

Spring 2003

## An Experimental Investigation of the Differences in Subjective Pilot Workload across Simulated and Real Flight Conditions

Todd V. Denning  
*Embry-Riddle Aeronautical University - Daytona Beach*

Follow this and additional works at: <https://commons.erau.edu/db-theses>



Part of the [Ergonomics Commons](#)

---

### Scholarly Commons Citation

Denning, Todd V., "An Experimental Investigation of the Differences in Subjective Pilot Workload across Simulated and Real Flight Conditions" (2003). *Theses - Daytona Beach*. 42.  
<https://commons.erau.edu/db-theses/42>

This thesis is brought to you for free and open access by Embry-Riddle Aeronautical University – Daytona Beach at ERAU Scholarly Commons. It has been accepted for inclusion in the Theses - Daytona Beach collection by an authorized administrator of ERAU Scholarly Commons. For more information, please contact [commons@erau.edu](mailto:commons@erau.edu).

AN EXPERIMENTAL INVESTIGATION OF THE DIFFERENCES IN SUBJECTIVE  
PILOT WORKLOAD ACROSS SIMULATED AND REAL FLIGHT CONDITIONS

by

Todd V. Denning

A Thesis Submitted to the  
Department of Human Factors & Systems  
in Partial Fulfillment of the Requirements for the Degree of  
Master of Science in Human Factors & Systems

Embry-Riddle Aeronautical University  
Daytona Beach, Florida  
Spring 2003

UMI Number: EP32082

### INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI<sup>®</sup>

---

UMI Microform EP32082  
Copyright 2011 by ProQuest LLC  
All rights reserved. This microform edition is protected against  
unauthorized copying under Title 17, United States Code.

---

ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

AN EXPERIMENTAL INVESTIGATION OF THE DIFFERENCES IN SUBJECTIVE  
PILOT WORKLOAD ACROSS SIMULATED AND REAL FLIGHT CONDITIONS

by

Todd V Denning

A Thesis Submitted to the  
Department of Human Factors & Systems  
in Partial Fulfillment of the Requirements for the Degree of  
Master of Science in Human Factors & Systems

Embry-Riddle Aeronautical University  
Daytona Beach, Florida  
Spring 2003

AN EXPERIMENTAL INVESTIGATION OF THE DIFFERENCES IN SUBJECTIVE  
PILOT WORKLOAD ACROSS SIMULATED AND REAL FLIGHT CONDITIONS

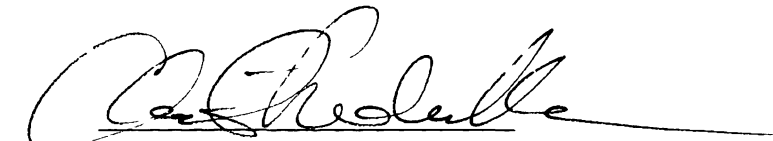
by

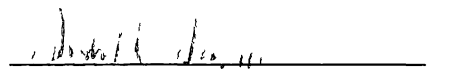
Todd V. Denning

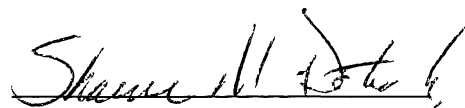
This thesis was prepared under the direction of the candidate's thesis committee chair, Steve Hall, Ph.D., Department of Human Factors & Systems, and has been approved by the members of the thesis committee. It was submitted to the Department of Human Factors & Systems and has been accepted in partial fulfillment of the requirements for the degree of Master of Science in Human Factors & Systems.

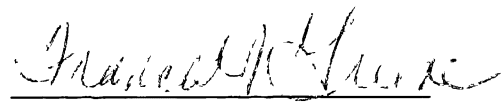
THESIS COMMITTEE:


  
Steve Hall, Ph.D., Chair

  
Christina Frederick-Recascino, Ph.D., Member

  
Gerald Gamache, Ph.D., Member

  
MS HFS Program Coordinator

  
Department Chair, Department of Human Factors & Systems

  
Assistant Dean of Student Academics

## ACKNOWLEDGEMENTS

The author would like to thank his committee members, Dr. Steve Hall for constant encouragement, editing skills, and statistics knowledge, Dr. Christina Frederick for being available to bounce ideas off of, and Dr. Gerald Gamache for opening my eyes to the wonderful world of human factors. He also wishes to thank the SATS team members, Steve Hampton, and Bob Peak, for getting the project off the ground, and Lawrence Landa and Susan Karkman for data collection and support.

Many thanks go out to the entire faculty at Embry-Riddle for providing a truly rewarding educational experience. Finally, thanks also to GOD, my wife Jennifer, and my family and friends for their patience, encouragement and strength.

## ABSTRACT

Author: Todd V Denning  
Title: An Experimental Investigation of the Differences in Subjective Pilot Workload Across Simulated and Real Flight Conditions  
Institution: Embry-Riddle Aeronautical University  
Degree: Master of Science in Human Factors & Systems  
Year: 2003

An investigation was undertaken to determine the difference in workload between simulated and real flight conditions. The results from the Modified Cooper-Harper and NASA-TLX did not show significance, however, the theoretical implications from the NASA-TLX subscales were of interest. As this is the first study comparing these two environments utilizing subjective workload measures, more research needs to take place in order to provide reliable and valid findings.

## TABLE OF CONTENTS

|   |      |
|---|------|
| ACKNOWLEDGEMENTS.....                         | iii  |
| ABSTRACT.....                                 | iv   |
| List of Tables.....                           | vii  |
| List of Figures.....                          | viii |
| INTRODUCTION.....                             | 1    |
| Statement of Problem.....                     | 3    |
| REVIEW OF LITERATURE.....                     | 4    |
| Properties of Workload Measures.....          | 6    |
| Subjective Measures of Workload.....          | 7    |
| Psychophysiological Measures of Workload..... | 9    |
| Performance Measures of Workload.....         | 10   |
| Analytic Measures of Workload.....            | 12   |
| Choosing Workload Measures.....               | 13   |
| NASA-Task Load Index (NASA-TLX).....          | 16   |
| Modified Cooper-Harper (MCH).....             | 20   |
| Workload in Aviation Systems.....             | 22   |
| Summary of Review of Literature.....          | 24   |
| Statement of Hypothesis.....                  | 25   |
| METHOD.....                                   | 26   |
| Participants.....                             | 26   |
| Apparatus.....                                | 26   |
| Design.....                                   | 27   |



|                                      |    |
|--------------------------------------|----|
| Procedures.....                      | 27 |
| RESULTS.....                         | 31 |
| Data Analysis.....                   | 31 |
| Discussion.....                      | 35 |
| References.....                      | 39 |
| APPENDIX A.....                      | 45 |
| NASA-Task Load Index (NASA-TLX)..... | 45 |
| Modified Cooper-Harper (MCH).....    | 47 |

## LIST OF TABLES

|          |  |    |
|----------|--|----|
| Table 1. | NASA-TLX Rating Scale Definitions.....                   | 18 |
| Table 2. | Frequency Chart for Modified Cooper-Harper.....          | 31 |
| Table 3. | NASA-TLX t-scores for Overall Workload and Subscales.... | 34 |
| Table 4  | Estimates of Effect Size, Cohen's <i>d</i> .....         | 34 |

## LIST OF FIGURES

|           |   |    |
|-----------|---|----|
| Figure 1. | Conceptual Framework for variables that influence Human Performance and Workload..... | 16 |
| Figure 2. | Modified Cooper-Harper frequency of ratings .....                                     | 31 |
| Figure 3. | Mean Ratings for overall NASA-TLX and subscales .....                                 | 32 |
| Figure 4. | Average by group for Mental Demand subscale.....                                      | 36 |

## INTRODUCTION

In the past 25 years, new technology has significantly lowered the physical workload pilots are exposed to when flying modern aircraft. For example, autopilot and flight management systems (FMS) can literally fly aircraft from one point to another and even control the throttles. However, the pilot must program and monitor such systems resulting in an increase in mental or psychological workload, in the aviation domain, increasing the amount of workload a pilot must undertake will eventually lead to a decrease in performance. Since the introduction of glass cockpits, which replaced old analog gauges with new streamlined multi-purpose and multi-function displays, and new FMS, pilots have reported increases in mental workload and decreases in physical workload, as pilots are reduced to operators and passive monitors of the cockpit systems (Mouloua, Deaton, & Hitt, 2001). For years, researchers have been seeking the proper tradeoff between technology and the amount of mental workload pilots can endure safely.

