

LITERATURE REVIEW

Workload Measures

Document ID: N01-006
Author: Sarah Miller
Date: August 2001

**National Advanced Driving
Simulator**

2401 Oakdale Blvd.
Iowa City, IA 52242-5003
Fax (319) 335-4658

TABLE OF CONTENTS

1. INTRODUCTION	4
2. MEASURES	5
2.1. Physiological Measures	6
2.1.1. Cardiac.....	6
2.1.1.1. Heart rate.....	7
2.1.1.2. Heart rate variability.....	7
2.1.1.3. Blood pressure.....	8
2.1.2. Respiratory.....	9
2.1.2.1. Respiratory rate.....	9
2.1.2.2. Volume and concentration of carbon-dioxide in air flow.....	9
2.1.3. Eye.....	10
2.1.3.1. Eye blink rate and interval of closure.....	10
2.1.3.2. Other vision measures.....	11
2.1.4. Speech Measures.....	11
2.1.5. Brain Activity.....	11
2.1.5.1. Electroencephalogram (EEG).....	12
2.1.5.2. Electrooculogram (EOG).....	12
2.1.6. Other Measures.....	13
2.1.7. Conclusions.....	13
2.2. Subjective Measures	14
2.2.1. Unidimensional Scales.....	15
2.2.1.1. Modified Cooper-Harper Scale (MCH).....	16
2.2.1.2. Overall Workload Scale (OW).....	16
2.2.2. Multidimensional Scales.....	16
2.2.2.1. NASA Task Load Index Scale (NASA-TLX).....	17
2.2.2.2. NASA-RTLX or RNASA-TLX.....	17
2.2.2.3. Subjective Workload Assessment Technique (SWAT).....	18
2.2.3. Other Subjective Measures.....	18
2.2.4. Conclusion.....	21
2.3. Performance Measures	21
2.3.1. Primary Task Performance.....	21
2.3.2. Secondary Task Performance.....	22
2.3.3. Conclusion.....	23
3. CONCLUSION/DISCUSSION	23

4. CONSIDERATIONS	27
5. RECOMMENDATIONS	32
6. SCENARIOS	32
7. WORKS CITED	33
APPENDIX A	38
Appendix A1: Modified Cooper-Harper (MCH)	39
Appendix A2: Overall Workload (OW).....	40
Appendix A3: NASA Task Load Index (NASA-TLX)	41
Appendix A4: Subjective Workload Assessment Technique (SWAT)	53
Appendix A5: Instantaneous Self Assessment (ISA)	54
Appendix A6: Rating Scale Mental Effort (RSME).....	54
Appendix A6: Rating Scale Mental Effort (RSME).....	55
Appendix A7: Activation Scale	56
Appendix A8: The Verbal Online Subjective Opinion (VOSO)	57
Appendix A9: Cooper-Harper.....	58
Appendix A10: Bedford Workload Scale	59
Appendix A11: Honeywell Cooper-Harper.....	60
Appendix A12: Equal-Appearing Intervals	61
Appendix A13: Driving Activity Load index (DALI).....	62
Appendix A14: Multi-Descriptor (MD)	63
Appendix A15: Analytical Hierarchy Process (AHP)	64
Appendix A16: The Workload/Compensation/Interference/Technical Effectiveness (WCI/TE).....	65

1. INTRODUCTION

Workload is becoming an increasingly important topic in our society. The study of workload is not new; it has been discussed and researched since man united with machine. While businesses are concerned with maximizing profit, the worker is focused on minimizing workload. The solution to this issue is to determine a way to accurately measure workload and determine what levels of workload are excessive. When the concept of workload was first coined, the concentration was on physical workload. Today, the world is a different place. Most physical work has been replaced with machines that do the heavy lifting and moving. Presently, studies involving workload are focused more on other types of workload, including psychomotor, perceptual, or communication workload (Wierwille, 1985). Driving combines many types of workload, but the most important are mental and, to a smaller extent, visual.

Since so much time and energy is spent on developing ways to optimize workload, it would seem that everything related to workload and workload measurement would be well defined and accepted. However, even though this topic is well known, it is very ambiguous. Over 20 years ago, 400 published studies were found that were devoted to measurement of mental workload (Hicks, 1979). At that time, there was no clear way to measure workload; today nothing has changed except there are more papers and more measures. There is still no universally accepted definition for mental workload. One proposed definition is: "Mental workload is a hypothetical construct that describes the extent to which the cognitive resources required to perform a task have been actively engaged by the operator" (Gopher, 1986). Another definition of mental workload proposed by Verwey (2000) is that "mental workload is related to the amount of attention required for making decisions." Just defining the concept of workload is not enough; there must also be a way to measure it. Since there is not even an accepted definition of workload, it is not surprising that there is not a single way to measure it either. There are almost as many ways to measure workload as there are jobs needing measurement.

Numerous articles have been devoted to various topics in mental workload. These topics range from explaining how electrochemical processes predict workload to how motivation is related to effort. This variety explains some of the difficulty associated not only with the exact definition of mental workload, but with the abstract nature of the topic. One of the difficulties associated with mental workload is the multidimensional nature of the topic, but the importance of finding an accurate estimate in many situations is what makes it worth examining. Measurement of workload is only one of the important concepts that need to be examined. "A workload measure is one by which the latter differences are expressed in relation to the overall ability of the human processing system to process information and generate responses" (Gopher, 1984). More important than how workload is measured is how the measurement is used. If workload is only studied in an experimental environment and is not applied to real-life situations, then the research is useless. It is easy to get a count of eye blinks per minute, but applying that to exactly

how much workload a pilot flying a bombing mission is experiencing may be quite difficult. This difficult task is essentially the most important aspect of workload because it is the reason workload is studied. It is also important to determine which method of measurement is best for the topic being studied. To determine the best method, examples will be given and recommendations will be made.

2. MEASURES

There are three main classifications for measurement of workload: physiological, subjective, and performance-based measures. Physiological measurement of workload is a factually based concept that relies on evidence that increased mental demands lead to increased physical response from the body (Moray, 1979). Physiological workload measures are devoted primarily to continuous measurement of the physical responses of the body. These changes are measured in cardiac activity, brain activity, respiratory activity, speech measures, and eye activity. Subjective measurement of levels of workload is based on the use of rankings or scales to measure the amount of workload a person is feeling. Subjective workload measures are devoted primarily to the intermittent question-answer type response to varying levels of workload. The two main types of scales used to measure subjective workload are unidimensional and multidimensional scales. Performance measurement of workload relies on examining the capacity of an individual by means of a primary or secondary task. By measuring how well a person performs on the task, or how their performance worsens with increasing workload, an estimate of mental workload can be determined.

When examining the different measures of mental workload, several criteria must be examined to determine the relative importance of a particular task. The most important criterion is sensitivity. The measure must be able to detect changes in levels of workload (Casali, 1983; De Waard, 1996; Derrick, 1988; Wierwille, 1993). Validity and reliability are also important to consider when choosing a workload measure (Crabtree, 1984; De Waard, 1996; Derrick, 1988; Kantowitz, 1992; Muckler & Seven, 1992; Rokicki, 1995; Tattersall & Foord, 1996). If the measure actually measures what it proposes to measure, it is valid. A reliable measure is consistent; it will always yield the same results for the same level of workload. Intrusion is another important requirement to consider before choosing a measure of workload (Crabtree, 1984; De Waard, 1996; Hill et al., 1992; Wierwille, 1993). An intrusive measure will cause a change in performance. Cost, both of implementation and administration, may also be a consideration (Crabtree, 1984; Derrick, 1988; Hill et al., 1992; Rokicki, 1995). If a measure requires expensive equipment or specialized workers to oversee a portion of the test, then it may not be cost-effective (Crabtree, 1984; Rokicki, 1995). Other things to consider are time needed for the test to be run (Derrick, 1988; Hill et al., 1992; Rokicki, 1995), interval of collection, operator acceptance (Hill et al., 1992), and ease of collection, processing, and analysis (Derrick, 1988; Rokicki, 1995; Tattersall & Foord, 1996).

The three classifications of measurement have many benefits and drawbacks, so finding one perfect form of measurement is nearly impossible. It is important to look at all areas of research to decide which measure is applicable for a given situation. It is common, and most say necessary, for researchers to use more than one method from at least two of the categories to get the most accurate measurement of mental workload.

2.1. Physiological Measures

Physiological measures use the physical reactions of the body to objectively measure the amount of mental work a person is experiencing. It would seem an objective measurement would be the most exact and therefore the best way to find workload because it does not require a direct response from the person, unlike subjective measures (De Waard, 1996). This rationale is not always supported because the body also responds physiologically to things other than mental workload. For example, the body responds to increased physical workload with increased physical responses (De Waard, 1996). When an increase in mental task difficulty is coupled with increased physical workload, the results may be skewed. Each of the physiological workload measures must be examined individually to find the relation between the physiological responses due to physical activity and mental activity.

Most research focuses on five physiological areas to measure workload: cardiac activity, respiratory activity, eye activity, speech measures, and brain activity. Cardiac activity is measured through heart rate, heart rate variability, and blood pressure. Respiratory activity measures the amount of air a person is breathing in and the number of breaths in a given amount of time. Eye measures mainly include horizontal eye movements, eye blink rate, and interval of closure, but there are several other less accepted measures. Speech measures take pitch, rate, loudness, jitter, and shimmer into account when determining workload. To measure brain activity, either the electroencephalograph (EEG) or electrooculogram (EOG) are usually used.

2.1.1. Cardiac

The most common physiological measurement of mental workload is cardiac monitoring. Cardiac measures are often used because they are easy to evaluate and are considered a fairly reliable indicator of workload. Cardiac measures can also be used in real-world environments because the measurements are unobtrusive and continuously available (Wilson, 1992).

2.1.1.1. Heart rate

Heart rate measurement is considered the most common and reliable measure of workload by cardiac means. Heart rate is an exact measurement because the signals can be measured in the form of beats. Generally, as workload increases heart rate increases (Costa, 1993; Hankins & Wilson, 1998; Jorna, 1993; Roscoe, 1993; Veltman & Gaillard, 1996; Wilson, 1993; Wilson, Fullenkamp, & Davis, 1994). Although this generalization is widely accepted, not all studies agree with the findings. Some articles are critical of the use of heart rate to measure workload because of the various psychological, environmental, and emotional factors that can influence the response (Jorna, 1992; Lee, 1990; Roscoe, 1992). For example, “feelings of uncertainty and anxiety can significantly raise heart rate” (Jorna, 1993). G-forces may also affect heart rate (Roscoe, 1993; Wilson, 1992). Since high G-forces are not usually encountered while driving, this is not a real concern. Other research has determined that heart rate “does not appear to be of value as a sole measure of pilot workload but it can be strongly recommended as a technique to augment a good subjective rating scale” (Roscoe, 1992).

Heart rate does not measure absolute levels of workload, only relative levels (Roscoe, 1992, 1993). This may be a benefit in real-world rather than simulated situations because there is less control over variables (Roscoe, 1993). Heart rate measurement is not intrusive, it is continuously available, and it is sensitive to changes in workload (Wierwille, 1993). It is important to remember that mean heart rate varies by individual. When using heart rate as a measurement tool, it is necessary to find a baseline measurement of the heart rate to compare two or more variables (Roscoe, 1992). While analyzing the data on heart rate, it is also important to determine the amount of physical work along with the mental workload because increasing physical load will lead to an increase in heart rate as well (Jorna, 1993).

2.1.1.2. Heart rate variability

Another cardiac measure of workload is heart rate variability (HRV). HRV measures the inter-beat intervals of the heartbeat over time. This measure is not used as extensively as heart rate, but many studies focus on the use of HRV to study workload because it is a fairly new and promising area of research. Heart rate variability, however, is not a widely accepted means of measuring mental workload. Some studies have found three different frequency bands useful for measuring HRV (Jorna, 1992; Veltman, 1998), others have determined there are over 26 different measures available (Wilson, 1992). When the spectrum is divided into three parts, there is a lower band that is associated with regulation of body temperature from .02 to .06 Hz., a middle band that is associated with blood pressure regulation from .07 to .14 Hz., and a high band associated with respiration from .10 to 50 Hz. (Jorna, 1992; Veltman, 1998). There is no one acceptable method of measuring variability; the most common and “convenient method for scoring HRV is to calculate the standard deviation or variance of the interbeat intervals over a

given time, or for a given number of beats" (Roscoe, 1992). Conversely, it was found that averaging the bands is not useful because of reliability problems (Jorna, 1992). This difference in opinion may lead to some discrepancies in the judged usefulness of HRV. Some research has determined that an increase in mental workload leads to a decrease in HRV (Jorna, 1993; Roscoe, 1992, 1993; Wilson, 1993), while other research determined that increased workload does not lead to a decrease in HRV (Brookings, Wilson, & Swain, 1996; Hankins & Wilson, 1998; Veltman & Gaillard, 1996; Wilson et al., 1994).

A reason for the dissociation may be physical and respiratory activity (Jorna, 1992). Some data shows that speech, respiration (Wilson, 1992), muscle activity, body position, physical fitness, and age (Jorna, 1992) can influence the results of HRV. When there is more respiration, the differences in mental effort measured with HRV are underestimated (Veltman & Gaillard, 1996). Also, "the respiratory component is not a stable one as changes in respiratory frequency alter its location in the heart rate spectrum" (Jorna, 1992).

