USAARL-TECH-TR--2020-038

UNITED STATES ARMY AEROMEDICAL RESEARCH LABORATORY

A Case for a Measurement of Incremental Workload

Jon Vogl, Amanda Hayes, Christopher Aura, Lana Milam, Leonard Temme, & Paul St. Onge

Notice

Qualified Requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Fort Belvoir, Virginia 22060. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of Address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
25-08-2020	Technical Report	

4. TITLE AND SUBTITLE	
A Case for a Measurement of Incremental Workload	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Jon Vogl, Amanda Hayes, Christopher Aura, Lana Milam, Leonard Temme, & Paul St. Onge	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army Aeromedical Research Laboratory P.O. Box 620577 Fort Rucker, AL 36362	USAARL-TECH-TR--2020-038
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
U.S. Army Aeromedical Research Laboratory P.O. Box 620577 Fort Rucker, AL 36362	USAARL
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

Goldbelt Frontier, LLC

14. ABSTRACT

This paper contains a review of literature pertaining to mental workload and mental workload measurement techniques. Each technique discussed also includes an assessment of the viability of use in an operational environment.

15. SUBJECT TERMS

Mental workload, cognitive workload, workload

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Loraine St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	42	19b. TELEPHONE NUMBER *(Include area code)* 334-255-6906

This page is intentionally blank.

Table of Contents

List of Figures

Introduction

The landscape of combat is rapidly changing. In 2017, the Army released the Army Modernization Strategy, which emphasized an increase in research and technology development efforts to ensure continued over-match of our advisories to discourage conflicts (Milley & McCarthy, 2017). This strategy contains six priorities including, "long-range precision fires, next generation combat vehicles, future vertical lift (FVL), Army network, air and missile defense capabilities, and Soldier lethality" (Milley & McCarthy, 2017). The strategy stresses that future engagements will likely take place in multiple domains. Multi-domain operations (MDO) include traditional land, air, and sea battle spaces and extend into space and the virtual world with reduced communications. To be effective, Soldiers in MDO will be operating such that they will be using cognitive resources more than ever before as they will be required to travel over longer distances and remain alert for longer periods of time while using extremely complex systems of systems.

To support the Army Modernization Strategy's initiative of Future Vertical Lift, the Army is currently developing requirements for the future attack reconnaissance aircraft (FARA) and the future long-range assault aircraft (FLRAA). The FARA and FLRAA platforms will allow for longer periods in combat in addition to hosting a suite of advanced technologies and weaponry. These aircraft will likely be the most advanced and complex systems of systems the Army has in its arsenal. This means that these aircraft are likely to require pilots to multi-task on a level far above current helicopters. Because of the increasing demands placed on the pilots while flying these vehicles, the need to monitor pilot cognitive workload, health, and well-being in real-time has become integral to mission accomplishment. With real-time physiological monitoring, it will be possible to track and understand the degree of task cognitive demand and associated Mental Workload (MWL) placed upon the pilots throughout the various phases of the MDO mission sets. These data will inform leadership and team members, as well as provide critical feedback to the individual operators. These data will also inform key decision points for the cockpit layouts specific to human system interaction. However, much work remains, as unknowns exist regarding which measures are most effective at capturing and quantifying MWL, how best to deploy those sensors within the cockpit, and how to quantify the data such that the results can easily be interpreted in real-time to aid decision-making.

In order to support the expanded FVL mission, research is ongoing at the U.S. Army Aeromedical Research Laboratory (USAARL; Fort Rucker, AL) to develop physiological measures that can readily and reliably capture and determine the cognitive state of pilots and crew. Although over a decade old, Cain (2007) provided a summary of the available measures of MWL, however, a knowledge gap exists regarding which measures best match the needs of the U.S. Army. Further, many different operational definitions of MWL have emerged over the years. MWL has generally been identified as the interaction between the cognitive resources **required** to complete a task and the resources **available** to complete a task (Wickens & Tsang, 2015). Below we will expand on this definition to address operational needs. The present report provides information about the theory behind MWL and the three most common tools used to assess it: subjective measures, performance measures, and psychophysiological measures.

Mental Workload and its Assessment

Interest in the concept of MWL and its assessment has increased over the last 50 years. In recent decades, advancements in technology have afforded researchers the ability to push the limits of data acquisition systems to capture more data per unit time with increased precision. The speed and miniaturization of computational hardware allows for many of these sensors to have a smaller footprint both on the person and within the environment. Wearable sensors and the associated hardware to drive the systems have shrunk in size by multiple log units over the last 50 years. Given the assessment tools now available, relating performance metrics to subjective and physiological responses requires less physical space and power than ever before. This data integration allows researchers to begin to conceptualize how these assessment tools fit together in practical applications. MWL assessment is currently ongoing at USAARL with the end goal of transitioning a real-time metric into flight trials to aid in FVL human system integration key decision points. This technical report is not an extensive evaluation of the current MWL literature. Rather, it is intended to introduce the basic concept of MWL, some of the foundational models that frame the topic, and the methodologies reported in the literature that may be employed to assess MWL in Soldiers and pilots. A more systematic review of the recent literature of MWL assessment in expert-trained populations is in preparation at USAARL.

Mental Workload Definition and Terminology

Van Acker, Parmentier, Vlerick, and Saldien (2018) proposed a formal, modern definition of MWL:

"*Mental workload is a subjectively experienced physiological processing state, revealing the interplay between one's limited and multidimensional cognitive resources and the cognitive work demands being exposed to.*"

This definition, based on a concept analysis of the workload literature, provides a stable framework to begin conceptualizing what the abstract idea of what MWL encompasses relative to the variety of definitions found in the literature (Cain, 2007). Three aspects of MWL elucidated by this definition will be discussed:

1. MWL is induced by the interaction of a specific operator's capabilities (i.e., experience level, personality factors, mental capacity, etc.) and the specific demands of the task(s), which can be further confounded by the environment in which the task occurs.
2. MWL is a subjective experience that results from physiological processes taking place while engaged in a task (e.g., neurophysiological activity, heart rate, etc.).
3. Cognitive resources utilized to perform a task are limited and multidimensional, and the availability of these resources influence the experience of MWL.

Therefore, MWL is a physiologically based subjective experience that can be operationally measured to gain insight regarding the operator's ability to accurately and efficiently continue a task or take on new tasks.

Mental workload as a task-resource interaction.

The first aspect highlights MWL as the reaction of an operator to specific task demands in a specific environment. Tasks invoke different levels of MWL depending on the characteristics of the specific task, the subjective response of the operator, and the environment. Thus, MWL emerges from the interaction of an individual with a task in a specific environment. To discuss this interaction, it is essential to differentiate the characteristics of the task from those of the environment and from those of the individual. This point, while obvious, is subtle. For example, Hancock and Matthews (2019) point out that the MWL literature frequently is confused because of inconsistencies in the use of terminology to describe the characteristics of the task, the environment, and operator's response to them. Early on, de Waard (1996) identified terms often associated incorrectly with MWL characteristics and task-environment interactions. To minimize such confusions, the present report uses de Waard's terminology.

Tasks come in a great variety of sizes, shapes, and complexities. Norman and Bobrow (1975) divided MWL tasks into those that are data-limited and those that are resource-limited. As the names imply, performance on **data-limited tasks** is limited by the incoming data presented to the operator, whereas performance on **resource-limited tasks** is limited by the capabilities of the operator for performing the task. This is an important distinction for MWL. When trying to perform a data-limited task, recruiting more mental resources to perform the task yields at best only marginal benefits since the increased effort does not deal with the shortcomings of the data signal; for example, when trying to read the small, blurry letter on a visual acuity chart, working harder will not make the text bigger nor will working harder un-blur it. On the other hand, recruiting more mental resources for resource-limited tasks may increase task performance. This is achieved by trying harder, focusing attention, and reducing distraction. For example, more effort must be expended to hold nine digits in working memory versus four digits. Additionally, with enough consistent practice with a resource-limited task, the task shifts to a more 'automatic' one, in a sense transforming the task from a resource-limited to a data-limited task (Norman & Bobrow, 1975).

Norman and Bobrow's (1975) performance-resource dichotomy can be expanded to the multitasking domain to predict the interference that will occur between tasks. When two resource-limited tasks are performed simultaneously, each task competes for an operator's limited resources. This results in performance decrements in both tasks, relative to each task being performed in isolation. Typically, this type of task pairing is to be avoided, but researchers can exploit resource competition between tasks to determine how much demand an operator can handle or to simulate real-world environments. Conversely, when a data-limited task is paired with a resource-limited task, there is little to no resource competition and no decrease in performance. Product designers can take advantage of this non-competitive task-pairing scenario to increase efficiency, safety, and user experience ratings.

It is important to note that many tasks have both data-limited and resource-limited processes associated with their performance. Norman and Bobrow's (1975) performance-resource functions detail these combinations, as seen in Figure 1. Data-limited processes are indicated by flat regions (i.e., expending resources does not improve performance). Resource-limited processes are indicated by curves with positive slopes (i.e., expending resources improves performance). Initially, a data-limited process will require some amount of resources via sensory processing (denoted by the initial slope on the data-limited curves outside the range of examined resource allocation). After sensory-related processing is complete, all that remains is the operator’s decision based on the quality of the perceived signal. A resource-limited process requires indefinite resource expenditure across the examined range of resource allocation. Thus, performance can continuously improve as resources are being spent (up to their limited resource capacity). The term transitional process has been ascribed to tasks that exhibit both resource-limited and data-limited functions in the performance-resource curve’s region of interest. Transitional processes encompass tasks that only require a finite amount of resource expenditure before a decision is to be made using the processed data.

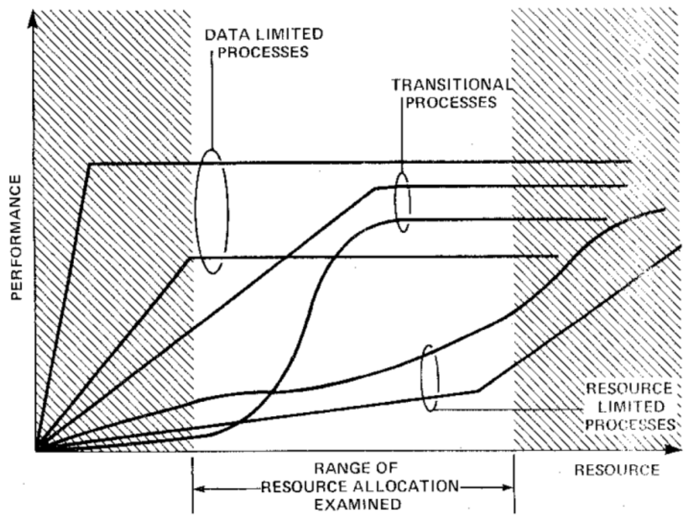

Figure 1. Norman and Bobrow's (1975) classes of performance-resource functions. Performance is represented on the y-axis and cognitive resource allocation on the x-axis. The middle section identifies the range evaluated by the researchers in their work. The region on the left represents tasks with fewer data-limiting processes that were below the levels tested. This depicts that data-limited processes have performance that rises quickly with little increase in resource expenditure, reaches a maximum, and cannot increase with more cognitive effort; thus, the flat horizontal lines emerge after the initial increase in resource allocation. The authors predict performance based on increased resource allocation in the right region. These resource-limited processes show gradual increase in performance as resource allocation increases. Transitional-processes are a combination of the data- and resource-limited processes in the region of resource allocation examined. The shape of the line changes based on the degree to which the tasks are dominated by each process. Two different combinations are identified, one linear, and one sigmoid.

The nature of the task is defined by the amount of resources expended to achieve a certain level of performance and determines several MWL factors. When examining the resources required to achieve a certain level of performance on a task, the concepts of task demand and task complexity require differentiation. **Task demand** represents the criterion resource capacity required to complete a task to the operator’s goal. Once this goal is set, the required resources are external and independent of the operator. For example, the ideal goal of achieving 100% on an exam represents a goal defined by the task, and yet, is external to the operators who take the exam. However, should an operator find that a score of 75% is acceptable; the operator sets a new subjective goal prior to engaging in the task. When comparing the demand between operators, the demand on the operator who seeks to reach the 100% goal will experience higher levels of task demand than the operator who finds 75% as acceptable. **Task complexity** is causally related to demand, and is characterized by the number of processing stages required to reach the goal. Intuitively, as the number of processing stages increase, the task complexity similarly increases. de Waard (1996) notes that demand and complexity are primarily external factors; they can be mediated by the subjective goals of the operator.

Another descriptor of the task can be seen from the point of view of the operator. **Difficulty** is defined as the amount of voluntarily mobilized resources required by a specific operator to keep up with the demand imposed by the task to reach the goal. Thus, factors intrinsic to the operator, such as individual capacity limits, strategy, state, and motivation begin to influence the perceived task difficulty (de Waard, 1996). As stated by Kantowitz (1987), the difficulty of a task is the result of the interaction between the operator and the task, while factors like a task's complexity looks at the task in isolation. Task difficulty is thus a perception whereas task complexity is a trait. Lastly, **effort** is the voluntary mobilization of an operator's resources in order to compensate for increased task demands. While effort is a reaction to task demand to maintain performance, they are not necessarily related. The structure of the task (data-limited vs. resource-limited) as well as practice effects, the operator's cognitive and physiological state, and the goals and strategies of the operator can affect how much effort is required to compensate for increasing task demand. Overall, a task cannot be defined by a specific level of workload due to individual differences. However, a task can be defined in terms of its processing type, criterion goals, complexity, and the attentional resources upon which the task places demand for criterion performance. On the other side of the interaction, the operator's perception of task difficulty is mediated by several factors driven by subjective decisions made by each unique operator. Rouse, Edwards, and Hammer (1993) state the idea that 'experienced load' may be a better term to encapsulate the idea that MWL is not just task-specific, but is also person-specific. The task-operator relationship will be expanded upon in the discussion of the region models of workload and performance.