Traditionally, in order to assess this tradeoff, workload has been measured during simulated missions (Wierwille & Casali, 1983; Casali & Wierwille, 1984; Battiste & Bortolussi, 1988; Kilmer, Knapp, Burdsal, Borresen, Bateman, & Malzahn, 1988; Wilson & Badeau, 1992; Aretz, Shacklett, Acquaro & Miller, 1995; Leino, Leppaluoto, Huttenen, Ruokonen, & Kuronen, 1995; Ylonen, Lyytinen, Leino, Leppaluoto, & Kuronen, 1997; Wilson, Skelly, & Purvis, 1999; Magnusson, 2002; Veltman, 2002). The simulators resemble the cockpit of the airplane, and scenarios are carried out to determine what amount of mental workload can be handled successfully, that is, without performance decrement. During the simulated missions, data are gathered and interpreted in order to determine the change in performance that has occurred as a result of the increase or decrease of automation or some other artificially injected apparatus. It is

these changes in task performance that are so critical and must be assessed in a simulated environment before system modifications are made in the cockpit. Any negative effects a pilot encounters as a result of system modification, or if the pilot exceeds his or her workload capacity on a particular aspect of the flight (e.g., navigation, monitoring, frustration, etc.), should be taken into consideration when determining the level of automation suitable for the cockpit. The amount of automation in an aircraft should not hinder the crew's ability to pilot the plane. If the workload in a simulated scenario is very high, or if some particular aspect of workload is very high, and results in decreased flight performance, chances are that when confronted with a similar scenario in actual flight conditions, the pilot and crew will pay the price for the high workload in the form of decreased performance and increasing chance of error. The key question is whether or not the workload in the simulator accurately reflects the workload in the live aircraft, quantitatively, and qualitatively. That is, are the workload components the same in both environments?

In all, the best way to find these performance decrements is to look at the workload encountered in both environments. There are many factors that may make the degree of workload experienced by the crew during simulation different from the real-world aircraft. First, the physical stimulation, such as the noise, vibration, motion, and physical fidelity of the simulation, is qualitatively and quantitatively different from real flight. Also, pilots flying in a simulator understand that the physical consequences of poor decision making or improper flight control input are non-existent. These factors may explain why some measures of workload, such as changes in heart rate, when measured in the simulated environment often do not correlate strongly with the same

measure taken during real flight (Wilson, Purvis, Skelly, Fullenkamp, & Davis, 1987). Some have suggested that the “absence of danger” does not explain this phenomenon, but that the difference is more pronounced with familiar and relatively simple simulators (Ylonen et al., 1997). Clearly there is no consensus on why the workload in the aircraft tends to be higher than the corresponding simulator, however, several studies suggest this to be the case (Leino et al., 1995; Wilson & Badeau, 1992; Wilson et al., 1999; Ylonen, et al., 1997). In all the above cases, physiological measures were used to assess the workload. If the task were undertaken utilizing self-report measures, it should further illuminate the possible difference between workload as experienced in simulation and live aircraft.

### **Statement of Problem**

Though simulations have numerous advantages over real-flight training and measurement scenarios, when it comes to workload there is no clear consensus as to whether or not the amount or the components encountered in a simulation are similar to that experienced in the cockpit. There is conflicting research pointing to the conclusion that mental workload and reactions may not be analogous across simulated and real flight environments (Leino et al., 1995; Wilson & Badeau, 1992; Wilson et al., 1999; Ylonen et al., 1997). Wilson and Badeau (1992) and Leino et al., (1995) have found that the amount of workload in a simulation is not the same amount of workload encountered in the cockpit. Wilson and Badeau (1992) found that using heart rate and eye blink, inferences based on laboratory data could not be generalized to actual flight, since that data is inherently different. Accordingly, Leino et al., (1995) found that plasma levels in

pilots measured after simulated flight were different in makeup than plasma levels measured after real flight. In both studies, the measured workload in the actual cockpit was higher than that measured in the simulation. The experiments conducted analyzing the amount of workload across simulated and real flight conditions thus far have been using physiological measures of workload, such as heartbeat, EKG, plasma levels, and Galvanic Skin Response (GSR). To date, no published studies have used subjective measures to compare the workload in these two environments. The present study seeks to measure the amount of workload in simulated and actual flight using subjective measures of workload. If the resulting data suggest that simulation induces either too much or too little workload for a given set of tasks as compared to real flight, then researchers should consider the extent to which simulation can be used to assess the level of workload experienced by a crew in a given situation, especially if the goal of the study is to find the combination of factors and events that produces some maximum workload component.

### **Review of the Literature**

Two overall methods of workload measurement have been extensively discussed in the literature: self-report and physiological (Charlton, 2002). Although once abandoned by professionals in favor of behaviorism, today, there is more reason than ever to admit cognitive states such as mental workload into the testing and evaluation of all performance based systems, especially those in the aviation domain. Charlton (2002, p. 98) defines mental workload as “the amount of cognitive or attentional resources being expended at a given point in time”. Other definitions of workload have also been

presented in the literature Roscoe (1978) defined workload as “the integrated mental and physical effort necessary to meet the demands of the flight task” Hart and Staveland (1988, p 140) see workload as “a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance”, and O’Donnell & Eggemeier, 1986 p 2, see workload as “that portion of the operator’s limited capacity actually required to perform a particular task”

The definition of workload used for the present study is a combination of the Hart and Staveland definition and the definition put forth by O’Donnell and Eggemeier. Inherent in these definitions lies the assumption that human operators have limited processing capability. Once the demand for this processing exceeds the limitations of the operator, performance decrements result. Having an indication of the expended workload or capacity would therefore be helpful in developing systems or introducing automation that will decrease workload. Once developers and engineers are able to grasp an accurate picture of the workload in a particular system, they can estimate the amount of additional workload that can be safely introduced into the system.

Although different definitions of workload abound, psychologists and engineers have developed four different measures that can be used to quantify workload, they are subjective, behavioral (performance), psychophysiological, and analytic. Subjective measures are designed to elicit the operator’s perceptions about the mental workload of the system, while performance measures are normally viewed as part of the system of interest (such as making it to a certain waypoint in an aircraft simulation). Similarly, psychophysiological measures record body changes related to the demands of the task being performed, while analytic measures are models mostly for predictive and



evaluative purposes (Tsang & Wilson, 1997). A review of these four measures follows as well as the reasoning behind choosing subjective measures for this study. First, however, because of the numerous situations in which workload measures have been implemented, the properties of each should be examined in order to determine that the appropriate measure is chosen.

### **Properties of Workload Measures**

There are eight properties of workload measures that have been investigated: sensitivity, diagnosticity, intrusiveness, validity, reliability, ease of use, and operator acceptance (Eggemeier, Wilson, Kramer, & Damos, 1991). Of these, the three most important are sensitivity, diagnosticity, and intrusiveness. Sensitivity refers to the “capability of a technique to detect differences in the levels of workload that are associated with performance of a task or system function” (Eggemeier et al., 1991, p. 208). Sensitivity is one of the most important properties of assessment techniques, because a measure must be able to discriminate between different levels of workload. Diagnosticity refers to the “capability of a measure to discriminate among different types of mental workload (p. 209). A measure is said to have diagnosticity if it distinguishes among different varieties of resources. Choosing a workload measure based on the degree of diagnosticity will depend entirely on the objective of the assessment. If the objective is to determine which aspect of workload contributes the most to overall workload, then a multi-dimensional workload measure should be considered. The third and most important property is the intrusiveness of the measure. Intrusiveness refers to “any disruption in ongoing primary-task performance that results from application of a

workload measurement technique” (p. 209). Intrusiveness has demonstrated to be a problem primarily in the administration of a secondary-task while utilizing performance based workload assessments. One other important property of any subjective workload measure is its reliability. There have been relatively few true evaluations of reliability (split-half, alternate forms, and test-retest); the reliability can be estimated by comparing results obtained from similar studies (Eggemeier et. al., 1991). Each of these properties of workload measures must be taken into account when deciding which measure would be most appropriate for an analysis.

### **Subjective Measures of Workload**

The key behind subjective measures lies in requiring the operator to rate the level of mental effort they feel is needed in order to accomplish a task. These methods take into account the context of the task, as well as the skill level of the operators. Frequently, rating scales are used to collect subjective workload data. In this vein, subjects are asked to select a term that best typifies their feelings or are asked to provide a number that represents the level of mental effort. Interviews, open-ended questions, and questionnaires can also be used to tailor the measure to a specific task environment (Tsang & Wilson, 1997). Subjective measures also differ in terms of the approach they take to the measurement.

First, a rating scale either asks for a rating on a uni-dimensional scale representing overall workload, or a multi-dimensional scale representing different dimensions that comprise workload. Because multi-dimensional ratings break workload down into individual variables, they are the only subjective ratings that can be used diagnostically to

pinpoint a certain aspect of mental workload. (Tsang & Vidulich, 1994). Uni-dimensional ratings are, however, easier to obtain due to the reduction of statistical analyses that must be performed in effort to describe an overall workload score. The second variation between subjective instruments is the timing of the measure. Some instruments are used immediately after performing a task, or retrospectively after performing all tasks. Finally, the ratings are either absolute or relative. In the relative condition, operators are asked to compare the task to either a single standard or to multiple task conditions (Tsang & Wilson, 1997).

