AN ABSTRACT OF THE THESIS OF

Jeffrey Michael Stebel for the degree of Master of Science in Industrial Engineering presented on October 4, 2005.

Title: Determining the Validity of the Task Management Environment to Predict Pilot Concurrent Task Management Performance.

Redacted for privacy

Abstract approved: _____

Kenneth H. Funk, II

Concurrent Task Management (CTM) is the process by which human operators of complex systems (such as pilots, drivers, surgeons, and operators) allocate their attention among multiple concurrent tasks (Funk, 1991).

A more thorough understanding of the human operator's knowledge, skills, and abilities associated with CTM might help us to prevent disasters such as those in aviation, in the operating room, and nuclear power plants. But to develop that understanding, we must develop valid tools to measure the human operator's CTM performance. In this research, a validation study of a software game developed in previous research, the Task Management Environment (TME), was performed. Since the TME exhibits some face validity, it was hypothesized that it may have the potential to be useful for predicting human CTM performance in an airplane cockpit. However, it raises an important question as to whether or not it is a good enough tool to accurately measure CTM performance.

Since the Frasca 141 flight simulator has been recognized and certified by the Federal Aviation Administration (FAA) as a valid human performance assessment tool for pilots, it was used to determine if the TME has external validity through a comparison of CTM performance, as measured by the TME, with CTM performance observed in the Frasca 141.

Nineteen pilot participants from the Flight Technology Program at Central Washington University were tested for CTM performance on a Frasca 141 flight simulator and the TME. Performances were compared using correlation analyses to determine their relationship. The findings indicated that CTM performance in the TME

does not correlate significantly with CTM performance in the Frasca 141 flight simulator. In conclusion, the TME does not have external validity and it cannot be used as a research tool to generalize pilot CTM performance to the "real-world" without modification.

However, the limitations of this study may have caused sources of unwanted variability in the results. Future research using the TME might shed new light on validation.

Determining the Validity of the Task Management Environment to Predict Pilot Concurrent Task Management Performance

by

Jeffrey Michael Stebel

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented October 4, 2005
Commencement June 2006

Master of Science thesis of Jeffrey Michael Stebel presented on October 4, 2005

APPROVED:

Redacted for privacy

Major Professor of Industrial Engineering

Redacted for privacy

Head of the Department of Industrial and Manufacturing Engineering

Redacted for privacy

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Redacted for privacy

Jeffrey Michael Stebel, Author

ACKNOWLEDGEMENTS

During this challenging research project, I have received great support from my professors, family, friends, and peers. Their encouragements and influences have been the sole motivation for "staying the course." With their support, this research project has been a great learning experience. I would like to give my gratitude to those who have made a positive impact on me during this journey.

First and foremost, I would like to thank my major professor Dr. Kenneth Funk. I will be forever grateful for meeting Dr. Funk. If not for him, I may not have had the chance to attend graduate school. His due diligence, knowledge, enthusiasm, guidance, and friendship has given me the opportunity to improve myself personally and professionally. I'd like to thank him for directing my focus and providing me with positive and negative feedback. I'd like to thank him for the opportunity to attend various professional conferences and to work with some of his professional contacts on a variety of projects.

Second, I would like to thank Dr. Anthony Veltri. His superior enthusiasm and motivating skills have given me a new found desire and confidence required to be innovative and successful in the business world. Mostly, I'd like to thank him for his unbiased views and realistic decision making strategies. I'll always remember to attack a problem strategically by building a pictorial model representation of my "mind's eye" using a systematic approach.

Also, I would like to give special thanks to my office mate, Javier Nicolalde, for all of his help, encouragement, wisdom, and vision. I've enjoyed working with him on many academic and professional projects. I'll miss the many creative and farsighted conversations that we have had in the office. I look forward to working with him in future business endeavors.

Additionally, I would like to acknowledge Colin Cowger and Mike Halbleib for their help on my experimental set up. Their aviation piloting knowledge, lingo, and aviation experiences have given me a greater interest in flying. I'll try to remember all my acronyms.

I would like to thank a number of people who have either led me to this point or who have supported me during this stage of my life. I'd like to thank my brother Todd Stebel for being a role model and allowing me to visit his home in Boise, Idaho.

I'd like to thank Neal Grandstaff for pushing me to pursue my passion for music through his enlightening and soulful training in classical guitar. Of which, I hope to maintain motivated, increase my knowledge, and continue to develop spiritually throughout my lifetime.