Workload is a subjectively experienced physiological processing state.

MWL can be conceptualized as a subjective experience allowing for the introspective analysis of the physiological processing state and performance levels experienced during the task. This concept is clear to an individual who is engaging in tasks that impose significantly higher levels of workload. As such, through introspective reflection of the experience of completing a task, an operator can rank their subjective experiences of MWL, either on an arbitrary scale or against other tasks of varying difficulty. While this ranking may be unique to the operator, generalizability may be possible across metrics with highly trained subjects. Additionally, evidence suggests that physiological changes occur while experiencing high levels of MWL. For review, see Lohani, Payne, & Strayner (2019). Coupled with task performance metrics, subjective responses and physiological measures offer researchers a window into the abstract experience of MWL.

O'Donnell and Eggemeier (1986) completed the foundational work that outlined the methodological criterion of workload assessment techniques. Using a set of five criteria, they outlined four methods of MWL assessment that interact with the subjectively experienced physiological-processing-state of the operator: primary task measures (performance-based), secondary task measures (performance-based), and physiological measures, and subjective measures. The two performance measures, primary task measures and secondary task measures, encompass strategies to relate a task's performance metrics to the level of MWL experienced by an operator. Physiological measures take advantage of subtle changes in the body that occur in response to increasing levels of task demand. Some obvious physiological changes perceived by an operator are increased heart rate, respiration rate, or muscle tension (Cain, 2007). Other physiological changes can occur without the operator's awareness. Dilation of the pupil, changes in visual scan paths, blink rate and duration, blood pressure, blood oxygen levels, heart rate variability, skin conductance, brain oxygenation, brain electrophysiological activity, as well as many other physiological signals have been measured in response to increasing MWL (Lohani, Payne, & Strayer, 2019). Lastly, self-report methods can also be employed as the experience of MWL is primarily a subjective one. Scales, such as the Instantaneous Self-Assessment of Workload (ISA), Subjective Workload Assessment Technique (SWAT) and the NASA Task Load Index (TLX), reliably assess workload through self-report (Cain, 2007). These various and continually advancing techniques, when used in combination (performance + physiological + subjective measures), can provide a more robust operational measure of MWL as experienced by an individual. However, concern arises regarding dissociations among the various performance, subjective, and physiological measures that can make workload assessment techniques difficult to interpret and utilize effectively (Hancock & Matthews, 2019).

Limited and Multidimensional Cognitive Resources

Lastly, a current definition of MWL requires a discussion of available mental resources that can be characterized as limited and multi-dimensional. Humans have limitations in regards to the amount of data that they can process. In fact, the very nature of the human attentional system is designed to serve as a filter for enormous amounts of data. Without this filtering, humans would exist in a constant state of overload. To more fully understand cognitive filtering, the terms "resource" and "capacity" warrant differentiation despite being used interchangeably in some literature. Wickens (1992) defined **capacity** as the upper limit of processing capability of a particular operator for a specific resource. **Resources** are defined as mental effort voluntarily allocated to improve processing efficiency (Wickens, 1992; Norman & Bobrow, 1975; de Waard, 1996). As such, capacity and resources vary depending on the individual differences between operators. Of note, that capacity is elastic, and has the ability to vary, given the task demands and the applied effort of the operator (Kahneman, 1973). However, once capacity is reached, the operator will 'hit their limit' for a particular resource and failure in the task will begin to occur.

Norman and Bobrow's (1975) data-limited vs. resource-limited dichotomy can explain why some tasks can be performed in parallel more efficiently than others. This dichotomy explains why effective time-sharing can occur between combinations of some resource-limited tasks (e.g., a visual and auditory task set) but not others (e.g., a visual and visual task set). Wickens (1984) proposed the Multiple Resource Theory (MRT) to answer this question. Instead of assuming a single resource exists that is shared across all methods of processing (i.e., the 'modal' view, as was initially commonly thought to be the case [Kahneman, 1973]), Wickens divides attentional resources into separate pools. These pools are based on four dichotomous dimensions: Stages (perceptual + cognitive/response), modalities (visual/auditory; can also expand to haptic), codes (spatial/verbal), and visual channels (focal/ambient). These dimensions can be seen in Figure 2, a depiction of the 'Wickens Cube' (Wickens, 2002).

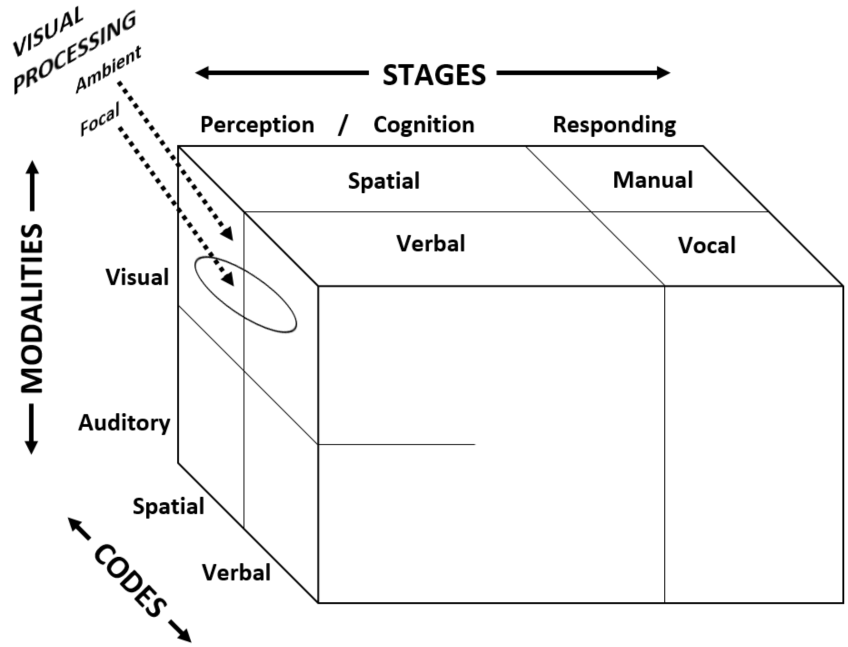

Figure 2. Visualization of the MRT model; the ‘Wickens Cube’ (Adapted from Wickens, 2002). This figure shows four dichotomous dimensions: Stages (x-axis), modalities (y-axis), codes z-axis), and visual channels (nested in the visual modality). Stages include the perception + cognition of data and the response selected and executed. Modalities are the channels used to deliver the information to be perceived and processed via cognition. The modality dimension represents sensory systems that can be efficiently time-shared and, in MRT, is often separated into visual and auditory modalities (but can be expanded to other senses, such as haptics). Codes represent the time-sharing efficiency between analogue/spatial and categorical/symbolic processes parts of perception, cognition and response that depend on separate resources divided across the two cerebral hemispheres (akin to Baddeley and Logie’s (1999) model of working memory which features a “visuo-spatial sketchpad” and “phonological loop” compartmentalizing the spatial/verbal dichotomy). Visual processing is divided into ambient and focal, with ambient representing unique aspects of peripheral vision and focal representing the information flow in central vision.

Each of the resource pools has its own capacity and an individual task may be able to push an operator beyond their particular capacity for a specific resource resulting in operator overload. However, proper task and product design can typically prevent operator overload. The MRT overcomes limitations of other human information processing models through its explanation of multitasking. Multitasking situations are defined as when one task is concurrently performed with another task (as is the case with secondary task procedures). Subsequently, the relative performance on each task is lower than the performance that could be achieved on each task in isolation (Wickens, 2002). This decrement in performance occurs due to the interference caused by the two tasks competing for the same resource. For example, driving in a busy urban area while navigating a global positioning system (GPS) display on the center console of the automobile will cause resource competition. In this multitasking situation, the GPS display task and the driving task would interfere with each other along the visual dimensions. Driving in the busy area and searching for the correct icons on the display requires simultaneous and constant visual surveillance of spatially distinct areas. Additionally, more interference may occur within the response-spatial (i.e., manual response) and cognition-spatial dimensions, as each task may require these resources. While MRT is most valuable when trying to account for performance in an “overload” situation, it can also be used as a framework in the design and development of products and research methods (Wickens, 2002; see computational MRT in Wickens, 2008).

Summary

Recalling the Van Acker et al. (2018) definition of MWL, continuing efforts exist to move towards a unified definition of the abstract concept: “*Mental workload is a subjectively experienced physiological processing state, revealing the interplay between one's limited and multidimensional cognitive resources and the cognitive work demands being exposed to.*" Three major aspects of MWL are highlighted in this definition. First, MWL is the reaction of a specific operator to the task demands in an effort to reach a criterion level of performance. The specific properties of the task (processing limitations, demand, and complexity) and the response of the operator (perceived difficulty, motivation, strategies, and state) represent the core features of MWL. Second, MWL is a subjective and physiological experience, allowing for different assessment techniques (primary task, secondary task, physiological, and subjective measures) to provide windows into the operators’ state. Third, understanding that mental resources are both limited **and** multi-dimensional is key to explaining human behavior in regards to task performance. MRT (Wickens, 2002) explains how a multi-dimensional set of resources can accurately account for a range of human experiences in regards to task performance. Putting these elements together, researchers can begin to define precise experimental procedures to continue the modelling and assessment of workload under specific dimensional levels of task demand.

Models of Mental Workload and Task Performance as a Function of Demand

Combining the previously defined concepts formulates a model of MWL and task performance as a function of task demand. A foundational outline of the relationship between task performance and task demand was outlined by Meister (1976) and was adapted by de Waard (1996). This relationship is illustrated in Figure 3 in the Region Model. Three regions separate into three discrete states that are characteristic of performance-workload relationships: Region A, Region B, and Region C. Region A defines a state of low demand and high performance. In this region, an increase in task demands will not lead to a decrease in performance, as the operator can increase their effort (i.e., the voluntary recruitment of available resources to task demands) to maintain performance. Region B defines a state of increasing task demands coupled with decreasing levels of performance. As such, Region B represents workload levels that exceed the capabilities (i.e., the capacity of a utilized resource) of the operator, and performance begins to approach catastrophic decline. Lastly, Region C defines a state of total failure in a task, where task demands are high and performance remains stagnant at low levels, regardless of any additional exertion of mental effort made by the operator.

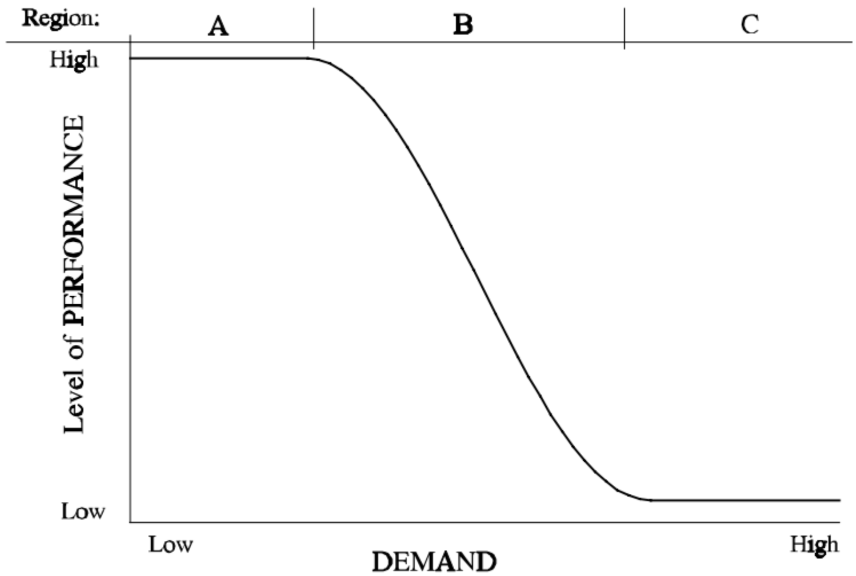

Figure 3. Region model. Relationship between task performance and task demand (Meister, 1976). Figure referenced from de Waard (1996). Three regions separate the three discrete states that are characteristic of performance-task demand relationships. Region A defines a state of low demand and high performance. Region B defines a state of increasing task demands coupled with decreasing levels of performance. Region C defines a state of total failure in a task, where task demands are high and performance remains stagnant at low levels.

This relationship determines where different workload assessment techniques can detect increases in capacity expenditure. For example, primary task measures are only sensitive to capacity expenditure (effort) in Region B, where actual performance begins to decline. Physiological measures, however, may discriminate levels of capacity expenditure in non-overload conditions, such as in Region A (O'Donnell & Eggemeier, 1986). Depending on the nature of the task(s), researchers can use this basic function to determine which workload assessment techniques would be most effective given the known constraints of the task.