It was also found that psychological factors like fatigue (De Waard, 1996) may also influence HRV. Although Jorna (1992) found that people "seem to select a particular psychological state they see fit for the task and seem to be unwilling to change that state," De Waard (1996) found that "HRV is reduced under increased complexity, but not as a result of increased complexity due to additional tasks." A reason for the discrepancy may be that some studies do not take the responses of other body systems into account or may choose the wrong measure of HRV.

Another problem with HRV measures is time considerations. Some "spectral analysis techniques require a minimum of three to five minutes of data to correctly resolve low frequency components" (Wilson, 1992). In driving, this may be a problem because the event of interest may not be long enough to get an accurate measure. HRV may be influenced not only by mental workload, but also by other factors, so it is important to find an accurate measuring tool.

2.1.1.3. Blood pressure

Blood pressure (BP) is usually a secondary measurement of workload. Not many studies examine the use of blood pressure to determine workload. BP is not widely used because it is a more obtrusive measure than heart rate or heart rate variability. Although blood pressure is found to increase as workload increases (Veltman & Gaillard, 1996), it does not provide any more detailed information about workload than heart rate. Veltman and Gaillard (1998) found BP to be "sensitive to the sympathetic branch of the autonomic activity only."

2.1.2. Respiratory

“Respiration is the physiological process primarily concerned with the interchange of oxygen and carbon dioxide between body tissues and the atmosphere” (Roscoe, 1992). There are several ways to use respiratory measures to find mental workload. Some can be used in real-world settings, while others can only be measured in a laboratory setting. The most common type of respiratory measurement is breathing rate. Other measures include monitoring the volume of air entering and exiting the lungs and measuring the amount of carbon dioxide in expired air (Roscoe, 1992).

2.1.2.1. Respiratory rate

Almost all research conducted on respiration uses rate to determine workload. Respiratory rate measures the number of breaths per given time period. Measuring breathing rate is a very easy and unobtrusive indicator of mental load. Measurement can take place in a real-world or controlled environment. It is generally agreed that an increase in respiratory rate is indicative of increased workload (Brookings et al., 1996; Fournier, Wilson, & Swain, 1999; Roscoe, 1992; Veltman & Gaillard, 1996; Wilson, 1992, 1993). Respiration is considered to be sensitive to changes in workload (Bucks & Seljos, 1994b). Respiratory rate has also been used “extensively as an indicator of emotional states, stress, arousal, and mental load” (Roscoe, 1992). The sensitivity of rate to factors other than increased mental workload may cause problems in reliability and consistency. Respiration rate influences other measures of workload, including heart rate variance, so it is necessary to measure respiration if measuring HRV to find a comparison between the two (Jorna, 1992; Veltman & Gaillard, 1996; Wilson, 1992).

One problem associated with the measurement of breathing rate is that it can be interrupted by speech (Brookings et al., 1996; Roscoe, 1992). Talking and breathing are interconnected in most real-world situations, so it is hard to apply respiratory rates to situations where speech is involved. Another problem with skewed results is physical activity. When mental workload increases, physical workload sometimes increases (Jorna, 1992, 1993). When the body exerts itself physically, there is an increase in respiration rate (Roscoe, 1992). Roscoe (1992) found that “physical activity causes an increase in rate and depth but emotional influences and increased arousal levels normally cause an increase in rate with a decrease in depth.”

2.1.2.2. Volume and concentration of carbon-dioxide in air flow

Measuring the volume flow of air and the amount of carbon dioxide expelled during breathing is not studied as extensively as simply measuring respiratory rate. One reason for not examining the effects of workload on these measures is that it is much harder to calculate the amount of air and carbon dioxide

flow than the number of breaths per unit time without being obtrusive. To measure airflow, a pneumotachograph can be used (Roscoe, 1992). "Indirect measurement techniques such as strain gauges, impedance pneumography and equipment that measures changes in air flow temperature, may be less intrusive, but these techniques are also less accurate" (De Waard, 1996). Most research supports the notion that volume decreases as workload increases (Veltman & Gaillard, 1996; Wilson, 1992). There is conflicting evidence that flow volume is not affected by changes in workload (Brookings et al., 1996).

2.1.3. Eye

Several measures use physiological changes in the eye to determine mental and visual workload. Although the eye is associated primarily with visual workload, it has been shown that some measures are able to accurately predict mental workload for some tasks as well (Van Orden, 1999). The main measures associated with the eye are horizontal eye activity (movement) (HEM), blink rate, and interval of closure. Other measures include eye fixation and pupil diameter. There is a brain activity measure, the electrooculogram (EOG), associated with visual activity that will be discussed in Section 2.1.5.

2.1.3.1. Eye blink rate and interval of closure

Eye blink rate is the number of eye closures in a given amount of time. Interval of closure (blink duration) "is defined as the time spent blinking" (East, 2000). Although measuring eye blink rate is easy, the results are mixed. It is generally accepted that eye blinks are good at measuring visual workload. Eye blinks and blink duration decrease with increasing visual workload (Brookings et al., 1996; De Waard, 1996; East, 2000; Hankins & Wilson, 1998; Van Orden, 1999; Veltman & Gaillard, 1996; Wilson, 1993). Most of the research was done in a flight or driving task where it is hard to separate visual workload from mental workload. Some research has separated visual and mental workload to determine that eye blinks (Van Orden, 1999) and eye blink duration (Sirevaag et al., 1993) are good at estimating some aspects of mental workload. Most do not mention anything other than visual workload, while others found that eye blinks are only good at measuring visual load (Brookings et al., 1996; East, 2000; Hankins & Wilson, 1998). Environmental changes may also influence eye blink and blink duration. When there are changes in light or air quality, eye blink rate may also change (De Waard, 1996).

It was found that blinking may provide more information than just an estimate of workload (Stern, 1984). Stern and Skelly (1984) examined the effects of both visual and auditory information processing on blinking. It was found that blinking is "inhibited during the 'taking-in' of information, whether such information is presented visually or auditorily. Once a decision is made whether it requires action or requires the inhibition of action, a blink is likely to occur. The non-inhibition of blinking during the above-

mentioned time periods is associated with a higher likelihood of occurrence of missed signals and erroneous responses.”

2.1.3.2. Other vision measures

Several other eye activity measures show promise in measuring visual and mental workload. The most promising is horizontal eye activity (movement) (HEM). HEM involves the “scanning eye movements (that) are used to acquire information from the instrument panel (Hankins & Wilson, 1998).” In a car, HEM would measure glances at the speedometer, side mirrors, or rear-view mirror. This measure was found to be a good indicator of visual and mental workload, but there are not many studies that examine HEMs to estimate mental workload. As workload increased, it was found that HEMs increase. Pupil diameter may be another good way of estimating mental workload under certain conditions. Pupil diameter is found to increase with increasing mental workload (Bucks, 1992; Beatty, 1982; Casali, 1983; May, 1990). Eye fixations are another measure used to estimate mental workload. Fixations measure the amount of time the eye spends “looking” at a selected object. Eye fixations are related to performance measures and are only considered diagnostic (De Waard, 1996).

2.1.4. *Speech Measures*

Speech measures are rarely studied as tools for measuring workload. One possible reason for not using speech to find workload is that it is difficult to take exact measures of different aspects of speaking. The six measures most often used to measure speech are pitch, rate, loudness, jitter, shimmer, and a derived speech measure (Brenner, Doherty, & Shipp, 1994). It was found that the three speech measures affected by workload are pitch, loudness, and rate. These three measures all increase as task difficulty increases. The derived speech measure was found to have the most correlation to workload demands (Brenner et al., 1994). Since not much research is devoted to using the voice to measure workload, the information is not corroborated. This may lead to problems when exact measures are needed to determine workload.

2.1.5. *Brain Activity*

All the previous physiological means for measuring workload employ indirect means to gather data. Cardiac, respiratory, eye, and speech activities are all influenced by signals the brain sends when experiencing different amounts of mental load. “The brain is responsible for processing information, making decisions and initiating actions on the external environment” (Brookings et al., 1996). It is generally agreed that the most precise measurement of mental workload comes directly from measuring the activity of the brain. “An advantage of using brain event-related activity to infer workload is that it

provides good temporal resolution of cognitive activity” (Fournier et al., 1999). Some other benefits of measuring brain activity are that they are continuously available and do not interfere with the task (Gevins et al., 1995). Although brain activity measurement does not directly interfere with the task, the gathering of the data may be distracting and intrusive. A major problem with measuring brain activity is that specialized equipment is needed. The equipment requires special training to operate and to interpret the data. The most common type of brain wave used for workload studies is the electroencephalogram (EEG). Another measure sometimes used is the electrooculogram (EOG).

2.1.5.1. Electroencephalogram (EEG)

The electroencephalogram (EEG) is by far the most studied and accepted form of workload measurement that uses brain activity. De Waard (1996) defines an EEG as “a recording of electrical activity made from the scalp.” EEG signals are generally classified into four bands: up to 4Hz. (Delta waves), 4 to 8 Hz. (Theta waves), 8 to 13 Hz. (Alpha waves), more than 13 Hz. (Beta waves) (De Waard, 1996). East (2000) adds another band from 31-42 Hz. (Ultra Beta). When there is an increase in mental workload, the EEG shows that the Alpha waves disappear and are replaced by Beta waves (Sabbatini, 1997). Generally, as mental workload increases, theta increases and alpha decreases (Hankins & Wilson, 1998). Beta waves were associated with changes in complexity. Delta and Alpha waves are affected differently by complexity and volume changes. (Brookings et al., 1996). Brookings *et al.* (1996) found EEG measures to be useful in finding and “evaluating the relative contributions of workload variables that are not detected by other indexes.” Physical movements may cause problems in the analysis of the EEG. Another problem with using the EEG as a measurement is the intrusiveness, and the cost of implementation. Although the EEG is a good predictor of workload most studies use the EEG to measure driver state.

2.1.5.2. Electrooculogram (EOG)

The electrooculogram is primarily used for measuring saccadic eye movements (Galley, 1993). This measure is another form of measuring eye blink rate and eye closure interval. Not much research is presently being conducted on the benefits of using electrooculogram for workload measurement. This may be due to the intrusiveness of the measure. Galley (1993) measured the velocity of the saccadic eye movements to find workload. This type of measure seems to be a good indicator of visual workload, but there are not enough studies to determine whether the electrooculogram results agree with other types of visual workload measures like eye blink rate. The EOG is intrusive, and it is expensive to implement this test.

2.1.6. Other Measures

There are other physiological measures of different kinds of workload that may be new or not widely studied. These measures may have some potential for measuring mental workload in driving situations. There are several complicated measures that measure different outputs by the body that would be hard to use in practical settings. They tend to be useful but extremely intrusive, requiring expensive machinery or tests for measurement.

There are several brain activity measures that are not studied as extensively. These measures may hold some promise in the area of mental workload measurement. The use of the Electromyogram (EMG) is a new but promising measure of mental workload. The EMG measures 'task irrelevant' facial muscles that are not required in the motor performance of a task (De Waard, 1996). Different facial muscles are found to be differentially sensitive to changes in mental workload. De Waard (1996) identifies the frontalis and the corrugator as muscles that have been studied. Another brain activity that shows promise is the ElectroCardioGram (ECG). The ECG is related to cardiac measures, specifically HRV (De Waard, 1996). Event related potentials (ERPs) are related to fluctuations in the EEG (De Waard, 1996). The effects of increased mental workload on ERPs are not well documented, but it was found that an advantage of the ERP is "its high diagnosticity to perceptual/cognitive (mental) processing, and its insensitivity to response factors" (De Waard, 1996).

Other changes that deal with physiological changes in parts of the body other than the brain may hold some potential for measurement of workload. Electrodermal activity (EDA) measures electrical changes in the skin (De Waard, 1996). As workload increases, EDA was found to increase (De Waard, 1996). The measure is not very selective because many factors were found to affect EDA. Changes in hormone levels are related to extremely stressful situations (East, 2000). Hormone levels are usually used for long-term studies on workload (De Waard, 1996). Fixation Fraction (FF) deals with eye fixation time (Wierwille, 1985).

2.1.7. Conclusions

Physiological measures are good for continuous monitoring of workload levels. A few physiological measures have potential for use in a driving simulator or a real-world environment. The best way to measure visual workload is to use eye blink rate. Measuring brain activity by using an EEG machine is also beneficial, but it is hard to use in real-world situations. The most accurate form of cardiac measure is heart rate because it is both unobtrusive and sensitive to changes in workload. There is conflicting information on the benefits of measuring cardiac activities like heart rate, heart rate variability, and blood pressure. Cardiac measures of heart rate and blood pressure can be confounded by physical or

emotional changes, so this needs to be taken into consideration when using these measures. Because of the promise shown in some of the research, it may be beneficial to further study a few new measures like pupil diameter or horizontal eye movement to determine whether they are good indicators of mental workload as well as visual workload.

2.2. Subjective Measures

Subjective measures are used to reflect the amount of information used in working memory (Yeh & Wickens, 1988). A simplistic, but realistic, way to look at workload measurement is that if a person feels a lot of workload, there is a lot of workload (Johannsen, 1979). Although physiological measures of workload may be more precise, subjective measures are more practical. The subjective tests are flexible for different people with different capabilities. "Because subjective ratings take into account individual differences in ability, state, and attitude – differences that may be obscured in objective measures of performance until breakdown makes them obvious – they are valuable because of, not despite, their subjectivity" (Muckler & Seven, 1992). Even though subjective and objective measures of workload are very different, it has been shown that subjective measures correlate with physiological measures of workload such as heart rate variability (Tattersall & Foord, 1996).