Subjective measures of workload as a class typically have high face validity and high operator acceptance, mainly due to their ease of use and the format which provides most subjects a platform to give direct opinions. Also, because the unit of measurement is not task dependent, there is high transferability to new systems and tasks (Tsang & Wilson, 1997). Subjective measures are broken down into two groups, multi-dimensional measures, which analyze workload using an additive model in the belief that workload is a combination of several factors, and uni-dimensional measures, which only give an overall workload score. One advantage to the multi-dimensional measures over the others is their ability to diagnose a particular dimension of the task. These multi-dimensional measures can focus on the individual aspects of the task, thus breaking down the overall workload rating into distinct parts. Despite the lack of internal consistency-based indices of reliability, subjective measures have been shown to have acceptable test-retest reliability and correlate highly with performance-based measures of workload (Wierwille & Casali, 1983; Tsang & Velazquez, 1993; and Tsang & Vidulich, 1994).

There are disadvantages to subjective measures as well. The major drawback lies in the reporting of the measures. It has been found that subjective measures can be susceptible to memory problems if the measures are used upon task completion. Research on memory loss has, however, shown that a delay in reporting of 15-30 minutes after task completion will not significantly affect the ratings, while delays of more than 48 hours are problematic (Eggemeier & Wilson, 1991). Finally, the ego of the subject may also be a factor in cases of high workload and large systems where the subject is hesitant to admit the system may have been troublesome. These confounds can be overcome through the use of a well designed study.

### **Psychophysiological Measures of Workload**

Psychophysiological measures are sensitive to changes in the participants' body that are associated with cognitive demand. When the cognitive demand fluctuates as a result of workload, the level of mental activity associated with task performance is adjusted and then associated with changes in the physiology of the subject. These physiological reactions have been reported from the cardiac, respiratory, ocular, and brain systems (O'Donnell & Eggemeier, 1986; Wilson & Eggemeier, 1991).

Among physiological measures, heart rate has the longest history of use, with the first reported use in flight in 1917 (Tsang & Wilson, 1997). The aviation literature contains multiple studies that documented numerous incidences of increased heart rate correlating with increased mental workload (Hasbrook & Rasmussen, 1967; Nicholson, Hill, Borland, & Drzanowski, 1973; Roscoe, 1976). The variability of the heart rate has also been suggested as a measure; however, its utility has not been determined for real-

world applications (Wilson & Eggemeier, 1991) Eye blink rate has been associated with physiological measures of workload, which has been shown to decrease with the addition of high levels of workload However, Fogarty and Stern (1989) suggested that eye blink increases when scanning a cluster of instruments since eye blink is associated with the end of information intake Thus, when using eye blink to measure workload, the context of information input must be taken into account

Two types of brain activity have also been used to assess mental workload, electroencephalograph (EEG) and evoked potentials Ongoing EEG activity is recorded, spectral analysis is performed, and changes in the energy levels of EEG bands are analyzed Evoked potentials consist of the smaller signals present in EEG The potentials are associated with processing information and are collected through a process of averaging across stimuli in order to extract the smaller waveform (Tsang & Wilson, 1997) One disadvantage to EEG is its high sensitivity to real-world situations such as eye blink, head and body movements, muscle activity, and speech Also, because of the small size of the evoked potentials, several stimuli must be used to increase the signal to noise ratio to make them apparent These make recording of EEG and evoked potentials difficult in real-world situations Another disadvantage is the cost of the experts for analysis of the evoked potentials and EEG patterns, and the obvious intrusion issues as well Numerous data leads must be attached to the pilot in order to ascertain any physical workload levels

### **Performance Measures of Workload**

Performance measures of workload use the behavior of the operator to infer the amount of workload within a system. For instance, erratic behavior or decreased performance may be an indication that workload is extremely high. The framework for this comes from the assumption that humans have limited processing capability and once that limit is reached, decreased performance results (Tsang & Wilson, 1997). The first performance based measure was the primary task method. In this method the performance of the operator is observed and changes in performance are noted as the workload is increased. Although the primary task method is ideal for laboratory studies, its practical real-world use is limited for several reasons. First, most systems such as cars, ships, and airplanes do not have recording capability. Second, in tasks where the primary task is accomplished without sufficient load, a secondary task must be used to influence the first. Third, insufficient load may cause increased performance, thus inflating the performance score. When primary task information is not available or is insufficient, it may be plausible to use the secondary task method.

In the secondary task method, a second task is performed concurrently with the primary task. The subject is told that the importance lies in the primary task, thus the secondary task is performed only with excess processing capability (Eggemeier & Wilson, 1991). Since both tasks are competing for limited resources, changes in primary task demand should result in changes in secondary task performance as resources are used or made available. Selection is very critical however, since secondary task performance will only be a valid workload measure as long as both tasks compete for the same resources (Tsang & Wilson, 1997).

There are a few advantages and disadvantages to performance based measures of workload. The primary benefit is the exceptionally high face validity. Also, in the lab, it is possible to control and manipulate the amount and type of cognitive resources being used. The primary task method is the most objective and direct method for assessing workload available. Primary task performance is sensitive to numerous manipulations, except when workload is extremely low. A secondary task may be entered in these situations to increase workload. Also, because some primary task measures are specific to the system being evaluated, results are not generalizable (O'Donnell & Eggemeier, 1986).

The secondary task method has limited diagnostic capabilities, unlike the primary task method. Because the two methods compete for limited resources, the pattern of interference between these two can pinpoint the type of processes and resources in use by the primary task (Tsang & Wilson, 1997). Secondary resources are very sensitive for the same reason. The secondary task method is only valuable if it uses similar resources and capabilities as the primary task. One of the drawbacks of this method is that the introduction of a secondary task may alter the fundamental processes of the primary task. It may also add unseen workload. The practicality and feasibility of the secondary task method is limited when the primary task is a real-world activity such as flying. It is difficult to superimpose a secondary task on top of the primary flight task while maintaining a reasonable level of realism in the study. The last drawback of this method is that it requires the user to have a very considerable background in workload and the experience to properly conduct the secondary task evaluation. The use of a secondary

task to evoke workload responses is only beginning to gain acceptance as a real workload assessment methodology.

### **Analytic Measures of Workload**

Analytic methods incorporate mathematical, engineering and psychological models in order to model the workload. One of the primary goals of the analytic methods is workload prediction. These are used and designed to be both predictive and evaluative in nature. Analytic methods have a distinct advantage over other methods because each part in the model, each level, must be explicitly defined. This allows that specific predictions can be made based on testing and comparisons of each individual level, thus increasing its diagnosticity. Also, because of the level of definition, the findings are easily communicated. Analytic models are traditionally based solely on subject matter experts input, and therefore must be subject to rigorous validation (Tsang & Wilson, 1997). There are several different analytic methods that have been employed to measure workload in systems around the world.

TAWL (Task Analysis/Workload), TLAP (Time-Line Analysis and Prediction), and W/INDEX (Workload Index), all allow for multiple concurrent processes or resources. Unfortunately, because the analytic methods are not commonly used and have been developed only recently, there is little literature relating to their use. Also, most analytic models are not as easy to apply as subjective or performance based workload assessment. On the other hand, the equipment needed to run these analyses is kept at a minimum, besides software to run the simulations and models. Most of those, however, can be handled on a common computer.



### **Choosing Workload Measures**

The reasons for choosing subjective measures of workload over the other three are numerous. To begin with, Muckler and Seven (1992) contend that all measurements contain a subjective element as long as the human is part of the assessment process. Therefore no truly perfect measure exists and tradeoffs must be made in order to determine which measures of workload will be used. The trade-offs made for this study were made using the properties of the different measures set forth in the above paragraphs, and the logistical constraints put upon the researcher by the study already in progress. First, the chosen measures had to be non-intrusive. Also, the measures needed to be relatively easy and quick to administer. Lastly, the measures must be sensitive to low workload conditions, since a low workload flight task forms the basis of this study. While the simulator portion of the study was performed in a laboratory, which lends itself to more intrusive measures of workload, the flight portion of the study was performed in a small Cessna 172 aircraft, restricting the use of more invasive measures.

