, only 3 had a monotonic relationship with the load manipulation. In the communications (Casali and Wierwille 1983) and cognitive (mediation) studies (Wierwille, Rahimi et al. 1985), 7 of 16 measures showed significant sensitivity. In the perceptual study, (Casali and Wierwille 1984) 7 of 14 measures were sensitive to increasing perceptual loads. The specific techniques that were found sensitive are noted later in this document under "5.0 Recommending workload measures".

Wickens (1992) provides a brief overview of the general classes of techniques, noting some of the advantages and disadvantages of each class. de Waard and Farmer provide more recent reviews of workload measurement methods. de Waard (1996) provides an extensive as well as critical assessment of both general categories and specific measurement techniques that provides a more current perspective on the views of O'Donnell and Eggemeier. de Waard's thesis is available on-line[4] and is a valuable reference both for those starting in the field of workload as well as for researchers looking for potential workload measurement methods. Farmer and Brownson (2003) provide a concise, current review of workload measurement methods and offer professional recommendations on techniques suitable for use in human-in-the-loop simulations. Farmer and Brownson focus on commonly used methods, commenting on suitability for the Eurocontrol Integra program involving air safety and air traffic management.

Castor (2003) and Wilson (2004) provide some of the most current reviews of workload measurement techniques, although distribution of these reports may be restricted and difficult to obtain. Each of these reports assesses various techniques on a number of levels, providing guidance on the maturity, sensitivity, reliability, and usefulness of each method. Castor (2003) also provides an assessment process to help select which of the various measures may be best applied based on the phenomenon under study.

These notable reviews suggest that workload is on the minds of many practitioners, however, the researchers cited are rather few in number. The works of these researchers re-appear across the literature, and the current review is no exception. Observations in the current review are based on the literature rather than conclusions arising from experiments conducted. In the following pages, different methods from each of the three categories are reviewed and reported with references for the interested reader to form their own hypotheses that can be tested. The reviews of Gopher and Donchin (1986) and O'Donnell and Eggemeier (1986) are recommended as concise historical and technical overviews of the field, as well as that of de Waard (1996) for his assessment of specific methods.

3.0 WORKLOAD MEASUREMENT TECHNIQUE CATEGORIES

Workload measurement techniques are typically organized into three broad categories: self-assessment or subjective Rating scales; Performance measures (including subdivisions of primary and secondary task measures); and, Psychophysiological measures (Eggemeier, Wilson et al. 1991, p. 207). It has already been noted that different measures are sensitive to different aspects of workload and not all workload measures are

[4] http://www.home.zonnet.nl/waard2/mwlch1.htm 9Jan04.

assessing the same thing. Part of this confusion arises from the lack of an accepted definition of workload and the tendency to use the term workload to mean either the demands imposed on the user, the effort the user exerts to satisfy those demands, and the consequences of attempting to meet those demands (Huey and Wickens 1993, p. 55).

Questionnaires and interview techniques, while informative, are not considered here as methods of measuring workload as they are more verbal descriptions of what the operator is experiencing. Questionnaires are complex to design properly to avoid unwanted biases, awkward to validate, and difficult to generalize (although the latter is also true of primary task measures). Because of this, the decision was made to exclude them from the current review and focus on quantitative workload measures that can be validation empirically.

3.1 Rating Scales

It seems appropriate that mental workload be measured by subjective means, as it is a psychological construct. Jex (1988) states "In the absence of any single objective measure of the diffuse metacontroller's activity, the fundamental measure, against which all objective measures must be calibrated, is the individual's subjective workload evaluation in each task." Casali and Wierwille (1984, p. 1046) note that their findings and other's indicate "… that properly designed rating scales with associated instructions … are particularly sensitive measurement instruments, especially with highly-trained populations …". Gopher and Donchin (1986, p. 41-2), however, assert "… an operator is often an unreliable and invalid measuring instrument." Nevertheless, subjective measures such as rating scales have a long history for measuring feelings of workload, effort, mood, fatigue, etc. Subjective methods attempt to quantify the personal interpretations and judgements of their experienced demand.

Most subjective workload measures imply (if they do not explicitly state) that it is mental workload that is being measured and the effects of physical work associated with gross motor muscles are not considered. The NASA TLX technique discussed below does have a category for Physical Demand that could capture the demands associated with physical labour although the wording describing this dimension seems more directed towards fine motor skills. Other subjective and physiological scales exist to measure physical labour such as the Borg's[5] subjective scale of relative effort or oxygen consumption (VO_2) as a measure of metabolic rate. It seams reasonable that experiments could be devised to test the independence of various subjective mental workload measures from physical exertion.

The repeatability and validity of such quantitative subjective techniques are sometimes uncertain and data manipulations are often questioned as being inappropriate. While the ordinal nature of the ratings is seldom questioned, use of interval or ratio arithmetic operations on the data has been the topic of much debate, with no definitive outcome in sight (Annett 2002). The issue is one of both philosophy and pragmatism. In many cases, there is no evidence that the data are anything but ordinal and as such, not amenable to arithmetic manipulation or parametric statistics, yet these are the everyday tools one is taught to use in science and engineering in order to support the conclusions and decisions made. On the other hand, subjective workload rating data may very well be interval – there is insufficient evidence to support or contradict this position. Other tools exist that could be used to explore this problem, such as non-parametric statistics, yet they have not found favour in practice. The typical approach seems to be to ignore possible violations of mathematical axioms in favour of convenience, accepting the risk that conclusions may not be justified given the data used.

[5] Borg, G. (1977) Simple rating methods for estimations of perceived exertion. Proceedings of the First International Symposium on Physical Work and Effort. Wenner-Gren Center, Stockholm.

To address this issue, non-parametric statistical analysis approaches are recommended over parametric statistics in such cases where the data should be considered ordinal. If this is undesirable, then parametric analyses should be validated through additional non-parametric analyses. This approach requires additional forethought on test designs to ensure appropriate plans for the questions to be answered since the analysis procedures are typically not as powerful. A mathematical assessment of the hazards of using parametric analyses with subjective methods that use ratings such as 5 or 7 point Likert scales would be useful guidance or even case-by-case assessment of conclusions based on parametric analysis compared to a corresponding non-parametric analysis. Alternative methods such as fuzzy logic may skirt the issue, providing a calculus to combine subjective data to form conclusions, although validation is still a difficult issue to resolve.

Meshkati and Lowewinthal (1988, p. 257) note many researchers feel that "… unless the subjective measurement technique is properly structured, they may serve only as gross indicators of stress level and have little diagnostic value, i.e., they may not indicate the source or type of workload involved." Meshkati, Hancock et al. (1995, p. 756) go further and note that some researchers believe that subjective feeling of difficulty is essentially dependent on the time stress involved in performing the task *for time-stressed tasks only*. This does not seem to address the source of workload in self-paced situations where time stress is low, but the number of tasks or task complexity is high, resulting in a sense of being overwhelmed by the reasoning or planning process required to complete the task.

While subjective measures have high face validity, their interpretation and ability to predict performance is uncertain. Vidulich (1988), Yeh and Wickens (1988) describe numerous instances where a dissociation between subjective and performance measures has been found. Vidulich concludes that subjective measures tend to be more sensitive to processing load in working memory while having low sensitivity to response execution demands. The hypothesis is that subjective workload is only sensitive to manipulations that are well represented consciously, so that varying demands in skill based tasks (or alternatively subject disinterest and inattentiveness) will not change in subjective ratings substantially. This suggests that subjective measures are highly suited to assessments of modern technologies that aid judgement and decision making, but are less suited to assessing physical or mechanical aids for repetitive or highly learned tasks. Others (Brookhuis and de Waard 2002) embrace dissociation as a natural constraint on all measures and focus on the picture presented by an assortment of measures that are differentially sensitive to the array of factors that contribute to workload.

Wierwille (1988, p. 320) suggests analytical techniques that average workload over time are inappropriate, despite indications that this is what subjective measures seem to be capturing. Instead, Wierwille argues momentary workload values represent the appropriate measure, particularly for design analyses. This suggests that the usual subjective measures on their own are insufficient to adequately characterize workload and some additional means of capturing instantaneous or momentary work overload is necessary. This seems to support more analytical, workload estimation approaches for applied assessments.

Self assessments involve rating demands on numerical or graphical scales, typically anchored either at one or two extrema per scale. Some subjective techniques use scales that are categorical, with definitions at every level, such as the Modified Cooper-Harper scale. Other techniques use an open-ended rating with a "standard" reference task as an anchor and subjects rate other tasks relative to the reference task. Hart and Wickens (1990) subdivide rating scale methods into uni-dimensional ratings such as Overall Workload (Vidulich and Tsang 1987), hierarchical ratings such as Modified Cooper-Harper or Bedford scales (Wierwille and Casali 1983), and multidimensional ratings such as SWAT (Reid and Nygren 1988) and NASA TLX (Hart and Staveland 1988).

Unidimensional and hierarchical measures mentioned above have good face validity, are easy to understand and use, and generally have good user acceptance. While useful in their own right, they also may serve as standards to assess more complex, multidimensional measures discussed below. What they lack is a calculus to combine ratings for predicting workload in different situations involving similar tasks. While not all multi-dimensional workload scales have a predictive mode, several do. The rest of this review of subjective measures will concentrate on measures that have, or could have, predictive capability through constructive modelling.