When determining what type of measure to use for workload, two of the main concerns are ease of use and effect on performance (Tattersall & Foord, 1996). Subjective measures are considered the easiest method of assessing workload (Yeh & Wickens, 1988). According to Hill *et al.* (1992), citing the results of Sheridan (1980), opponents of physiological measures argue that subjective measures are an accurate indicator of workload, and increasing numbers of studies have found operator ratings to be a more direct indicator of workload than physiological measures. Subjective measures are considered to be the least intrusive, most flexible, most convenient, least time consuming, and least expensive form of evaluating workload (Yeh & Wickens, 1988).

One drawback to subjective measures is that they do not provide a continuous form of measurement (Yeh & Wickens, 1988). Measurements can be taken during the task, but they do not have to be. These measurements should not affect performance if taken at the right time, but may interfere with primary task performance if taken at the wrong time (Tattersall & Foord, 1996). According to Wickens (1984), as cited in Tattersall and Ford (1996), it is necessary to allow time for completion of the subjective scale, as not to violate Wicken's Multiple Resource Theory. For example, if a person needs to complete a NASA-TLX while they are driving over a difficult section of road, the Multiple Resource Theory may be violated. It was found that it is not necessary to subjectively interview a person during or immediately following the difficult section. Delays of up to 15 minutes in reporting do cause some differences in reporting of scores in administration of the SWAT test, but these delays are not significant (Eggemeier, 1983, 1984; Wierwille,

1993). A reason that delays may not have been recommended in the past is that people may forget the amount of workload they were feeling during a particular segment of the task if the delay is excessive. It was found that “longer delay intervals requiring performance of additional intervening task conditions would result in significant effects on target task workload ratings” (Eggemeier, 1984). De Waard (1996) indicated that delays of up to 30 minutes only affected complex multiple-task performance.

Another problem to be concerned with when using a subjective form of measurement is the rating scale. “The context surrounding an evaluation may strongly influence the results” (Colle & Reid, 1998). When workload is measured on a subjective scale, judgments about ratings tend to be based on the entire scale. The categorical intervals tend to be used equally often (Colle & Reid, 1998). When using an entire scale and not an entire range of workload, it is necessary to give representative numbers for different amounts of workload in examples. For example, the administrator could give different situations that would result in different scores on a workload scale so the person can get an idea of the different meanings on the scale. Another way to account for these discrepancies is to use multiple subjective scales.

Problems may also occur with familiarity. As a person becomes more comfortable performing a task, perceived workload may decrease (Hicks, 1979). This may be a problem while running multiple trials of the same experiment, or testing a person who is extremely familiar with the environment. Also, people sometimes have problems differentiating between mental and physical workload (Hicks, 1979)

Subjective measures can be divided into two categories: unidimensional and multidimensional ratings. Unidimensional rating scales are considered the simplest to use because there are no complicated analysis techniques. The unidimensional scale has only one dimension. Generally, the unidimensional scale is more sensitive than the multidimensional scale (De Waard, 1996). The multidimensional workload scale is considered to be a more complex and more time consuming form of measurement, and has from three to six dimensions. The multidimensional scale is generally more diagnostic (De Waard, 1996). Various unidimensional and multidimensional scales will be discussed in the following sections

2.2.1. Unidimensional Scales

Unidimensional rating scales have not been given much validity in past research. They are often considered to be too simple to measure the complexity of workload. Upon further analysis, the one-dimensional scale has been given some validity (Byers, 1989; Gopher, 1984; Hendy, Hamilton, & Landry, 1993; Hill et al., 1992)[Vidulich, 1987 #106]. Unidimensional scales have even been found to outperform multidimensional scales [Vidulich, 1987 #106]. “(People) appear to be able to use a single scale to evaluate all tasks, despite their huge diversity in modalities, mental operations, and response modes”

(Gopher, 1984). Univariate scales are often the easiest and least time-consuming to both measure and analyze of the two rating scales.

2.2.1.1. Modified Cooper-Harper Scale (MCH)

The Modified Cooper-Harper (MCH) scale is “a 10-point unidimensional rating scale that results in a global rating of workload” (Hill et al., 1992). This study was developed to be a change from the psychomotor Cooper-Harper scale and “increase the range of applicability to situations commonly found in modern systems” (Wierwille, 1983). The MCH scale is used to measure perceptual, cognitive, and communications workload (Wierwille, 1983). There is contradictory evidence on the validity and sensitivity of this scale. Generally the MCH was found to be a good estimator of overall mental workload (Casali, 1983; Wierwille, 1983, 1993, 1985). Conversely Hill *et al.* (1992) found the MCH to be of little value. It was hard to complete, not accepted or liked, was not sensitive, and had a poor description of workload.

2.2.1.2. Overall Workload Scale (OW)

The Overall Workload scale utilizes a unidimensional scale from 0 to 100. Zero represents very low workload and 100 represents very high workload (Hill et al., 1992). “A single, 20-step bipolar scale is used to obtain this global rating. A score from 0 to 100 (assigned to the nearest 5) is obtained” (Hill et al., 1992). The OW scale was found to be an excellent way to measure workload on a unidimensional scale (Byers, 1989; Hill et al., 1992). It has even been found produce results comparable to the NASA-TLX [Vidulich, 1987 #106]. The Overall Workload scale doesn’t take much time to complete and is easy to use. This scale takes little time to learn how to administer, prepare for, or analyze. It is also considered to be almost as sensitive as the multidimensional scales (Byers, 1989; Hill et al., 1992). Hill *et al.* (1992) suggest that the Overall Workload Scale could be “useful as a screening tool to identify potential chokepoints of workload.”

2.2.2. Multidimensional Scales

The multidimensional form of measuring workload is the most widely used and accepted way to assess workload by subjective means. There are currently two main multidimensional measures being used in the real-world and simulated environment, the NASA-Task Load Index (NASA-TLX) scale, and the Subjective Workload Assessment Technique (SWAT). There are also several other scales that are less well known. The multidimensional natures of the scales provide a more in-depth analysis of the many aspects of workload, where the one-dimensional scales cannot.

Generally, the multidimensional form of measurement takes more time to complete, so it is hard to use a multidimensional scale during the study because of the violation of the Multiple Resource Theory as explained in the introduction to subjective scales. Not only is the gathering of results time-consuming, the analysis takes time too. Therefore, someone who is trained how to use that particular multidimensional scale must analyze the results. Recently there has been conflicting evidence that multidimensional scales are not as valuable as once thought to accurately assess workload (Hendy et al., 1993). Hendy *et al.* (1993) asserts that the NASA-TLX and SWAT scales use weight calculations that are “superfluous.”

2.2.2.1. NASA Task Load Index Scale (NASA-TLX)

“The NASA Task Load Index uses six dimensions to assess workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. Twenty-step bipolar scales are used to obtain ratings for these dimensions. A score from 0 to 100 is obtained on each scale” (Hill et al., 1992). This scale uses a weighting process that requires a paired comparison task. The task requires the operator to choose which dimension is more relevant to workload for a particular task across all pairs of the six dimensions. The workload scale is obtained for each task by multiplying the weight by the individual dimension scale score, summing across scales, and dividing by the total weights (Hill et al., 1992).

Generally, the NASA-TLX is an extremely good multidimensional scale for measuring mental workload (Byers, 1989; Hill et al., 1992). Hill *et al.* (1992) found that the TLX was well liked, sensitive to changes in workload, and had high diagnosticity. One drawback is the time needed to complete and analyze the test. Another drawback with the TLX scale, as with any other subjective scale, is consistency. Hankins & Wilson (1998) reported that the TLX ratings “lacked internal consistency from the effort and frustration levels reported to the performance scale.”

2.2.2.2. NASA-RTLX or RNASA-TLX

Recently a different type of TLX scale was developed called the NASA Raw Task Load Index (NASA-RTLX). This scale was developed because the collection and analysis of the original TLX scale was cumbersome and labor intensive (Byers, 1989). The RTLX computes a score by taking the sum of the TLX test and dividing it by six. This new way to score the NASA-TLX was found to be almost equivalent to the original TLX scale $R = .977$ ($p < 10^{-6}$) (Byers, 1989), with far less time involved for analysis. In a driving situation, Park & Cha (1998) found that the RTLX scale was the more sensitive to mental demand and difficulty in driving than the TLX.

2.2.2.3. Subjective Workload Assessment Technique (SWAT)

The Subjective Workload Assessment Technique uses three levels – low, medium, and high – for each of the three dimensions of time load, mental load, and physiological stress load to assess workload (Hill et al., 1992). The SWAT technique scales the measurement scores to produce a single rating scale with interval properties (Hill et al., 1992). This multidimensional test uses three steps to complete and analyze workload. Hill *et al.* (1992) outlines the test in the following step method. The first step is scale development, which combines all the possible combinations of the three dimensions in 27 cards. The person sorts the cards into a ranking that reflects his or her perception of increasing workload. The rankings are used to develop a scale with interval properties. The second step is rating the workload. The third step is to convert the scores into a 0 to 100 scale using the scale developed in step one.

The theory behind the SWAT technique is that it “(gains) insight into the mechanism of human information processing resources, together with the notion that it is possible to derive a model, by some rational procedure, that has greater validity than that of an arbitrarily chosen model” (Hendy et al., 1993). Some studies indicate that the SWAT scale proves to be useful in estimating changes in mental workload (Colle & Reid, 1998; De Waard, 1996; Eggemeier, 1983; Wierwille, 1993). In contrast, it was found that the three dimensions “lack subjective orthogonality” (Boyd, 1983). For example, high levels of time-load will also artificially inflate the level in the mental workload category.

When the SWAT scale is compared to the NASA-TLX, the TLX scale is generally considered to be the better scale for measuring mental workload (Hill et al., 1992; Park, 1998). Conversely Colle and Reid (1998) found that the SWAT was more sensitive to changes in difficulty and context, and Wierwille and Eggemeier (1993) found SWAT to potentially be able to identify “cognitive mechanisms affecting mental workload.”

An interesting observation was reported by Hill *et al.* (1992) – that when following the SWAT procedure outlined in the user’s guide, there was a 43% failure rate on the first attempt to perform the sorting step for the SWAT. Experienced operators encountered this high failure rate, so it is suggested that the failure rate would be much higher for inexperienced operators (Hill et al., 1992).

2.2.3. Other Subjective Measures

There are many less known subjective scales used for measurement of various types of workload. In general, these measures usually were developed for use in the aviation industry, and they have not made the crossover into driving research. Some of these measures may hold promise for studying mental workload in driving; others are only useful for the specific task they were designed for.

The Instantaneous Self Assessment technique is a unidimensional scale that uses five different ratings for perceived workload: excessive, high, comfortable, relaxed, and under-utilized (Tattersall & Foord, 1996). This test uses a visual prompt, and the rating is pressed on a keypad (Tattersall & Foord, 1996). The ISA is a newly developed technique that had not been used in any studies before Tattersall and Foord (1996). In Tattersall and Foord (1996), it was found that the ISA was comparative to other subjective measures of workload. ISA ratings most closely correlated with the SWAT test. One of the problems with the ISA technique is that it competes for attentional resources with the primary task, which violates the Multiple Resource Theory.

The Rating Scale Mental Effort (RSME) scale is a unidimensional scale. "Ratings of invested effort are indicated by a cross on a continuous line. The line runs from 0 to 150 mm, and every 10 mm is indicated" (De Waard, 1996). This scale rates invested effort of the task, not explicitly mental effort. De Waard (1996) found the RSME could distinguish between the "task-load situation and baseline." This does not seem to be a mainstream test because no other studies could be found that use the RSME to measure mental workload.

The Activation scale is comparable to the RSME. "The scale has a range from 0 to 270 and is scored by measuring the distance from the origin to the mark in millimeters" (De Waard, 1996). The reference points are based on general tasks like "I am reading the newspaper" (De Waard, 1996). People mark their estimated workload, comparing it to the general tasks. The sensitivity and diagnosticity of this test is not documented.

The Verbal Online Subjective Opinion (VOSO) and the Subjective Opinion via Continuous Control (SOCC) scales are unidimensional ratings. The VOSO is much like the Overall Workload scale with people providing a verbal estimate of mental workload on a scale between 0 and 10. The SOCC scale is estimated with a hand-control to a minimum, medium, or maximum point. These two scales are very sensitive to short periods of mental load (Wierwille, 1993).

The Cooper-Harper rating scale is a unidimensional measure primarily for measurement of psychomotor workload in pilots (Rehmann, 1995). This scale has little use in driving situations or for measuring mental workload.

The Bedford Workload Scale is another modification of the Cooper-Harper rating scale. It is a unidimensional scale that ranks whether it was possible to complete the task, if workload was tolerable for the task, and if workload was satisfactory without reduction (Rehmann, 1995). The Bedford scale was developed for pilots, but it could be used for drivers. This scale is about as useful as the MCH.

The Honeywell Cooper-Harper Rating scale is another unidimensional modification of the Cooper-Harper scale. This scale is used for overall workload measurement, so it is not directly related to mental workload. The Honeywell scale was developed for pilot workload (Rehmann, 1995).

The Dynamic Workload Scale is a unidimensional scale used primarily for aircraft certification by Airbus Industries (Rehmann, 1995). This scale is not extremely useful for measuring mental workload while driving.

In the Equal-Appearing Intervals, “subjects rate the workload in one of several categories using the assumption that each category is equidistant from adjacent categories” (Hicks, 1979). This is a unidimensional scale.