I'd like to thank the first person who gave me a legitimate chance of opportunity in my athletic pursuits, Coach Andy Gray. His respect and trust has made a major impact in my life. For that, I will always strive to be the best leader at whatever I do.

I'd like to thank Levi Herman, Mitch Wylie, Brian Lohrer, Brian Farber, Kimberly Stebel, Grandma Dorothy Ryan, Grandpa "Spud" Stebel, Dr. James Bauer, Dr. Toni Doolen, Dr. Dennis Caplan and all of my friends, family, and peers for their friendships, good times, and jovial conversations. These relationships have made my time in Corvallis, Oregon easier and very enjoyable.

I'd like to thank Phyllis Helvie, Denise Emery, and Jean Robinson in the Industrial Engineering Department at Oregon State University. I've enjoyed their guidance, professionalism, and friendship while managing academic affairs.

Finally, saving the "best" for last, I would like to thank my parents Anne and Darryl Stebel for their unconditional love and support throughout this long distance journey. Their experiences, strengths, and words of encouragement have made me strive to become the person that I can be personally and professionally. I hope that I've made them proud.

As Ozzie Smith (Hall of Fame, 2002) once said, "Nothing is good enough if it can be made better and better is never good enough if it can be made best." I will use these words as encouragement to live my life with purpose.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Chapter Page

TABLE OF CONTENTS (Continued)

Chapter Page

TABLE OF CONTENTS (Continued)

Chapter Page

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 United Airlines Flight 173

On December 28th in the year of 1978, United Airlines Flight 173 (a McDonnell-Douglas DC-8), was on its final approach to Portland International Airport (PDX) when it crashed into a wooded suburb of Portland, Oregon. The aircraft had been delayed southeast of the airport at a low altitude for about one hour while the flight crew coped with a landing gear malfunction and prepared the passengers for the possibility of a landing gear failure upon landing. The plane crashed about 6 nautical miles southeast of Portland International Airport after running out of fuel. The aircraft was completely destroyed. Of the 181 passengers and 8 crew members aboard the aircraft, 10 people were killed (8 passengers, 1 flight engineer, and 1 flight attendant), while 21 passengers and 2 crew members were seriously injured (NTSB, 1979).

Figure 1.1 United Airlines Flight 173 Crashed Near Portland, OR on 12-28-78.

The National Transportation Safety Board (NTSB) performed an accident investigation and determined that the probable cause of the accident was the captain's failure to effectively monitor and attend to the aircraft's fuel state and failure to

effectively attend to crew member advisories regarding the fuel state. This resulted in the exhaustion of fuel to all engines. The captain's preoccupation with a landing gear malfunction and preparations for a possible emergency landing was a result of the captain's inattention, poor multi-tasking performance, and/or failure to prioritize between operating tasks. Failure of the 2 other flight crew members, either to fully comprehend the criticality of the fuel state or to successfully communicate their concern to the captain, contributed significantly to the accident (NTSB, 1979). Perhaps this accident could have been prevented with the existence of a valid research tool that could predict how well a pilot might manage multiple concurrent tasks.

This accident, although unique in its own respect, is not unique in nature. About two-thirds of all commercial air transportation accidents are caused by, at least in part, by "pilot error." Directly related, a significant number of these errors, like the example above, can be classified as Concurrent Task Management (CTM) errors. CTM is the process by which operators (or pilots for example) selectively attend to tasks so as to safely, effectively, and efficiently complete a mission. A CTM error is caused by inappropriate attention to less important tasks while neglecting more important tasks. Previous CTM research set out to develop and evaluate means for training pilots to manage concurrent tasks more effectively so as to avoid errors. This included the modeling of CTM, understanding task prioritization, learning more about cognitive abilities, and the development and validation of new tools for human performance measurement of CTM. An imperative need for valid tools to study CTM has emerged.

Following previous CTM studies on theory, modeling, training, aiding, and cognitive abilities, the main goal of this research study was to evaluate and determine the validity of a tool called the Task Management Environment (TME) to determine if it can be used to predict pilot CTM performance.

1.2 Overview

This thesis begins with a review of existing literature related to Concurrent Task Management (CTM), presented in Chapter 2. It attempts to introduce and define CTM and the theories of psychology and engineering with respect to CTM. Then, a

fundamental research question emerges: How do we measure CTM from a human performance point of view? This will introduce an aviation simulation tool that is used in industry and a developmental task management research tool, define the validation process, and introduce significant validation studies.