However, Meister's (1976) region model does not take into consideration the domain of underload (i.e., boredom), as is common in vigilance tasks. The idea that underload affects performance stems from a classic study performed by Yerkes & Dodson (1908), the cloudy genesis of the colloquially known 'inverted U' performance curve. The initial work of Yerkes and Dodson (1908) revealed that exposure to medium strength electric shocks were more likely to elicit learning of a habit in mice, compared to both low and high strength electric shocks. While Yerkes and Dodson (1908) focused on the relationship between state and learning, the idea has been expanded to include the effects of arousal (i.e., stress, motivation, task demand, etc.) on performance (Hebb, 1955; Kahneman, 1973; Teigen, 1994). This expansion to performance is commonly referred to as the Yerkes-Dodson Law. Optimal performance sits at moderate levels of arousal, but lowering or raising arousal from these levels can have a detrimental effect.

de Waard (1996) expanded the region model to include the domain of underload and the 'inverted-U' function. As seen in Figure 4, the extended region model depicts how the inverse relationship between MWL and performance varies as a function of task demand. The addition of Region D, for ‘Disengagement’, encompasses the domain of underload. Region D defines a state of low task demands and low performance corresponding to higher levels of experienced MWL caused by a reduction in maximum capacity (see Malleable Attentional Resources Theory (MART) of Young & Stanton, 2002). The capacity bars in Figure 4 show that maximum resource capacity is severely reduced in Region D. As such, even small amounts of task demand utilize a large percentage of available resource capacity. Due to this higher rate of resource capacity expenditure, MWL is perceived by the operator as being high.

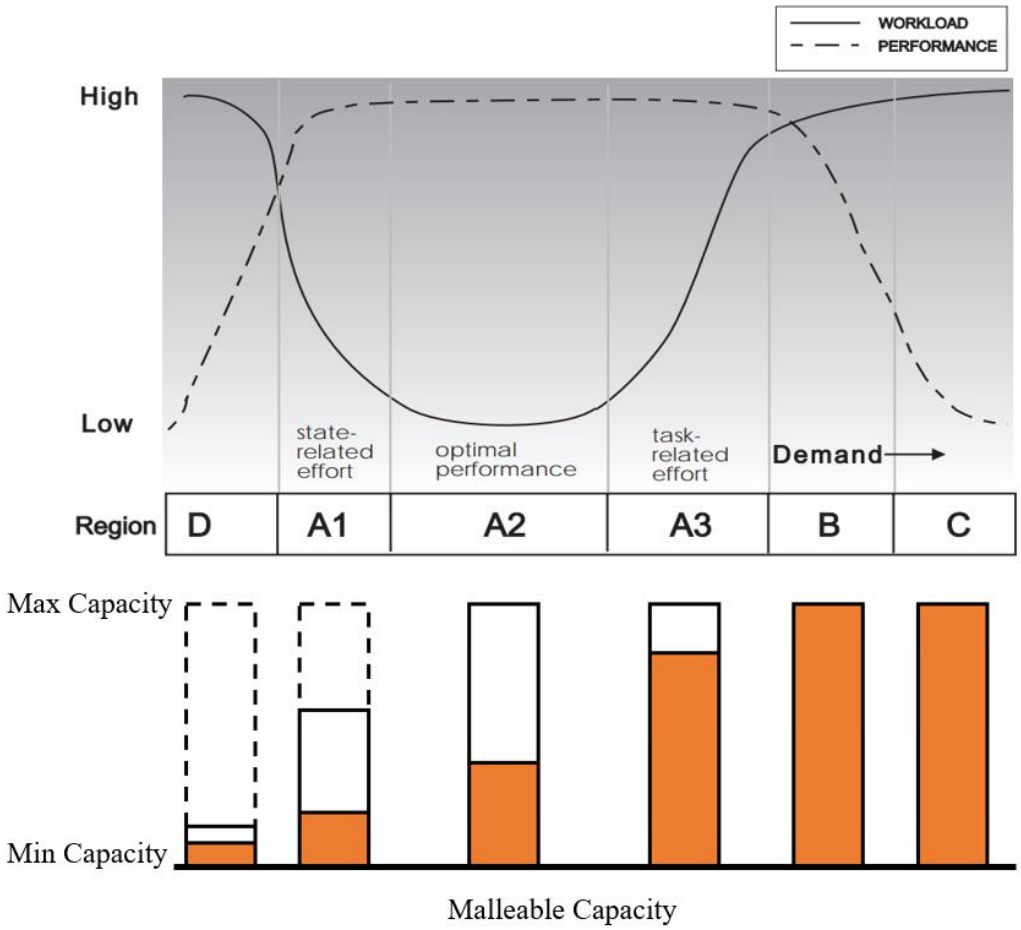

Figure 4. Extended region model (adapted from de Waard, 1996). This model highlights the domain of underload (Region D), and separates the traditional Region A into three parts to identify changes that occur at the workload level. The addition of Region D completes the ‘inverted U’ model of performance. The concepts of malleable capacities and resource expenditure are depicted for each region below the plot.

As task demand increases, the operator becomes more engaged with the task and enters Region A, represented by high performance and moderate task demand. However, de Waard partitions Region A to display variations in MWL within this region. In Region A1, a reduction in MWL is observed. Using state-related effort, an operator becomes more engaged with the moderately demanding task and commits more resource capacity to the task. This expansion reduces the overall percentage of utilized resource capacity (relative to Region D), which, in turn, reduces MWL. Region A2, then, represents the ideal state of moderate task demands, high performance, and low MWL. Region A3 encompasses the idea that human operators are flexible in their reactions to increasing task demands. In Region A3, an operator uses task-related effort to maintain high performance even as MWL and task demands increase. The capacity bar below Region A3 in Figure 4 shows how task-related effort increases resource capacity expenditure (i.e., the increasing orange bar) and, ultimately, MWL. Regions B and C represent maximum resource capacity expenditure with increasing diminished performance until task failure and/or abandon.

In the context of the region model, MWL is interpreted in a multidimensional fashion. The different workload dimensions occupy different regions of the model simultaneously, as would be accounted for by MRT (Wickens, 1984). For example, consider a driving task where an operator must both drive a vehicle safely and engage in a conversation via text messaging. Visual and manual response resources are heavily taxed by both tasks individually and the performance of both tasks concurrently will likely push a driver into Region B or C for each resource. This is problematic, as driving-related performance decrements associated with Regions B and C are likely to lead to an accident. To remedy this situation, the visual and manual response demand of text messaging can be offloaded to the less taxed auditory and verbal response domains through the use of hands-free telephone operation. This offloading would result in lower levels of demand spread across multiple mental resources. Thus, the demand spread across multiple resources would result in the operator being in Region A2 or A3 for each resource, rather than remaining in the regions negatively affecting performance (i.e., Regions B or C).

Using the region model, the ideal limits of experienced workload are determined by the point at which an operator begins to be overloaded. This threshold is typically referred to as the ‘redline’ (Colle & Reid, 2005). Traditionally, the redline is a threshold set between the boundary of region A and region B in the original region model, or in de Waard’s (1996) expanded region model, Regions A3 and B. Therefore, the region to the left of the redline is referred to as the ‘reserve capacity’ region, and the region to the right of the redline is referred to as the ‘overload’ region (Young, Brookhuis, Wickens, & Hancock, 2015). Depending on the task, and concerns for the well-being of the operator, adjusting the redline to occur between regions A2 and A3 may be ideal. This adjustment could mitigate the long-term effects of prolonged sessions of high MWL experienced in a work setting, such as long-term stress (Zijlstra & Mulder, 1989; de Waard, 1996) or hypertension (Johnson & Anderson, 1990). Likewise, working towards the domain of underload, an ideal redline would occur between regions A1 and A2, where enough state-related effort is utilized to keep the operator actively engaged. A definitive placement of a redline for the ‘underload’ region remains elusive given the current state of understanding (Young et al., 2015). When one resource capacity approaches the redline due to increasing task demand, diverting some of that demand to another resource capacity will improve performance (Dixon & Wickens, 2005). This diversion could be implemented through informed design of the task or product or even through strategies utilized by the operator.

Workload Measurement Techniques

While MWL may be an abstract concept not entirely grounded by a representational architecture in the brain, there are tangible responses elicited by the human body that can provide a basis for quantifying MWL. These responses to variable levels of MWL manifest themselves in three domains: subjective experience, the performance of primary and secondary tasks, and physiology. As O'Donnell and Eggemeier (1986) describe, each of these domains has specific advantages and disadvantages. For example, physiological workload assessment techniques offer relatively sensitive, sometimes diagnostic (depending on the measurement used), and generally non-intrusive methods to measure MWL. On the other hand, secondary task performance measures provide sensitive, highly diagnostic, but intrusive methods to examine MWL. This section reviews some of the specific MWL assessment techniques used in the subjective, performance, and physiological modalities as framed by the aviation domain and criteria set forth by O'Donnell and Eggemeier (1986).

Subjective Measures

One of the major points put forward by the Van Acker, et al. (2018) definition of MWL is that workload is a *subjectively* experienced physiological processing state. Generally, when an operator is experiencing high levels of MWL, they are aware that they are being stressed and pushed toward their limits in terms of resources, time, attention, etc. Using introspection (i.e., reflecting on one's experience), an operator can provide a report of their consciously experienced levels of MWL. As such, many attempts at capturing this introspective report have been made. Typically, subjective assessments of MWL rely upon the use of rating scales that identify how much effort or capacity was used during task completion. These rating scales are used to generate a unidimensional assessment of workload (as in the Crew Status Survey, Instantaneous Self-Assessment of Workload, Malvern Capacity Estimate, Modified Cooper-Harper, and the Rating Scale Mental Effort) or a multidimensional assessment of workload (as in the Subjective Workload Assessment Technique, NASA Task Load Index, and the Workload Profile). Such subjective workload assessments typically are obtained after the task is completed, relying on the operator's memory of their experiences during the task. However, some rating scales can be administered during the task (as in the Instantaneous Self-Assessment of Workload) as a way to provide a more reliable measure of workload through different phases of the task but at the cost of potentially interfering with task performance. Rating scales are of a subjective nature and may be affected by factors outside of the task, thus the use of rating scales alone is often seen as unreliable (Cain, 2007).

Crew Status Survey (CSS).

The Crew Status Survey (CSS) is a seven-point unidimensional MWL assessment scale that was validated and verified through the testing of trained pilots and aircrew members (Ames & George, 1993). Operators are asked to assess their subjectively perceived level of MWL using two responses on the seven-point scale (1 = low workload; 7= high workload) as shown in Figure 5. The first response requires the operator to assess the maximum level of MWL that was experienced during the task by selecting one of the seven levels. The second response asks operators to identify the statement that best describes the average level of MWL experienced during the task. Addressing both the maximum and average values allows for more strategic analysis of the data compared to assuming operators are not biased towards maximum workload inducing events that took place throughout the task. This quick assessment offers experimenters the ability to assess subjective MWL during the task and/or following the task.

The CSS is a revision of the School of Aerospace Medicine (SAM) Form 202 MWL estimate that offered many advantages, such as ease of use, minimal training effort, and scale steps that were anchored in absolute terms. However, the SAM Form 202 lacked technical verification of the continuous underlying MWL dimension being tested. The development of the CSS involved revising the SAM Form 202 using a pair comparison and rank order estimation test. The results of Ames and George's (1993) efforts yielded a revised scale (adopted as the CSS) with verified ordinal steps and nearly equal psychological intervals between steps. Based on this set of assumptions the authors argue that the CSS can be considered an interval scale; thus, allowing more types of statistical analyses to be performed with the resulting data than an ordinal scale. The potential usefulness of the CSS in flight test applications is supported by the many advantages of the SAM Form 202, the absolute MWL score output (compared to relative workload score output), and the fact that the revisions were performed with trained pilots and aircrew members.

Level	Description
1	Nothing to do; No system demands.
2	Light Activity; minimal demands.
3	Moderate activity; easily managed considerable spare time.
4	Busy; Challenging but manageable; Adequate time available.
5	Very busy; Demanding to manage; Barely enough time.
6	Extremely Busy; Very difficult; Non-essential tasks postponed.
7	Overloaded; System unmanageable; Essential tasks undone; Unsafe

Figure 5. Anchor number and description used in the Crew Status Survey (CSS).

Instantaneous Self-Assessment of Workload (ISA).

The ISA was developed specifically to assess MWL of Air Traffic Controllers (ATC) as they performed their tasks. The ISA was designed to distract minimally from the primary ATC task (Brennan, 1992; Jordan, 1992). The ISA requires the individual to rate the level of MWL on a scale from one (very low workload) to five (very high workload) (scale shown in Figure 6) following the illumination of a red light emitting diode (LED) set to flash every two minutes. The ISA response is then entered on a specialized keyboard pad or as a verbal response to a tone. This measurement technique has been validated against the NASA-TLX and is thought to be one of the least distractive methods of gathering self-assessed MWL during a task (Casner & Gore, 2010).

Level	Workload	Spare Capacity	Description
5	Excessive	None	Behind on tasks; losing track of the full picture
4	High	Very Little	Non-essential tasks suffering. Could not work at this level very long.
3	Comfortable Pace	Some	All tasks well in hand. Busy but stimulating pace. Could keep going continuously at this level.
2	Relaxed	Ample	More than enough time for all tasks. Active on UAS monitoring task less than 50% of the time.
1	Under Utilized	Very Much	Nothing to do. Boring.

Figure 6. Instantaneous Self-Assessment Workload Scale (Kirwan et al., 1997).

Malvern Capacity Estimate (MACE).

The MACE was developed for use in evaluation of ATC, along with the ISA. The MACE is a self-report subjective scale that requires users to identify how much spare capacity they have available to complete secondary tasks while they are engaged in a primary task. Particularly, Goillau & Kelley (1996) utilized the MACE to determine how many more planes an ATC could direct at a given time. Following the completion of the task, participants were asked to use the scale to estimate on average how many planes they could handle per hour. They were also asked to rate how many planes they could handle at peaks of activity. The MACE was developed to "derive a direct estimate of capacity" (Goillau & Kelly, 1996, p. 7). See Figure 7.

-100%	-75%	-50%	-25%	0%	+25%	+50%	+75%	+100%
Less traffic than this run				Same traffic as this run				More traffic than this run

Figure 7. Malvern Capacity Estimate.