The Driver Activity Load Index (DALI) is related to the NASA-TLX. Not much information is available on this measure.

The Multi-descriptor scale (MD) is a multidimensional rating scale originally developed for aviation. There are seven descriptors: attentional demand, error level, difficulty, task complexity, mental workload, stress level, and overload level. This scale was found to be insensitive to changes in workload [Casali, 1983 #91; Wierwille, 1985 #102].

The Analytical Hierarchy Process (AHP) is a multidimensional scale that “uses the method of paired comparisons to measure workload” (Rehmann, 1995). The comparisons are made in matrix form, and the eigenvector is calculated to weight each of the conditions [Vidulich, 1987 #106]. This rating scale has potential to be a sensitive and reliable indicator of mental workload, but the analysis of the data is complicated. Vidulich and Tsang (1987) found the AHP to outperform the NASA-TLX and the OW scales with regard to validity and reliability. The article expressed concern with regard to the similarity between the tasks assessed. It is suggested that if the tasks were less similar, other scales may estimate workload better.

The Workload/Compensation/Interference/Technical Effectiveness (WCI/TE) scale is a multidimensional matrix used to rate subjective workload. This technique requires complex analysis to compute a score of 0-100 (Rehmann, 1995).

2.2.4. Conclusion

Subjective measurement of workload is good for determining how much workload a person “feels.” In the past, multi-dimensional measures were considered the best form of subjective measurement of workload. Recently, however, there is some evidence that unidimensional ratings of workload could be just as good as the multidimensional scales. For simple tasks, or while performing a task, a unidimensional rating scale is very good because it is fast, easy, and not distracting. The overall workload scale has been shown to be a good unidimensional rating scale. At the end of the test, it may be beneficial to use a multidimensional scale to gain a more exact estimate of workload. The best multidimensional measurement is the NASA-TLX. Although it takes a long time to complete, it has been shown to be very accurate. Time is not as big a consideration at the end of an experiment as in the beginning.

2.3. Performance Measures

“Performance may be roughly defined as the effectiveness in accomplishing a particular task” (Paas & Vanmerrienboer, 1993). The two main ways to measure workload by means of performance are primary and secondary measures. The basis for using primary and secondary tasks to measure workload is based on the assumption that people have limited resources (Yeh & Wickens, 1988). Derrick (1988) explains how the “tasks that demand the same resource structure will reveal performance decrements when time-shared and further decrements when the difficulty of one or both is manipulated.” This means that workload can be estimated by measuring the decrease in performance by either the primary or secondary tasks. The primary task measure is a more direct way to measure workload than the secondary task measure, but both are used and at least moderately accepted.

2.3.1. Primary Task Performance

Primary task performance measurement measures the workload based on the capability to perform the main task (Rehmann, 1995). This is a direct and nonintrusive form of measurement. “Primary task measures are ideal in that they provide an indication of both operator and system performance” (Sirevaag et al., 1993). Primary tasks have to be individually determined for each situation (Hicks, 1979), but may include measuring steering wheel movements (De Waard, 1996; Hicks, 1979), lane-keeping behavior, speed control (Wierwille, 1993), and Time-to-Line Crossing (TLC) (De Waard, 1996) in driving situations.

It was found that steering wheel movements, particularly wheel reversals, are sensitive to changes in workload (Hicks, 1979). The measurement of steering wheel reversals does not require much specialized equipment (De Waard, 1996). Lane-keeping behavior, like deviation from the centerline or lateral deviation, is not shown to be sensitive to changes in workload (De Waard, 1996; Hicks, 1979). A reason

for this is because “lane-keeping in experienced drivers ... is automatic” (De Waard, 1996). It is also hard to measure lane deviation without specialized equipment (De Waard, 1996). Speed has been shown to decrease as workload increases (Wierwille, 1993). Speed is a sensitive measure, but can be disrupted by changes in traffic (De Waard, 1996). Speed can easily be measured by determining the time it takes to complete a course (De Waard, 1996). Time-to-Line Crossing is defined as “the time required for the vehicle to reach either the center or edge line of the driving lane if no further corrective steering wheel movements are executed” (De Waard, 1996). As mental workload increases, TLC increases.

One of the problems associated with strictly using the performance on a primary task is that it does not take into account spare mental capacity (Sirevaag et al., 1993). For example, two tasks may be performed equally, but one person’s mental capacity may be pushed to its limits while another person’s mental capacity is not pushed at all (De Waard, 1996). Another problem with using primary performance measures to estimate workload is motivation. When people are more motivated, their workload may increase, but their performance might not increase to the same extent (Vidulich & Wickens, 1986). It is also hard to measure changes to performance due to workload, unless the workload is very high. Changing from a low to medium level of workload probably will not produce a change in performance even though workload is increasing. Another problem with using primary task performance is that the measures are not easily transferred from one task to another (Sirevaag et al., 1993). For example, if there are several different trials on different courses, the primary performance measure must be separately chosen for each course.

2.3.2. Secondary Task Performance

The secondary task is an additional measure to the primary task. “The basic idea of a secondary task is that it measures the difference between the ‘mental capacity’ consumed by the main task, and the total available capacity” (Mulder, 1979). The basis for this measurement ties in with the Multiple Resource Theory because primary task performance takes a certain amount of resources, so the remaining resources are theoretically used on secondary task performance (Wickens, 1998). “Poor dual-task performance would suggest competition for many of the same resources, whereas efficient dual-task performance would suggest little resource competition” (Derrick, 1988). An advantage of secondary measurement over primary measurement is that it is able to determine if there is any spare mental capacity (Sirevaag et al., 1993). Some examples of secondary tasks relating to driving are car following, mirror checking, and addition tasks (De Waard, 1996).

Car following and mirror checking are known as embedded secondary tasks because they naturally occur while driving (De Waard, 1996). The problem with embedded tasks is that they may not be considered to be less important than the primary task. For a secondary task to be used, less importance must be placed

on it than the primary task. De Waard (1996) explains how it is not known whether less importance is placed on the car following task than the lane keeping task as both are necessary for safe driving. Artificial tasks like addition or simple mathematical computations are considered good secondary tasks. The problem with artificial tasks is the intrusion factor. Artificial tasks may intrude on other workload measures.

The major problem that may occur when secondary tasks are used to measure workload is that they may disrupt primary task performance (Colle & Reid, 1999; Sirevaag et al., 1993). Some people may not perform the primary task before they perform the secondary task. This causes problems for measurement of changes in performance of the secondary task. When determining the validity of different measures of workload, it is imperative that the primary and secondary tasks use the same resource. For example, a primary measure that is primarily visual must be coupled with a secondary measure that is also visual to achieve the best measure of performance (De Waard, 1996). Many of the discrepancies between and within the different measures of workload are due to inaccurate or poor collection of the data. It is important to keep safety in mind when choosing a secondary task because the performance may be degraded to the point that it becomes dangerous when workload is too high. Another problem with secondary task measurement is that it is only theoretical because presently there is no accepted index of measurement (Colle & Reid, 1999). As with the primary task, a separate secondary task must be chosen for each separate situation.

2.3.3. Conclusion

Most performance measures are able to estimate changes in high levels of workload. If the task is too easy, the performance measures do not indicate levels of workload accurately because performance is not lacking in any area. Primary task performance is easier to measure than secondary performance, and it has been more extensively studied for accuracy. Therefore, primary task performance should be used if performance is to be measured to find mental workload.

3. CONCLUSION/DISCUSSION

“In principle, all three approaches represent alternative paradigms to the study of the same phenomenon; that is the relationship between the demands imposed on the human by the task (the environment) and the human’s ability to cope with them. In practice, however, there is little knowledge on the way in which measures obtained by one method are related to those obtained by another. Furthermore, there seems to be considerable disagreement among proponents of each method as to which provides a better or more valid estimate of the underlying limitations” (Gopher, 1984). Table 1 presents the benefits and drawbacks of each method.

Table 1 Conclusions

Measure	Result of Increased Workload	Benefits	Drawbacks
<u>Physiological</u>		Continuous	
<u>Cardiac</u>			
Heart Rate	Increases	Widely accepted and studied, easy to measure	May not be completely reliable Doesn't measure absolute levels of work
Heart Rate Variability	Decreases	Some studies indicate better accuracy than HR	Not widely studied or accepted Influenced by respiration, equipment req.
Blood Pressure	Increases	Can be used to calculate modulus	Not widely studied or accepted No more information than HR or HRV
<u>Respiratory</u>			
Respiratory Rate	Increases	Easy, unobtrusive, sensitive, reliable	Influenced by emotion, stress, speech
Volume per breath	Decreases		Hard to calculate, obtrusive, not studied
<u>Brain Activity</u>			
Electroencephalogram	Alpha waves replaced by Beta waves	Extremely accurate, reliable, catches changes other measures may miss	Obtrusive, requires special equipment and training, may not be cost effective
Electrooculogram	Less jumps in data (Related to gaze and blink rate)	Accurate for visual measures	Not widely used, obtrusive, requires special equipment and training to use
Eye Blink Measures	Rate decreases; pupil diameter increases	Most accurate for visual workload	Not as accurate for other work measures
Speech Measures	Pitch, loudness, and rate Increase	Can be used to determine influence on HRV	Not studied, speech not important for all applications
<u>Subjective</u>			
<u>Unidimensional</u>			
Modified Cooper-Harper Scale	Higher rating	Easy to give during an experiment Fast, fairly accurate	Conflicting opinions, considered hard to use
Overall Workload Scale	Higher rating	Fast, accurate for unidimensional scale easy to administer, prepare for and analyze	Mainly used for identifying "chokepoints"
General Unidimensional	Higher rating	Fast, accurate at measuring overall workload	Only one dimension, not many dimensions
<u>Multidimensional</u>			
NASA Task Load Index Scale	Higher rating	Accurate, valid	Takes a long time to administer and analyze
Subjective Workload Assessment Technique	Higher rating	May be more sensitive to increases in difficulty than TLX	Reports of high failure in analysis of results No agreement on accuracy or sensitivity
<u>Performance</u>			
Primary	Decreases	Accurate to changes in workload	Not accurate when low levels of workload
Secondary	Decreases	Finds spare mental capacity better than Primary	May interfere with task, not accepted

To study workload, it is first necessary to define exactly what kind of workload is to be estimated. Measures of workload do not always agree. Most studies do not attempt to explain this disagreement or explain it as one measure being more accurate than another. This may be true in some cases, but arguments of accuracy can be given for any given workload measure. Most of the time, the problem with workload measurement is not each measure's validity, it is what kind of workload is being measured. Even if a measure is specifically targeting mental workload, there are several ways of going about it. Some measure relative or overall measures of workload, while others try to find absolute levels.

Since it has been determined that it is necessary for more than one measure to be used when estimating mental workload, it is necessary to compare between measures of workload. The reports that attempt to find agreement between measures usually stick to the same category. For example, one would compare heart rate, heart rate variability, and eye blinks. Not as many attempt to explain differences between physiological and performance measures.

It is easier to find studies that compare within each measure of workload than to find studies that compare between measures of workload. For example, the studies that examine heart rate as a measure of workload also examine the effect of workload on respiration or blink rate. There are not as many studies that involve a physiological measure and a subjective measure in the same experiment. Thus it is hard to determine how the correlation between different measures of workload is related to actual workload.

There are comparisons made between physiological and subjective measures, subjective and performance measures, and physiological and performance measures. The most comparisons have been made between physiological and subjective because there is disagreement on the effectiveness of some of these measures. Proponents of physiological measures believe subjective measures are lacking, and supporters of subjective measures suggest physiological measures are not accurate estimators. When there is comparison between different physiological and subjective measures, most of the studies also involve the comparison within the measures. For example, Haskins *et al.* (1998) compares physiological measures such as heart rate, heart rate variability, eye blinks and electroencephalogram measures with the subjective NASA-TLX scale. So while determining if physiological indices match subjective indices, there is also a comparison within the different physiological measures.

When comparing the three types of workload, most studies examine subjective vs. physiological measures. A reason why a large number of studies use this comparison is that the method of

gathering information and the type of information gathered is very different between the two measures. It is hard to draw conclusions about the comparisons between different studies because most of the studies compare different methods of measurement. Although it is difficult to determine which measure is best for workload, some conclusions may be drawn from these comparisons.

When comparing subjective vs. physiological measures of mental workload, it is necessary to first determine if the measures come to the same conclusion about the level of mental effort used. If there is agreement, then it is likely that all the measures are good at determining workload. If a disagreement occurs, it is likely that one of the measures is not an accurate indicator of the type of changes in mental workload being examined. One reason the measures may not agree is how the measure is taken. Roscoe (1993) found subjective and heart rate responses to agree, but dissociation occurred when a person used only part of the rating scale or rated only part of the task rather than the whole scale or the whole time period. Roscoe (1993) indicates that heart rate and subjective ratings indicate relative differences in workload, not absolute levels, but 20 percent of the subjects show a disagreement between heart rate and other measures. Other studies also found some agreement between subjective and physiological ratings (Backs, Ryan, & Wilson, 1994a; Brookings et al., 1996; Roscoe, 1992; Tattersall & Foord, 1996; Wilson, 1992, 1993). It was found that high heart rate, low blink rate, short blink duration, high respiration rate, low breath amplitude, and high subjective ratings are all connected to increases in mental workload (Wilson, 1993).