Subjective methods have been chosen for the present study because of their relative ease of use, high operator acceptance, reliability, sensitivity and low amount of intrusiveness. Diagnosticity of the instruments was also a consideration; however, because the major component of the study is not which aspects of workload are most sensitive, but rather how workload differs from one task configuration to another, a method with low diagnosticity will suffice (Eggemeier et al., 1991). The use of psychophysiological measures were also taken into consideration; however, this study was part of an ongoing research project, and intrusiveness had to be kept at a minimum. The flight operations used in the present study were performed using a Cessna 172 where

space is at a premium and the presence of a “backseat” operator to monitor physiological measures in real time was not an option. Furthermore, the presence of such equipment and the attachment of the required electrodes to the participants may have consequences for flight safety; very few institutions are equipped to deal with potential compromises in flight safety for the sake of research. Finally, the reality of psychophysical measures of workload is that the interpretation of such data requires a great deal of knowledge and experience and is beyond the capacity of most researchers.

Performance based measures could also have been used; however, because the primary task method has been found to have low sensitivity, it is not necessarily conducive to situations of low workload as is the present study (Tsang & Wilson, 1997). The secondary task method can be used to increase the sensitivity of the primary task; still, the knowledge needed to correctly apply the task and then evaluate the method has not been gleaned by the author. Again, applying these secondary task measures would violate the intrusiveness constraint put on the study. For example, it would be very logistically difficult to introduce a secondary task, such as a card sort or detecting a stimulus, while operating the aircraft. Not only would this violate the intrusiveness conditions set forth by the parent study, but also raises some serious safety concerns for both experimenter and participant. These factors led to the decision not to use performance based measures.

Finally, analytic methods were ruled out immediately because of cost, and because they were not easily obtained. Also, the author does not have the skills to effectively model the workload in an analytical workload study. By delineating each measure’s applicability to the properties necessary, subjective workload methods were

chosen. Two subjective measures, one uni-dimensional (Modified Cooper-Harper) and one multi-dimensional (NASA-TLX), were chosen based mostly on availability and sensitivity to low workload conditions. A review of these two measures will be the subject of the next section.

### NASA Task Load Index (NASA-TLX)

The basic premise underlying the NASA-TLX is that workload is a multi-dimensional phenomenon that encompasses several domains. These domains must be considered when attempting to find an overall workload score for a specific task. The framework behind the TLX is that “workload is a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance” (Hart & Staveland, 1988, p. 140).

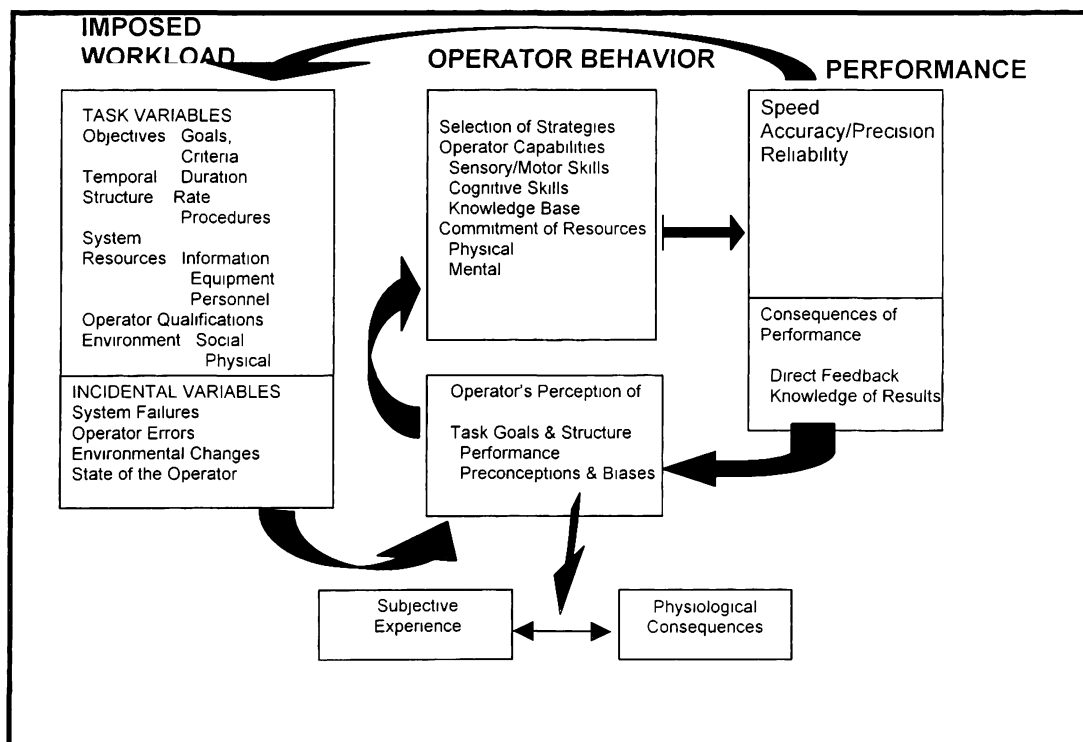


Figure 1. Conceptual Framework for variables that influence Human Performance and Workload.

The developers of the TLX did not consider workload a property, but something that emerges out of task requirements, the way in which it is performed, and the skills, experiences and behaviors of the operator. The conceptual framework adopted by Hart and Staveland (1988) showed how different sources of workload are combined and related.

In Figure 1 (Hart & Staveland, 1988 pg.140), imposed workload refers to the actual task the operator is attempting. The demand of the task is created by its objectives, duration, structure, and by the resources provided. The operators are guided by the demands imposed by the task, but also by their own perceptions of the task. Performance is usually measured through physical effort, however, due to automation, this is becoming less of the case.

Instead of physical effort, mental effort serves as an intervening, yet hard to quantify variable. Eventually, the feedback provides the operator information about the success or failure of their task; this allows the operator to change his or her behavior accordingly (Hart & Staveland, 1988). Finally, the operator feels the effects of the task in both physical and mental ways; it is that subjective experience upon which the NASA-TLX bases its ratings.

The NASA-TLX consists of six component scales, within which, an average of each scale is weighted, and combined to make up the overall workload score. Three behavior related domains were chosen, physical demand, mental demand, and performance; as well as three subject-related domains, temporal demand, effort, frustration (see Table 1) as representing the components of workload.

**Table 1**  
**NASA-TLX Rating Scale Definitions**

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

Once the task is completed, participants rate themselves on a bi-polar scale of each component of workload. Participants are then instructed to make 15 paired comparisons to provide a weighting of the importance of each component (see Appendix A for NASA-TLX) for that subscale. Totaling the number of times one domain is chosen over another gives the weight for that subscale. The resulting weighting is then multiplied by the raw subscale score from each subscale and is used to calculate overall workload scores for each subject and each segment (if the task is broken down in this fashion). The use of the weighted scores over unweighted averages serves to reduce between subject variability (Hart & Staveland, 1988; Hancock, 1996). These weights account for differences in the definition of workload between subjects, and the differences in the sources of workload between tasks. However, there have been several studies advocating not using the traditional TLX's weighting procedure (Nygren, 1991;

Byers, Bittner, & Hill, 1989). These studies contend that the raw scores obtained through the bi-polar scales on each of the six dimensions of workload offered tend to be of equal statistical power when compared with the same ratings after being weighted. When compared with traditional TLX scores, the new raw TLX scores correlated highly, with  $r$ s ranging from 0.96 to 0.98 (Nygren, 1991). There is still much disagreement about the weighting procedures and values; however, most research uses the weighting procedures to indicate overall workload.

Due to its ease of use, the NASA-TLX is very applicable in operational environments. Three different versions, verbal, paper and pencil, and computerized methods provide additional adaptability. The three methods were correlated highly with computer vs. verbal = .96, computer vs. paper/pencil = .94, and verbal vs. paper/pencil = .95. Test- retest reliability estimates ranging from .77 (Battiste & Bortolussi, 1988) to .83 (Hart & Staveland, 1988).

Because of its adaptability in the operational environment, there are numerous tasks for which the NASA-TLX has been employed successfully. It has been used in computer monitoring tasks (Fuld, Liu, & Wickens, 1987), target tracking tasks (Hancock & Caird, 1993), car stereo evaluations (Jordan & Johnson, 1993), noise studies (Becker, Warm, Dember, & Hancock, 1995), and flight simulations (Battiste & Bortolussi, 1988; Kilmer et al., 1988; Wilson & Badeau, 1992; Aretz et al., 1995). Hill, Iavecchia, Byers, Bittner, Zakland, and Christ (1992) conducted a systematic review of four workload measures. Their conclusion indicated that the NASA TLX is a useful tool under operational testing conditions and was sensitive to different levels of workload, thus providing content validity evidence. It also had the highest validity as measured using

Jackknife Principal Component Analyses (PCA's) of four different scales based on a field test of a remotely piloted vehicle (RPV) (Byers et al., 1989). Vidulich and Tsang (1986) showed the NASA-bipolar method consistently showed an increase in subjective workload as the difficulty of a tracking task was increased. This provides another line of construct validity evidence.