The VACP method (Visual, Auditory, Cognitive, Psychomotor: Aldrich and McCracken 1984; McCracken and Aldrich 1984; Aldrich, Szabo et al. 1989) is an early attempt at a simple, diagnostic workload measurement tool that, when coupled with task network modelling, can be used for predictive workload assessment. Subjects assess task demands according to a standardized, categorical list in each of the four dimensions. Each dimension's levels were assessed to create interval rankings that could be used in predictive, constructive simulations. A simple summation process was proposed with a rather arbitrary limit placed on the maximum value any dimension could attain before operator overload would occur. Despite objections to the lack of validation and inappropriate aggregation of ordinal data, it has proven useful in system design to identify where systems might over-burden the users. The VACP method remains a frequently used metric and can be used as demand-estimates in more complex workload prediction models.

W/Index (Workload Index: North and Riley 1989) was developed to formalize Wickens' concept of contention among multiple resources in workload calculations. VACP ratings of individual task demands have been suggested as input data to W/Index. While possibly an improvement on the VACP method for predicting workload, W/Index did not have an overload criterion. Further, neither VACP, W/Index nor the analysis tools that used them had mechanisms to reallocate or delay tasks to form coping strategies as human operators do, resulting in unrealistic predictions of workload.

Boles and Adair (2001) attempt to assess the cause of workload through the Multiple Resource Questionnaire (MRQ) and suggest that this method correlates well with other subjective methods. Whether MRQ is a better method, or a practical workload measurement technique, remains to be seen, but linking MRQ with hierarchical measures such as the Modified Cooper Harper (MCH) scale may provide analysts with the power they require: an easy to use, validated workload measurement that has diagnostic support.

Of all the subjective techniques, SWAT seems to be the most common technique reported in the literature (see Appendix 1). Multidimensional or multi-scale assessment techniques often have aggregation procedures to produce an overall workload rating. SWAT seems to be alone in proposing an aggregation procedure that has a sound metrological basis, conjoint scaling, although NASA TLX also makes some claim to aggregation legitimacy with a paired-comparison weighting scheme. Other workload estimation techniques, such as VACP (Aldrich, Szabo et al. 1989) and W/Index (North and Riley 1989) make cruder assumptions about aggregation, although perhaps, an approximation appropriate for the level of precision incurred with subjective assessment techniques.

Reid and Nygren (1988) describe the SWAT technique and the theoretical assumptions underlying it. They note that "... performance measures cannot, of themselves, describe workload ..." because operators may vary effort to maintain a constant performance level, but the perceived workload involved will vary commensurately with effort. SWAT rates experiences on three dimensions (Time Load, Mental Effort, and Psychological Stress), each with three integer or categorical levels. SWAT addresses the principal complaint about subjective measures: the guarantee that ratings are interval or ratio scaled and not just ordinal rankings. This is accomplished through conjoint measurement and scaling of the subjective ratings provided

by the subjects (that may be only ordinal before scaling) through a comparison of the relative importance of all levels over all three dimensions (27 distinct comparisons of each level and dimension with every other combination).

Hart and Staveland (1988) describe the NASA Task Load Index (TLX) workload measurement technique and present the empirical validation supporting it. They assert that "... subjective ratings may come closest to tapping the essence of mental workload and provide the most generally valid and sensitive indicator." They also note that subjective techniques are subject to operator biases and preconceptions, and can only be based on what is remembered or abstracted from an experience. They suggest, however, that subjects are unlikely to remember experiences sufficiently well, making absolute judgements of workload or even relative measures for different tasks meaningless. This is a bit at odds with some of the literature that shows assessments are reasonably stable when elicited by detailed after-action review, but ties in with Colle and Reid's (1998) concept that context plays a significant role in workload measurement.

There are two concerns with the formal NASA TLX method. One is the process of scoring "Own Performance"; the other is the scaling procedure based on the paired comparisons. The first is a concern of how the scales are presented. Each dimension is presented as a graduated, horizontal line, anchored at each end. Hart and Staveland (1988) validated the dimensions through an extensive series of experiments and analyses, developing the anchors such that the rating on each scale corresponds linearly to their contributions to overall workload. In 5 of the 6 scales, the ratings range from Low on the left to High on the right; this is reversed for Own Performance, which goes from Excellent to Poor. While this makes sense if one considers each dimension as a contributor to workload, the Own Performance scale can cause confusion among subjects. Subjects may tend not to think of the scales as contributors to workload so much as independent measures for the trial. Periodically reversing the direction of successive scales is advocated by experts in scale development to reduce the tendency of subjects filling them in with little consideration of their meaning. This was not the purpose of the scale reversal in the NASA TLX, where the rating scales were arranged such that each dimension contributed positively to the overall workload score. It is a trivial process to reverse the Own Performance scale presentation, but the implications of this change on the validation of the method have not been assessed. Castor (2003, p. 36) notes that the Own Performance dimension does not need to be part of the workload assessment portion of the NASA TLX algorithm as multidimensional scaling suggests Own Performance is quite separate from the other five dimensions that seem to cluster on a single factor.

The second concern is how NASA TLX aggregates ratings by summing weighted ratings from each scale. The weights are determined by a paired comparison of the relative importance of each scale to workload. In this process, the lowest ranked scale can receive a weighting of 0; in other words, it may not contribute to the computed composite workload measure. Hart and Staveland (1988) imply that the workload contribution weights associated for each of the scales are context dependent and so, such an occurrence is appropriate. If it is indeed the case that only relative workload estimates within the same tasks are feasible, this is not a problem. If, however, one wishes to compare different tasks or experiences, then allowing the weights to vary between experiences significantly complicates the comparison. It seems sensible that different tasks will produce different scale ratings, and that different subjects may perceive the importance of each scale differently, subjectively combining each scale to result in personal assessment of the workload. It does not follow that subjects will change their perception of the contribution of each scale to workload and it seems more sensible that the weights would be largely invariant with task type for an individual. No studies were found that examine how subjects' assessments of scale weights change with context, although Hart and Staveland (1988) suggest that the within-subject weights are largely stable across studies, which suggests that studies of the stability of weights versus individual differences would be feasible.

Colle and Reid (1998) assessed SWAT and NASA TLX scales in a series of experiments, finding SWAT somewhat more sensitive to the various experimental manipulations than was NASA TLX. The pattern of results was similar between the two methods and the differences were too small to state that one technique was better than the other. This is at odds with some other studies that found NASA TLX more sensitive, particularly at low workload levels (Hart and Staveland 1988; Nygren 1991; Hill, Iavecchia et al. 1992). Byers, Bittner et al. (1988) and Hill, Iavecchia et al. (1992) found NASA TLX somewhat more sensitive than SWAT to the experimental manipulations, however, Hart and Wickens (1990, p267) also report that of several subjective measures, NASA TLX proved to correlate best with performance measures while displaying low intersubject variability and good user acceptance. Conversely, Whitaker, Hohne et al. (1997) found SWAT more sensitive than NASA TLX. Both NASA TLX and SWAT are usually reported to be more sensitive than the unidimensional scales such as Overall Workload (OW) and Modified Cooper-Harper (MCH) although the early work by Wierwille and colleagues found the MCH method more sensitive. The small number of levels within each SWAT dimension has been criticized as not providing adequate sensitivity; more dimensions would make the tedious card-sorting process impracticable. While NASA TLX offers greater precision within each subscale, it is difficult to say that it offers greater diagnosticity or greater accuracy, if such a concept is applicable here. Obviously, there is room for improvement, debate on measurement procedures and the interpretation of results.

NASA TLX seems to have higher user acceptability than SWAT because of the shorter scale development phase. The card-sorting procedure of the SWAT method seems to be a significant factor in lower acceptance by users; subjects find even the paired comparison procedure of NASA TLX an imposition, despite taking only a few minutes to complete. Further, it is unclear whether the extra burden of performing scaling procedures for aggregation are of much value (Hendy, Hamilton et al. 1993), It is likely that simple averaging of all scales (if even that is mathematically permissible) would tend to overestimate the true workload, since some of the scales may overlap in their assessment. Some practitioners have found the SWAT scale development somewhat onerous and several have suggested alternatives that they claim are as sensitive if not more sensitive, although lacking the mathematical rigour of the original SWAT. Luximon and Goonetilleke (2001) explored 5 variants on the SWAT method, finding that the simpler approaches were more sensitive and less time consuming to perform, although they lacked the mathematical rigour of the standard SWAT. Biers and Masline (1987; Biers 1995) have explored alternative formulations that, in the limited studies conducted, performed as well as SWAT, but with less work.

Should SWAT, or other ratings based on conjoint scaling, prove too onerous for practical use, it may nevertheless serve as a scale-development standard; that is, the more onerous method may provide an error estimate associated with approximate methods. For instance, the 6-scale NASA TLX procedure, with 10 or 20-points per scale (potentially infinite), would be prohibitively time consuming for practical conjoint methods. Subjects might be induced to willingly suffer a large, tedious, card sorting process (or a subset thereof) if there was high likelihood of developing a universal scale at the end, but if the process proves to be individually specific, there may be little advantage to this approach. The similarity between SWAT and NASA TLX empirical results would seem to suggest that there is hope for interval or ratio data resulting from subjective ratings, which has been a problem for predictive models of workload (Reid and Nygren 1988)[6].