Modified Cooper-Harper (MCH).

The Modified Cooper-Harper (MCH) was originally developed to evaluate the handling abilities of aircraft (Cooper, 1957). It was later modified to evaluate handling abilities as well as serve as an indirect measure pilot MWL (Cooper & Harper, 1969; Casali & Wierwille, 1983). The scale consists of a decision tree where an individual must rate the handling of the aircraft system from one (excellent) to ten (major deficiencies). A rating of one would suggest low MWL and a rating of ten would suggest a high MWL. See Figure 8. The MCH has acted as a long-term aviation industry standard that has seen further modifications to more specific niches, such as unmanned vehicle systems (Cummings, Myers, & Scott, 2006),

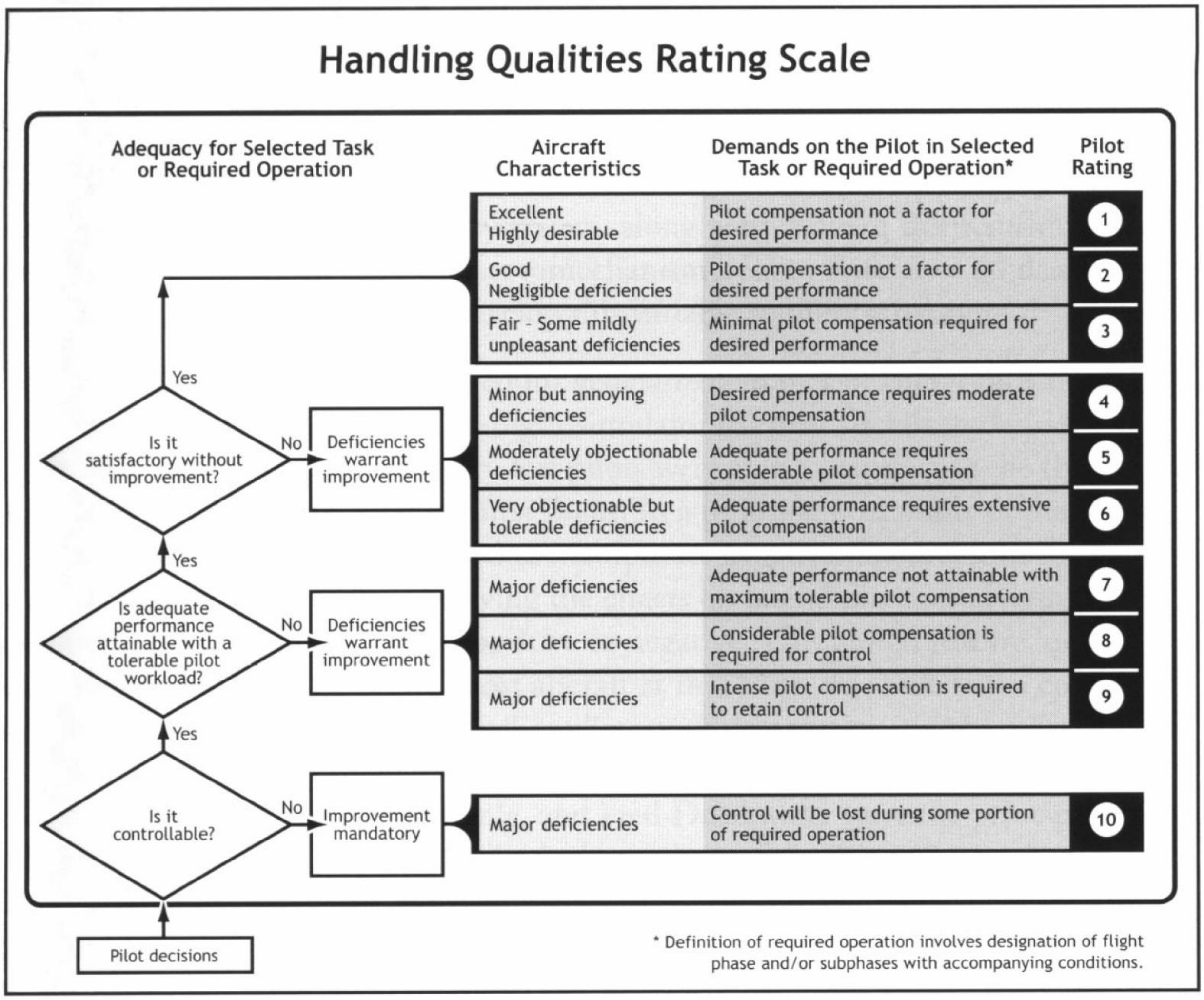

Figure 8. Modified Cooper-Harper (Cooper & Harper, 1969). Figure from Cummings, Myers, and Scott (2006).

Bedford Workload Rating Scale (BWRS).

The Bedford Workload Rating Scale (BWRS) is another modification of the original Cooper-Harper Aircraft-Handling Qualities rating scale, both of which were developed for use in the aviation domain (Roscoe, 1984). The BWRS requires that operators identify their levels of spare capacity while completing a task. To aid operator assessment, the BWRS juxtaposes a hierarchical decision tree with a ten-point scale with descriptors at each level, as seen in Figure 9 (Bachelder & Godfroy-Cooper, 2019). Answers to the questions anchored in the decision tree provide information about whether it was possible to compete the task, whether MWL was tolerable, and if the levels of MWL experienced were satisfactory. Beyond those decision points, segmented sections of the ten-point scale offer a finer resolution of MWL assessment in terms of spare capacity from one (“Workload insignificant”) to ten (“Task abandoned. Pilot unable to apply sufficient effort."). The primary advantage of the BWRS lies in the descriptions tied to each of the available ratings (Casner & Gore, 2010), as each rating provides an interpretation of the rating itself. However, a fundamental issue with the BWRS lies in asking operators to evaluate their "spare capacity," an ambiguous term with several interpretations (such as additional mental capacity, a free hand, extra time, etc.) that are not clearly defined in the scale (Casner & Gore, 2010).

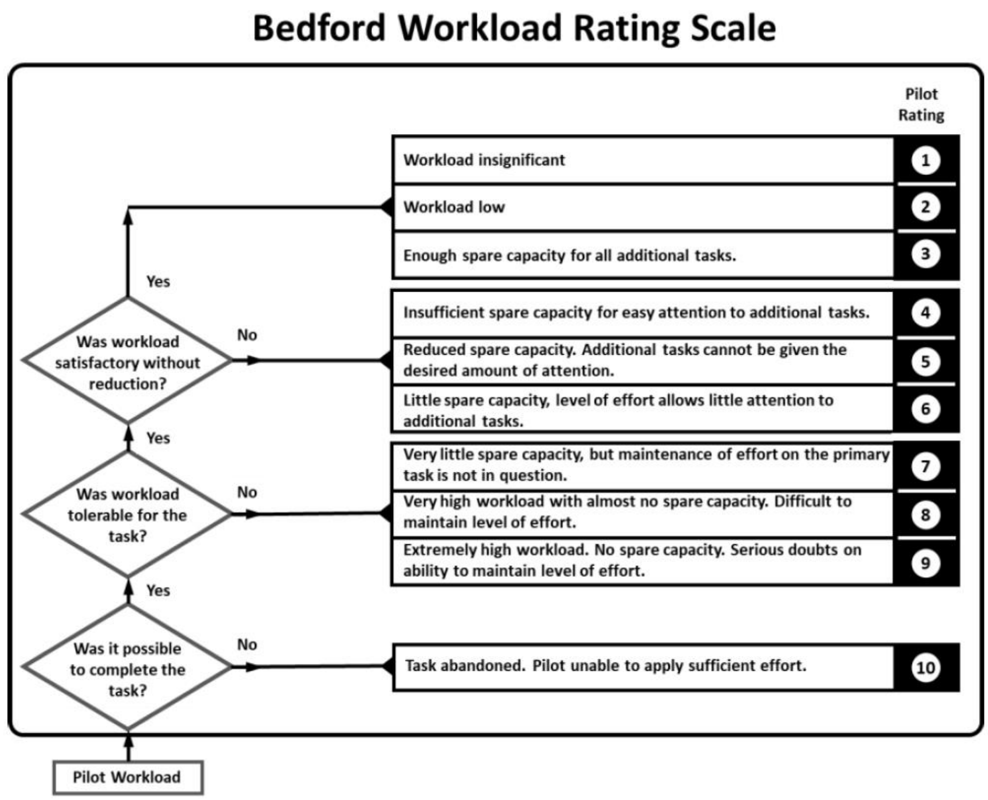

Figure 9. The Bedford Workload Rating Scale (figure from Bachelder & Godfroy-Cooper, 2019).

Rating Scale Mental Effort (RSME).

The Rating Scale Mental Effort (RSME) is a measure of perceived MWL which asks operators to mark their level of mental effort on a scale of 0 (no effort) to 150 (extreme effort). The RSME is a visual analogue scale using a 150 millimeters (mm) long line with anchors set at every 10 mm. The operators mark on the scaled line where they perceive their level of effort to be during the task. Some anchor points set along the line are labeled with descriptions of invested effort, such as 'some effort' around 37 mm and 'extreme effort' around 112 mm, among others. Final scoring of the RSME is the measurement, in mm, from the origin on the line to the point marked by the operator. Zijlstra (1993) developed and validated the RSME in a series of experiments demonstrating its sensitivity to task load and correlations with physiological measures of MWL. de Waard (1996) notes that the operator's reflection on "invested effort" is more attainable subjectively than reflections on more abstract concepts of MWL, such as reflections on "mental demand" as required by the NASA TLX. See Figure 10.

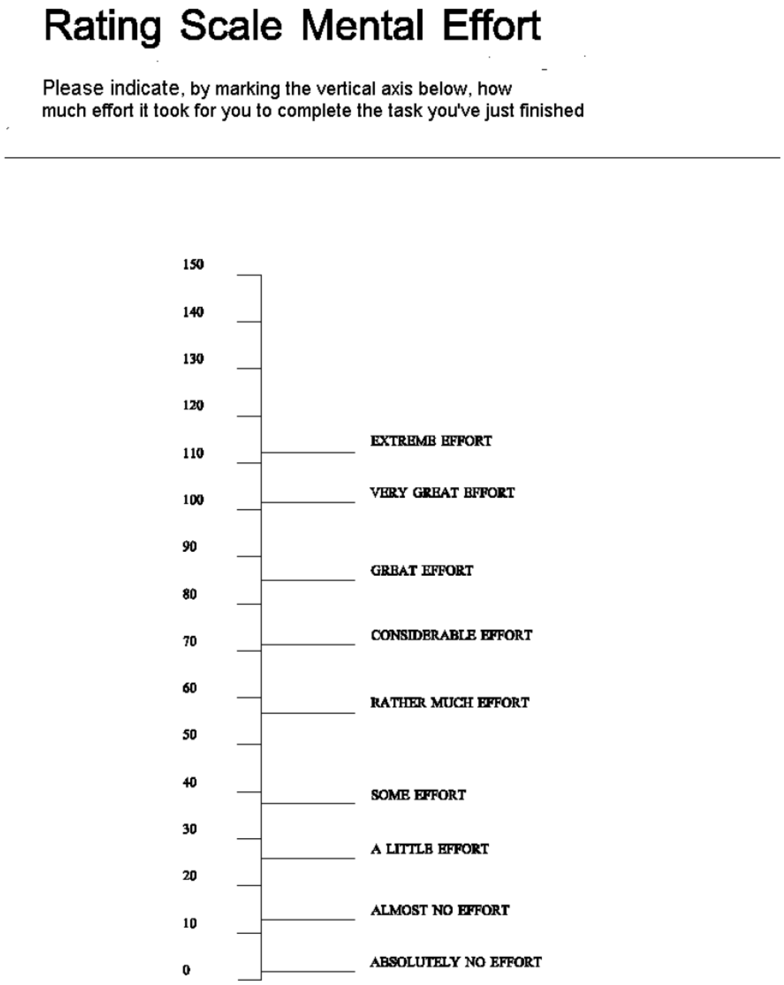

Figure 10. Rating Scale Mental Effort (RSME), developed by Zijlstra (1993), figure referenced from de Waard (1996). *Note.* Figure is not to scale.

NASA Task Load Index (NASA TLX).

The NASA TLX was developed as a multi-dimensional scale, rating MWL along six dimensions that can be combined into a unitary MWL score (Hart & Staveland, 1988). The six dimensions are mental demand, physical demand, temporal demand, effort, performance, and frustration. Each dimension has a complex description that is provided to the individual prior to using the scale (Figure 11). To complete the NASA TLX, the individual rates each dimension individually from low to high on a visual analogue scale. Each visual analogue scale consists of a horizontal line with twenty-one tick marks on which the individuals indicate perceived effort specific to each of the six dimensions. The individual dimensions are then weighted based on the individual's own determination of each dimension's importance. The reported values on each scale are tallied according to their respective weight to derive a unitary summary score. Alternatively, the summary score can be calculated without using the weighting technique, which provides a simpler method of calculating the summary score (Hart, 2006).

Rating Scale Definitions

Title	Endpoints	Descriptions
Mental Demand	Low/High	How much mental and perceptual activity was required (e.g., thinking deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exact or forgiving?
Physical Demand	Low/High	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low/High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	Low/High	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

Figure 11. NASA TLX Rating Scale Definitions (Hart & Staveland, 1988).

Subjective Workload Assessment Technique (SWAT).