There is not such an extensive comparison between subjective vs. performance and physiological vs. performance measures. Not many comparisons are made with performance measures because not many studies use performance measures to evaluate workload. The studies that do make comparisons between performance and subjective measures often find dissociation. Yeh (1988) makes a comparison between performance and subjective measures. He explains that a "dissociation between performance and subjective measures (can occur) when a pair of dual-task configurations differ in the degree of competition for common resources." Another explanation for dissociation is motivation (Vidulich & Wickens, 1986). Vidulich (1986) also cites the work by Wickens and Yeh (1982) that suggests "increased motivation will improve performance but will also increase workload ratings." The few studies in this category indicate the need for many measures to get an accurate estimate of workload. Derrick (1988) explains how "someone who relies solely on subjective data may likely favor the system that has serious performance limitations." Some agreement was found between secondary tasks and subjective measures of workload (Colle & Reid, 1999).

4. CONSIDERATIONS

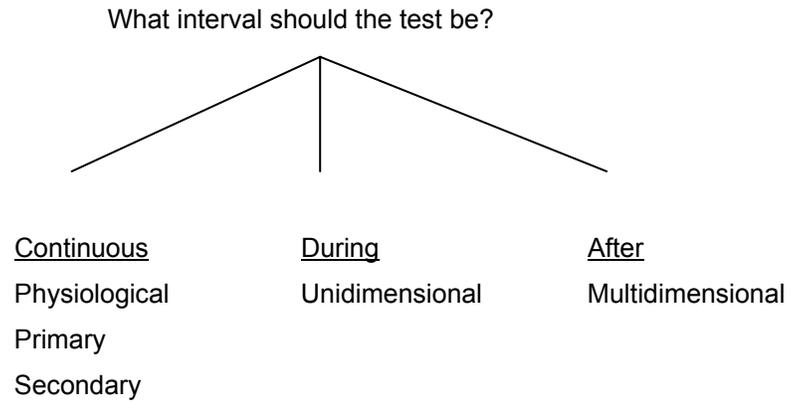
When determining which measurement to use, several considerations should be made. The relative importance of these considerations may change as circumstances change. Some of the most important considerations in any circumstance are the accuracy, reliability, validity, and predictability of the measure. These are very important to consider because the measure must be useful in determining workload or else it is a waste of time. Another important consideration is determining the type of data that needs to be gathered. This is used to see if it is necessary to have continuous data collection or if it is only necessary to collect data at certain intervals. The data gathered may need to be of an auditory, visual, or written nature. Other semi-important considerations are the ease of collecting, processing, and analyzing the results. These considerations may or may not be important in a given study, but they must be considered in real-world and experimental studies. Other less important considerations to examine are cost, user acceptance of the measure, intrusiveness, and time needed for taking the measurement. These factors are usually not as important in an experimental environment, but may become more important in real-world situations. Table 2 describes the different considerations for determining the best test for measuring workload.

To determine which measures of workload should be used in a particular situation, it is necessary to rank the importance of each consideration. The most important considerations should be looked at before less important considerations to pick one or several measures of mental workload that best suit the task. After the considerations are ranked in order of importance, it is necessary to follow the decision tree to decide which of the measures fit the conditions for the task. To determine the best overall measure for the task, it is necessary to find which measures fit the criteria designated. The best measure will be the one that meets the necessary conditions required by the experiment. If the measure doesn't fit all the conditions, it may be necessary to use the ranking of each condition to determine the importance of the criteria to pick the best measure. Listed in Figure 1 are the decision trees.

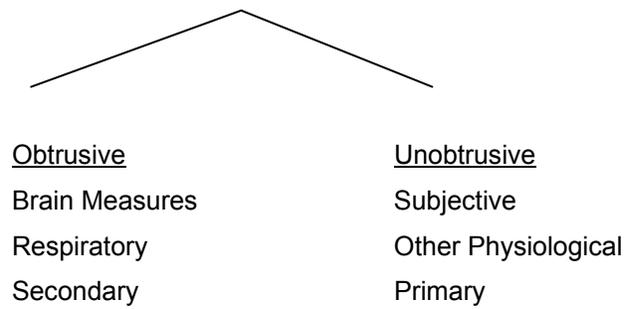
Table 2 Considerations

Consideration	Explanation
Time	Depends on when the measure is taken
Reliability	Must predict workload each time measure is used
Validity	Must be dependent on workload, not due to independent factors
Accuracy	Must mirror changes in workload
Predictability	Can determine workload from measure
Sensitivity	Must detect changes in workload
Intrusiveness	May be distracting or uncomfortable for user
Operator Acceptance	Person must accept the measure
<u>Interval of Collection</u>	Measure during, after, or continuously throughout experiment
Continuous Collection	
Interval Collection	
<u>Form of Gathering Data</u>	Different types of data may be gathered at different points
Auditory	
Written	
Machine	
Ease of Collection	Untrained Experimenter or Time Considerations
Ease of Processing	Untrained Experimenter or Time Considerations
Ease of Analysis	Untrained Experimenter or Time Considerations
Cost of implementation	Requirement of extra equipment or time increases cost
Type of Equipment Needed	Requirement of extra equipment

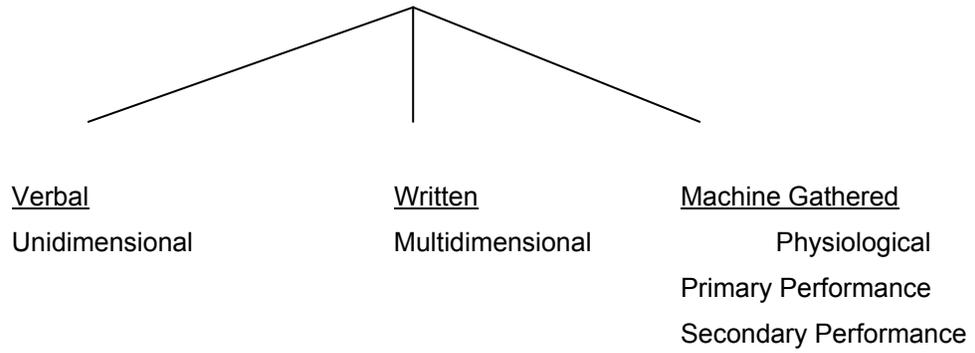
Figure 1 Decision Trees



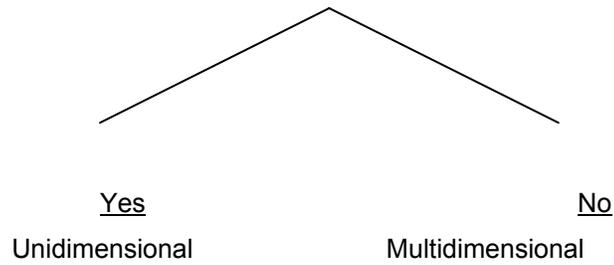
Is the test obtrusive or unobtrusive?



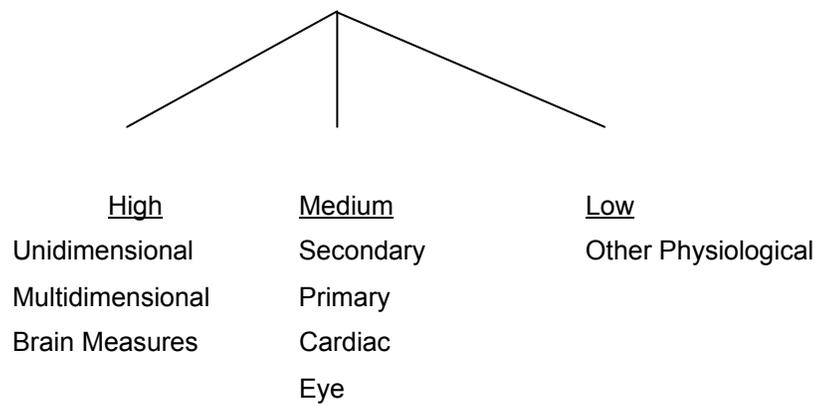
What is the form of the test?



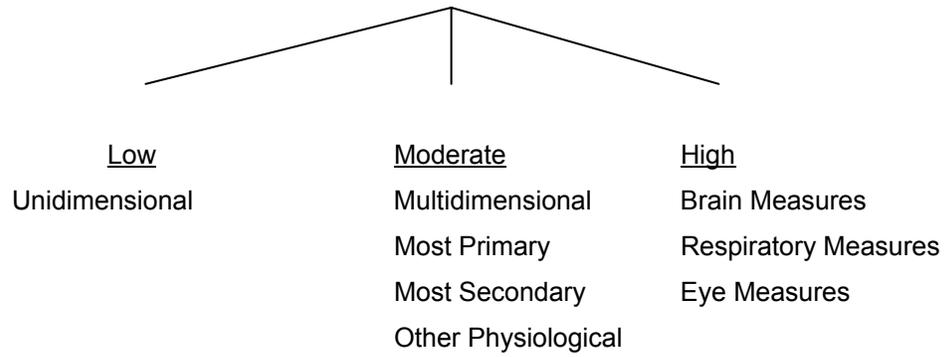
Is time a consideration?
(only applies to subjective measures)



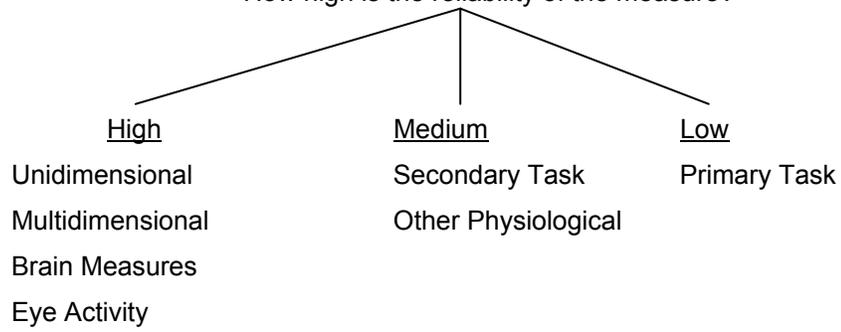
What kind of sensitivity is important?



What is the cost of implementation and analysis?



How high is the reliability of the measure?



5. RECOMMENDATIONS

For maximum accuracy, it is necessary to use multiple measures in combination. This provides more than one estimate of workload, thus cutting down on mistakes in measurement or incomplete data. It is recommended that one measure be continuous, one measure be taken at intervals during the experiment, and one measure be taken after the experiment. To do this, a physiological measure and a subjective measure must be used. Performance measures can also be used in combination with physiological or subjective measures. By taking multiple measures of workload, the chance for error due to selecting the wrong test is reduced. For example, measuring the number of eye blinks is good for measuring increases in visual workload. If this test is used for a non-visual task, then the data gathered will not be as accurate.

6. SCENARIOS

Scenario: There is a situation where there are several different subjects who drive a course in a driving simulator. They take a break and drive the same closed-circuit course in the real world. It is necessary to estimate the workload during several different small scenarios both during the simulated driving and on the closed course. It is also necessary to estimate the overall workload for the entire drive for both courses, to make a comparison of workload levels

Measure Choice: To estimate workload for this situation, it is first necessary to rank the considerations in order. Reliability is a very important measure in this situation because the estimate of workload is to be compared between the different courses. The first decision tree question should be the one about reliability. Under the high-reliability branch of the decision tree are multidimensional measures, unidimensional measures, brain measures, and eye measures. Another important consideration is time. Right after each scenario, it is necessary to give a quick estimate of workload, and there is lots of time between the different courses. The time consideration decision tree should be used next. During or right after each scenario, a unidimensional measure or a physiological measure can be used. Between the courses, a multidimensional measure should be used. It is always important to use several different measures. Right now, the decision on which measures to use during the test are brain measures, eye measures and/or unidimensional measures. After each course, a multidimensional measure should be used. Since cost of implementation and analysis is important and there is no brain measurement equipment available, brain measures cannot be used. This means an eye measure should be a continuous form of measurement; a unidimensional measure should be used after

each scenario, and a multidimensional measure should be used following the completion of each course. The best unidimensional measure is a verbal single-score rating. Some of these scales are the OW, RSME, or any other single-number estimation. The OW is recommended because there is the most information available on the validity of that scale. The best multidimensional scale is the NASA-TLX or NASA-RTLX. These are the same scales, the only difference is the scoring. If cost of analysis is a big factor, the RTLX can be used without much change in validity.

Scenario: There is a very long simulated driving scenario where there are several different episodes where workload needs to be measured. Measuring change in workload is important during each situation. Workload also needs to be estimated at the end of the entire scenario to determine overall workload. There are several different subjects, but they only drive the course one time.

Measure Choice: There are no major differences between this scenario and the last one, but now sensitivity is the most important consideration. The order of importance for the rest of the considerations does not change. Sensitivity is the most important because detecting changes in levels of workload between normal driving and the episodes is now the most important issue. Measures that are sensitive to changes in workload include unidimensional, multidimensional, and brain measures. The recommendations do not change for this scenario compared to the other one.

7. WORKS CITED

- Backs, R. W., Ryan, A. M., & Wilson, G. F. (1994a). Psychophysiological Measures of Workload During Continuous Manual Performance. *Human Factors*, 36(3), 514-531.
- Backs, R. W., & Seljos, K. A. (1994b). Metabolic and Cardiorespiratory Measures of Mental Effort - the Effects of Level of Difficulty in a Working-Memory Task. *International Journal of Psychophysiology*, 16(1), 57-68.
- Backs, R. W., Walrath, L.C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics*, 23(4), 243-254.
- Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91, 276-292.
- Boyd, S. P. (1983, October 10-14, 1983). *Assessing the validity of SWAT as a workload measurement instrument*. Paper presented at the Proceedings of the Human Factors Society - 27th Annual Meeting, Norfolk, Virginia.