### **Modified Cooper-Harper (MCH)**

The Cooper-Harper scale, a 10-point scale utilizing a decision tree, was designed in 1969 and was mainly used for the evaluation of aircraft handling qualities. The original scale represented a handling qualities/workload rating scale, which was then adapted by Wierwille and Casali (1983) in order to develop a useful workload scale (the Modified Cooper-Harper- MCH). This scale was associated with cognitive functions, such as perception, monitoring, evaluation, communications, and problem-solving. Mainly minor changes in terminology were used in order to accomplish this task. These changes included changing the rating scale end points to “very easy” and “impossible”; asking the operator to rate mental workload and not controllability; and emphasizing difficulty and not deficiency. The validity of the MCH was assessed in three different experiments by Wierwille and Casali (1983). Analysis of human operators in systems concluded that activities can be grouped into four categories: perceptual, mediational (cognitive), communications, and motor. Since the original Cooper-Harper could be used for most motor applications, experiments were conducted at making the scale applicable to the first three groups. Three experiments, a perceptual, a cognitive, and a communications experiment, were used to assess the validity of the MCH in measuring

workload. Each experiment differed in terms of low, medium, and high load levels placed on subjects. In all three experiments, significant results were reached for at least two of three load levels, that is, the MCH successfully distinguished between high/medium/low loads (Wierwille & Casali, 1983). Additionally, in all three experiments, the score means increased monotonically with each load level, providing evidence of construct validity. Wierwille, Rahimi, and Casali, (1985) performed experiments in a fixed base flight simulator; Skipper, Rieger, and Wierwille (1986) conducted studies in a moving-base flight simulator; Byers et al., (1988) conducted workload analyses utilizing MCH in a field test of a remotely piloted vehicle; and Kilmer et al., (1988) compared the MCH with the Subjective Workload Assessment Technique (SWAT) using a psychomotor dual-task experiment; and have all reported similar reliable results. In addition, the MCH has been used effectively to assess workload in a number of environments, (e.g., a remotely piloted vehicle (Byers et al., 1988); flight simulations (Wierwille et al., 1985; Skipper et al., 1986) and tracking tasks (Casali & Wierwille, 1984)). The MCH has demonstrated itself to be a reliable and valid measure of overall workload in each of these domains. Because of its adaptability, the results can be generalized to different populations.

Several other paper and pencil measures of workload were considered in addition to the NASA-TLX and MCH. The Subjective Workload Assessment Technique (SWAT) (e.g., Reid & Nygren, 1988) requires a time consuming card-sort procedure prior to the task. Other workload scales, the Bedford Workload Scale, the Overall Workload Scale, and the NASA Bi-Polar Rating Scale, were not available at the outset of this research, and thus were not explored for use.



## **Workload in Aviation Systems**

When looking at the aviation literature, it is plain to see the myriad changes that have come about as a result of increased automation. However, none are more disturbing than the fact that regardless of the use of automation, pilots are still required to monitor the same systems they use to control. This increase in the mental workload of the pilots usually leads to a reduction in the performance of the pilot. This relationship has been studied through the years from Lindbergh to Boeing and Airbus, and aviation has changed dramatically. The aviation industry has gone from three pilots to two pilots in the cockpit, increased automation causing decreased situation awareness and degradation of manual flight skills, and an increase in mental workload. In the future, we can expect to see airplanes that carry 800 or more passengers, and fly 25% faster than the Concorde (over 1,500 mph) (Mouloua et al., 2001). With these kinds of advances it is necessary now more than ever to be sure pilots are not caused more stress and strain than is necessary. In current systems, automation has turned pilots into constant monitors of the aircraft system, taxing the mental workload of the pilots. Most commercial aircraft manufacturers have taken a step in the right direction and now employ workload studies as part of the verification process (and marketing process) of all their aircraft (Mouloua et al., 2001). Not only does this practice make new aircraft easier to fly, because of the constant monitoring of workload conditions, but also makes them more attractive to the airline industry. Increased automation without increased mental workload, while maintaining proper situation awareness is the goal of all automation experts. Besides performance issues, increases in workload have numerous safety considerations.

With increased workload, pilots' performance may decrease, potentially impacting safety and efficiency. In the glass cockpit systems, when there is an automation failure, pilots must be able to quickly ascertain what and where the problem is, and how to take correct and accurate control of the aircraft. It has been demonstrated time and again how this is not the case with current flight deck design. Funk, Lyall, and Niemczyk (1997) examined hundreds of documents containing citations of flight deck automation problems and analyzed the reports from Aviation Safety Reporting Systems across the country; they found over 1,800 incidents involving automated systems. These and other findings emphasize the need for more human centered design alternatives and the constant need for workload monitoring (Mouloua et al., 2001). As previously mentioned, the easiest, safest, and most economical way to monitor the workload is through the judicious use of simulation.

Workload assessments have been used widely in simulation studies. However, rarely have there been comparisons of these workloads from the simulators to the cockpits. When these comparisons have been utilized, the results are somewhat unsettling. Wilson and Badeau (1992) recorded psychophysiological data during flight and in a laboratory setting. The results showed that direct transference of the laboratory findings to the flight environment is not possible in most cases because enough flight data does not exist to properly interpret these findings. Leino et al., (1995) also found, through psychophysiological measures, that the psychological workload in a BA Hawk MK 51 flight simulator did not correspond and was significantly lower than the workload in the corresponding jet trainer for 10 Finnish Air Force pilots. In a study by Wilson et al., (1987) the heart rates of pilots changed in actual flight, however, differences were

hardly noticeable in simulation. While it is true that experienced pilots perform better in familiar and simple simulators, the workload experienced in the simulators should at least approach that of the actual aircraft. If this transference cannot and does not take place, how can we be sure that workload is not too high, and when confronted with serious and life threatening situations in the actual aircraft, humans will be able to cope? On the other hand, Magnusson (2002) studied the psychophysiological reactions in simulated and real missions and found analogous results in both simulator and actual flight. Similarly, Ylonen et al., (1997) found no significant differences in heart rate during real and simulated Hawk MK 51 Flight. Clearly more research is needed to determine the extent to which the amount and type of workload induced during simulation is equivalent to the workload induced by actual flight. This study seeks to add to the depth and breadth of these simulator studies by employing subjective workload measures to assess the workload in a simulator and then the comparable aircraft.

### **Summary and Thesis Question**

The literature review has shown that workload is an important factor in the aviation domain for safety, performance, and efficiency reasons (Mouloua et.al., 2001). Several studies cited in the preceding pages have shown that workload assessed by psychophysiological measures in a simulator may not accurately reflect the experienced workload in actual flight. Most of the workload studies thus far have utilized psychophysiological measures when making such a comparison. In addition, there are no published studies comparing workload in the simulator to that of the comparable aircraft utilizing subjective evaluations of workload. The present study seeks to find out if

measures of subjective workload as measured during simulated flight are similar to measures of subjective workload as measured during real flight

### **Statement of Hypothesis**

The objective of the present study is to investigate whether or not workload levels measured following simulated and real flights differ. It is often assumed that the quantity of workload imposed on pilots during simulated and real flights is similar, but it is hypothesized for the current study that simulated flights impose less workload than real flights. Specifically, workload scores measured using the NASA-TLX and the MCH in a simulated environment are hypothesized to be lower than workload scores measured in a real-flight environment. Beyond assessing whether or not a quantitative difference likely exists, a confidence interval around the observed mean difference will also be constructed for the NASA-TLX scores (MCH scores are ordinal and thus do not lend themselves to traditional confidence interval construction). Beyond this, the diagnostic properties of the NASA-TLX should allow the researcher to gain insight as to which specific subscales comprise the majority of the load in each environment and to qualitatively compare the workload across the two environments. Load levels within each sub-scale will be examined within each environment and compared across the simulation and real flight environments. It is hypothesized that differences between the mental demand sub-scale will exist in favor of with higher loading in the simulator. However, there should be opposite differences on the physical demand subscale, showing the real flight environment eliciting higher physical demand ratings in the actual aircraft than in the simulator.

## METHOD

### **Participants**

There were 50 participants in the study, all selected on a volunteer basis from Embry-Riddle Aeronautical University, Daytona Beach campus. Participants were required to have an Instrument Rating and less than 300 flight hours at the beginning of the study. The average number of flight hours was 199.7 hours. Additionally, the average Cessna 172 time was 132.4 hours, and the average simulator time (in a Frasca type simulator) was 22.8 hours. Because the study had been designed by the time of this proposal, there was limited opportunity to influence sample size. Fortunately, this may be cause for little concern as a search of the literature found that when performing workload studies with both simulation and flight, sample sizes larger than 10 per group are rare. Similar previous studies have used only 5-10 participants. For example, Ylonen et. al (1997) used 10 male subjects; Leino et. al (1995) also used 10 male subjects; Wilson et. al (1999) used 8 subjects; Magnusson (2002) used 5 pilots; and another study by Wilson et. al (1987) also used 8 subjects.