The SWAT was developed to measure MWL along three domains, time load, mental effort load, and psychological stress load (Reid and Nygren, 1988). The SWAT User's Guide defines time load as the amount of pressure related to time that the operator felt when completing the task. “Mental Effort Load is the amount of attention or concentration that is required to perform a task; and Psychological Stress Load is the presence of confusion, frustration, and/or anxiety associated with task performance” (Reid, Potter, & Bressler, 1989, p. 11). The SWAT is typically completed in two steps. The first step is the “scale development phase,” in which users are required to sort 27 cards (each with a rank-ordered statement concerning each of the three domains tested) in order of importance so that the scale is tailored to their understanding of the terms associated with MWL (Reid et al., 1989). Following this step, the individual then completes a series of similar tasks with differing workload levels. Next, the individual rates the amount of effort required to complete each task using the tailored scale. The result is a personalized MWL score for each task. More information on the scaling procedures used in the SWAT can be found in Reid et al. (1989). The SWAT has been adapted into a computer program and is more complicated and time consuming than the other subjective workload scales discussed here, therefore it is unlikely to be useful in an operational setting. Despite this potential operational drawback, the SWAT provides MWL personalized to the individual operator.

Workload Profile (WP).

Subjective MWL assessment techniques are taken to be a global assessment of MWL; that is, they typically lack diagnosticity in terms of identifying which mental resources are being taxed by a task (O’Donnell & Eggemeier, 1986). However, Tsang and Velazquez (1996) challenged that idea by creating and validating a multidimensional instrument called the Workload Profile (WP). The WP was designed using Wickens’ (1984) MRT as its foundation. After completing the tasks, individuals rate the proportion of attentional resources expended during task engagement along the four MRT dimensions and sub-concept clustering (Figure 12). The rating scale for each dimension in the WP ranges from 0, the task placed no demand on a resource, and 1, the task required maximum attention of a specific resource (Tsang & Velazquez, 1996). As seen in Figure 11, tasks (m2, m2s1, m2s3, etc.) are listed as rows in any order for each operator’s assessment. While this direct assessment of workload and mental resources does provide insight into the specific resources being utilized by an operator, it does require the operator to learn how to differentiate the multi-dimensional nature of the attentional resources at their disposal. However, if provided a thorough explanation of each resource, operators can generally make this assessment without extensive training (Tsang & Velazquez, 1996; Rubio, Diaz, Martin, & Puente, 2004). Rubio et al. (2004) found this scale more sensitive and diagnostic than both the NASA TLX and SWAT. Overall, the WP presents a simple method for combining the diagnostic benefits of secondary task measures with the low intrusiveness of subjective workload assessment techniques.

Workload Dimensions								
	Stage of processing		Code of processing		Input		Output	
Task	Perceptual/ Central	Response	Spatial	Verbal	Visual	Auditory	Manual	Speech
m2								
m2s1								
m2s3								
m4								
m4s1								
m4s3								
s1								
s3								

Figure 12. Workload Profile rating sheet (Rubio et al., 2004).

Concluding note on subjective measures.

All of these measurement techniques are identified as subjective, which is to be expected since the definition of MWL used here (Van Acker et al., 2018) identifies MWL as a subjective experience of physiological processing states. Each of the subjective MWL scales described above depends on and reflects an individual's sensitivity to changes in their own physiology. Thus, these subjective scales rest on some form of introspection, which etymologically means to look inward. The implication is that some sensory processes analogous to vision, hearing, touch, olfaction, etc. are involved with introspection. All these MWL scales may rest on a common, shared, or similar form of introspection; in which case, the major differences among the scales derive from the different methods they use to structure the process of introspection. Thus, for all their differences, these MWL scales use the same basic strategy to assess MWL, introspection, a point worth noting since introspection continues to play an immensely important role in philosophy, quantitative experimental psychology, contemporary cognitive sciences, and, of course, daily life.

Introspection can be defined as a process by which we observe our current or recent mental processes, which is completely consistent with the goal of measuring MWL. Introspection has a number of characteristics that have been identified and discussed in the context of other disciplines but which have not received much, if any, attention in the MWL literature. For example, while it is obvious that introspection addresses mental events, events that occur in the conscious mind with its associated brain states, unless the events can be called into consciousness, they are not amenable to introspection, and so do not figure into these subjective measures of MWL. Despite this, the border between introspective and non-introspective events is certainly soft, probably as soft as the border between the conscious and the unconscious mind. This must be the case since so many factors determine what can be made conscious or brought to mind at any one time. This malleability of recall seems to underlie the variety of subjective MWL scales since each scale structures the process of introspection differently. Regardless of their differences, it seems important to emphasize that all of them measure something that occurs strictly in the mind as distinct from the physical world. In other words, these MWL scales reflect changing conditions in the physical world only to the extent that such physical changes impact the mind in some way that introspection can access (Huemer, 2019).

Another important point is that any statement based on introspection describes the individual making that statement, the person reporting their introspection. It reflects the individual's immediate direct access to the contents of their mind. It is private knowledge and pertains only to that individual's mind or brain states at that moment. While this may be obvious, it has consequences that are subtle and important. For example, the individual can make whole classes of statements that are infallible and cannot be doubted or corrected unless deceit, deception, or some form of abnormality is discovered. Thus, only the individual can know without doubt what they are thinking, and assertions to the contrary are absurd. Introspection establishes special relations among the self, truth, falsehoods, and classes of knowledge (Schwitzgebel, 2019).

The different subjective MWL scales implicitly acknowledge the importance time has for introspection. Judgements about one's mental state or brain status are dramatically altered by time. One can describe the status of one's mind at the moment or immediately in the past or possibly even in the imminent future. Over time, one's introspection slips from current status to memory and recall. It is possible that the minimum interval between current status and the increasing role of recall is related to the psychological present moment. Thus, for example, one may recall that one was thinking of something else just a moment ago so that present introspection queries the recall of a past time. It is the recognition of the importance of this distinction that underlies the brevity of the ISA.

Additional aspects of the process of introspection include the assumption that one's mental state is immediately accessible. Even if that state is a jumble of confusion that needs sorting out, still the state is readily at hand, part of the immediate present, more like a phone call than a letter. In addition, introspection requires a change from one mental state to another. For example, with the ISA, the mind quickly shifts from the ATC task to the introspection necessary for self-assessment and the keypad response. The steps that are necessary for the ISA response all require effort, which is another characteristic of introspection; it requires effort and is to some extent necessarily intrusive.

By the mid-nineteenth century, introspection had developed from philosophy to being an essential component of experimental psychology and physiology (Boring, 1950; 1953). As a discipline, introspection was most successful dealing with sensory psychophysics and the refinement of what have become standard psychophysical procedures, some of which may be applicable to the study of MWL. The major shortcoming or reservation about introspection serving as a research tool has been that the very process of introspection itself interferes with the mental act or brain state that it is deployed to assess. August Comte raised these objections at the very beginning of the development of introspection as a research tool. "But as for observing in the same way *intellectual* phenomena at the time of their actual presence, that is a manifest impossibility. The thinker cannot divide himself into two, of whom one reasons whilst the other observes him reason. The organ observed and the organ observing being, in this case, identical, how could observation take place? This pretended psychological method is then radically null and void" (Comte, 1830 – William James translation 1890, quoted in Schwitzgebel, 2019).

Granting the justice of Comte's objection, which many others have voiced in different ways over the years, introspection has developed into an immensely powerful research tool in numerous areas, including the sensory sciences, which may serve as a model for the eventual development of an MWL psychophysics (Trnka & Smelik, 2020).

Performance Measures

One of the most obvious side effects of high levels of MWL is the reduction of performance in a task, usually through metrics concerning accuracy and speed. As such, an operator's natural performance on a task can serve as an indicator of experienced MWL or spare capacity. Two types of performance measures are available: primary task measures and secondary task measures (O'Donnell & Eggemeier, 1986; Cain, 2007). Each of these measures has advantages and disadvantages in terms of sensitivity, diagonosticity, and task interference (see workload assessment criteria set forth by O'Donnell & Eggemeier, 1986), but both lend themselves to the assessment of MWL in the aviation domain.

Primary task performance measures.

The most intuitive way to measure the effects of MWL on performance is to assess the metrics that define performance in the task itself. That is, to obtain a primary task performance measure. The details of these metrics are typically defined by a specific task, but general metrics such as speed of performance (e.g., reaction time or latency) and number of errors have been shown to be sensitive to workload manipulations (O'Donnell & Eggemeier, 1986). Referring back to the performance-workload curve detailed by de Waard (1996), task performance begins to fall only as task demands push an operator from Region A3 to Region B. As such, while primary task performance metrics may be a good reflection of an operator's efforts, primary task performance measures lack sensitivity beyond the ability to discriminate overload from non-overload conditions (O'Donnell & Eggemeier, 1986). More specifically, with increasing task demands, changes in performance only occur in Region B, meaning primary task performance measures are only sensitive in this region. As this would suggest, primary task performance measures have no sensitivity to increasing workload in Regions A1, A2, and A3, where performance is steady even with changing levels of workload (in Regions A1 and A3), as well as Region C where an operator's capacity is continuously overloaded and performance is low.

Along with their relative insensitivity, primary task performance measures lack diagnosticity, which refers to the assessment technique's ability to diagnose which set of resources, as defined by MRT, are affected by the task (O'Donnell & Eggemeier, 1986). Observing changes in reaction time or error rate in a primary task assessment will not provide a researcher insight into what resources are being utilized by the task. Instead, primary task performance measures provide a 'global' assessment of workload. As such, primary task performance measures provide a general picture of the workload elicited by a task and can guide decisions on the determination of overall system effectiveness.

The sensitivity and diagnosticity limitations of primary task performance measures paint a relatively unexciting picture as far as MWL assessment is concerned, but there are benefits to their use. First, primary task performance metrics provide the most direct access to the moment when workload has exceeded a user's capacity (i.e., entered Region B or crossed the redline). As such, while it would be ideal to predict this approach to Region B ahead of time (as may be possible with other techniques), primary task performance measures provide a solid back up dataset for adaptive systems to initiate assistance. Second, while providing a demarcation between overload and non-overload conditions, primary task measures are in no way intrusive to the operator's performance (O'Donnell & Eggemeier, 1986). Lastly, primary task metrics are data that are collected in most laboratory tasks regardless of whether they are being used to assess workload. Today, implementation requirements for these simple metrics are both efficient and affordable, even in operational environments. Overall, primary task performance measures provide researchers an MWL assessment technique that is directly tied to the capacity of the resources demanded by a task, but they lack the diagnosticity afforded by other performance measures.

Secondary task performance measures.

Examining performance changes when an operator is engaged in two tasks simultaneously provides another method of MWL assessment. This approach is referred to as a secondary task performance measure. In this multitasking arrangement, the main task of interest is defined as the primary task, while the additional task being performed is referred to as the secondary task. The concurrent performance of both the primary and secondary tasks provides an estimate of the *primary task* MWL experienced by the operator. This estimate is derived from performance on the secondary task, which serves as a measure of the spare capacity an operator has while engaged in the primary task (O'Donnell & Eggemeier, 1986). Performance on the secondary task begins to falter as the demands of the primary task begin to rise, allowing the decreasing performance on the secondary task to serve as an estimate of reserve capacity. Indeed, a secondary task can be designed in a way that forces it to weigh on or avoid specific resources (as defined by MRT) to assess their individual capacities.

Using secondary task performance measures requires preparation steps to administer the dual task experimental configuration. First, it requires that both the primary task and secondary task have their performance assessed independently to serve as a baseline (O'Donnell & Eggemeier, 1986). This baseline allows for the comparison of performance between the tasks in isolation and the concurrent performance of the tasks. Second, either the primary or secondary task performance needs to be emphasized to operators. That is, the operators need to be instructed to focus on error-free performance and/or consistent performance on one of the tasks even at the detriment of the other task. As such, there are two different secondary task paradigms, a **subsidiary task paradigm** requires ideal performance on the primary task, allowing performance on the secondary task to falter; while a **loading task paradigm** requires ideal performance on the secondary task, allowing performance on the primary task to falter. Third, safety considerations must be addressed, given the nature of the tasks. For example, using a primary task of driving a vehicle in busy traffic may dictate a subsidiary task paradigm to ensure primary task performance is maintained, as well as carefully designing a secondary task to not weigh on already stressed resources (such as requiring visual identification or manual responses).

While there are traditionally designed secondary tasks in the literature, it would be more efficient to describe how one could go about designing a secondary task using the framework of MRT (Wickens, 2008). Using task analysis, both the primary and secondary tasks can be defined by a **task demand vector** that highlights which task, and how strongly each task weighs on each of the resources defined by MRT. Then, depending on the application, sources of interference between the tasks can shift accordingly throughout the design process. For example, a researcher may wish to avoid overloading an operator's visual and motor responses, as is in the case of a primary task of driving, while weighing heavily on their cognitive-spatial resources to observe changes in eye scan patterns with the increased workload. As such, an ideal secondary task would avoid visual resources by being administered via the auditory modality. Likewise, to avoid interfering with the manual processing of vehicle controls, an ideal secondary task would avoid manual responses in favor of verbal responses. Lastly, to invoke the interference and higher levels of workload for the cognitive-spatial resource, the task should require mental processing of visual stimuli. Rather than picking an arbitrary task, this secondary task should be built with mental resources in mind, affording the secondary task performance measure a high level of diagnosticity (to make the overloaded resource identifiable and the stage of the task that caused the overload). Additionally, using a battery of secondary tasks, as there are no universal secondary tasks that work for all primary tasks, a researcher can begin to diagnose which resources are being overloaded through iterative testing with subsidiary tasks with different task demand vectors.