[6] Other SWAT references of interest:

Reid, G. B., C. A. Shingledecker, et al. (1981). *Application of conjoint measurement to workload scale development*. Proceedings of the Human Factors Society – 25th Annual Meeting, Rochester, New York.

Reid, G. B., C. A. Shingledecker, et al. (1981). *Development of multidimensional subjective measures of workload*. Conference on Cybernetics and Society sponsored by IEEE Systems, Man and Cybernetics Society, Atlanta, Georgia.

Reid, G. B., F. T. Eggemeier, et al. (1982). *An individual differences approach to SWAT scale development*. Proceedings of the Human Factors Society – 26th Annual Meeting.

Choosing between the methods is largely a matter of preference: if an interval scale is desired, then SWAT is preferred; if ease of use is desired, while likely maintaining a close correlation to SWAT, then NASA TLX seems a viable option.

While the NASA TLX method can be performed with paper and pencil, computerized versions have also been developed. The original DOS™ version may still be available from NASA or from the HSIAC website[7], however, a more recent implementation for Windows™ has been developed[8] to support experimental investigations at DRDC Toronto. This DRDC version can use the original NASA TLX format or a variant that include a univariate Overall Workload rating (not validated) as well as reversing the Own Performance scale. Experimental designs can be specified and the results exported as comma-separated text files that can be imported into spreadsheet programs for further data analysis.

While many subjective methods may be useful for assessing the workload associated with a task, job or experience, most are not useful for predicting the workload associated with combinations of tasks or activities. It seems reasonable that such a predictive version could be developed for some of the methods described above. Prediction using the VACP and W/Index methods were early such attempts, however, there is little theoretical or practical evidence that the approaches used were valid.

The DRAWS (Defence Research Agency Workload Scale) measurement technique (Farmer, Belyavin et al. 1995; Farmer, Jordan et al. 1995) asks subjects to rate their perception of the Input, Central, and Output task demands. DRAWS can be used for assessing single tasks, or for assessing experiences with multiple, concurrent tasks. There is a fourth category called time pressure that is also rated. The scale is nominally from zero (no load) to 100 (fully loaded), although subjects are permitted to record values greater than 100. The DRAWS ratings are thought to represent the time pressure associated with each stage of Input, Central and Output processing of the task. No information concerning validation of the scale was found, however, the POP (Prediction of Operator Performance) model was developed as a predictive form of DRAWS. POP integrates subjective DRAWS ratings of individual task demands when those tasks are performed in together by the operator. POP uses a Markov process to model the interference for contending tasks, and aggregates the individual task DRAWS ratings into an overall workload or time pressure for the operator. The POP model has been validated against a selection of laboratory studies with good predictive ability of the multitask DRAWS ratings; it has not been validated against field studies. This suggests that DRAWS ratings and the POP model make a suitable measurement and prediction scheme, although further validation would be appropriate.

The IP model (Hendy and Farrell 1997) is one of a few methods that claims to relate workload to observable aspects of task execution, in this case time pressure, reflecting individual capabilities and task demands as a single, measurable quantity (processing time required divided by the environmentally imposed time available). The IP model does not measure workload itself, although it postulates that workload equates to time pressure and error rates are relationships that depend on time pressure alone. Although the IP model can be considered a uni-dimensional scale, it considers other factors usually associated with workload (such as strategy selection and individual differences) and uses these factors to moderate time pressure, either by increasing the execution time or by decreasing the amount of time available. Validation of the IP model has been limited to studies on a simplified air traffic control task, although its assessment on tasks similar to the POP validation tasks is planned.

[7] DOS version of NASA TLX: http://iac.dtic.mil/hsiac/products.htm

[8] Windows version of NASA TLX: For information contact Brad Cain, brad.cain@drdc-rddc.gc.ca.

The POP and IP models have similar fundamental assumptions, and are being integrated into a computational model for predictive analysis, taking advantage of the strengths of each[9]; this product, POPIP, is being developed within the Integrated Performance Modelling Environment (IPME), although its validity has to be established. If this integration proves successful, the combination of DRAWS and POPIP should provide a useful means of measuring, modelling and predicting workload.

3.2 Performance Measures

Performance measures of workload can be classified into two major types: primary task measures and secondary task measures. In most investigations, performance of the primary task will always be of interest as its generalization to in-service performance is central to the study. In secondary task methods, performance of the secondary task itself may have no practical importance and serves only to load or measure the load of the operator. As Lysaght, Hill et al. (1989, p. 67) point out, "A statement about operator performance is meaningless unless system performance is also acceptable. Accordingly, there is a need to measure both." Thus, there is a necessary precondition that system performance be acceptable; operator workload is not a sufficient measure for assessments.

In order to have primary task measures that are reliable, tests must have appropriate context, relevance, representation, and time-on-task training. Despite the relevance of the primary task to operational activities, it is often not possible to assess the cost of performing the primary task by performance measures alone because of changes in "strategic reallocation of mental capacity". Wilson (2004) notes, "Because of the protective (compensatory) effect of increased effort, it is clear that measuring performance is not sufficient to assess the state of the operator. The level of performance does not provide information about the costs involved in the adaptive response to stress. Under conditions where there is no discernible breakdown of performance under stress, physiological and subjective measures of operator functional state mainly reflect the amount of mental effort (strain) required to maintain task performance." Thus, while primary task measures may be considered a necessary measure of workload, they should not be considered sufficient on their own. This is supported by the apparent dissociation of performance and demand noted in the previous section.

Although operational performance measures are easy to justify, they often lack scientific rigour, making interpretation of the results difficult. Uncontrolled and perhaps unknown factors may dominate results rather than the intended manipulations in the trials. Conversely, laboratory tasks provide more experimental control, but lack the ecological validity of operational task measures. A combination of experimental and operational assessment is often the best approach. Advances in simulator technologies are creating experiences with a greater sense of presence such that simulators should become increasing accurate estimates of operational performance before real world measurements are undertaken. Nevertheless, task performance measures are key for postulating predictive models based on other operator-state factors that can be evaluated in constructive simulations with many replications; virtual simulations are typically restricted to a few replications and so can consider commensurately fewer conditions.

Primary task measures attempt to assess the operator's performance on the task of interest directly, and this is useful where the demands exceed the operator's capacity such that performance degrades from baseline or ideal levels. Speed, accuracy, reaction or response times, and error rates are often used to assess primary task performance. Primary tasks measures are thought to be "... globally-sensitive and provide an index of variations in load across a variety of operator information processing functions" (Eggemeier, Wilson et al. 1991, p. 209).

[9] Dr. A.J. Belyavin, QinetiQ, Plc., Farnborough, Hants, UK. Personal communication.

Wickens (1992, p. 392) notes that primary task measures may not be sufficient or adequate:

1) If the variability in the task demands are insufficient to result in observable primary task performance changes (no information on remaining capacity can be inferred);
2) If alternative primary tasks use different modalities (it may be the reserve capacity available for other secondary-tasks that is important to determine);
3) If the primary task demands cause incidental effects such as fatigue or stress that become important performance shaping factors in longer exposures (short evaluations may not show a difference between contending designs); and
4) If other factors, such as strategy, affect performance and workload differently (giving rise to a perception of dissociation between the two).

Hart and Wickens (1990) note that while primary task measures are important in workload assessment, they are more a measure of what the system can achieve rather than an estimate of the cost of operator achievement and that a dissociation between workload and primary task performance is frequently observed (Yeh and Wickens 1988).

Secondary task measures provide an index of the remaining operator capacity while performing primary tasks, and are more diagnostic than primary task measures alone. The characteristics of the secondary task are used to infer the interaction between the primary and secondary tasks and this approach is frequently used when the operator can adapt to demand manipulations such that primary-task performance is apparently unaffected. The secondary-task paradigm can be further classified into Auxiliary Task and Loading Task methodologies, but the intent of both is to increase operator load to the point where changes in effort or strategy are no-longer able to compensate for changes in the demand manipulation.

In auxiliary task methods (the more common of the secondary task approaches), operators are instructed to maintain consistent performance on the primary task regardless of the difficulty of the overall task. The variation of performance on the secondary auxiliary-task is measured as an indicator of the operator's reserve capacity, serving as a surrogate workload measurement under the various loading conditions.

The loading task approach deliberately causes degradation of the primary task, requiring consistent performance on the secondary task. This shifts the primary task performance into a region where it is sensitive to the demand manipulation. The performance decrement of the primary task is measured as the loading task difficulty is increased.

Whether a non-intrusive (auxiliary) or an intrusive (loading) secondary task approach is adopted, there is still a multitasking issue that should be considered. Williges and Wierwille (1979, p. 558) note that subjects probably expend more mental effort during dual task performance than the sum of the effort required to perform each task alone, even if the tasks do not interfere. This overhead is attributed to the management and scheduling of the two tasks by some form of metacontroller within the cognitive system. This hypothesis is difficult to determine without a formal psychological model of the workload process. A major premise of these dual or multitask environments is that the workload is inherently different from a single task condition, regardless of the single task level of difficulty. Fracker and Wickens (1989) argue against that generality, noting in some cases, dual-task performance cannot be distinguished from a more complex single task. They do not consider the possibility that a single, complex task might be treated as a collection of task elements, and processing of these elements in unison requires coordination. The POPIP model (noted elsewhere in this text), if successful, may present an opportunity to test these hypotheses formally in a manner that can be validated empirically.