### **Apparatus**

Two measures of workload were used for the study: the NASA TLX and the Modified Cooper-Harper (MCH). The TLX was selected based on its availability, ease of administration, and sensitivity, especially in low workload conditions. The MCH was selected primarily on its ease of use and availability. It is not as sensitive as the NASA TLX, but it is quickly and easily administered.

The simulator used in the study was the Elite iGATE system and was configured to fly like the actual Cessna 172 test aircraft. The Elite Cessna 172 flight model, when used with the Elite iGATE simulator hardware is certified by the FAA as a flight training device. The aircraft used for the real flight portion of the study was a standard Cessna 172 equipped with a standard CDI instrument package.

### **Design**

The experiment utilized a between subjects design where flight condition (simulated and real flight) was manipulated and workload was measured using the NASA-TLX and the MCH.

### **Procedure**

There were two different scripts followed for the conditions. For the simulator condition, the participant was welcomed into the lab and given the consent form and asked to fill out a demographic data sheet. Next, the experimenter briefed the participant on the operation of the simulator and gave the participant a five minute practice session in order to familiarize him/her with the controls and settings. The participant also flew one practice approach under Visual Flight Rules (VFR) conditions prior to the beginning of the experiment. To start the simulation, the aircraft was set up on a 30 degree intercept for the ILS localizer two miles outside the outer marker. Data collection began upon reaching the outer marker and ended upon reaching decision height (232 MSL). Once decision height was reached, the simulation scenario was reset and another approach was made until four approaches were completed. Once the four approaches were completed,

the participant was asked to complete the two workload measures. Presentation order of the workload measures was randomized by the SATS researcher to counterbalance any order effects. The participant was then thanked for his/her help and cooperation.

The real-world flight session was conducted in a similar fashion to the simulator portion. The session began with the completion of a demographic form and an initial study briefing by a SATS safety officer. Then the participant was introduced to the research pilot and an extensive safety briefing was completed. In order to induce similar amounts of workload under the simulated and real flight conditions, the research pilot maintained control of the aircraft for most of the flight, turning over control of the aircraft only several minutes prior to intercepting the localizer on approach. The participant then completed three ILS approaches at runway 7L at Daytona Beach International Airport. Under some circumstances, participants may have completed their approaches on runway 36 at Space Center Executive airport in Titusville. Data collection proceeded in the same fashion as in the simulated flight. Upon return to DAB, the participants were debriefed by the SATS safety officer and asked to complete the two workload measures. Under most circumstances, this debriefing occurred within 5 minutes of the completion of the flight. As in the simulator portion of the study, the ordering of the workload tests was randomized by the experimenter.

It is important to note that all participants completed the simulator portion of the study. However, only half of the participants were given the workload tests after flying in the simulator. The other half of the participants were given the workload tests after flying in the aircraft.

## RESULTS

**Data Analysis**

The objective of the present study was to investigate the extent to which the quantity and quality (based on subscales of the NASA-TLX) of workload measured following simulated flight differs (if at all) from that measured following real flight in the aircraft. Due to the fact that the NASA TLX produces interval-level data and the MCH produces ordinal-level data, separate statistical analyses were performed. Analyses of the NASA-TLX scores were performed using an independent *t*-test. A confidence interval around the observed mean difference was computed as well as a standardized estimate of effect size, Cohen's *d*.

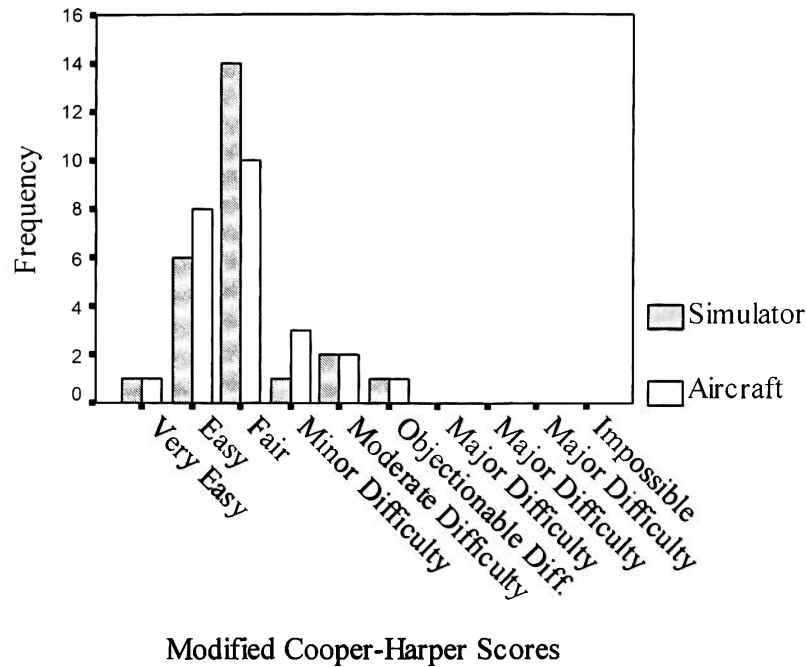
For the Modified Cooper Harper, the results of the Mann-Whitney U and z scores indicated means were not significantly different across conditions,  $U(50) = 307.50$ ,  $Z = .104$ ,  $p = .917$ . Frequency information across conditions is reported in table 2.

Table 2  
*Frequency Chart for Modified Cooper-Harper*

| Numerical Rating | Rating Description                          | Simulator | Aircraft |
|------------------|---|-----------|----------|
| 1                | Very easy, highly desirable                 | 1         | 1        |
| 2                | Easy, desirable                             | 6         | 8        |
| 3                | Fair, mild difficulty                       | 14        | 10       |
| 4                | Minor but annoying difficulty               | 1         | 3        |
| 5                | Moderately objectionable difficulty         | 2         | 2        |
| 6                | Very objectionable but tolerable difficulty | 1         | 1        |

*Note.* No subject indicated workload ratings above 6, therefore ratings 7-10 are not displayed.

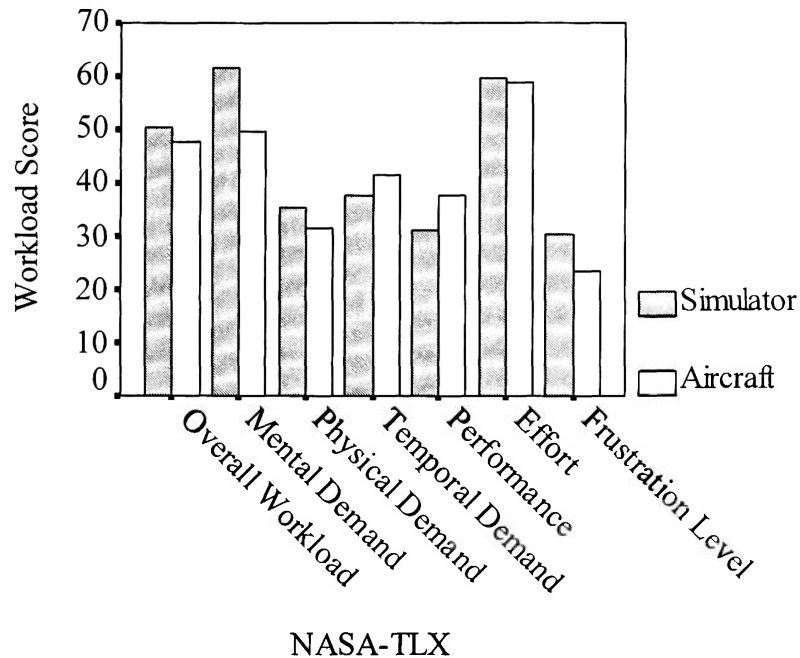




Modified Cooper-Harper Scores  
 Figure 2. Modified Cooper-Harper frequency of ratings.

The average rank frequency in the simulator condition (25.70) was not different than the aircraft condition (25.30). Figure 2 shows the frequency of each rating on the Modified Cooper-Harper.

For the NASA Task Load Index, a comparison of the means of the overall workload scale and the subscales was undertaken (see figure 2). Neither the analysis of the overall workload scale nor the subscales presented statistical differences (see Table 3).



**Figure 3.** Mean Ratings for overall NASA-TLX and subscales

Overall, this investigation found that participants did not rate the workload higher in the simulator ( $M = 50.20$ ,  $SD = 10.73$ ) significantly higher than the aircraft ( $M = 47.54$ ,  $SD = 16.32$ ) as measured by the NASA-TLX,  $t(48) = .682$ ,  $p = .498$ , *ns*. An estimate of effect size was performed on the independent  $t$ -test for overall workload on the overall workload score, Cohen's  $d = .197$  indicating a negligible difference.