The subsidiary secondary task paradigm is the more intuitive and more frequently used approach in the literature. In the subsidiary task paradigm the operator is instructed to maintain primary task performance at the expense of secondary task performance. As such, when the primary task demands increase, there are less resources available for the secondary task performance, and performance of the secondary task begins to decrease. This provides a notion of how much additional work can be performed by an operator at varying levels of task demand set forth by the primary task. Thinking in terms of the region model, the subsidiary task paradigm forces a shift in total workload from Region A to Region B, causing performance decrements to occur in the secondary task but not in the primary task. This concept is depicted in Figure 13 (O'Donnell & Eggemeier, 1986 adapted from Brown, 1964) with different levels of primary task demand changing the available reserve capacity for the secondary task. Compared to a single primary task performance measure, the subsidiary task paradigm prevents an operator from compensating for decreased performance on the primary task with increased performance on the secondary task, permitting for a more sensitive measure of capacity expenditure.

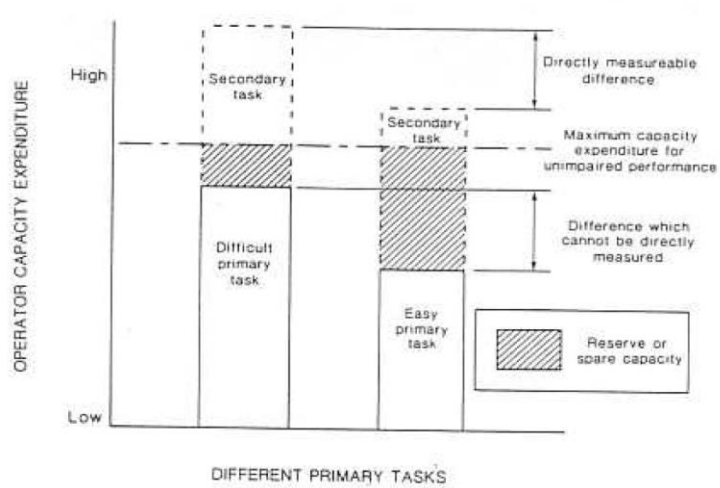

Figure 13. Representation of subsidiary task paradigm (O'Donnell & Eggemeier, 1986; adapted from Brown, 1964).

In contrast to the subsidiary task paradigm, in the loading task paradigm, performance on the secondary task is to be maintained at an error-free performance level while the primary task performance is allowed to falter. This does not mean that the labels of the primary and secondary tasks change because the secondary task performance is emphasized. The task of interest is still the task defined as the primary task. Like the subsidiary task paradigm, the loading task paradigm pushes the operator from Region A to Region B, except, instead of the secondary task falling in performance, the primary task performance is negatively affected. As such, the loading task paradigm allows for the simulation of a more complex operational environment. Increasing the base level of load on a specific resource with a secondary loading task may be more indicative of the additional task demands presented in an operational environment. The loading task paradigm can be used in the evaluation of multiple types of secondary tasks, as one secondary task may push a primary task into performance degradation, while the other does not. Therefore, the task that does not induce primary task performance degradation would be an ideal choice.

Concluding note on performance measures.

Performance measures provide a simple method to measure MWL by providing tangible evidence of changes to resource capacity based on performance outcomes on either the primary or secondary task, depending on the approach used. Using primary task measures, a direct measurement of operator performance is obtained and used to make decisions about how much an operator can handle during a specific task. Additionally, the use of performance measures to estimate MWL allows for an easy-to-implement assessment of MWL that maintains ecological validity. For example, using an embedded task such as radio communications as a secondary task during a flight simulation allows for measurement of MWL while utilizing a task that would be more natural during actual flight.

Psychophysiological Measures

Referencing again the definition of MWL: "*Mental workload is a subjectively experienced physiological processing state, revealing the interplay between one's limited and multidimensional cognitive resources and the cognitive work demands being exposed to.*" MWL produces characteristic physiological responses, which are easily identified using various metrics and modalities, which have been found to correlate with subjective measurements (e.g., workload questionnaires such as those mentioned above) (Lohani et al., 2019). With few exceptions, the modality of these measures often gravitate around the physiological measurement of the tone of the autonomic nervous system (ANS), and are thus not under the operators' volitional control. Measures such as blood pressure, heart rate, heart rate variability, pupil diameter, electrodermal activity, and central and peripheral oxygenation are all heavily dependent on an individual's level of arousal, which is directly reflected in the "tone" of the ANS (i.e., the balance between the sympathetic and parasympathetic nervous systems). While these measures are sensitive to workload, they are also heavily influenced by fitness, stress, fear, and fatigue. Thus, there is a great amount of inter-individual variability, and the assessment of MWL requires an accurate accounting of those influences. However, there are other physiological metrics that rely on the measurement of more volitional activity. Eye movement analyses and blinking activity patterns are good examples of this. However, if the operator is intrinsically monitoring this behavior, the individual influence on these measures can be increased. Regardless, given the appropriate consideration when interpreting these data, all of these measures can provide invaluable insight into the psychophysiological state of the operator.

Oculometrics.

Researchers have been using the eyes as "windows to the soul" for a few hundred years and have developed several means to recording oculometric activity. Eye movements that are believed to be related to workload include blink rate and duration, pupil diameter, and saccadic eye movement. Ahlstrom and Friedman-Berg (2006) cite that when workload is high, the number of blinks and the duration of blinks decrease, pupil diameter increases, and saccadic eye movement activity decreases. Ahlstrom also reports that eye movement has become a popularized measure of workload since it is non-invasive, objective, and because the study tasks need not be interrupted to measure current or incremental workload changes. There are multiple techniques used to capture oculometric data for workload assessment.

Electrooculography (EOG) involves recording changes in the corneo-retinal standing potential between the front and back of the eye. When electrodes are placed around the eye (typically above and below) these changes can be mapped to the position of the eye in the orbit and provide a coarse designation of where a person is looking and the eye movements involved. This method is far less accurate and more invasive than modern eye tracking solutions. However, EOG collected using electrodes above and below the eye are very effective at capturing blinking activity in a manner that is more reliable and robust than video-based tracking systems, so it is still widely employed.

Advances in digital camera technology have led to a host of entirely non-invasive eye tracking solutions that pair well with MWL studies. An infrared camera focused on the eyes creates a digital recording that is analyzed in real-time to identify the pupils and compares that position to another reference on the face (e.g., the nose) or the eye, itself (e.g., corneal reflection). This method has the distinct advantage of allowing accurate estimates on the size of the pupil in real-time (often at sample rates of hundreds of Hertz [Hz]), as well as a gaze-position-tracking-accuracy of less than a visual degree. Further, the cameras can be mounted to the platform or the operator, providing flexibility to collect in operational and near-operational environments. As easy as it is to collect, the quality of the data is heavily dependent on the environment in which the data is collected and the quality of the calibration.

Pupillometry.

It has been observed that changes in MWL also induce small, rapid, changes in pupil diameter. Aura, Temme, & St. Onge (2020) found that increasing task difficulty on a visual search task resulted in significant increases in pupil diameter during the execution of that task. Further, using a simplified approach, Aura et al. (2020) were able to extract a sinusoidally modulating luminance signal from the pupil diameter and still detect a significant increase in pupil diameter during high MWL variations of a continuous recall task (N-Back). However, the level of filtering used in Aura et al. is far from that which will be necessary to measure these variables in the constantly changing luminance of the operational environment. Fortunately, there are promising analytics in development that may effectively address this issue.

A dilation or constriction of the pupil evoked by changes in environmental luminance levels can complicate capturing and interpreting MWL changes. The Index of Cognitive Activity

(ICA) was developed by Marshall (2000; 2002; 2007) to separate pupillary activity caused by the pupil's response to light from that evoked by changes in MWL by examining sudden discontinuities of pupillary activity. The ICA has been demonstrated to be an effective metric in operational environments such as the surgical suite (Nguyen, Chen, Marshall, Ghodoussipour, Chen, Gill, & Hung, 2020; Richstone, Schwartz, Seideman, Cadeddu, Marshall, & Kavoussi, 2010) and in driving tasks (Dlugosch, Conti, & Benglar, 2013; Platten, 2012).

Duchowski et al. (2018) referenced the proprietary nature of the ICA and, developed a similar metric to the ICA called the Index of Pupillary Activity (IPA). Additionally, the authors provide the IPA as open-source Python code. Both use a wavelet analysis approach, but the IPA differs in the choice of wavelet thresholding approach and the use of the modulus maxima. Application of the IPA revealed its sensitivity to task difficulty, corroborated through performance and subjective measures, and its independence from working memory capacity.

Electrocardiography (ECG) and cardiovascular measures.

ECG is used to collect the small electrical changes that result from the depolarization and repolarization of the cardiac muscles. From the resulting electrocardiogram, psychophysiological measures including heart rate and heart rate variability can be derived. Data collection using ECG requires the application of electrodes on the operator and an associated amplifier and data recording mechanism. There are a number of electrode placement option available that have proven to be effective. One of the most commonly used electrode placement techniques is the Lead II configuration (Figure 14).

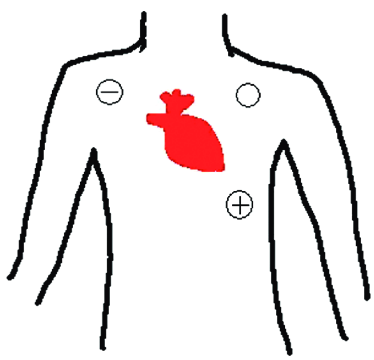

Figure 14. Lead II configuration of ECG electrode placement (Christensen & Wright, 2014). Depicted is positive, negative, and reference electrode placement.

Heart rate refers to the number of heart beats in a specified amount of time and is commonly measured as the number of heart beats that occur in one minute, or beats per minute (bpm) (Hughes, Hancock, Marlow, Stowers, & Salas, 2019). Heart rate has been found to be a sensitive measure of changes in MWL. By way of specific example, Lahtinen, Koskelo, Laitinen, and Leino (2007) found that heart rate was particularly sensitive to workload changes in a combat flight test of experienced and inexperienced pilots. Further, results indicated that not only was there a significant difference in heart rate between all phases of flight compared to baseline, but that the greatest change in heart rate occurred when the pilots were required to complete a tactical maneuver, which was designed to be the most cognitively tasking phase of flight.

Heart rate variability refers to the analysis of changes in the inter-beat-interval, or R-R interval (Figure 15), over time. Heart rate variability can be examined along two domains of measurement, the frequency-domain and the time-domain. Successful measures common in the frequency domain include low frequency/high frequency (LF/HF) ratios and specific frequency bands (high, mid, and low frequency bands) (Tao et al., 2019). An increase in LF/HF ratio indicates higher MWL, while trends of decreasing frequency in each of the high, mid, and low frequency bands are correlated with higher levels of MWL. Among the time-domain measures, inter-beat-interval is the most commonly reported heart rate variability measure, with the inter-beat-interval decreasing with increasing MWL (Tao et al., 2019). A meta-analysis conducted by Hughes et al. (2019) identified that heart rate variability is most effective in studies that have tasks that take longer periods, and tasks in which workload levels are manipulated continuously, especially in contrast to measures like blood pressure, which showed little to no effect from MWL manipulations.

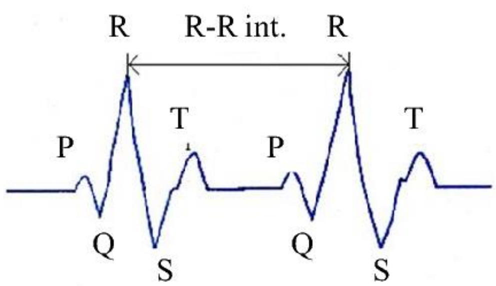

Figure 15. R-R Interval of PQRS complex of normal healthy heart beats. (Murai, Hayashi, Maenaka, & Hyguchi, 2015).

Respiration.

Respiration rate is a widely used measure of MWL (Tao et al., 2019). Respiration rate is often measured using a transducer affixed to a belt that is placed around the torso of an individual where inspiration and exhalation movement is most obvious (Figure 16). Cain (2007) explains that respiration rate often increases when the individual experiences stress or higher than normal workload. In Tao et al.'s (2019) review, respiration rate proved a sensitive workload metric across multiple types of tasks, from air traffic control and simulated aviation tasks to mental arithmetic and continuous memory tasks. Cain (2007) also notes that while respiration is a valuable piece of information in determining MWL, it would not be appropriate to infer workload based on respiration data alone.

Figure 16. Individual outfitted with BioNomadix respiration belt (biopac.com).

Catecholamines and hormonal responses.

The psychophysiological responses to MWL and stress due to an anticipation of increased workload, either physical or mental, cause a release in catecholamines (e.g., adrenaline [aka epinephrine] and noradrenaline) and cortisol into the bloodstream (Leino, Leppaluoto, Ruokonen, & Kuronen, 1999). These chemicals are neurotransmitters, which influence an individual's heart rate, heart rate variability, and respiration rate, as well as other physiological stress responses. All of these responses are measureable functions of the nervous system. Testing for increases in the presence of these hormonal chemicals in individuals is often carried out through salivary samples or through blood draws. The samples provided then must be analyzed in a laboratory to determine the probable MWL level of the individual and are therefore less feasible in an operational environment (Cain, 2007).

Electrodermal activity (EDA).

EDA refers to the electrical conductivity of the skin due to the human body's autonomic response of sweating. EDA is often referred to in the literature in a number of ways, for instance, Mehler, Reimer, Coughlin, and Dusek (2009) refer to the use of EDA as skin conductance level whereas Nourbakhsh, Wang, Chen, and Calvo (2012) use the term "galvanic skin response". It is believed that when workload increases, the EDA increases. EDA is often measured by placing special electrodes on the hand or finger to measure the electrical conductivity (Figure 17). This conductivity is generally caused by the opening of the pores of sweat glands. There are multiple problems that have been identified with using EDA alone as a measurement of MWL. For instance, Shaffer, Combatalade, Peper, & Meehan (2016) explain that environmental conditions (e.g., humidity) can greatly affect measurements of EDA and cause misleading measurements.