Chapter 3 briefly covers the research objectives and hypotheses for this study.

Chapter 4 documents the experimental methodology of this study including experimental setup, methods, sample population, equipment, and measures. The study was performed using a Flight Training Device (FTD) called Frasca 141 and a software program called the Task Management Environment (TME).

Chapter 5 documents the statistical results of the data collected in this study. It includes demographic, objective, and subjective information. It gives significant implications towards external validation of the TME.

Chapter 6 summarizes the findings (especially from a correlation analyses), discusses why the results were surprising, gives implications toward CTM/TME research, discusses the limitations of this study, introduces recommendations for future use of the TME, and gives a brief summary and conclusions.

2 Literature Review

2.1 Concurrent Task Management

2.1.1 Introduction

In recent history, air transportation has been recognized as a statistically safe means of travel, although aircraft incidents and accidents still occur. On January 1, 2005, the Aviation Safety Network released accident statistics for the year 2004 showing a record low total of 425 airliner accident fatalities overall, as a result of 26 accidents. In contrast, the second-safest year which was 1955, recorded 572 fatalities. With regards to the number of accidents, just one year was safer: 2003, when 25 accidents occurred. The decreasing number of accidents aligns with the downward trend that started in 1989 (Ranter, 2005). However, due to the large demand for air travel, there is still a major cause for concern with respect to incidents and accidents. Boeing's statistics of (commercial jet) aircraft accidents reveal that flight crew errors accounted for 70% of the 149 hull loss accidents (destroyed aircraft) in worldwide commercial fleets through the period of 1988-1997 (Boeing, 2004).

Although there has been a decrease in the number of aircraft incidents and accidents, aircraft safety has remained an important issue to the aviation community and the public. Current aircraft are equipped with sophisticated technology to help aid in achieving safe flight, but accidents still occur with disastrous consequences. About 60 to 90 percent of accidents and incidents in complex systems (such as an aircraft cockpit) are attributed to *human error* (Reason, 1990; Wickens, 1998).

In relation, the aircraft cockpit has become a complicated and advanced system. This has caused new problems concerning the ability of a pilot to attentively and effectively manage this complex environment. In flight simulator studies of pilot human error, vigilance, and decision-making, it was found that the pilot's lack of or inability to effectively manage cockpit resources and instruments was a common cause of human errors (Ruffel-Smith, 1979).

In the cockpit, a pilot performs multiple concurrent tasks to accomplish a flight mission. For example, the pilot may have to simultaneously level the aircraft at an

assigned altitude and switch communication frequency to talk with Air Traffic Control (ATC) while approaching a landing. The pilot acts as a human systems manager. His or her job is to monitor systems, allocate resources to systems, and make educated decisions to optimize the management of these tasks to the best of his or her ability. This role of managing tasks, performed by a pilot, was initially labeled *Cockpit* Task Management (CTM) by Funk (1991). Its name was later changed to *Concurrent* Task Management, since the management of multiple concurrent tasks is not unique to the aviation domain and the appropriate change allowed the maintaining of the same mnemonic, CTM. A CTM error occurs whenever a person attends to a task of less importance when there is a higher priority task present that requires attention (Funk, Colvin, Bishara, Nicolalde, Shakeri, and Chen, 2003).

Funk (1991) developed a Normative Theory of CTM to describe pilot activities in the cockpit. This theory included the definition of a task management agenda, assessment of the situation, activation of tasks, assessment of progress and status of active tasks, termination of tasks, assessment of task resource requirements, prioritization of tasks, allocation of resources, and updating the task management agenda. Pilots must continuously assess, prioritize, execute, monitor, and terminate tasks to the best of their ability, often in dynamic situations. Not only will pilot errors occur while executing these tasks, they will also occur while managing tasks in the cockpit. Since the number of tasks in the cockpit may exceed the limited capacity of pilots' resources, effective CTM requires the proper allocation of attention resources to tasks most crucial for the safety of the flight (Bishara, 2002).

The following sections of this chapter attempt to introduce and describe CTM, psychology and engineering theories relevant to CTM, a need for CTM measuring tools, and measuring human performance. Furthermore, this will set up the validation needs of a task management environment to measure CTM, which is the main topic of this research.