The NASA-TLX subscales were also analyzed. An independent  $t$ -test was performed on all six subscales. Table 3 shows the  $t$  scores and confidence intervals for each subscale. There were no statistical differences between the groups. Estimates of effect size utilizing Cohen's  $d$  were also found for overall workload and the subscales.

Table 3  
*NASA-TLX t-scores for Overall Workload and Subscales*

| Domain            | <i>t</i> | <i>p</i> | Mean<br>Difference<br>(Sim – AC) | 95% CI |        |
|-------------------|----------|----------|----------------------------------|--------|--------|
|                   |          |          |                                  | LB     | UB     |
| Overall Workload  | .682     | .498     | 2.665                            | -5.191 | 10.522 |
| Mental Demand     | 1.801    | .078     | 11.92                            | -1.39  | 25.23  |
| Physical Demand   | .653     | .517     | 4.00                             | -8.33  | 16.33  |
| Temporal Demand   | -.637    | .527     | -4.12                            | -17.13 | 8.89   |
| Performance       | -1.261   | .213     | -6.48                            | -16.81 | 3.85   |
| Effort            | .100     | .921     | .56                              | -10.73 | 11.85  |
| Frustration Level | 1.209    | .233     | 6.80                             | -4.51  | 18.11  |

Table 4  
*Estimates of Effect Size, Cohen's *d**

| Domain            | Cohen's <i>d</i> |
|-------------------|------------------|
| Overall Workload  | .197             |
| Mental Demand     | .519             |
| Physical Demand   | .186             |
| Temporal Demand   | -.184            |
| Performance       | -.364            |
| Effort            | .029             |
| Frustration Level | .349             |

## DISCUSSION

Previous research comparing workload in a simulator and the corresponding aircraft has had mixed results. Thus far, no consensus exists in the literature as to which environment, simulator or aircraft, should produce higher measured workload levels. However, where differences have existed, lower measured workload in the simulator seems to be most common (Wilson & Badeau, 1992, Leino et al., 1995, and Wilson et al., 1999). In addition, previous published studies have relied mainly on comparing workload in these two environments using psychophysiological measures. It is in this regard that the present study departs. The results of this study have tremendous impact. Most important, they add to the body of literature making direct comparisons between simulator and comparable aircraft. Second, this study opens a new chapter in simulation evaluation, comparing workload ratings from subjective workload measures where only psychophysiological workload measures have been used previously. Finally, the results provide a framework for future research.

Among the results, the Modified Cooper-Harper failed to find significant differences in group mean ratings across the simulator and aircraft groups. The MCH, while a good indicator of overall workload, may not be sufficiently sensitive to low workload conditions. While the study utilized participants flying instrument approaches, these may not produce enough workload to warrant mean differences between the two groups.

An analysis of the difference between the means using the NASA-TLX yielded similar results. The ratings from the simulator group ( $M = 50.2$ ) were higher than the ratings obtained from the aircraft group ( $M = 47.5$ ), indicating participants felt the

workload in the simulator was higher than the aircraft, but not significantly so. While this result was surprising, it was not only by how much would the groups differ, but how would the groups differ, that was among the questions to be answered.

The analysis of the subscales of the NASA-TLX did not indicate significance between any of the groups (see Table 3). However, there were interesting results to come from the analysis. First, the mean difference between the simulator ( $M = 61.40$ ,  $SD = 20.58$ ) and the aircraft ( $M = 49.48$ ,  $SD = 25.92$ ) on the mental demand subscale did approach significance  $t(48) = 1.801$ ,  $p = .078$ .

This finding implies that participants felt more mental and perceptual activity was required to complete the simulator task than the aircraft task (figure 4). The effect size for this group,  $d = .519$ , could be classified as “medium”, and could be the shape of a definite trend (Cohen, 1988). This finding could be explained by looking at the demographics for the participants. The amount of time spent in a Cessna 172 was almost six times the amount of time spent in a comparable simulator. This suggests that students would be more comfortable and more familiar with the Cessna 172 aircraft than with the simulator.

The other subscales of the TLX failed to find any difference between the means. Based on the results however, this should be expected. Participants rated their temporal demand in the simulator and in the aircraft as roughly equal (37.60 vs. 41.72), as well as the amount of physical effort (35.4 vs. 31.40) and frustration level (30.44 vs. 23.64) in both environments. Participants also rated themselves as performing equally well in both conditions (31.24 vs. 37.72).

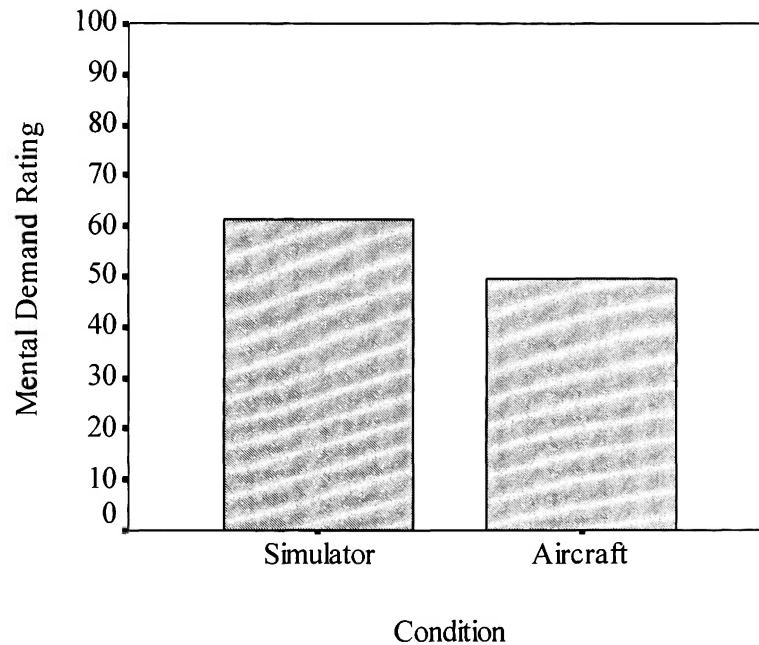


Figure 4. Average by group for Mental Demand subscale

What exactly this means for overall simulation testing and evaluation remains unclear. Unfortunately, there is no previous research to indicate which environment should record more workload. Should workload studies still be undertaken in a simulator and still be seen as a valid indicator of the workload that is represented in the aircraft? The data from the present study suggests the answer to this question is yes. While a quantitative difference in terms of significance was not obtained, the information about the specific aspects of workload contained within each environment is invaluable. Knowing what particular aspect of workload, mental demand, frustration, etc. is causing the most or the least load in each environment can aid engineers in designing aircraft that can be safely and adequately handled by the crew. Additionally, the use of workload evaluation tools to evaluate automation before, during and after implementation should remain a necessary protocol.

### **Future Recommendations**

In the future, it may be advantageous to complete the research with an analysis of different demographic variables, such as experience level, or experience using some sort of higher end simulation tool, such as the Elite iGATE system. Additionally, counterbalanced repeated-measures designs may provide better insight into this issue and allow researchers to examine workload ratings from an alternate-forms test reliability perspective.

## REFERENCES

- Arretz, Anthony J., Shacklett, Scott F., Acquaro, Phil L., & Miller, Derek. (1995). The Prediction of Pilot Subjective Workload Ratings. *Proceedings of the 39<sup>th</sup> Annual Meeting of the Human Factors and Ergonomics Society*. San Diego, CA: 94-97.
- Battiste, V., & Bortolussi, M.R. (1988). Transport Pilot Workload: A comparison of Two objective techniques. *Proceedings of the 32<sup>nd</sup> Annual Meeting of the Human Factors Society*. Santa Monica, CA: pp. 150-154.
- Becker, A.B, Warm, J.S., Dember, W.N., & Hancock, P.A. (1995). Effects of Jet Engine Noise and Performance Feedback on Perceived Workload in a Monitoring Task. *International Journal of Aviation Psychology*. 5(1): 49-62.
- Byers, J.C., Bittner, A.C. Jr., & Hill, S.G. (1989). Traditional and Raw Task Load Index (TLX) correlations: Are Paired Comparisons Necessary? In A.Mital (Ed.) *Advances in Industrial Ergonomics and Safety* (Vol. 1). London: Taylor & Francis.
- Byers, J.C., Bittner, A.C., Hill, S.G., Zaklad, A.L., & Christ, R.E. (1988). Workload Assessment of a Remotely Piloted Vehicle (RPV) System. *Proceedings of the 32<sup>nd</sup> Annual Meeting of the Human Factors Society*. Santa Monica, CA: pp.1145-1149.
- Casali, J.G., & Wierwille, W.W. (1984). On the Measurement of Pilot Perceptual Workload: A comparison of Assessment Techniques Addressing Sensitivity and Intrusion Issues. *Ergonomics*, 27: 1033-1050.