Colle and Reid (1999) note that secondary task (and presumably primary task) performance measures, while of conceptual and theoretical interest, are less practical than subjective measures, particularly in operational assessments. They note that globally sensitive secondary task measures need to be developed, inferring that secondary tasks can be of use in more diagnostic applications. Unfortunately, secondary tasks based on rigid, laboratory tasks are often too constrained or contrived. This can lead to measurements that do not adequately reflect the operator's ability to dynamically shift tasks to accomplish them in a timely manner. An aspect of performance measures that makes them awkward is their inherent lack of generalization, although sometimes extrapolation from one domain to another seems plausible. Different performance measures are often required for specific primary tasks in different applications, making standardization across domains difficult (Meshkati and Lowewinthal 1988).

Selection of secondary tasks must not be done "will-he, nil-he"; there is a need to match the secondary tasks to the primary tasks such that the operator is loaded appropriately and context-specific sensitivity is captured (Wickens 1977). The choice of secondary task can a profound effect on performance and, hence, on the outcome of the experiment. Damos (1991) suggests that a number of considerations for selecting secondary tasks and gives further references to other reviews. In particular, Damos (1991, p. 114) notes that practice on each task alone is not sufficient to ensure optimal dual task performance; the tasks must be also practiced within the context of the dual task experience, an aspect related to the metacontroller's role described by Jex (1988). Important features of the dual task training are the appropriateness and timeliness of feedback to guide the subject's selection of strategies.

Meshkati, Hancock et al. (1995, p. 754) note that some researchers are concerned that secondary task measures might produce an undesired change of strategy, distorting performance on the primary task by affecting strategies. Since performance on either the primary task or the secondary task must be held constant over all manipulation levels for this technique to be useful, an embedded secondary task (one that is a normal component of the operator's responsibilities, albeit of a lower priority than the primary task) is more appropriate, gaining both operator acceptance and ecological validity. Using artificial secondary task methods in operational or even training assessments presents problems because the intrusiveness of the secondary task into the primary task may present safety hazards or adversely affect the training objectives (Meshkati and Lowewinthal 1988). Presumably, undesirable intrusion is less of an issue when the secondary task is a natural, embedded task that is part of the operator's normal routine. This leads to an ecologically valid, secondary task measure that often gains greater operator acceptance.

Many of the workload reviews examined discuss specific secondary tasks, however, none included a table of tasks that described the key secondary task characteristics and identified families of primary tasks with which they would be appropriate to pair. Secondary task selection criteria that were identified in these reviews include:

1) Non-interference with primary task (consume similar resources, but not interact with the primary task);
2) Easily learned; and
3) Self paced (easily interrupted or delayed).

Typical variables for secondary task measures are:

1) Reaction time;
2) Time estimation variance;
3) Accuracy and response time (to mental arithmetic or memory search);

4) Signal detection rates;
5) Tracking performance (such as RMS error or control reversals);
6) Number of concurrent tasks in an interval; and
7) Percentage of time occupied.

Secondary tasks include, but are no means limited to:

1) Rhythmic tapping;
2) Random number generation;
3) Probe reaction time (Sternberg memory search task, Bakan task);
4) Verbal shadowing;
5) Spatial reasoning;
6) Time estimation and time production (retrospective estimate of elapsed time);
7) Critical instability tracking task; and
8) Compensatory or pursuit tracking tasks.

There is a major caveat with current dual or multitask workload measurement paradigms: most give little or no thought to formal, measurement theory when evaluating secondary task workload measurement (Colle and Reid 1997). While other fields of psychology and human performance have embraced formal methods, the performance oriented mental workload community seems to be remiss. The definition of mental workload equivalency curves to characterize the demands of tasks appears to be lacking in the performance measurement approach to workload assessment.

3.3 Psychophysiological Measures

"The goal of psychophysiological applications in the assessment of mental workload is to develop measures with well known properties that can be applied in specific situations. This goal has come about from the complex nature of the mental workload construct and the acceptance that there is no golden yardstick of mental workload" (Neuman 2002, p. 601). Much of the psychophysiological literature focuses on determining the aspects of workload that particular methods are sensitive to. While sensitivity and relevance are obviously important factors in selecting any method, they seem more so with physiological measures because these measures tend to be general, systemic indicators of stress. Lysaght, Hill et al. (1989, p. 137) note that "... the use of an inappropriate technique may be misleading; it may be a good technique in some instances, but (may be) the wrong tool for the question at hand."

The principal attractions of psychophysiological measures are continual and objective measurement of operator state. Psychophysiology attempts to interpret the psychological processes through their effect on the body state, rather than through task performance or perceptual ratings. If successful, this approach would have a number of advantageous applications, however, as Wickens (1992, p. 199) notes, psychophysiological measures are "... one conceptual step removed from the inference that the system designers would like to make." This requirement to infer workload is an issue both for researchers seeking to assess workload as well as for designers of automated-support systems that attempt to assess operator state and provide assistance accordingly. Meshkati, Hancock et al. (1995, p. 757) suggest that "… physiological methods do not measure the imposed load, but rather they give information concerning how the individuals themselves respond to the

load and, in particular, whether they are able to cope with it." Psychophysiological measurements may be particularly useful when subjective methods or performance measures become insensitive due to covert changes in operator strategies, or the applied level of effort lead to an apparent dissociation among subjective and performance measures.

A requirement of most psychophysiological measures is for reference data that establishes the operator's unstressed background state. Such background states are subject to many factors and may change markedly over time so an operational baseline state is often used. The operational baseline state is measured when the subject is not under any specific stress, but it will reflect systemic stresses incurred from being "in-theatre" or reflect the day-to-day changes in a subject's life. Thus, as with subjective and performance measures, contextual issues should be considered in the evaluation of the results. Further, the baseline values as well as the operational values may vary considerably between individuals, so psychophysiological-measurement systems often need to be tailored to each individual rather than using group norms, making interpretation more involved. Another practical consideration is the availability of a suitable test environment that is related in part to context. Physiological measures are of little use early in systems design since it is unlikely there will be a physical mock up or simulator to provide the appropriate stimuli (Mitchell 2000).

In the past, physiological measures often entailed cumbersome, invasive equipment, unsuitable for most applied settings. This has changed dramatically in the past decade as advances in technology have made the equipment much more portable and capable. There is still a significant degree of invasiveness with some techniques that users may object to, making in-service use awkward. There should be a clear advantage demonstrated to operational personnel expected to use these procedures if there is any serious hope that operational users will endure the invasive measurements.

While most of the negative issues associated with psychophysiological measures from the user's perspective are technological (and hence susceptible to improvements in hardware and methods), "… the lack of a strong conceptual link from the physiological measures to performance …" is its greatest weakness from the analyst's perspective (Kramer 1991). Wilson and Eggemeier (1991, p. 351) reiterate the need for a better understanding of the links among the various physiological responses and workload; this would no doubt be well served by formal measurement theory. Wilson and O'Donnell (1988) lament the failures of early attempts to find the essential link between objective, non-intrusive, physiological measures and mental workload. They note that attempts to use individual physiological measures, while being sensitive to specific demands, were unlikely to be generally sensitive because of the multi-facetted nature of workload. Instead, they suggest a battery of psychophysiological measures would be necessary, along with a suitable interpretation scheme. Wilson and O'Donnell critically review several of the more common psychophysiological measures and describe their Neurophysiological Workload Test Battery (NWTB), although it is not known if this is consistent with Colle and Reid's formal measurement theories. Kramer (1991), Wilson and Eggemeier (1991) present critical reviews, although both these assessments are over a decade old.

NATO RTO has recently published a review of operator functional state assessment that includes an overview of psychophysiological techniques (Wilson 2004, Chp4). The report provides a brief description of a number of techniques, assesses the advantages and disadvantages of each technique, and documents the requirements for use. Numerous references are provided for each technique, and this document seems a good starting point for entry into psychophysiological-measurement techniques. Fahrenberg and Wientjes (1999, pp. 111-112) note a number of references that detail measurement techniques for various psychophysiological approaches as well as references to dealing with measurement problems and artefacts. The Society for Psychophysiological Research maintains a web site[10] that contains several guidelines for specific psychophysiological measurement techniques.

[10] http://unix.wlu.edu/~spr/

Corwin, Sandry-Garza et al. (1989) studied a number of workload measurement methods in a simulator at NASA Ames for a several commercial airline flight scenarios. In the first study, they found that (pp. 96-99):

1) Subjective measures (NASA TLX and SWAT) demonstrated validity and reliability.
2) Physiological measures were generally disappointing:
   a) Eye movement and eye blink were insensitive to experimental manipulations;
   b) Inter-beat heart rate interval was thought to be reliable and valid, but too prone to other effects; and
   c) Heart rate variability and blood pressure spectral analysis were insensitive to workload manipulations.
3) Primary flight control measures were thought to be good discriminators of workload, but secondary tasks were not good discriminators of workload.