2.1.2 Concurrent Task Management

The term Concurrent Task Management (CTM) was coined by Funk et al (2003) to refer to that process whereby a pilot (or operator) manages multiple, concurrent tasks

that must be performed to operate a complex environment like an airplane. CTM is not limited to the domain of aviation, although it has been applied to aviation in this research. An operator is anyone that works within a human-machine system and manages multiple concurrent tasks, such as drivers, surgeons, cooks, etc. CTM involves the initiation, monitoring, prioritization, interruption, resumption, and termination of tasks in such a way as to safely, effectively, and efficiently complete a mission within a complex environment.

Complex environments (or systems), such as an aircraft cockpit, an operating room, or an automobile cockpit, impose a high level of resource demand on human operator capacity. This capacity is limited by the amount of multi-tasking workload that an operator can manage. Technological advancements and improvements in these complex environments (or systems) have led to the support, complement, and extension of human operators' behavioral capabilities.

The process by which people attend to multiple concurrent tasks has been studied for nearly a century. McQueen (1917) identified the mechanisms by which people cope with more than one task presented at the same time. He stated,

> "When two disparate operations are simultaneously performed, introspective evidence obtained under experimental conditions is brought forward in proof of the occurrence of all four possibilities: (1) the attention alternates from one to the other; (2) one of the tasks becomes automatic and drops out of consciousness, or both; (3) the processes combine into a psychical fusion, an easy form of fusion being through the medium of rhythm, though the rhythmisation may be unconscious; (4) there may be simultaneous attention to the two tasks as separate."

McQueen argued that attention to two tasks simultaneously is very rare. He argued that it is more likely that conscious attention is directed to only one of the tasks while the other is being performed automatically.

2.1.2.1 Multi-Tasking

Simply, CTM could be thought of as the management of multi-tasking. Psychologists and human factors scientists (or engineers) have recognized multi-tasking

performance to be very important. In fact, it has been thought of as the key to successful performance in a complex system. To understand this phenomenon, researchers have developed a number of theories to define and measure multi-tasking behaviors. A few significant studies have investigated and defined multiple task performance, mental workload, and multi-task behavior.

Multiple Task Performance is a major component of multi-tasking. There is an abundance of literature on this mechanism. Damos (1991) set up experiments where the subject performed two abstract tasks simultaneously. She studied task difficulty and other factors to determine the effects on tracking accuracy, arithmetic response time and accuracy, and other performance measures.

The concept of *Mental Workload* has emerged from the literature on multi-tasking. Although workload is loosely defined, it refers to the resources required by a set of concurrent tasks. There is a vast amount of research as to how to measure this but it has been widely accepted that there is always an optimal level of workload in performance, and an individual's composite task performance will degrade if the demand is above the optimal level of workload (Funk, Colvin, Bishara, Nicolalde, Shakeri, and Chen, 2003).

Multi-Task Behavior encompasses the concept of dual-task behavior. Many multi-tasking theories of human behavior are based on the analogy of multi-tasking in a computer operating system where mental resources are likened to computer memory and processor time. Tasks are likened to processes and some sort of executive routine allocates resources to a task. Theories and models based on engineering methods have been developed to describe these behaviors (Pattipati and Kleinman, 1991; Rubenstein, Meyer, and Evans, 2001).

2.1.2.2 Attention

An operators multi-tasking performance is dependent on his or her ability to focus his or her attention on the tasks at hand. *Attention* is the mechanism that allows us to concentrate all of our cognitive conscious resources on a specific event or task. Attention is the limiting factor that determines how many tasks we can attend to concurrently. For example, the driver of an automobile needs to remain attentive in order to successfully

attend to driving tasks such as adjusting the speed, changing direction, communicating with other drivers and pedestrians, adjusting the stereo, and talking on the cell phone. When a driver is in a situation of high resource demand, such as when he or she approaches high traffic, it is more difficult to focus on all of the tasks previously mentioned. In order to remain safe, the driver would need to interrupt less important tasks, such as talking on the cell phone, and focus his or her attention on operating the automobile successfully.

2.1.2.3 Multiple Resource Theory

To account for the human ability to perform more than two tasks simultaneously, Wickens (1980) proposed a *Multiple Resource Theory* (MRT). In Wicken's MRT, human mental capacity can be viewed as a collection of limited, differentiated resources that must be allocated among competing tasks. If a task receives a full allocation of required resources, performance (e.g., speed and accuracy) will be good. If resources are withdrawn from a task, performance will deteriorate. Multiple resource theory accounts for most of the dichotomies explored in the dual-task experiments in which performance in the primary task is not affected by increasing difficulty in the secondary task. However, in the context of multiple concurrent task environments, how does multiple resource theory explain performance in more than two simultaneously occurring tasks? In this special instance, the human operator cannot be conceived as an infinite channel of information processing. Instead, the limited resources have to be managed.