- Charlton, Samuel C. (2002). Measurement of Cognitive States in Test and Evaluation.  
In: Charlton, Samuel G., and O'Brien, Thomas G. (Eds.). *Handbook of Human Factors Testing and Evaluation*. London: Lawrence Erlbaum Associates.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Eggemeier, F.T., & Wilson, G.F. (1991). Performance-based and Subjective Assessment of Workload in Multi-task Environments. In D.L. Damos (Ed.). *Multiple Task Performance*. London: Taylor & Francis.
- Eggemeier, F.T., Wilson, G.F., Dramer, A.F., & Damos, D.L. (1991). Workload Assessment in Multi-Task Environments. In D.L. Damos (Ed.). *Multiple Task Performance*. London: Taylor & Francis.
- Fogarty, C., & Stern, J.A. (1989). Eye Movements and Blinks: Their Relationship to Higher Cognitive Processes. *International Journal of Psychophysiology*, 8:35-42.
- Fuld, R., Liu, Y., & Wickens, C.D. (1987). Computer Monitoring vs. Self-Monitoring: The Impact of Automation on Error Detection. (NASA-87-4). Champaign: University of Illinois.
- Funk, K.H., Lyall, E.A., & Niemczyk, M.C. (1997). Flightdeck Automation Problems: Perceptions and Reality. M. Mouloua & J.M. Koonce (Eds.). *Human-Automation Interaction: Research and Practice*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hancock, P.A., & Caird, J.K. (1993). Experimental Evaluation of a Model of Mental Workload. *Human Factors*. 35(3): 413-419.

- Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock, Peter A., and Meshkati, Najmedin (Eds.). *Human Mental Workload*. Amsterdam: North Holland.
- Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C. Jr., Zaklad, A.L., & Christ, R.E. (1992). Comparison of four Subjective Workload Rating Scales. *Human Factors*, 34(3): 429-439.
- Jordan, P.W., & Johnson, G.L. (1993). Exploring Mental Workload via TLX: The Case of Operating a Car Stereo Whilst Driving. In A.G. Gale, I.D. Brown, C.M. Halsegrave, H.W. Krusse, and S.P. Taylor (Eds.). *Vision in Vehicles – IV*. Amsterdam: North Holland.
- Kilmer, Kevin J., Knapp, Robert, Burdsal, Charles Jr., Borresen, Robert, Bateman, Robert, and Malzahn, Don. (1988). Techniques of Subjective Assessment: A Comparison of The SWAT and Modified Cooper-Harper Scales. *Proceedings of the 32<sup>nd</sup> Annual Meeting of the Human Factors Society*. Santa Monica, CA: pp. 155-159.
- Leino, T., Leppaluoto, J., Huttenen, P., Ruokonen, A., & Kuronen, P. (1995). Neuroendocrine Responses to Real and Simulated BA Hawk MK 51 Flight. *Aviation, Space, and Environmental Medicine*. 66: 108-113.

- Magnusson, S. (2002). Similarities and Differences in Psychophysiological Reactions Between Simulated and Real Air-to-Ground Missions. *The International Journal of Aviation Psychology*, 12(1): 49-61.
- Mouloua, Mustapha, Deaton, John, & Hitt, James M. II. (2001). Automation and Workload in Aviation Systems. In: Hancock, Peter A, and Desmond, Paula A. (Eds.). *Stress, Workload, and Fatigue*. London: Lawrence Erlbaum Associates.
- Muckler, F.A., & Seven, S.A. (1992). Selecting Performance Measures: “Objective” versus “subjective” measurement. *Human Factors*, 34(3), 441-455.
- Nygren, Thomas E. (1991). Psychometric Properties of Subjective Workload Measurement Techniques: Implication for Their Use in the Assessment of Perceived Mental Workload. *Human Factors*, 33(1): 17-33.
- O’Donnell, R.D., & Eggemeier, F.T. (1986). Workload Assessment Methodology. In: Boff, K.R., Kaufman, L., and Thomas, J. (Eds.). *Handbook of Perception and Human Performance: Volume II. Cognitive Processes and Performance*. New York: John Wiley.
- Reid, G.B., & Nygren, T.E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. In P.A. Hancock, & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland.

- Rueb, J.D., Vidulich, M.A., & Hassoun, J.A. (1994). Use of Workload Redlines: A KC-135 Crew-Reduction Application. *The International Journal of Aviation Psychology*, 4: 47-64.
- Scerbo, Mark W. (2001). Stress, Workload and Boredom in Vigilance: A problem and an Answer. In: Hancock, Peter A., and Desmond, Paula A. (Eds.). *Stress, Workload, and Fatigue*. London: Lawrence Erlbaum Associates.
- Skipper, J.H., Rieger, C.A., & Wierwille, W.W. (1986). Evaluation of Decision-tree Rating Scales for Mental Workload Estimation. *Ergonomics*, 29: 585-599.
- Tsang, P.S., & Velazquez, V.L. (1993). Subjective Workload Profile. *Proceedings of the 7<sup>th</sup> International Symposium on Aviation Psychology*. Columbus, OH: Ohio State University Association of Aviation Psychologists.
- Tsang, P. S., & Vidulich, M. A. (1994). The Roles of Immediacy and Redundancy in Relative Subjective Workload Assessment. *Human Factors*, 36(3): 503-513.
- Tsang, P.S., & Wilson, G.F. (1997). Mental Workload. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics*. New York: Wiley.
- Veltman, J.A. (2002). A Comparative Study of Psychophysiological Reaction During Simulator and Real Flight. *The International Journal of Aviation Psychology*, 12(1): 33-48.
- Vidulich, M.A., & Tsang, P.S. (1986). Techniques of Subjective Workload Assessment: A Comparison of SWAT and the NASA-bipolar methods. *Ergonomics*, 29: 1385-1389.

- Wierwille, W.W., & Casali, J.G. (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): 623-642.
- Wierwille, W.W., & Eggemeier, F.T. (1993). Recommendations for Mental Workload Measurement in a Test and Evaluation Environment. *Human Factors*, 35(2): 263-282.
- Wierwille, W.W., Rahimi, M., & Casali, J.G. (1985). Evaluation of 16 Measures of Mental Workload Using a Simulated Flight Task Emphasizing Mediatlional Activity. *Human Factors*, 27(5): 489-502.
- Wilson, Glenn F., Badeau, Albert. (1992). Psychophysiological Measures of Cognitive Workload in Laboratory and Flight. *Sixth Annual Workshop on Space Operations Applications and Research (SOAR)*, Houston, TX.
- Wilson, Glenn F., Purvis, B., Skelly, J., Fullenkamp, & Davis, I. (1987). Physiological Data Used to Measure Pilot Workload in Actual Flight and Simulator Conditions. *Proceedings of the 31<sup>st</sup> Annual Meeting of the Human Factors Society*. Santa Monica, CA: pp.779-783.
- Wilson, Glenn P., Skelly, June, & Purvis, Bradley. (1999). Reactions to Emergency Situations in Actual and Simulated Flight. *AGARD Conference Proceedings*. The Hague, The Netherlands.
- Ylonen, Hannamajja, Lyytinen, Heikki, Leino, Tuomo, Leppaluoto, Juhani, & Kuronen, Pentti. (1997) Heart Rate Responses to Real and Simulated BA Hawk MK 51 Flight. *Aviation, Space, and Environmental Medicine*. 68(7):601-605.

APPENDIX A  
Subjective Workload Tests

Appendix A  
Subjective Workload Tests

NASA TLX Rating Sheet

Instructions: On each scale, place a mark that represents the magnitude of that factor in the task(s) you just performed.

|-----|  
LOW MENTAL DEMAND HIGH

|-----|  
LOW PHYSICAL DEMAND HIGH

|-----|  
LOW TEMPORAL DEMAND HIGH

|-----|  
EXCELLENT PERFORMANCE POOR

|-----|  
LOW EFFORT HIGH

|-----|  
LOW FRUSTRATION HIGH

NASA-TLX  
Pairwise Comparison of Factors

Instructions: Circle the member of each pair that provided the most significant source of variation in the task(s) that you just performed.

PHYSICAL DEMAND / MENTAL DEMAND

TEMPORAL DEMAND / MENTAL DEMAND

PERFORMANCE / MENTAL DEMAND

FRUSTRATION / MENTAL DEMAND

EFFORT / MENTAL DEMAND

TEMPORAL DEMAND / PHYSICAL DEMAND

PERFORMANCE / PHYSICAL DEMAND

FRUSTRATION / PHYSICAL DEMAND

EFFORT / PHYSICAL DEMAND

TEMPORAL DEMAND / PERFORMANCE

TEMPORAL DEMAND / FRUSTRATION

TEMPORAL DEMAND / EFFORT

PERFORMANCE / FRUSTRATION

PERFORMANCE / EFFORT

EFFORT / FRUSTRATION



Modified Cooper-Harper