- Brenner, M., Doherty, E. T., & Shipp, T. (1994). Speech Measures Indicating Workload Demand. *Aviation Space and Environmental Medicine*, 65(1), 21-26.
- Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, 42(3), 361-377.
- Byers, J. C., Bittner, A.C., Hill, S.G. (1989). *Traditional and raw task load index (TLX) correlations: are paired comparisons necessary?* Paper presented at the International Industrial Ergonomics and Safety Conference, Cincinnati, Ohio.
- Casali, J. G., Wierwille, Walter W. (1983). A Comparison of Rating Scale, Secondary-Task, Physiological, and Primary-Task Workload Estimation Techniques in a Simulated Flight Task Emphasizing Communications Load. *Human Factors*, 25(6), 623-641.
- Colle, H. A., & Reid, G. B. (1998). Context effects in subjective mental workload ratings. *Human Factors*, 40(4), 591-600.
- Colle, H. A., & Reid, G. B. (1999). Double trade-off curves with different cognitive processing combinations: Testing the cancellation axiom of mental workload measurement theory. *Human Factors*, 41(1), 35-50.
- Costa, G. (1993). Evaluation of Workload in Air-Traffic-Controllers. *Ergonomics*, 36(9), 1111-1120.
- Crabtree, M. S. (1984, October 22-26, 1984). *Benefits of using objective and subjective workload measures*. Paper presented at the Proceedings of the Human Factors Society - 28th Annual Meeting, San Antonio, Texas.
- De Waard, D. (1996). *The Measurement of Drivers' Mental Workload.*, University of Groningen, Groningen.
- Derrick, W. L. (1988). Dimensions of Operator Workload. *Human Factors*, 30(1), 95-110.
- East, J. A. (2000). *Feature Selection for Predicting Pilot Mental Workload.*, Air University, Wright-Patterson Air Force Base, Ohio.
- Eggemeier, F. T., Crabtree, M.S., La Pointe, P.A. (1983, October 10-14, 1983). *The effect of delayed report of subjective ratings of mental workload*. Paper presented at the Proceedings of the Human Factors Society - 27th Annual Meeting, Norfolk, Virginia.
- Eggemeier, F. T., Melville, B.E., Crabtree, M.S. (1984, October 22-26, 1984). *The effect of intervening task performance on subjective workload ratings*. Paper presented at the Proceedings of the Human Factors Society - 28th Annual Meeting, San Antonio, Texas.
- Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. *International Journal of Psychophysiology*, 31(2), 129-145.
- Galley, N. (1993). The Evaluation of the Electrooculogram as a Psychophysiological Measuring Instrument in the Driver Study of Driver Behavior. *Ergonomics*, 36(9), 1063-1070.

- Gevins, A., Leong, H., Du, R., Smith, M. E., Le, J., Duroseu, D., Zhang, J., & Libove, J. (1995). Towards Measurement of Brain-Function in Operational Environments. *Biological Psychology*, 40(1-2), 169-186.
- Gopher, D., Braune, R. (1984). On the Psychophysics of Workload: Why Bother with Subjective Measures. *Human Factors*, 25(5), 519-532.
- Gopher, D., Donchin, E. (1986). Workload - An examination of the concept. *Handbook of perception and human performance*, 2, 41-49.
- Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation Space and Environmental Medicine*, 69(4), 360-367.
- Hendy, K. C., Hamilton, K. M., & Landry, L. N. (1993). Measuring Subjective Workload - When Is One Scale Better Than Many. *Human Factors*, 35(4), 579-601.
- Hicks, T. G., Wierwille, W.W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors*, 21(2), 129-143.
- Hill, S. G., Lavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., & Christ, R. E. (1992). Comparison of 4 Subjective Workload Rating-Scales. *Human Factors*, 34(4), 429-439.
- Johannsen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., Wickens, C. (1979). Final Report of the Experimental Psychology Group. In N. Moray (Ed.), *Mental Workload* (Vol. 8). New York: Plenum Press.
- Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload measure. *Biological Psychology*, 34, 237-257.
- Jorna, P. G. A. M. (1993). Heart-Rate and Workload Variations in Actual and Simulated Flight. *Ergonomics*, 36(9), 1043-1054.
- Kantowitz, B. H. (1992). Selecting Measures for Human-Factors Research. *Human Factors*, 34(4), 387-398.
- Lee, D. H., Parks, K.S. (1990). Multivariate analysis of mental and physical load components in sinus arrhythmia scores. *Ergonomics*, 33(1), 35-47.
- May, J. G., Kennedy, R.S., Williams, M.C., Dunlap, W.P., Brannan, J.R. (1990). Eye movement indices of mental workload. *Acta Psychologica*, 75, 75-89.
- Moray, N. (1979). Models and Measures of Mental Workload. In N. Moray (Ed.), *Mental Workload* (Vol. 8). New York: Plenum Press.
- Muckler, F. A., & Seven, S. A. (1992). Selecting Performance-Measures - Objective Versus Subjective Measurement. *Human Factors*, 34(4), 441-455.
- Mulder, G. (1979). Mental load, mental effort and attention. In N. Moray (Ed.), *Mental Workload: Its Theory and Measurement*. New York and London: Plenum Press.

- Paas, F., & Vanmerrienboer, J. J. G. (1993). The Efficiency of Instructional Conditions - an Approach to Combine Mental Effort and Performance-Measures. *Human Factors*, 35(4), 737-743.
- Park, P., Cha, D. (1998). *Comparison of Subjective Mental Workload Assessment Techniques for the Evaluation of In-Vehicle Navigation System Usability*. Suwon, Korea: Ajou University.
- Rehmann, A. J. (1995). *Handbook of Human Performance Measures and Crew Requirements for Flightdeck Research* (DOT/FAA/CT-TN95/49).
- Rokicki, S. M. (1995). Psychophysiological Measures Applied to Operational Test and Evaluation. *Biological Psychology*, 40(1-2), 223-228.
- Roscoe, A. H. (1992). Assessing Pilot Workload - Why Measure Heart-Rate, Hrv and Respiration. *Biological Psychology*, 34(2-3), 259-287.
- Roscoe, A. H. (1993). Heart-Rate as a Psychophysiological Measure for in-Flight Workload Assessment. *Ergonomics*, 36(9), 1055-1062.
- Sabbatini, R. M. E. (1997). *The History of the Electroencephalogram*. Brain and Mind Magazine. Available: <http://www.epub.org.br/cm/n03/tecnologia/historia.htm> [2001, 02-27-01].
- Sirevaag, E. J., Kramer, A. F., Wickens, C. D., Reisweber, M., Strayer, D. L., & Grenell, J. F. (1993). Assessment of Pilot Performance and Mental Workload in Rotary Wing Aircraft. *Ergonomics*, 36(9), 1121-1140.
- Stern, J. A., Skelly, J.J. (1984, October 22-26, 1984). *The eye blink and workload considerations*. Paper presented at the Proceedings of the Human Factors Society - 28th Annual Meeting, San Antonio, Texas.
- Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5), 740-748.
- Van Orden, K., F., Makeig, Scott, Jung, Tzyy-Ping, Limbert, Wendy. (1999). *Eye activity correlates of workload during a visuospatial memory task* (Special Documents 186). San Diego: Defense Technical Information Center.
- Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, 42(3), 323-342.
- Veltman, J. A., Gaillard, A.W.K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5), 656-669.
- Vidulich, M. A., & Wickens, C. D. (1986). Causes of Dissociation between Subjective Workload Measures and Performance - Caveats for the Use of Subjective Assessments. *Applied Ergonomics*, 17(4), 291-296.
- Wickens, C. D., Gordon, S.E., Liu, Y. (1998). An Introduction to Human Factors Engineering.,, 392-395.

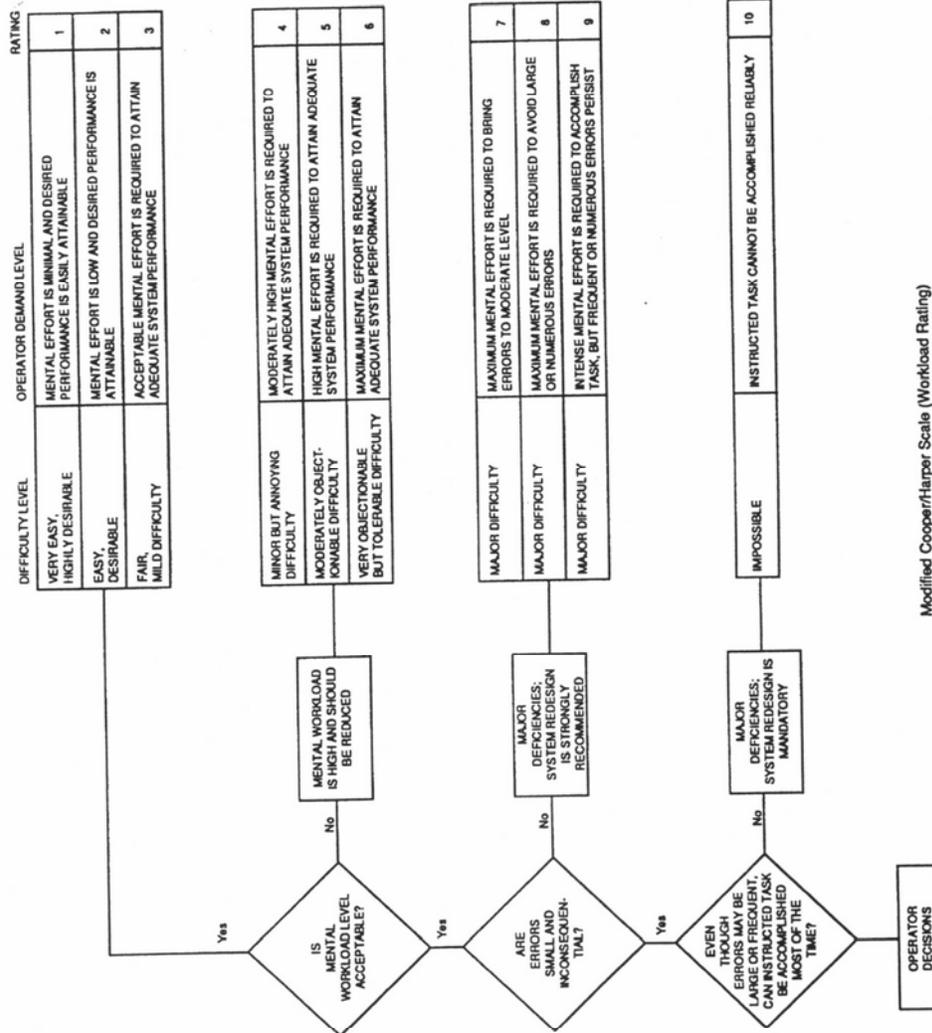
- Wierwille, W. W., Casali, J.G. (1983, October 10-14, 1983). *A validated rating scale for global mental workload measurement applications*. Paper presented at the Proceedings of the Human Factors Society - 27th Annual Meeting, Norfolk, Virginia.
- Wierwille, W. W., Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35(2), 263-281.
- Wierwille, W. W., Rahimi, M., Casali, J.G. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human Factors*, 27(5), 489-502.
- Wilson, G. F. (1992). Applied Use of Cardiac and Respiration Measures - Practical Considerations and Precautions. *Biological Psychology*, 34(2-3), 163-178.
- Wilson, G. F. (1993). Air-to-Ground Training Missions - a Psychophysiological Workload Analysis. *Ergonomics*, 36(9), 1071-1087.
- Wilson, G. F., Fullenkamp, P., & Davis, I. (1994). Evoked-Potential, Cardiac, Blink, and Respiration Measures of Pilot Workload in Air-to-Ground Missions. *Aviation Space and Environmental Medicine*, 65(2), 100-105.
- Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. *Human Factors*, 30(1), 111-120.

APPENDIX A

- A1: Modified Cooper-Harper (MCH)**
- A2: Overall Workload (OW)**
- A3: NASA Task Load Index (NASA-TLX)**
- A4: Subjective Workload Assessment Technique (SWAT)**
- A5: Instantaneous Self Assessment (ISA)**
- A6: Rating Scale Mental Effort (RSME)**
- A7: Activation Scale**
- A8: The Verbal Online Subjective Opinion (VOSO)**
- A9: Cooper-Harper**
- A10: Bedford Workload Scale**
- A11: Honeywell Cooper-Harper**
- A12: Equal-Appearing Intervals**
- A13: Driving Activity Load index (DALI)**
- A14: Multi-Descriptor (MD)**
- A15: Analytical Hierarchy Process (AHP)**
- A16: The Workload/Compensation/Interference/Technical Effectiveness (WCI/TE)**

Appendix A1: Modified Cooper-Harper (MCH)

Study # _____
 Trial # _____
 Study Date _____
 Participant # _____



Modified Cooper/Harper Scale (Workload Rating)

Appendix A2: Overall Workload (OW)

Task or Mission Segment: _____

Please mark an "X" on the line which best corresponds to how you rate your Overall Workload.