2.1.2.4 The Normative Theory of CTM

As mentioned previously, Funk (1991) developed a normative theory of CTM to help define CTM and characterize it. CTM is described as a procedure that is executed by the flight crews to manage cockpit tasks. A statement of the theory depends on definitions for several key terms.

Behavior was defined as a collection of a system's input, state, and output values over time. A system exhibits a behavior if observed values of input, state, and output values match those of the behavior. For instance, by increasing the throttle settings (input), the aircraft accelerates to rotation speed (state), and begins to fly (output). The

aircraft (system) exhibits the flying behavior by matching the inputs, state, and outputs of the flying behavior.

A *Goal* was defined for a system, such as an aircraft, as a set of desired behaviors. If one of the behaviors is realized, then the goal is achieved. Otherwise, the goal is not achieved. For a commercial air transport mission, the primary goal might be to transport people over large distances in a time-effective, comfortable, economic, and safe manner.

A *Task* was defined as a process that is completed to cause a system to achieve the goal. The execution of a task to achieve a goal requires resources. For example, to prepare an aircraft for departure, resources from the human system may be necessary. Generally speaking, tasks require resources to achieve the goal.

A *Resource* was defined as something that may take the form of an equipment or aid, such as radios, displays, controls, or autopilot. Human resources include people, such as a captain, first officer, or flight engineer. Given some key definitions about CTM in a system, it is also important to understand that tasks can be in a latent or active state.

Psychologists, pilots, and human factors scientists have come to recognize that it is not only difficult to successfully accomplish tasks in a multi-task environment, but it is often even more challenging to manage them. To address this challenge, many pilots try to use a simple prioritization technique known as the Aviate, Navigate, and Communicate (ANC) acronym. It is based on categories of cockpit tasks. *Aviate* is defined as keeping the airplane in the air and on a proper heading. *Navigate* is defined as determining where to go and how to get there. *Communicate* is defined as talking with air traffic control personnel and flight crew.

However, implementation of the normative theory of CTM can be used to prioritize tasks more specifically. In order to accomplish this, Funk described a procedure that begins with the assessment of a current situation and ends an updated agenda. The procedure works in a continuous cycle and can be portrayed as follows:

- **Assess the Current Situation**
- **Activate Tasks** (whose initial events have occurred)
- **Assess Status of Active Tasks**
- **Terminate Tasks** (with achieved or unachievable goals)
- **Assess Task Resource Requirements**
- **Prioritize Active Tasks**
- **Allocate Resources** (to tasks in order of priority)

- **Initiate New High Priority Tasks**
- **Interrupt Low Priority Tasks** (if necessary)
- **Resume Interrupted Tasks** (when possible)
- **Update Agenda**

Colvin (1999) developed a visual representation (Figure 2.1) from Funk's description of the CTM process while trying to accomplish a mission goal. Given a hierarchy of goals to accomplish during a mission, the first step in this theory proposes the creation of an initial agenda.

This agenda consists of two factors. The agenda must have a task, in order to achieve each goal, and an initial event. Once an agenda has been established, the process of agenda management begins and continues until the mission goal is achieved or has been determined unachievable. If the mission is determined to be unachievable, then the process should end only after the aircraft and its subsystems reach a safe state.

Here, the pilot must assess the current situation. The states of all relevant aircraft systems and subsystems must be considered to determine if significant events have occurred. When initial events occur, the pilot must activate tasks that are contingent upon those events. Then the pilot must assess the status of active tasks to determine if satisfactory progress is being made toward achieving the task's goal. Now, the current status of each task should be forecasted to determine the likelihood that the goal will be achieved. A tasks' status may be declared satisfactory if its goal is achieved. Based on this assessment, the pilot should terminate tasks with achieved or unachievable goals. Due to changing circumstances, task goals may become irrelevant. Termination of tasks will reduce the competition for resource allocation. Then the pilot should assess remaining task resource requirements to determine what resources are required to complete them, by prioritizing active tasks. This includes assessing the urgency of a task goal, importance and urgency of other task goals, current and projected status of the task, and the current and projected status of other tasks. Some research suggests that CTM requires a strategy.