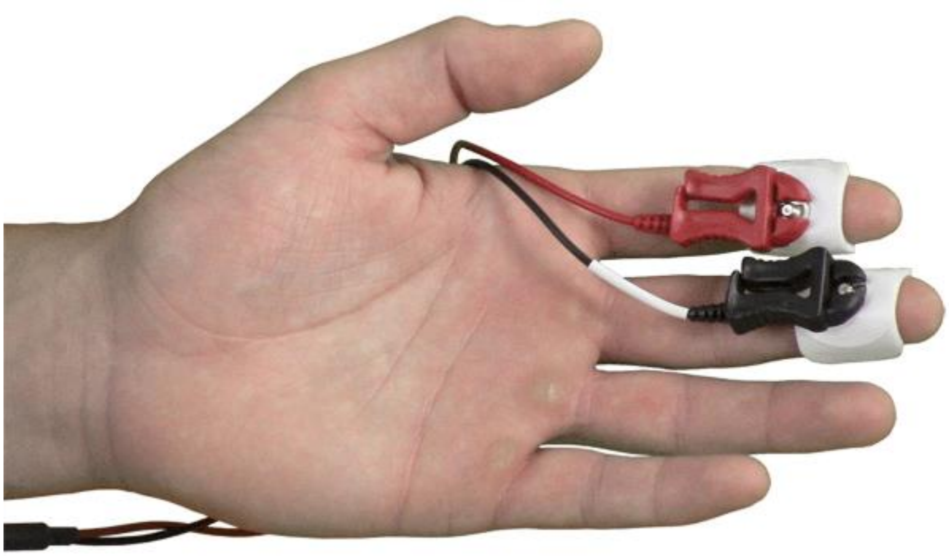

Figure 17. EDA measurement hook-up (biopac.com).

Electroencephalography (EEG).

EEG refers to the amplification of electrical signals put out by the pyramidal cells of the brain (Lohani, Payne, & Strayer, 2019). These electrical signals are transmitted through electrodes placed on the scalp of a subject, which is connected to an amplifier (Figure 18). Data is often analyzed by looking at the relative power in several frequency bands, including Delta (less than 4 Hz), Theta (4 - 8 Hz), Alpha (8 - 14 Hz), and Beta (14 - 30 Hz). Lohani et al. (2019) cites that "…mental workload increases theta power and reduces alpha power activity (Mun, Whang, Park, & Park, 2017), whereas fatigue increases alpha power (Käthner, Wriessnegger, Müller-Putz, Kübler, & Halder, 2014)" (p. 3). Cain (2007), in his extensive review of MWL measurement techniques, quotes Wickens (1992) that EEG is not the measurement of how hard someone is working cognitively on a given task or during a period of time, rather EEG is able to measure the capacity that remains within the individuals' cognitive ability. Cain (2007) further notes that EEG data can be affected by environmental factors during data collection and requires complicated analysis and interpretation practices.

Figure 18. B-Alert EEG data collection system (advancedbrainmonitoring.com).

Electromyography (EMG).

EMG refers to the measurement of the preliminary activity of a muscle or muscle group prior to movement in response to some prompt and is generally used in MWL studies as a measure of response time (D'Addario, Donmez, & Ising, 2014). EMG measurement is often completed by placing special electrodes on the muscle or muscle groups of interest. D'Addario et al. (2014) explains that aside from response time of preliminary muscle movement as a measure, researchers often focus on electromechanical delay, which is the average time between the preliminary activity of the muscle or muscle group and the actual response or movement of the muscle to respond to a particular prompt. Gaetan et al. (2015) used EMG measurement of the right finger of helicopter pilots. Pilots were first assessed on their flight experience and then were tasked with completing missions of differing MWL levels in a helicopter simulator. Their physiological stress responses were measured using the muscle movement as well as EDA of their right finger. It was found that EMG in conjunction with personalized physiological and psychological profiles were effective in predicting the individual pilots' MWL.

Functional near-infrared spectroscopy (fNIRS).

Functional Near-Infrared Spectroscopy (fNIRS) is a diffuse optical method for measuring oxygen-dependent metabolism processes occurring in localized regions of the brain (von Lühmann, Herff, Heger, & Schultz, 2015). Like functional magnetic resonance imaging (fMRI), fNIRS provides a localized functional response of how oxygen supply in the blood changes for a specific brain region. However, fNIRS does not require an operator to be confined in a small space, like fMRI, and can potentially be used within an operational setting (Herff, Heger, Fortmann, Hennrich, Putze, & Schultz, 2014). To provide this measurement, a Near-Infrared (NIR) light is emitted into a specific region of the head. The light is then scattered and absorbed by the various tissues (skin, muscle, bone, etc.) it is transmitted through, attenuating its overall magnitude. The light is then propagated back towards the surface where it can be detected and measured by a NIR-sensitive photodetector (von Lühmann et al., 2015). As such, fNIRS offers researchers the ability to observe resource consumption (in this case, oxygen) by a specific region of the brain pertaining to a specific function, which, in turn, suggests the fNIRS approach can be favorable in terms of diagnosticity.

Research in MWL has successfully utilized fNIRS technology, especially during the last ten years, as it has become more reliable and affordable. Discrimination between different levels of MWL induced by working memory tasks of varying complexity has been performed with high degree of accuracy. Especially, when using fNIRS to measure hemodynamic activity in the prefrontal cortex (Herff et al., 2014; Aghajani, Garbey, & Omurtag, 2017; Sassaroli et al., 2008). Additionally, fNIRS-related MWL level discrimination has seen success in workload assessment with more complex tasks, such as remotely operated vehicles (Durantin, Gagnon, Tremblay, & Dehais, 2014), air traffic control (Ayaz, Shewokis, Bunce, Izzetoglu, Willems, & Onoral, 2012), and driving (Foy & Chapman, 2018). However, there are limitations to consider when using fNIRS, especially in operational environments, such as head and face movement induced artifacts, ambient light and sound, muscle movements, and the slow timing of the hemodynamic response (Girouard et al., 2010). Luckily, much of this noise can be removed with proper data filtering, an area of continued research (von Lühmann et al., 2015).

Concluding note on physiological metrics.

The measurement of physiological responses to workload in aircrew during flight has traditionally been difficult. The flight environment of rotary-wing aircraft in particular contributes to this difficulty, namely due to the high vibration patterns. However, increases in physiological technology and processing techniques have made it possible record usable signals. Recent studies at USAARL using a UH-60 full-motion simulator have assessed all of the aforementioned measures, except fNIRS and EMG. The EEG, EOG, and ECG have provided the cleanest and most usable data, while the EEG and ECG data have most frequently shown associations with performance and workload ratings (Feltman, Kelley, Bernhardt, Britt, & Mathews, 2019; Feltman, Bernhardt, & Kelley, 2020). Several eye tracking systems have been used in simulator studies as well (e.g., Feltman et al., 2018; Feltman et al., 2020; Hayes, Aura, & Feltman, 2020), but have resulted in difficulties in interpreting and using the data. For example, changes in the visibility in the cockpit (e.g., night conditions, snowing) impacts the light available which causes an automatic change in pupil dilation. Thus, use of pupillometry is difficult under varying lighting conditions common in the flight environment. Ongoing work at USAARL continues to examine the diagnositicity of physiological metrics in identifying the source of workload.

Associations, Insensitivities, and Dissociations of Workload Measures

Using performance, physiological, and subjective measures, researchers can characterize an operator's MWL in terms of direct performance decrements, objectively measured reflexive indicators, and subjectively perceived levels of mental effort. As is often the recommendation, multiple measurement types should be combined to accurately assess MWL. However, with this combination, the potential for inconsistencies across measurement types appears to be a common problem in the literature (Hancock & Matthews, 2019). In general terms, the three categories of workload measures can have one of three responses: increasing, decreasing, or insensitive (i.e., stable) responses. When all of the measures respond in the same manner, they are said to experience **association** with each other (i.e., all measures show increasing values when under higher levels of MWL). Conversely, when measures disagree they demonstrate **dissociation**. An extreme example of dissociation (e.g., a double dissociation) would be when a performance measure elicits no change in performance (insensitivity), while the measured physiological activity decreases relative to baseline, and subjective responses show heightened levels of workload. This situation would make interpretation of the collected MWL measures very difficult.

In an effort to conceptualize this problem, Hancock and Matthews (2019) demonstrated how the three primary measurement methods could relate to each other in Figure 19. The combination of the three matrices yields a cube that represents 27 distinct outcomes that can occur with combinations of subjective, performance, and physiological measures. The regions marked A+ and A- in Figure 19 indicate outcomes where all three measures share the same trend of either increasing or decreasing, respectively. These outcomes are termed double associations and are ideal for interpretation, as all three measures paint the same picture of the operator's experienced MWL. Indeed, single associations can occur when two measures agree with each other while the third measure is either insensitive or dissociates with the pair. Recalling the region model of workload and performance, if an operator is not pushed into Region B where performance begins to falter, primary task performance measures will show insensitivity. However, even though performance remains stable throughout Regions A1-3, subjective and physiological measures may likely show an association as the operator approaches Region B through means of task-related effort. As such, in de Waard's modified region model (1996), Region A3 presents this outcome (insensitive primary task measures potentially associating physiological and subjective measures) as a common occurrence. Associations, insensitivities, and dissociations may function as a feature that could identify the region in which an operator is functioning. Using this conceptual model, the relationship between measures can be plotted for specific measures, tasks, and operators.

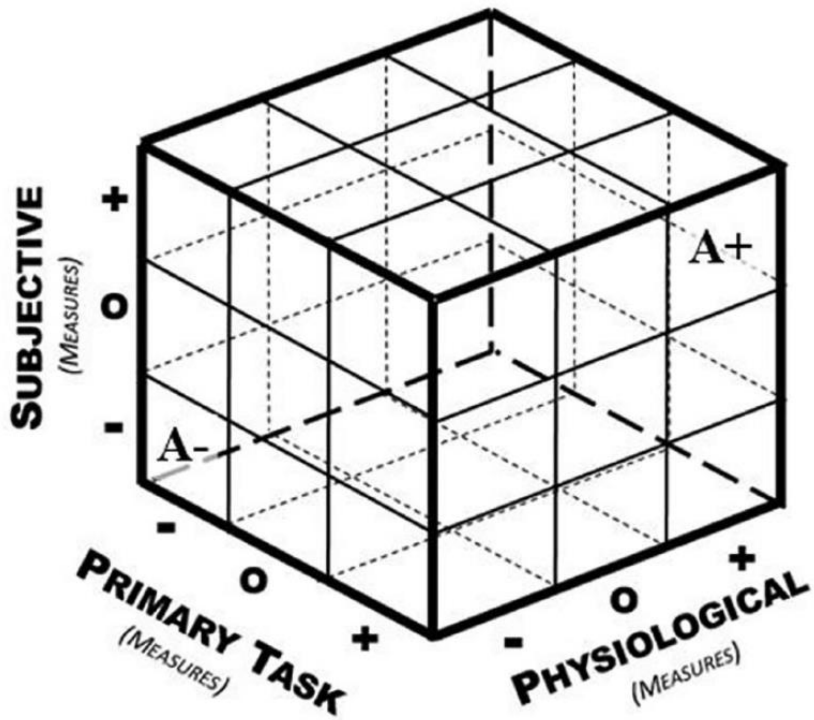

Figure 19. Hancock and Matthews' (2019) depiction of the associations, insensitivities, and dissociations of MWL measurement techniques (performance, physiological, and subjective measures). Twenty-seven distinct outcomes are possible. Outcomes marked by A+ and A- indicate areas of double association (i.e., where all three measures change with a similar trend of increasing or decreasing).

Conclusions

In a study using physiological, psychological, and experience profiles of helicopter pilots to predict workload, Gaetan et al. (2015) determined that individualizing workload baseline profiles is an important factor in the future of real-time workload monitoring, but that the field is promising and could be impactful in determining the fitness of pilots in real-time in the near future. In order to support the FVL mission through enhanced operations, it will be critical to develop physiological sensors that will directly and constantly measure the cognitive state of pilots and crew. It will become paramount that leadership monitor individual physiological data undertaking missions in FVL aircraft. The conceptual review provided here serves as an exploration of the available and widely used MWL measurement techniques for consideration in future research.

References

Aghajani, H., Garbey, M., & Omurtag, A. (2017). Measuring mental workload with EEG+ fNIRS. *Frontiers in Human Neuroscience*, *11*, 359.

Ahlstrom, U., & Friedman-Berg, F. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics, 36*, 623-636.

Ames, L., & George, E. (1993). *Revision and verification of a seven-point workload estimate scale* (Report No. AFFTC-TIM-93-01). Edwards Air Force Base, CA: Air Force Flight Test Center.

Aura, C., Temme, L., & St. Onge, P. (2020). *A hands-free, visual search task featuring scalable cognitive challenge based upon the U.S. Army Trail Making Test (TMT)* (Report No. 2020-07). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage, 59*(1), 36-47.

Bachelder, E. N., & Godfroy-Cooper, M. (2019). Pilot Workload Estimation: Synthesis of Spectral Requirements Analysis and Weber's Law. In AIAA Scitech 2019 Forum (p. 1228).

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple component model. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 28-61). New York: Cambridge University Press.

Boring, E. G. (1950). *History of experimental psychology (2nd Edition)*. New York: Appleton-Century-Crofts, Inc.