Overall Workload:	
	Very Low Very High

Appendix A3: NASA Task Load Index (NASA-TLX)

SUBJECT INSTRUCTIONS SOURCES-OF-WORKLOAD EVALUATION

Throughout this experiment the rating scales are used to assess your experiences in the different task conditions. Scales of this sort are extremely useful, but their utility suffers from the tendency people have to interpret them in individual ways. For example, some people feel that mental or temporal demands are the essential aspects of workload regardless of the effort they expended on a given task or the level of performance they achieved. Others feel that if they performed well the workload must have been low and if they performed badly it must have been high. Yet others feel that effort or feelings of frustration are the most important factors in workload; and so on. The results of previous studies have already found every conceivable pattern of values. In addition, the factors that create levels of workload differ depending on the task. For example, some tasks might be difficult because they must be completed very quickly. Others may seem easy or hard because of the intensity of mental or physical effort required. Yet others feel difficult because they cannot be performed well, no matter how much effort is expended.

The evaluation you are about to perform is a technique that has been developed by NASA to assess the relative importance of six factors in determining how much workload you experienced. The procedure is simple: You will be presented with a series of pairs of rating scale titles (for example, Effort vs. Mental Demands) and asked to choose which of the items was more important to your experience of workload in the task(s) that you just performed. Each pair of scale titles will appear on a separate card.

Circle the Scale Title that represents the more important contributor to workload for the specific task(s) you performed in this experiment.

After you have finished the entire series we will be able to use the pattern of your choices to create a weighted combination of the ratings from that task into a summary workload score. Please consider your choices carefully and make them consistent with how you used the rating scales during the particular task you were asked to evaluate. Don't think that there is any *correct* pattern: we are only interested in your opinions.

If you have any questions, please ask them now. Otherwise, start whenever you are ready. Thank you for your participation.

RATING SCALE DEFINITIONS

Title	Endpoints	Descriptions
MENTAL DEMAND	<i>Low/High</i>	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, exacting, or forgiving?
PHYSICAL DEMAND	<i>Low/High</i>	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	<i>Low/High</i>	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	<i>Good/Poor</i>	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	<i>Low/High</i>	How hard did you have to work (mentally and physically) accomplish your level of performance?
FRUSTRATION LEVEL	<i>Low/High</i>	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

Effort or Performance	Temporal Demand or Frustration
Temporal Demand or Effort	Physical Demand or Frustration
Performance or Frustration	Physical Demand or Temporal Demand

<p>Physical Demand or Performance</p>	<p>Temporal Demand or Mental Demand</p>
---	---

Frustration or Effort	Performance or Mental Demand
Performance or Temporal Demand	Mental Demand or Effort
Mental Demand or Physical Demand	Effort or Physical Demand

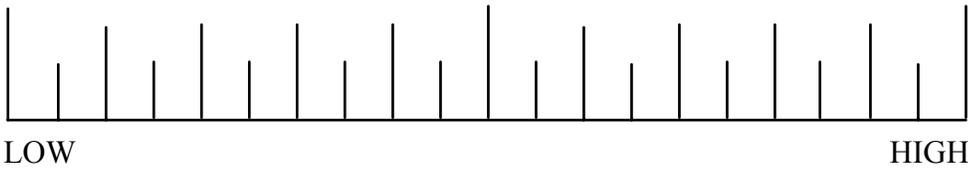
Frustration
or
Mental Demand

Study _____
Study Date _____

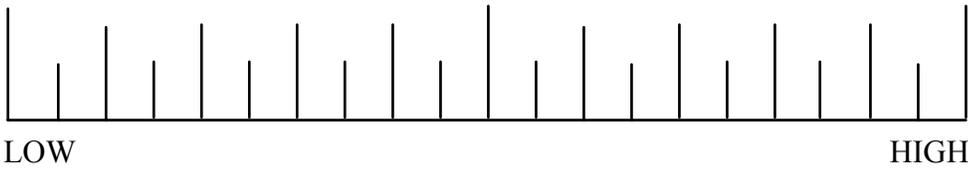
Trial# _____
Participant # _____

RATING SHEET

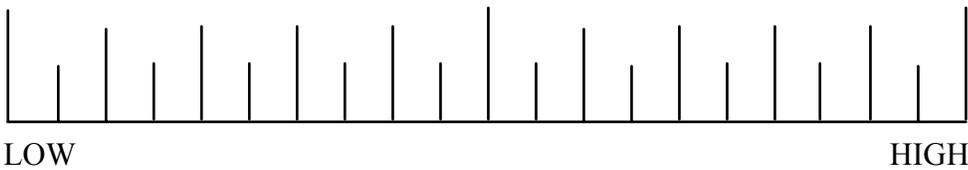
MENTAL DEMAND



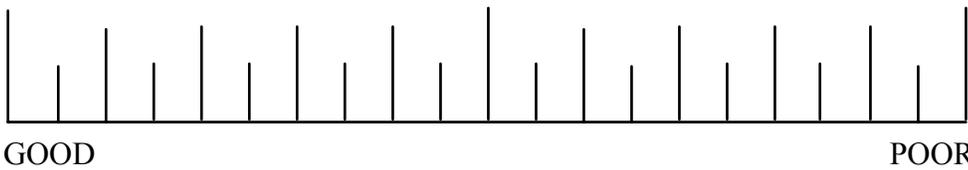
PHYSICAL DEMAND



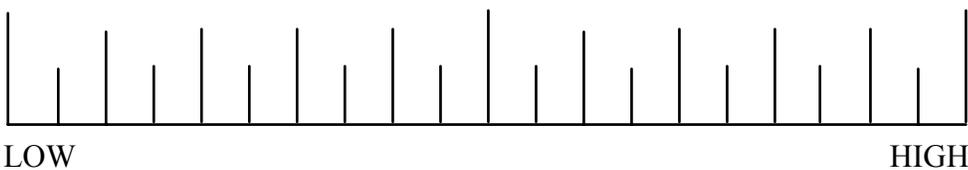
TEMPORAL DEMAND



PERFORMANCE



EFFORT



FRUSTRATION



NASA Task Load

Scoring Instructions

- 1) In the tally column, record a mark for each time a participant chose a scale on the evaluation cards (e.g., each time the participant circled "Mental Demand" on a comparison card, the experimenter puts a mark on the "Mental Demand" row of the tally column).
- 2) Sum the number of tally marks for each scale in the tally column, and record the number of marks in the weight column. Weights cannot equal more than 5.
- 3) Sum all weights and record this number in the "Total Count" box. The total count must equal 15. If it does not equal 15, a miscalculation has occurred.
- 4) In the Raw Ratings column, record the responses from the Rating Sheet for each scale. The Rating Sheet provides a vertical line anchored at 0 and 100 and divided into intervals of 5 for each scale. To determine the number associated with a response, count the number of intervals from the left assuming that the left most bar is NOT counted, and multiply by 5 (e.g., if the participant marked an "X" on the fourth interval bar from the left, as below, the score would be $4 \times 5 = 20$).



If a participant marks between two interval bars, the value of the right bar is used (i.e., round up). The maximum Raw Rating for any one scale is 100.

- 5) Multiply the Raw Rating by the Weight for that scale. Record this number in the Adjusted Rating column.
- 6) Sum the Adjusted Ratings and record the total in the Sum "Adjusted Rating" box.
- 7) Divide the number in the Sum "Adjusted Rating" box by 15 to obtain the overall weighted workload score. Record the resulting quotient in the **WEIGHTED RATING** box.

Study _____
 Study Date _____

Trial# _____
 Participant # _____

SOURCE-OF-WORKLOAD TALLY & WEIGHTED RATING WORKSHEET				
Scale	Tally	Weight	Raw Rating	Adjusted Rating (Weight x Raw)
MENTAL Demand				
PHYSICAL Demand				
TEMPORAL Demand				
PERFORMANCE				
EFFORT				
FRUSTRATION				
Total Count =				
Note: Total Count is included as a check. Total Count must equal 15 or a miscalculation has occurred. Also, no weight can have a value greater than 5.				
			Sum "Adjusted Rating" Column =	
			WEIGHTED RATING = (Sum of Adjusted Ratings)/15	

Appendix A4: Subjective Workload Assessment Technique (SWAT)

SWAT Dimensions and Levels

TIME STRESS:

- 1) Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.
- 2) Occasionally have spare time. Interruptions or overlap among activities occur frequently.
- 3) Almost never have spare time. Interruptions or overlap among activities are very frequent or occur all of the time.

MENTAL EFFORT:

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

PSYCHOLOGICAL STRESS:

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

Appendix A5: Instantaneous Self Assessment (ISA)

Ratings: Indicate Level

2.1.1.1. Excessive

2.1.1.2. High

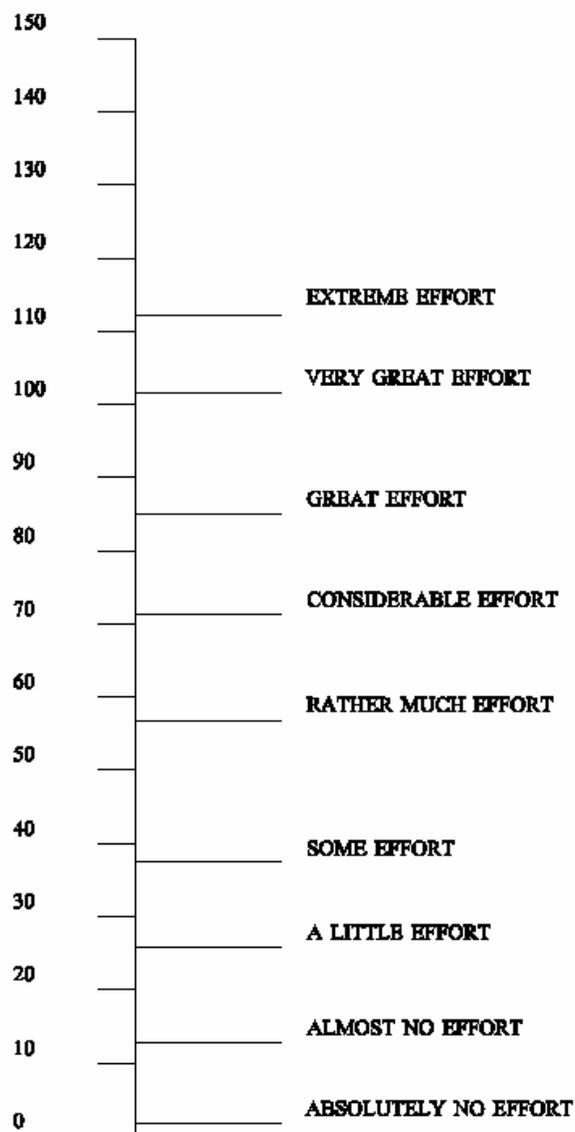
2.1.1.3. Comfortable

2.1.1.4. Relaxed

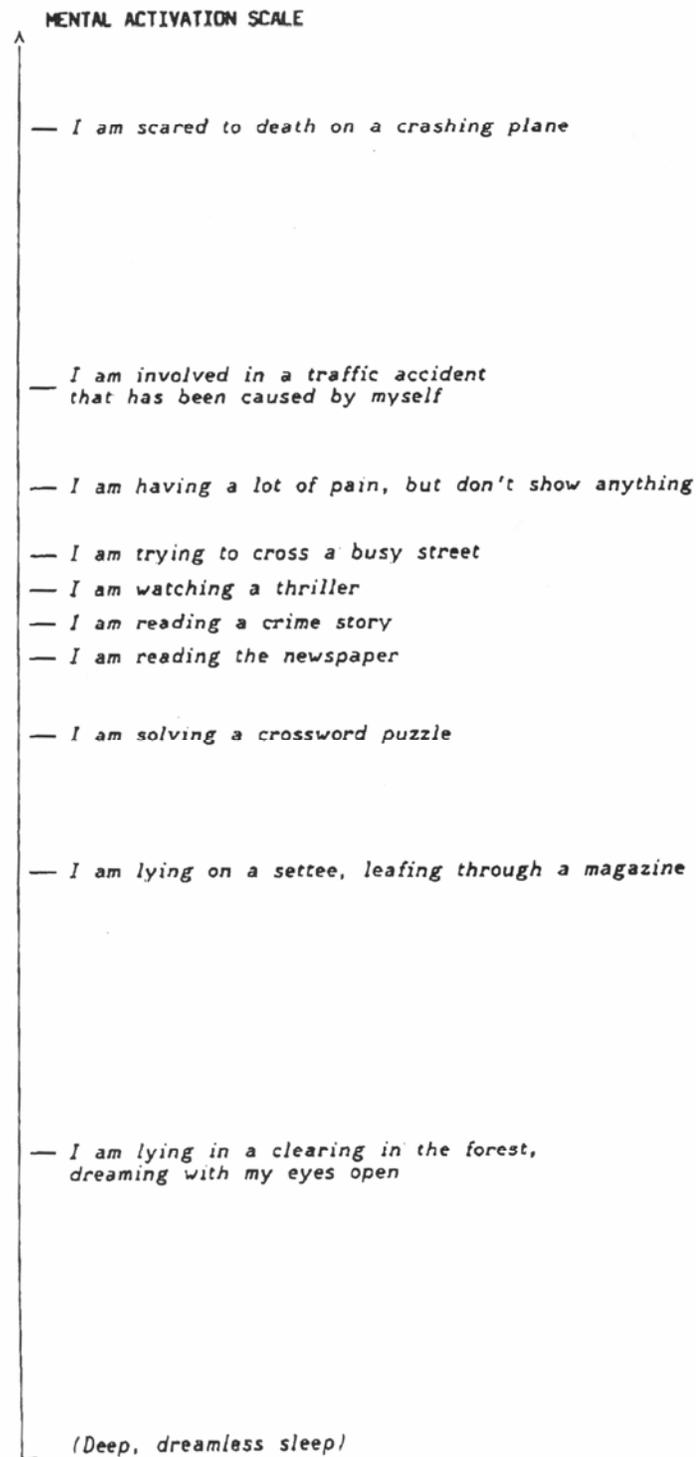
Under-Utilized

Rating Scale Mental Effort

Please indicate, by marking the vertical axis below, how much effort it took for you to complete the task you've just finished