In a second simulator study by the same authors (pp. 159-163), subjective methods again proved valid and reliable. There was some question about whether the simpler, unidimensional Bedford Scale coupled with the Pilot Subjective Evaluation information elicitation technique would not be of greater value for evaluating new aircraft flight deck workload than the more general and data-intensive NASA TLX and SWAT techniques. Again, none of the physiological measures were found to be sensitive to the workload manipulations. In this study, secondary tasks tended to be ignored by the subjects, so only primary flight task (control input) performance was analyzed. The authors felt that the primary-task measure was a useful indicator of workload despite the subjects' tendency to ignore the competing task. In their final report, Corwin, Sandry-Garza et al. (1989) note that workload measurement was still immature, but they recommended the following subjective techniques: Bedford/Modified Cooper-Harper; NASA TLX, SWAT. They also recommended both primary flight control task performance as well as embedded and ecologically appropriate secondary task performance as suitable workload measures for assessing new aircraft. Of the physiological measures studied, only heart rate was suggested as a physiological measure and that was qualified by noting it may be more of a measure of arousal, reflecting stress effects not directly attributable to workload. These results seem typical of evaluations of workload methods. The lack of a clear, positive result for the psychophysiological methods must be disheartening for their advocates.

Most researchers in the field would agree that several psychophysiological measures correlate reasonably well with various aspects of workload and hold promise for objective workload measurement, however, considerable research remains to be done to properly classify and characterize these methods so they can be applied appropriately and assembled into a battery of measurement techniques that could provide a general assessment method. Boucsein and Backs (1999, Table 1.1, p. 9) suggest that a few measures are driven principally by mental strain, although most measures seem sensitive to stress in general rather than just the stress of workload.

Stress related hormones may provide useful, long term measures in either laboratory or field settings, however, they may not be sufficiently sensitive to provide real time interventions (Eggemeier, Wilson et al. 1991). Currently, psychophysiological experts recommend that psychophysiological approaches be applied as a battery of measures to isolate mental workload contributions. The measures selected must be appropriate for the task and the aspect of workload or strain that is of interest; it is not prudent to measure only one or many indiscriminately (Gaillard and Kramer 1999). This is due in part to recognition that the operator's state "... should be regarded as the result of many physiological and psychological processes ..." (Gaillard and Kramer 1999, p. 32). The Boucsein and Backs monograph contains several papers on psychophysiological measurement approaches and the first chapter lists a number of measures with application references and general indicators of the effect (Boucsein and Backs 1999).

Some of the more common psychophysiological methods are noted briefly in the following paragraphs.

3.3.1 Electroencephalography

Measurement of brain activity through electroencephalography (EEG) is used in many fields, not only in workload assessment, and technology is making it practicable for some operational settings. Castor (2003), Boucsein and Backs (1999) note, however, that EEG approaches are prone to artefacts and so have not been used often in field studies. Technical reasons also preclude the use of brain imaging in the field, and probably most human factors laboratories as well. The EEG data are complex waveforms that require sophisticated signal processing equipment. The waveform spectrum is typically divided into a number of frequency bands and workload assessments is made on the power within these bands or on time shifts of event related potentials (ERP).

Freude and Ullsperger (1999) note that movement related readiness potential (BP) and Preparatory Slow Brain Potentials (SP) seem to be complementarily sensitive to attention, demand, and decision making. It is not clear whether the changing amplitudes in these measures can be correlated with workload or performance in a class of tasks to create stable, general predictions of performance changes in practice. Wickens (1992, p. 398) notes that evoked brain potential is better thought of as measure of residual capacity than as a measure of effort, and the reduced P300 amplitude is sensitive to central processing demands, but not response demands, providing an unobtrusive measure of mental workload. Wilson and O'Donnell (1988) note that the P300 amplitude ERP may index the degree of surprise or mismatch of a stimulus with expectation while the P300 latency is more related to the difficulty of a task. Meshkati, Hancock et al. (1995) feel that the Evoked Response Potential family of measures holds the most promise of the physiological measures of workload, despite the technological and interpretation hurdles.

3.3.2 Eye Movement

Measurements of eye activity can be used unobtrusively and much of the technology may already be in place to support these measurements. For example, helmet mounted sighting systems for fighter pilots or proposed helmet mounted display information systems for the infantry may provide a means to obtain these data without further intrusion on the subject. If a stable, helmet mounted display is available for operational reasons, then a number of measures of eye activity can be made unobtrusively, such as: horizontal and vertical eye movement (extent and speed), blink activity (duration, latency and frequency), fixation duration, point of regard and pupil diameter.

Although ocular measures are sensitive to mental demands, they are also sensitive to other factors; in particular, they are sensitive to fatigue. The literature contains contradictory findings that may be due to differences in experimental methods (Sirevaag and Stern 1999). Further, the measurements often require a stable sensor capable of detecting small movements, something difficult to achieve in the field or even in the laboratory at times because of movement of the sensor on the head as the body moves.

Blink measures can be context dependent. Blink rate has been observed to decline with increased workload resulting from processing visual stimuli, however, it has been observed to increase with increased load resulting from memory tasks (Wilson 2004) and the connection between blink rate and workload seems tenuous (Castor 2003). Blink closure duration appears to decrease with increased workload resulting from visual stimuli or gathering data from a wide field of view while blink latency increases with memory and response demands (Castor 2003).

Pupil diameter appears to be sensitive to a number of demands and emotional states, making it less diagnostic, however, the measurements need to be quite precise (on the order of tenths of a millimetre) making application difficult in environments with vibration or requiring considerable eye and head movement. Wilson (2004) notes that pupil diameter generally increases with higher cognitive processing levels and it is sensitive to rapid changes in workload, however, when overload occurs, pupil diameter can become unresponsive to changes or even reverse its response. Nevertheless, research continues (Marshall, Pleydell-Pearce et al. 2002) and a new technique has emerged showing promising results, the Index of Cognitive Activity[11], although no details describing the method or its validation were found.

3.3.3 Heart Rate

Various heart rate measures (such as the rate, its variability, and resulting blood pressure) have been reported to be sensitive to workload. These measures are relatively easy to employ unobtrusively both in the laboratory and in the field. Heart rate measures suffer from interactions with respiration, physical work and emotional strain, and so would likely require unique measures to isolate mental workload contributions. Wilson (2004, p. 4-7) notes that there are numerous coupled control mechanisms and feedback loops in the cardio-vascular system, making definitive interpretation difficult. Meshkati (1988) states that heart rate variability is probably the most used physiological measure in workload measurement and references other literature, noting the varied effectiveness of heart rate variability in workload assessment. Reliable measurement of heart rate and its variability require at least 30 seconds, but not more than 5 minutes for optimal sensitivity with concurrent measurement and correction of respiration effects (Castor 2003).

Mulder, Mulder et al. (1999) note that heart rate measures (particularly heart rate variability in the 0.07 – 0.14 Hz range) are sensitive to task complexity and compensatory effort resulting from stressors (fatigue, noise, etc.), but that cognition and emotion may be too tightly coupled to distinguish effect. Mulder, Mulder et al. (1999, p. 144-145) report that there have been problems with reproducibility of results suggesting more work is required before a comprehensive, formal method can be recommended and work to this end was underway throughout the 1990s. Despite the difficulties associated with the heart rate measures, heart rate variability continues to be studied and commonly used in conjunction with respiratory measures to assess operator state and mental workload.

Blood pressure has been found to correlate with mental demand, however, it does not appear to be very sensitive and it is prone to exercise artefacts. While easily and often measured, blood pressure does not appear to be a principal candidate for workload measurement (Castor 2003).

3.3.4 Respiration

Wilson (2004) notes that respiration is not simply a factor for adjusting heart rate measures; respiration measures offer their own valuable information on operator state. There are several measures that can be recorded, such as the time for inspiration or expiration, the complete cycle time, the volume and flow rate. Several of these measures may be measured or inferred with little intrusion. Respiration rate has been observed to increase while respiration volume decreases as stress and mental workload increase, but it is also highly dependent on physical activity. This suggests that while it provides useful information about operator state, it is not a suitable workload measure on its own (Castor 2003). Nevertheless, because it is a necessary measurement for correcting heart rate measures, it remains a candidate for supplying part of the workload measurement picture.

[11] http://www.sci.sdsu.edu/cerf/darpa/darpa.htm

3.3.5 Future Psychophysiological Measure Developments

Formal, coupled models relating various psychophysiological measures and workload need to be developed. Often, physiological phenomena interact with one another such that, although it may be possible to correlate these measures *post hoc*, their independent use as a predictor of workload levels seems quite limited. Nevertheless, this class of methods has potential for unobtrusive, objective measurement of mental workload, particularly for embedded, automated aiding applications, however, few are sufficiently practicable or understood sufficiently for military field operations in their current state of development.

Although bulk and weight have been reduced, the electronic apparatus, sensors and wires associated with psychophysiological methods are seldom acceptable in the workplace, and sometimes not practicable in the laboratory. Modern technology continues to improve devices for a number of different measures and unobtrusive ambulatory recorders with a vast array of sensors are now available that were impracticable a decade ago.

The science of psychophysiological measurement is not static, resulting in better understanding with improved techniques, and advances in technology extending many methods from the laboratory into operational environments more likely. Wilson (2001) reported on several in-flight physiological measures, finding high repeatability among several methods. Some methods did not correlate well with subjective measures; heart rate appeared more sensitive to physical demands of the task rather than mental workload. Others such as electrodermal activity and electroencephalogram measurements showed good correlation with variation in the task cognitive demands while blink rates were found to correlate well with visual demands. Other studies support blink measures, particularly startle eye blink, as being sensitive to workload (Neuman 2002).