Boring, E. G. (1953). A history of introspection. *Psychological bulletin*, *50*(3), 169-189.

Brennan, S.D. (1992). *An experimental report on rating scale descriptor sets for the instantaneous self assessment (ISA) recorder*. (DRA Technical Memorandum (CAD5) 92017). Portsmouth: DRA Maritime Command and Control Division.

Brown, I. D. (1964). The measurement of perceptual load and reserve capacity. *Occupational medicine*, *14*(1), 44-49.

Cain, B. (2007). *A review of the mental workload literature* (Report No. RTO-TR-HFM-121-Part-II). NATO Research and Technology.

Casali, J., & Wierwille, 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*, p. 623-641.

Casner, S., & Gore, B. (2010). *Measuring and evaluating workload: A primer* (Report No. NASA/TM-2010-216395). Washington, DC: National Aeronautics and Space

Administration.

Christensen, S., & Wright, H. (2014). Quantifying effort in individuals with aphasia invest in working memory tasks through heart rate variability. *American Journal of Speech-Language Pathology, 23,* S361-S371.

Colle, H. A., & Reid, G. B. (2005). Estimating a mental workload redline in a simulated air-to-ground combat mission. *The International Journal of Aviation Psychology*, *15*(4), 303-319.

Cooper, G. (1957). Understanding and interpreting pilot opinion. *Aeronautical Engineering Review*, *16*(3), 47-51.

Cooper, G., & Harper, R. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (Report No. ASD-TR-76-19). Moffett Field, CA: National Aeronautics and Space Administration.

Cummings, M. L., Myers, K., & Scott, S. D. (2006). Modified Cooper Harper evaluation tool for unmanned vehicle displays. In *Proceedings of UVS Canada: Conference on Unmanned Vehicle Systems Canada*.

D'Addario, P., Domnez, B., & Ising, K.W. (2014). EMG provides an earlier glimpse into the effects of cognitive distraction on brake motor response. *Proceedings of the Human Factors and Ergonomics Annual Meeting, 58*(1), 2200-2204.

de Waard, D. (1996). *The measurement of drivers' mental workload*. Netherlands: Groningen University, Traffic Research Center.

Dixon, S. R., & Wickens, C. D. (2005). Control of multiple unmanned aerial vehicles: A workload analysis. *Human Factors*, *47*(3), 479-487.

Dlugosch, C., Conti, A., & Benglar, K. (2013). Driver distraction through conversation measured with pupillometry. *Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 198-204.

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., ... & Giannopoulos, I. (2018). The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-13.

Durantin, G., Gagnon, J. F., Tremblay, S., & Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural brain research*, *259,* 16-23.

Feltman, K. A., Bernhardt, K. A., & Kelley, A. M. (2020). Measuring the Domain Specificity of Workload Using EEG: Auditory and Visual Domains in Rotary-Wing Simulated Flight. *Human Factors*, 0018720820928626.

Feltman, K.A., Curry, I., Bernhardt, K., Hayes, A., Kirby, S., Davis, B., Sapp, J., Durbin, D., &

Hartnett, G. (2018). Integrated cueing environment: Simulation event four. (USAARL Report No. 2018-16). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

Feltman, K.A., Kelley, A. M., Bernhardt, K., Britt, T., & Mathews, C. (2019). Real-time operator state monitoring during periods of high and low workload across four workload domains. (USAARL Report No. 2019-12). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

Feltman, K.A., McAtee, A., St. Onge, P., Aura, C., Stewart, J., Hartnett, G., Hicks, J., Durbin, D., Godfroy-Cooper, M., & Miller, J. (2020). Pilot cueing for 360° obstacle awareness during DVE mission. (USAARL Report No. 2020-26). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied ergonomics*, *73*, 90-99.

Gaetan, S., Dousset, E., Marqueste, T., Bringoux, L., Bourdin, C., Vercher, J., & Besson, P. (2015). Cognitive workload and psychophysiological parameters during multitask activity in helicopter pilots. *Aerospace Medicine and Human Performance*, *86*(12), 1052-1057.

Girouard, A., Solovey, E. T., Hirshfield, L. M., Peck, E. M., Chauncey, K., Sassaroli, A., ... & Jacob, R. J. (2010). From brain signals to adaptive interfaces: using fNIRS in HCI. In *Brain-Computer Interfaces* (pp. 221-237). Springer: London.

Goillau, P.J., & Kelly, C.J. (1996). Malvern Capacity Estimate (MACE) - A proposed cognitive measure for complex systems. *Engineering Psychology and Cognitive Ergonomics, 1*(1), 219-226.

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human factors*, *61*(3), 374-392.

Hart, S., & Staveland, L. (1988). Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & Meshkati (Eds.), *Advances in Psychology, 52. Human mental workload* (pp.139-183). New York: North Holland.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage publications.

Hayes, A., Aura, C., & Feltman, K. (2020). Pupillometric workload measurement in the 360 degree cueing environment (ICE). *Proceedings of the Vertical Flight Society's 76th Annual Forum and Technology Display*.

Hebb, D. O. (1955). Drives and the CNS (conceptual nervous system). *Psychological review*, *62*(4), 243.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task quantified in the prefrontal cortex using fNIRS. *Frontiers in* *human neuroscience*, *7*, 935.

Huemer, W. (2019). Franz Brentano. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/brentano/

Hughes, A., Hancock, G., Marlow, S., Stowers, K., & Salas, E. (2019). Cardiac measures of cognitive workload: A meta-Analysis. *Human Factors, 61*(3), 393-414.

Johnson, A. K., & Anderson, E. A. (1990). Stress and arousal. In J.T. Cacioppo & L.G. Tassinary. *Principles of psychophysiology* (pp. 216-252). Cambridge: Cambridge University Press.

Jordan, C.S. (1992). *Experimental study of the effect of an instantaneous self assessment workload recorder on task performance* (Report No. DRA Technical Memorandum (CAD5) 92017). Portsmouth: DRA Maritime Command and Control Division.

Kahneman, D. (1973). *Attention and effort*. New Jersey: Prentice Hall.

Kantowitz, B. H. (1987). 3. Mental workload. In *Advances in psychology* (Vol. 47, pp. 81-121). New York: North-Holland.

Käthner, I., Wriessnegger, S., Müller-Putz, G., Kübler, A., & Halder, S. (2014). Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain-computer interface. *Biological Psychology, 102*, 118-129.

Kirwan, B., Evans, A., Donahoe, L., Kilner, A., Lamoureaux, T., Atkinson, T., & MacKendrick, H. (1997). *Human factors in the ATM system design life cycle*. FAA/Eurocontrol ATM R&D Seminar, 16-20 June 1997, Paris, France.

Lahtinen, T., Koskelo, J., Laitinen, T., & Leino, T. (2007). Heart rate and performance during combat missions in a flight simulator. *Aviation, Space, and Environmental Medicine, 78*(4), 387-391.

Leino, T. K., Leppaluoto, J., Ruokonen, A., & Kuronen, P. (1999). Neuroendocrine responses to psychological workload of instrument flying in student pilots. *Aviation, space, and environmental medicine*, *70*(6), 565-570.

Lohani, M., Payne, B., & Strayer, D. (2019). A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in Human Neuroscience, 13*(57), 1-27.

Marshall, S. P. (2000). *U.S. Patent No. 6,090,051*. Washington, DC: U.S. Patent and Trademark Office.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. *Proceedings of the IEEE 7th conference on Human Factors and Power Plants*, 7-7.

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviation, space, and environmental medicine*, *78*(5), B165-B175.

Mehler, B., Reimer, B., Coughliun, J., & Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Journal of the Transportation Research Board, 2138*, 6-12.

Meister, D. (1976). *Behavioral foundations of system development*. New York: Wiley.

Milley. M., & McCarthy, R.D. (2017). *Modernization priorities for the United States Army*.

Mun, S., Whang, M., Park, S., & Park, M. (2017). Effects of mental workload on involuntary attention: A somatosensory ERP study. *Neuropsychologia*, *106*, 7-20.

Murai, K., Hayashi, Y., Maenaka, K., & Higuchi, K. (2015). *Basic study on evaluation of navigator's mental work by sticking plaster-type sensor.* 10th System of Systems Engineering Conference (SoSE), 17-20 May 2015, San Antonio, TX.

Nguyen, J.H., Chen, J., Marshall, S.P. Ghodoussipour, S., Chen, A., Gill, I. S., & Hung, A. J. (2020). Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. *World Journal of Urology, 38*, 1599-1605.

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

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. *Proceedings of the 24th Australian Computer-Human Interaction Conference, OzCHI 2012*, 420-423.

O'Donnell, R.D. & Eggemeier, F.T. (1986). Workload assessment methodology. In K. Boff, et al. (Eds.), *Handbook of perception and human performance. Volume II. Cognitive processes and performance* (pp. 42:1-42:49). New York: John Wiley and Sons.

Platten, F. (2012). *Analysis of mental workload and operating behavior in secondary tasks while driving*. (Unpublished doctoral dissertation). Chemnitz University of Technology, Chemnitz, Germany.

Roscoe, A. H. (1984). Assessing pilot workload in flight. ROYAL AIRCRAFT ESTABLISHMENT BEDFORD (UNITED KINGDOM) BEDFORD United Kingdom.

Reid, G. & Nygren, T. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology, 52*, 185-218.

Reid, G., Potter, S., & Bressler, J. (1989). *Subjective Workload Assessment Technique (SWAT): A User's Guide* (Report No. AAMRL-TR-89-023). Wright Patterson Airforce Base, OH: Armstrong Aerospace Medical Research Laboratory.

Richstone, L., Schwartz, M.J., Seideman, C., Cadeddu, J., Marshall, S., Kavoussi, L.R. (2010). Eye Metrics as an Objective Assessment of Surgical Skill, *Annals of Surgery, 252*(1), 177-182.

Rouse, W. B., Edwards, S. L., & Hammer, J. M. (1993). Modeling the dynamics of mental workload and human performance in complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(6), 1662-1671.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*(1), 61-86.

Sassaroli, A., Zheng, F., Hirshfield, L. M., Girouard, A., Solovey, E. T., Jacob, R. J., & Fantini, S. (2008). Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *Journal of Innovative Optical Health Sciences, 1*(02), 227-237.

Schwitzgebel, E. (2019). Introspection. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/win2019/entries/introspection/

Shaffer, F., Combatalade, D., Peper, E., & Meehan, Z. (2016). A guide to cleaner electrodermal activity measurements. *Biofeedback, 44*(2), 90-100.

Tao, D., Tan, H., Wang, H., Zhang, X., Qu, X., & Zhang, T. (2019). A Systematic Review of Physiological Measures of Mental Workload. *International journal of environmental research and public health*, *16*(15), 2716.

Teigen, K.H. (1994). Yerkes-Dodson: a law for all seasons. *Theory & Psychology, 4*, 525-547.

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358-381.

Trnka, R., & Smelik, V. (2020). Elimination of bias in introspection: Methodological advances, refinements, and recommendations. *New Ideas in Psychology*, *56*, 100753.

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology & Work*, *20*(3), 351-365.

von Lühmann, A., Herff, C., Heger, D., & Schultz, T. (2015). Toward a wireless open source instrument: Functional near-infrared spectroscopy in mobile neuroergonomics and BCI applications. *Frontiers in Human Neuroscience*, *9*, 617.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159-177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, *50*(3), 449-455.

Wickens, C., & Tsang, P. (2015). Workload. In D.A. Boehm-Davis, F.T. Durso, & J.D. Lee (Eds.), *APA Handbook of Human Systems Integration* (pp. 277-292). Washington, DC: American Psychological Association.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman and D.R. Davies (Eds.), *Varieties of attention* (pp. 63-102). London: Academic Press.

Wickens, C.D. (1992). *Engineering psychology and human performance*. New York: Harper Collins.

Yerkes, R.M., & Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459-482.

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human factors*, *44*(3), 365-375.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1-17.

Zijlstra, F.R.H. (1993). *Efficiency in work behaviour: A design approach for modern tools*. (Doctoral dissertation). Delft University of Technology. Delft, The Netherlands.

Zijlstra, F., & Mulder, G. (1989). Mentale belasting: theoretische gezichtspunten en overzicht van meetmethoden (mental workload: theoretical points-of-view and an overview of measurement methods). *Mentale belasting en werkstress. Een arbeidspsychologische benadering*, 21-41.

Appendix A. Acronyms

Acronym	Definitions
ANS	Autonomic nervous system
BWRS	Bedford Workload Rating Scale
CSS	Crew status survey
ECG	Electrocardiography
EDA	Electrodermal activity
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
FARA	Future attack reconnaissance aircraft
FLRAA	Future long-range assault aircraft
fNIRS	Functional near-infrared spectroscopy
FVL	Future vertical lift
GPS	Global positioning system
ICA	Index of cognitive activity
IPA	Index of pupillary activity
ISA	Instantaneous self-assessment of workload
MACE	Malvern capacity estimate
MART	Malleable attentional resources theory
MCH	Modified Cooper-Harper
MDO	Multi-domain operations
mm	Millimeter
MRT	Multiple resource theory
MWL	Mental workload

NASA TLX	NASA task load index
RMSE	Rating scale mental effort
SAM	School of Aviation Medicine
SWAT	Subjective workload assessment technique
TEPR	Task evoked pupillary response
USAARL	United States Army Aeromedical Research Laboratory
WP	Workload profile

U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama

All of USAARL's science and technical information documents are available for download from the Defense Technical Information Center.

https://discover.dtic.mil/results/?q=USAARL

Army Futures Command
U.S. Army Medical Research and Development Command