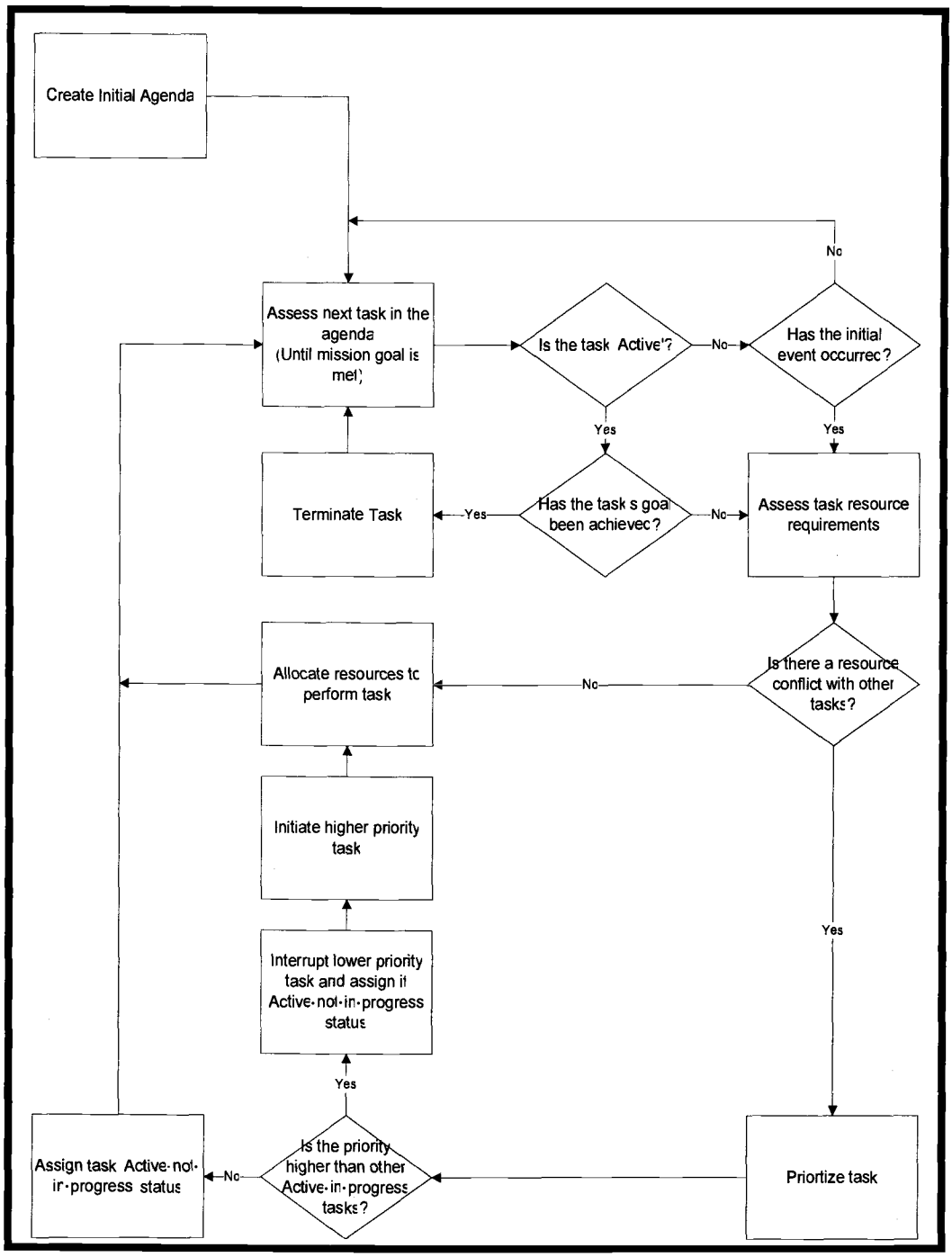


Figure 2.1 A Visual Representation of the Normative Theory of CTM.