Wilson (2004) also reports on several psychophysiological techniques that are being developed for medical applications. Oximetry, Near-Infrared spectroscopy, fMRI (functional magnetic resonance imaging) and stress hormone assessment, while perhaps not practical for human factors or field applications today, may prove useful in the future for workload measurement as technology advances, making these techniques easier to use and less intrusive.

4.0 IMPORTANT REMAINING WORKLOAD MEASUREMENT ISSUES

Regardless of the workload methods selected, formulation of a general theory of workload that can put measurements into context requires many and varied experiments; reports from a single experiment are insufficient (Wierwille 1988, p. 316). It is essential that the results be shown to create a theory that is generalizable. Wierwille notes that individual differences are key features missing from most measurement approaches.

Colle and Reid (1998) note that context has a significant bearing on the measurement of workload, and that this is likely a perceptual rather than judgement issue. This has implications for selecting the range of stimuli, since Colle and Reid indicate that restricted ranges of stimuli can bias the workload rating results. In evaluations with a range of task difficulties at the low end of the difficulty scale, subjects tend to overrate the demands at the high end of this range. Conversely, for a range of tasks at the high end of the demand scale, subjects tend to underrate the demands at the low end of the scale. They conclude that this "... threat to validity ..." necessitates including context as a major consideration in workload experimental design, preferably by presenting a very broad range of stimuli to avoid range bias and by not labelling (anchoring) the measurement scales. Presumably, a similar effect could be accomplished by using doubly-anchored scales in subjective methods where the limiting anchors reflect absolute ratings, but single anchor, relative ratings may also be useful in workload measurement (Vidulich and Tsang 1987).

5.0 RECOMMENDING WORKLOAD MEASURES

When selecting a workload measure, or a battery of measures, the analyst should consider what the objective of the assessment is. If several design options need to be ranked on workload, then perhaps a univariate measure such as an Overall Workload scale is sufficient. If more diagnostic information is required, and this cannot be obtained through interviews, then the NASA TLX measure may be more appropriate. Primary and embedded Secondary task measures relevant to the operational context in which workload measures are desired should also be used. Psychophysiological measures are not recommended for most field analyses at this time; psychophysiological measures require further research to develop formal relationships among the various factors before they will be of use to the general analysis community.

Farmer and Brownson (2003) recommend that a battery of workload measures be selected for simulation-based assessments and provide guidance on such a selection (although the criteria used to select appropriate methods is unclear):

1) Modified Cooper-Harper (MCH);
2) Instantaneous Self Assessment (ISA);
3) Primary and Secondary tasks;
4) Heart Rate;
5) Heart Rate Variability;
6) NASA TLX;
7) Defence Research Agency Workload Scale (DRAWS); and
8) Blink rate.

This does not mean analysts should apply any or all measures from such a list that **might** be useful, but that many should be considered for the insight they can provide. The lack of a formal model relating the various workload measures seriously complicates interpretation. If a shotgun selection of methods is adopted, the analyst might well end up with a bewildering and contradictory set of results. A careful assessment of the task under study and its context is necessary to select an appropriate battery of workload measurement methods. This battery should include at least one objective measure and make use of quantitative subjective assessments (rather than subjective pass/fail ratings).

Wierwille, Rahimi et al. (1985) made the following recommendations for workload measurement techniques based on a series of evaluations.

1) For studies that are predominantly psychomotor (Wierwille and Connor 1983), the recommend the Cooper-Harper scale, the WCI/TE scale[12], and control movements/unit time; two other sensitive, but non-monotonic techniques (time estimation standard deviation and mean pulse rate) could be used to support the other methods, but are not recommended for use alone.

[12] See: Donnell, M.L. (1979) The application of decision-analytic techniques to the test and evaluation phase of the acquisition of a major air system: Phase III. Technical Report PR79-6-91. McLean, VA: Decisions and Designs, Inc. – Article not reviewed.

2) In communications studies (Casali and Wierwille 1983), the following methods were found to be sensitive load: Modified Cooper-Harper (MCH), Multi-descriptor Scale[13]; time estimation; pupil diameter; errors of omission; errors of commission; and, communications response time. Of these, the Multidescriptor Scale has been little used and might be advantageously replaced by another, more common subjective scale such as NASA TLX or SWAT.

3) In cognitive (mediation) studies (Wierwille, Rahimi et al. 1985), MCH, WCI/TE, time estimation, fixation fraction (proportion of time on principal display), mean reaction time and mean error rate were judged sufficiently sensitive to be useful, although time estimation was judged to be rather intrusive on the primary task of calculation.

4) Casali and Wierwille (1984) suggested the following measures were sensitive to perceptual loads: Modified Cooper-Harper, Multidescriptor and Workload Compensation Interface/Technical Effectiveness scales; primary tasks (control inputs) and secondary tasks (time estimation variability and rhythmic tapping); respiration rate. They concluded that none of these measurement techniques intruded significantly on the primary task performance, although I would question the user-acceptability of these secondary task measures in many practical applications.

Casper, Shively et al. (1987) created a decision support tool, WC FIELDE (Workload Consultant for Field Evaluations), in the mid 1980s to help researchers select appropriate workload measurement techniques. A web search failed to find much information, although it was referenced on two sites[14] and while it may still be available through HSIAC[15], no additional information on WC FIELDE was found on the NASA web site.

Castor (2003) has presented a method for matching task characteristics to workload measurement methods, but the GARTEUR (Group for Aeronautical Research and Technology in Europe) tool could be elaborated to include more workload methods; similarly, the WC FIELDE tool could be expanded and updated. Lysaght, Hill et al. (1989, pp. 64-65 and Chapter 8) rate a number of techniques according to their sensitivity, cost and diagnosticity, then propose a "matching model" to help researchers select appropriate workload measures. This was to elaborate on the WC FIELDE work, adding an expert system shell and expanding the scope beyond aviation, however, no evidence was found to suggest that this proposal came to fruition in a practical implementation.

The Internet search results (Appendix 1) show a large effort in the psychophysiological arena. Resources such as the Society for Psychophysiological Research are key to providing summary advice as well as supporting evidence for the various methods. This should become a valuable resource for practitioners to keep abreast of developments in psychophysiological methods and how they might be successfully applied to workload measurement in the future.

6.0 CONCLUSION

In the past twenty years, the science of workload measurement has not progressed nearly as far as one might have hoped. Many of the issues and concerns of the early 1980s are with us today. The science is not

[13] See: Casali, J.G. (1982) A sensitivity/intrusion comparison of mental workload estimation techniques using a simulated flight task emphasizing perceptual piloting behaviors. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia. – Article not reviewed.

[14] WC FIELDE: http://softwaretechnews.com/stn4-2/stiac.html and http://www.manningaffordability.com/s&tweb/heresource/tool/tool_list.htm

[15] http://www.softwaretechnews.com/stn4-2/hsiac.html

completely static, particularly in the psychophysiological domain, and a critical review of the topic could fill a sizable report. There seems to be more literature appearing frequently and it would be impossible to keep up with all the developments that may be of use unless one specializes in workload measurement.

That mental workload is multidimensional is not seriously challenged today, but whether workload is a scalar or vector has yet to be resolved and may only be relevant to predictive modelling, when the analyst wishes to assess the workload associated with performing two novel tasks together. Attempts to build computational models of human behaviour that are moderated by workload may provide a useful testbed to augment experimental methods attempting to validate proposed measures.

Subjective workload measures that support predictive modelling, such as VACP and DRAWS, usually focus on task demand in multiple channels. When coupled with task duration in simulations, these approaches produce aggregate measures that are sensitive to both task difficulty and time. These results provide diagnostic information of where the high workload is developing in the system and can be used to validate models for other scenarios. Predictive modelling approaches that focus on an overall workload metric such as time pressure, confound task demands with the time available. If task demands and the resulting workload can be characterized by one parameter, then an overall subjective workload measure may be sufficient. In all cases, it is highly desirable to latch predicted objective performance to empirical measurements if suitable performance models are available as a step towards validating the overall simulation.

When measuring workload empirically, the current recommendations are largely the same as twenty years ago: select a variety of workload measurement techniques that seem appropriate to the application and are likely to provide insight; do not select too many redundant measures, as this could produce conflicting results simply by chance. Understanding of the problem under study may require a number of experiments or trials in converging operations to clarify why some the results dissociate among the measures. A number of researchers in the field have created tools to help guide the selection of measurement techniques. An open source, public domain version of these tools, with references to validation data, would be a useful addition to the human factors and research communities.

Based on this review of the literature, psychophysiological measures should not be recommended for applied problems until researchers can develop a formal, unifying theory that explains the interactions of various physiological phenomena and the relationship to workload, despite the recent technological advances made. As a general practice, a global, univariate workload measure is suggested in conjunction with NASA TLX, as well as contextually relevant primary and embedded secondary task measures. SWAT is an alternative to the NASA TLX, although it is more laborious.