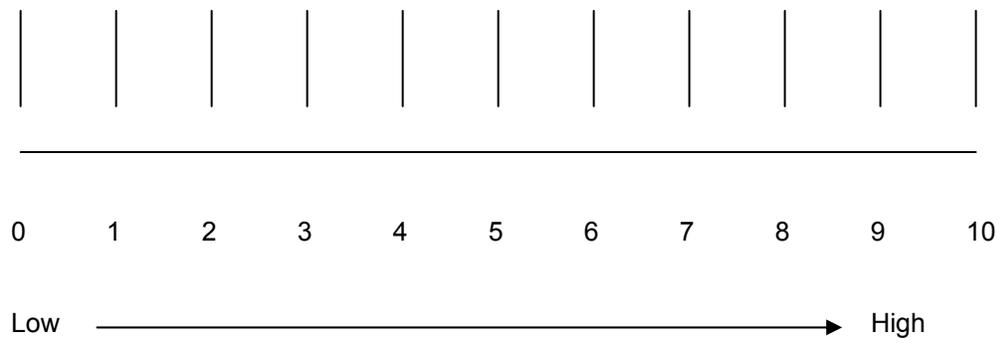


Appendix A7: Activation Scale

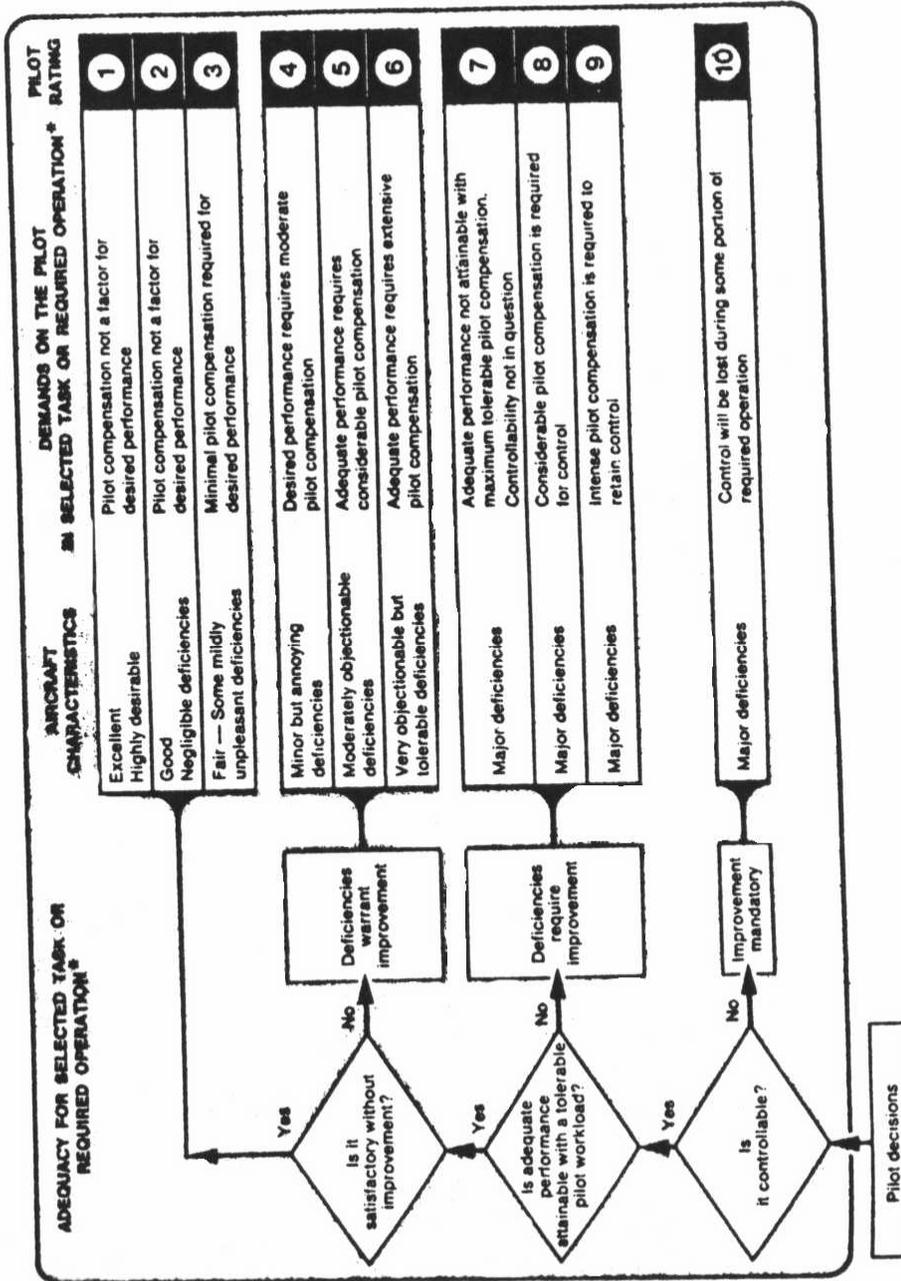


Appendix A8: The Verbal Online Subjective Opinion (VOSO)

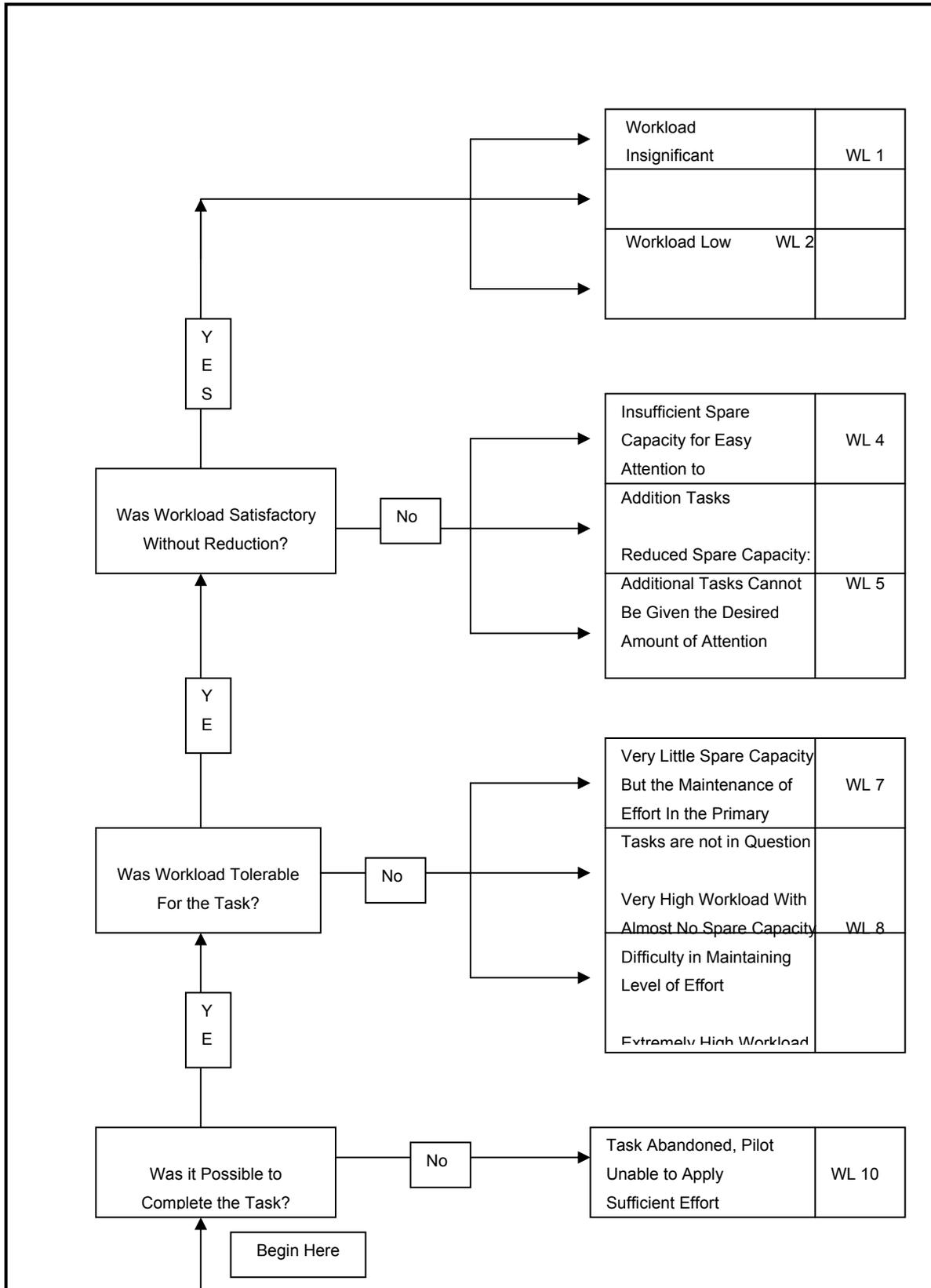
Level of Workload:



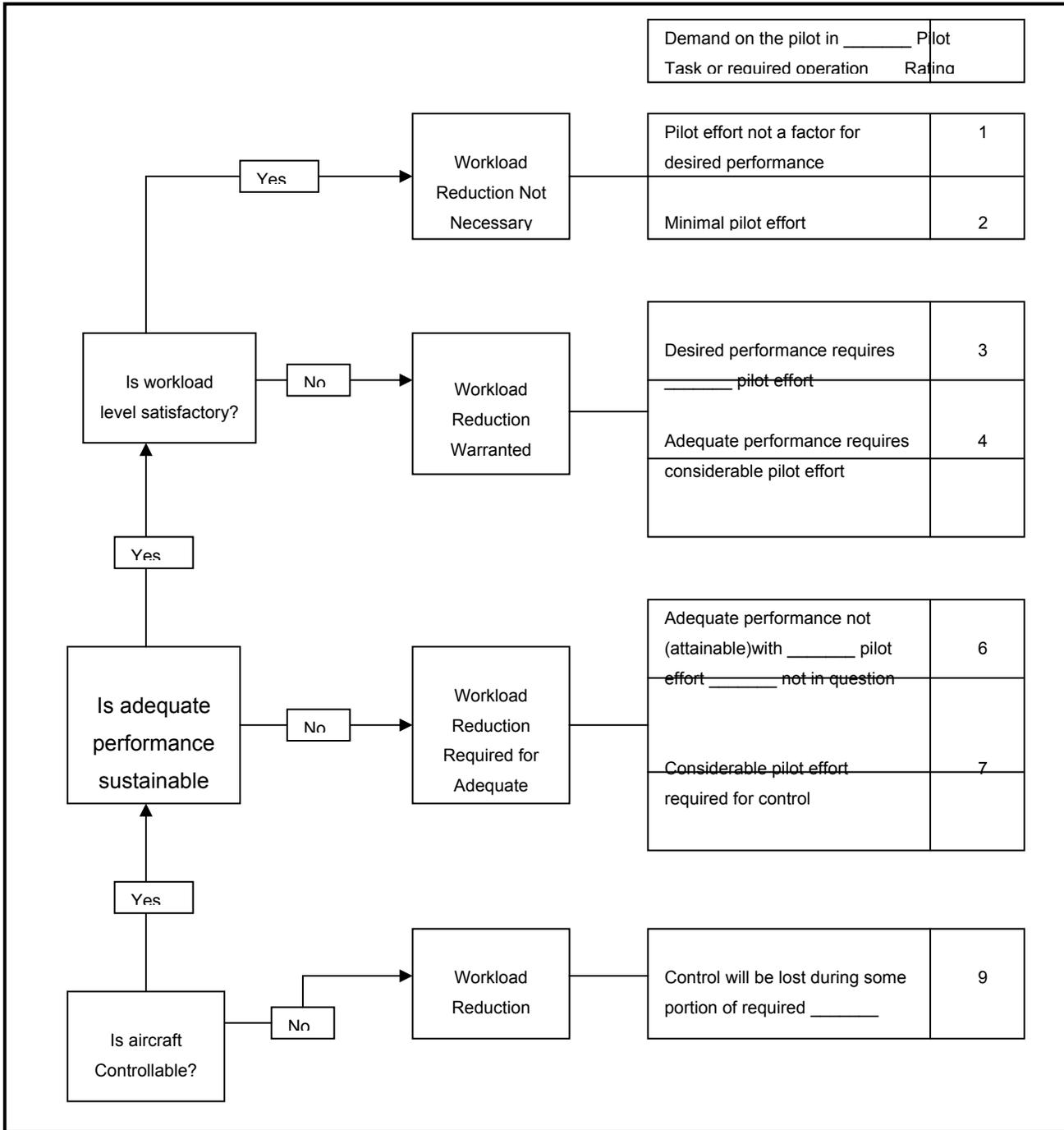
Appendix A9: Cooper-Harper



Appendix A10: Bedford Workload Scale



Appendix A11: Honeywell Cooper-Harper



Appendix A12: Equal-Appearing Intervals

(DO NOT HAVE)

Appendix A13: Driving Activity Load index (DALI)

(DO NOT HAVE)

Appendix A14: Multi-Descriptor (MD)

(DO NOT HAVE)

Appendix A15: Analytical Hierarchy Process (AHP)