7.0 ABBREVIATIONS

BP	Brain readiness Potential	Freude and Ullsperger 1999
DRAWS	Defence Research Agency Workload Scale	Farmer, Belyavin et al. 1995 Farmer, Jordan et al. 1995
DRDC	Defence Research and Development Canada	http://www.drdc-rddc.dnd.ca
EEG	Electroencephalography	
ERP	Event Related Potential or Evoked Response Potential	Wilson and O'Donnell 1988

fMRI	Functional Magnetic Resonance Imaging	
GARTEUR	Group for Aeronautical Research and Technology in Europe	http://www.nlr.nl/public/hosted-sites/garteur/rfc.html
HMD	Helmet/Head Mounted Display	
HRV	Heart Rate Variability	Mulder, Mulder et al. 1999
HSIAC	Human System Information Analysis Center	http://iac.dtic.mil/hsiac/
IP	Information Processing	Hendy and Farrell 1997
IPME	Integrated Performance Modelling Environment	http://maad.com
ISA	Instantaneous Self Assessment	http://www.eurocontrol.int/eec/public/standard_page/1996_note_10.html
MCH	Modified Cooper Harper	
MRQ	Multiple Resource Questionnaire	Boles and Adair 2001
NASA	North American Space Agency	
NASA TLX	NASA Task Load Index	Hart and Staveland 1988
NATO	North Atlantic Treaty Organization	http://www.nato.int/
NATO RTO	NATO Research and Technology Organization	http://www.rta.nato.int/
OW	Overall Workload	
POP	Prediction of Operator Performance	Farmer, Belyavin et al. 1995 Farmer, Jordan et al. 1995
POPIP	Prediction of Operator Performance Information Processing	
SP	Preparatory Slow Brain Potential	Freude and Ullsperger 1999
SWAT	Subjective Workload Assessment Technique	Reid, G. B., C. A. Shingledecker, et al. 1981 Reid, G. B., C. A. Shingledecker, et al. 1981 Reid, G. B., F. T. Eggemeier, et al. 1982
VACP	Visual, Auditory, Cognitive, Psychomotor	Aldrich and McCracken 1984 McCracken and Aldrich 1984 Aldrich, Szabo et al. 1989
WC FIELDE	Workload Consultant for Field Evaluations	Casper, Shively et al. 1987
W/Index	Workload Index	North and Riley 1989

8.0 REFERENCES

Aldrich, T.B. and McCracken, J.H. (1984). A computer analysis to predict crew workload during LHX Scout-Attack Missions Vol. 1. Fort Rucker, Alabama, US Army Research Institute Field Unit, Fort Rucker, Alabama.

Aldrich, T.B., Szabo, S.M., et al. (1989). The development and application of models to predict operator workload during system design. Applications of human performance models to system design. G.R.B. McMillan, D.; Salas, E.; Strub, M.H.; Sutton, R.; van Breda, L. New York, Plenum Press. 2: 65-80.

Annett, J. (2002). "Subjective rating scales: science or art?" Ergonomics 45(14): 966-987.

Biers, D.W. (1995). SWAT: Do we need conjoint measurement. Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting, San Diego, California.

Biers, D.W. and Masline, P.J. (1987). Alternative approaches to analyzing SWAT data. Proceedings of the Human Factors Society 31st Annual Meeting. New York City. The Human Factors Society.

Boff, K.R., Kaufman, L., et al. (1986). Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance., John Wiley and Sons, Inc.

Boles, D.B. and Adair, L.P. (2001). "Validity of the Multiple Resources Questionnaire (MRQ)." submitted to the Human Factors and Ergonomics Society.

Boucsein, W. and Backs, R.W. (1999). Engineering psychophysiology as a discipline: Historical and theoretical aspects. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 3-30.

Brookhuis, K.A. and de Waard, D. (2002). "On the assessment of mental workload and other subjective qualifications." Ergonomics 45(14): 1026-1030.

Byers, C.J., Bittner, A.C.J. et al. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society's 32nd Annual Meeting, Santa Monica, California, USA.

Casali, J.G. and Wierwille, W.W. (1983). "A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load." Human Factors 25(6): 623-641.

Casali, J.G. and Wierwille, W.W. (1984). "On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues." Ergonomics 27(10): 1033-1050.

Casper, P.A.L., Shively, R.J., et al. (1987). Decision support for workload assessment: Introducing WC FIELDE. Proceedings of the Human Factors Society 31st Annual Meeting, New York City, The Human Factors Society.

Castor, M.C. (2003). GARTEUR Handbook of mental workload measurement, GARTEUR, Group for Aeronautical Research and Technology in Europe, Flight Mechanics Action Group FM AG13: 164.

Colle, H.A. and Reid, G.B. (1997). "A framework for mental workload research and applications using formal measurement theory." International Journal of Cognitive Ergonomics 1(4): 303-313.

Colle, H.A. and Reid, G.B. (1998). "Context effects in subjective mental workload ratings." Human Factors 40(4): 591-600.

Colle, H.A. and Reid, G.B. (1999). "Double trade-off curves with different cognitive processing combinations: Testing the cancellation axiom of mental workload measurement theory." Human Factors 41(1): 35-50.

Corwin, W.H., Sandry-Garza, D.L. et al. (1989). Assessment of crew workload measurement methods, techniques and procedures. Volume II – Guidelines for the use of workload assessment techniques in aircraft certification. Dayton, OH, Wright-Patterson Air Force Base, Wright Research and Development Centre, Cockpit Integration Directorate: 51.

Corwin, W.H., Sandry-Garza, D.L., et al. (1989). Assessment of crew workload measurement methods, techniques and procedures. Volume I – Process, Methods and Results. Dayton, OH, Wright-Patterson Air Force Base, Wright Research and Development Centre, Cockpit Integration Directorate: 237.

Curry, R., Jex, H., et al. (1979). Final report of the control engineering group. Mental Workload: Its theory and measurement. N. Moray. New York, Plenum Press: 235-253.

Damos, D.L. (1991). Dual-task methodology: some common problems. Multiple-task performance. D.L. Damos. London, GB, Taylor & Francis: 101-119.

de Waard, R. (1996). The measurement of driver's mental workload. Traffic Research Centre (now Centre for Environmental and Traffic Psychology). Heran, NL, University of Groningen: 198.

Derrick, W.L. (1988). "Dimensions of operator workload." Human Factors 30(1): 95-110.

Eggemeier, F.T., Wilson, G.F., et al. (1991). Workload assessment in multi-task environments. Multiple task performance. D.L. Damos. London, GB, Taylor & Francis, Ltd.: 207-216.

Fahrenberg, J. and Wientjes, C.J.E. (1999). Recording methods in applied environments. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 111-135.

Farmer, E. and Brownson, A. (2003). Review of workload measurement, analysis and interpretation methods, European Organisation for the Safety of Air Navigation: 33.

Farmer, E.W., Belyavin, A.J., et al. (1995). Predictive Workload Assessment: Final Report. Farnborough, Hampshire, Defence Research Agency.

Farmer, E.W., Jordan, C.S., et al. (1995). Dimensions of Operator Workload: Final Report. Farnborough, UK, Defence Research Agency: 58.

Fracker, M.L. and Wickens, C.D. (1989). "Resources, confusions, and compatibility in dual-axis tracking: Displays, controls, and dynamics." Journal of Experimental Psychology: Human Perception and Performance 15(1): 80-96.

Freude, G. and Ullsperger, P. (1999). Slow brain potentials as a measure of effort? Applications in mental workload studies in laboratory settings. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 255-267.

Gaillard, A.W.K. (1993). "Comparing the concepts of mental load and stress." Ergonomics 36(9): 991-1005.

Gaillard, A.W.K. and Kramer, A.F. (1999). Theoretical and methodological issues in psychophysiological research. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 31-58.

Gopher, D. and Braune, R. (1984). "On the psychophysics of workload: Why bother with subjective measures?" Human Factors 26(5): 519-532.

Gopher, D. and Donchin, E. (1986). Workload – An examination of the concept. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. K.R. Boff, L. Kaufman and J.P. Thomas, John Wiley and Sons, Inc: 41-1:41-49.

Hancock, P.A. and Meshkati, N. (1988). Human Mental Workload. Amsterdam, North-Holland (Elsevier Science Publishers B.V.).

Hart, S. and Wickens, C.D. (1990). Workload assessment and prediction. MANPRINT: An approach to systems integration. H.R. Booher. New York, van Nostrand Reinhold: 257-296.

Hart, S.G. and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human Mental Workload. P.A.M. Hancock, N. Amsterdam, North-Holland: 139-183.

Hendy, K.C. and Farrell, P.S.E. (1997). Implementing a model of human information processing in a task network simulation environment. Toronto, Defence and Civil Institute of Environmental Medicine: 110.

Hendy, K.C., Hamilton, K.M., et al. (1993). "Measuring subjective workload: When is one scale better than many." Human Factors 35(4): 579-601.

Hill, S.G., Iavecchia, H.P., et al. (1992). "Comparison of four subjective workload rating scales." Human Factors 34(4): 429-439.

Huey, F.M. and Wickens, C.D. (1993). Workload transition: Implications for individual and team performance. Washington, DC, National Academy Press.

Jex, H.R. (1988). Measuring mental workload: Problems, progress, and promises. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 5-39.

Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. Multiple task performance. D.L. Damos. London, UK, Taylor & Francis, Ltd.: 279-328.

Kramer, A.F., Sirevaag, E.J., et al. (1987). "A psychophysiological assessment of operator workload during simulated flight missions." Human Factors 29(2): 145-160.

Luximon, A. and Goonetilleke, R.S. (2001). "Simplified subjective workload assessment technique." Ergonomics 44(3): 229-243.

Lysaght, R.J., Hill, S.G., et al. (1989). Operator workload: comprehensive review and evaluation of operator workload methodologies. Fort Bliss, Texas, U.S. Army Research Institute for the Behavioural and Social Sciences: 262.

Marshall, S.P., Pleydell-Pearce, C.W., et al. (2002). Integrating psychophysiological measures of cognitive workload and eye movements to detect strategy shifts. Proceedings of the 36th Hawaii International Conference on Systems Sciences (HICSS'03), IEEE Computer Society.

McCracken, J.H. and Aldrich, T.B. (1984). Analyses of selected LHX mission functions: Implications for operator workload and system automation goals. Fort Rucker, Alabama, US Army Research Institute Aircrew Performance and Training.

Meshkati, N. (1988). Heart rate variability and mental workload assessment. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 101-115.

Meshkati, N., Hancock, P.A., et al. (1995). Techniques in mental workload assessment. Evaluation of human work. J. Wilson. London, GB, Taylor & Francis, Ltd.

Meshkati, N. and Lowewinthal, A. (1988). An eclectic and critical review of four primary mental workload assessment methods: A guide for developing a comprehensive model. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 251-267.

Mitchell, D.K. (2000). Mental workload and ARL workload modeling tools. Aberdeen Proving Ground, MD, US Army Research Laboratory, Human Research & Engineering Directorate: 35.

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

Mulder, G., Mulder, L.J.M., et al. (1999). A psychophysiological approach to working conditions. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 139-159.

Neuman, D.L. (2002). "Effect of varying levels of mental workload on startle eye blink modulation." Ergonomics 45(8): 583-602.

North, R.A. and Riley, V.A. (1989). W/INDEX: A predictive model of operator workload. Applications of human performance models to system design. G.R.B. McMillan, D.; Salas, E.; Strub, M.H.; Sutton, R.; van Breda, L. New York, Plenum Press. 2: 81-90.

Nygren, T.E. (1991). "Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload." Human Factors 33(1): 17-33.

O'Donnell, R.D. and Eggemeier, F.T. (1986). Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. K.R. Boff, L. Kaufman and J.P. Thomas, John Wiley and Sons, Inc: 42-1:42-49.

Reid, G.B., Eggemeier, F.T., et al. (1982). An individual differences approach to SWAT scale development. Proceedings of the Human Factors Society – 26th Annual Meeting.

Reid, G.B. and Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. Human Mental Workload. P.A.M. Hancock, N. Amsterdam, Elsevier Science Publishers B.V. (North-Holland): 185-218.

Reid, G.B., Shingledecker, C.A., et al. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society – 25th Annual Meeting.

Reid, G.B., Shingledecker, C.A., et al. (1981). Development of multidimensional subjective measures of workload. Conference on Cybernetics and Society sponsored by IEEE Systems, Man and Cybernetics Society, Atlanta, Georgia.

Roscoe, A.H., Ellis, G.A., et al. (1978). Assessing pilot workload, NATO AGARD (Advisory Group for Aerospace Research and Development) 7 rue Ancelle, 92200 Neuilly-sur-Seine, France: 84.

Sirevaag, E.J. and Stern, J.A. (1999). Ocular measures of fatigue and cognitive factors. Engineering psychophysiology: issues and applications. R.W. Backs and W. Boucsein. Mahwah, NJ., Lawrence Erlbaum Associates, Inc.: 269-287.

Vidulich, M.A. (1988). The cognitive psychology of subjective mental workload. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 219-229.

Vidulich, M.A. and Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society – 31st Annual Meeting. New York City. The Human Factors Society.

Whitaker, L.A., Hohne, J., et al. (1997). Assessing cognitive workload metrics for evaluating telecommunication tasks. Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting.

Wickens, C.D. (1977). Measures of workload, stress and secondary tasks. Mental workload: Its theory and measurement. N. Moray. New York, Plenum Press: 70-99.

Wickens, C.D. (1992). Engineering Psychology and Human Performance. New York, HarperCollins Publishers Inc.

Wickens, C.D. and Hollands, J.G. (1999). Engineering Psychology and Human Performance. Upper Saddle River, New Jersey, Prentice Hall.

Wierwille, W.W. (1988). Important remaining issues in mental workload estimation. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North-Holland): 315-333.

Wierwille, W.W. and Casali, J.G. (1983). “A validated rating scale for global mental workload measurement applications.” Proceedings of the Human Factors Society – 27th Annual Meeting: 129-133.

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

Wierwille, W.W., Rahimi, M., et al. (1985). “Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity.” Human Factors 27(5): 489-502.

Williges, R.C. and Wierwille, W.W. (1979). “Behavioral measures of aircrew mental workload.” Human Factors 21(5): 549-574.

Wilson, G.C., et. al. (2004). Operator functional state assessment. Paris, FR, North Atlantic Treaty Organisation (NATO), Research and Technology Organisation (RTO) BP 25, F-92201, Neuilly-sur-Seine Cedex, France: 220.

Wilson, G.F. (2001). "An analysis of mental workload in pilots during flight using multiple psychophysiological measures." International Journal of Aviation Psychology 12(1): 3-18.

Wilson, G.F. and Eggemeier, F.T. (1991). Psychophysiological assessment of workload in multi-task environments. Multiple task performance. D.L. Damos. London, UK, Taylor & Francis, Tld.: 329-360.

Wilson, G.F. and O'Donnell, R.D. (1988). Measurement of operator workload with the neuropsychological workload test battery. Human Mental Workload. P.A. Hancock and N. Meshkati. Amsterdam, NL, Elsevier Science Publishers B.V. (North Holland): 63-115.

Yeh, Y. and Wickens, C.D. (1988). "Dissociation of performance and subjective measures of workload." Human Factors 30(1): 111-120.

Appendix 1 – Internet Workload Measurement Technique Search: Hits by Keyword

An Internet search using the GOOGLE search engine was conducted to gauge the frequency of use of various workload techniques. No attempt was made to eliminate duplicate hits.

		Google Search Results	
	General Search Terms > Specific Search Terms	workload	mental OR cognitive workload
Subjective Ratings			
	Activation Scale	8	8
	Bedford Scale	31	23
	Defence Research Agency Workload Scale (DRAWS, DSTL, QinetiQ)	59	27
	Information Processing/Perceptual Control Theory (IP/PCT)	16	14
	Instantaneous Self Assessment of workload (ISA)	14900	2200
	Malvern Capacity Estimate (MACE)	2600	542
	Modified Cooper-Harper (MCH)	2810	742
	Multiple Resources Questionnaire (MRQ)	84	11
	NASA Task Load Index (NASA TLX)	1600	934
	Observer Rating Form	23	17
	Prediction of Operator Performance (POP, DSTL, QinetiQ)	10	17
	Pro-SWAT	3	3
	Quantitative Workload Inventory (QWI)	60	20
	Rating Scale Mental Effort (RSME)	82	65
	Raw TLX (RTLX)	42	35
	Self report	5230	3750
	Subjective Workload Assessment Technique (SWAT)	3660	758
	VACP	50	41
	W/Index	71	25
Performance Measures			
	Dual task	801	712
	Embedded task	35	27
	Primary Task	4340	1940
	Reaction Time (RT)	32600	8250
	Secondary task	864	684
	Subsidiary task	84	65

		Google Search Results	
	General Search Terms > Specific Search Terms	workload	mental OR cognitive workload
Psychophysical Measures			
	Psychophysiological	2060	1800
	Eye movement measures		
	Blink duration	44	43
	Blink latency	9	7
	Blink rate	211	177
	Endogenous eye blinks (EOG)	606	317
	Eye blink	172	141
	Eye fixations	231	195
	Eye movement	2180	1670
	Glissadic saccades	1	1
	Oculographic activity	2	2
	Pupil diameter	198	172
	Saccade duration	11	11
	Saccadic velocity	15	15
	Cardio-vascular/respiratory measures		
	Blood pressure	43100	15800
	Heart Period (HP)	19	17
	Heart rate	23100	8240
	Heart rate variability	1510	676
	Inter-beat-interval (IBI)	530	156
	Respiration	6460	2770
	Respiratory Sinus Arrhythmia (RSA)	42	38
	Stress-related hormone measures		
	Adrenaline	5500	2060
	Catecholamines	2250	575
	Cortisol	2850	1770
	Epinephrine	2570	730
	Noradrenaline	962	368
	Pprolactin	576	325
	Vanillylmandelic acid	15	5

		Google Search Results	
	General Search Terms > Specific Search Terms	workload	mental OR cognitive workload
Psychophysical Measures			
	Electrical biosignals		
	Autonomic nervous system (ANS)	9390	2720
	Central nervous system (CNS)	9260	4020
	Electrocardiogram (ECG)	12800	4270
	Electrodermal activity (EDA)	3930	498
	Electroencephalogram (EEG)	4320	2900
	Electromyogram EMG	3540	1660
	Event related potentials	409	379
	Evoked cortical brain potential	2	2
	Evoked potential	361	275
	P300 amplitude	59	59
	P300 latency	36	36
	Parasympathetic nervous system	216	98
	Peripheral nervous system (PNS)	868	250
	Skin conduction response (SCR)	4960	698
	Skin resistance level SRL	2320	291
	Skin resistance response SRR	589	95
	Somatic nervous system	30	18
	Speaking fundamental frequency	4	3
	Speaking rate	93	69
	Sympathetic nervous systems (SNS)	1550	596
	Vocal intensity	11	9