2.1.2.5 Strategic Workload Management

The concept of Strategic Workload Management was initially introduced by Hart (1985) and has been studied and summarized by others (e.g., Raby and Wickens, 1994; Adams, Tenney, and Pew, 1991). Rogers (1996) expanded on Hart's work and Funk's

Normative Theory to propose a preliminary set of discrete flight deck CTM processes. These processes, according to Rogers, occur in the following order:

- **Assess Situation**
- **Identify Tasks**
- **Prioritize Tasks**
- **Assess Resources**
- **Allocate Resources**
- **Schedule Tasks**
- **Perform Tasks**
- **Monitor Tasks**
- **Manage Interruptions**

Rogers's enhancement of strategic workload management incorporated the concept of *Task Interruptions* and *Task Scheduling* in a more explicit way as opposed to Funk's theory, which includes them implicitly. The concepts of task interruptions and task scheduling are linked to other approaches of modeling human behavior in complex systems. For example, Shakeri (2003) proposed that tasks have to be scheduled and interrupted strategically and Pew and Mavor (1998) studied the area of strategic task switching.

Rogers presented three important conclusions with respect to the nature of CTM. First, he argued that CTM has two modes, strategic and tactical. *Strategic* CTM refers to phases of operation in which there is little time pressure. In this case, the operator creates a strategy and a plan to perform the tasks in order to avoid high workload moments. The *tactical* mode of CTM refers to moments of high workload and time pressure. In these cases, the operator acts in a more reactive manner where he or she extracts skill based knowledge from his or her memory in the case that a moment of high workload exists for the present tasks. In addition, the operator might opt to shed some of the tasks.

Rogers argued that CTM is time-driven, meaning that the primary task characteristic that influences task prioritization is urgency. *Urgency* was defined by Funk (1991) as the ratio of the time that it would take to complete the task or bring it to a satisfactory before its deadline.

Finally, Rogers argued that tasks are categorized into "discrete real-time tasks, discrete pre-planned tasks, and continuous or repetitive tasks." He argued that discrete tasks are ordered along a priority or time dimension and continuous tasks are interleaved

with discrete tasks but not explicitly ordered."

A thorough understanding of CTM has helped shed light on how operators manage multiple concurrent tasks in complex systems. The studies by Hart (1985), Funk (1991), and Rogers (1996) have made significant contributions to CTM research. A better understanding of CTM could help in the prediction of human performance. Thus, the prediction of human performance in a complex system could prevent CTM errors that have lead to incidents and accidents. Several studies described below have shown that CTM errors have contributed significantly to aircraft accidents and incidents.

2.1.2.6 CTM Error Taxonomy

The National Transportation Safety Board (NTSB) has an incident/accident database. This database is an official record of U.S. aviation incident/accident data, including causal factors. The NTSB recognizes an aircraft accident as an occurrence in which a person (occupant or non-occupant) receives fatal or serious injury or any aircraft receives substantial damage.

As part of a study of NTSB accident reports, Chou (1991) developed a CTM error taxonomy. Technically, each variable in the taxonomy (e.g. task initiation, task monitoring, task prioritization, etc.) is a process within CTM. They are not classes of errors. A specific class of error would be an error occurring in one of the processes. The CTM error taxonomy is as follows:

- **Task Initiation Error:** inappropriate initiation of a task (e.g. too early or too late).
- **Task Monitoring Error:** inappropriate assessment of task progress and status (e.g. satisfactory or unsatisfactory).
- **Task Prioritization Error:** inappropriate assignment of task priorities relative to their importance and urgency (e.g. attending to a low priority task before a high priority task).
- **Resource Allocation Error:** inappropriate assignment of human and machine resources to tasks (e.g. focusing cognitive attention or resource on a landing gear light malfunction rather than managing the direction/vertical speed/altitude of the airplane).
- **Task Interruption Error:** inappropriate suspension of lower priority tasks so that resources may be allocated to higher priority tasks (e.g. fail to shed a communicating task when an aviating task is higher priority).

- **Task Resumption Error:** inappropriate resumption of interrupted tasks when priorities change or resources become available (e.g. failure to resume task when it becomes a high priority).
- **Task Termination Error:** inappropriate termination of tasks including those that have been completed, that cannot be completed, or that are no longer relevant (e.g. continue to monitor gyroscope while landing).

After studying the NTSB incident/accident reports, Chou identified 80 CTM errors in 76 of the 324. His findings conclude that CTM errors played a significant role in 23% of the accidents reviewed. This evidence shows that CTM is a significant factor in flight safety.

To further support Chou's conclusion, he conducted a flight simulator study to observe CTM errors similar to those identified in the incident/accident analyses. The participants performed several flight scenarios using a low-fidelity flight simulator. Meanwhile, their CTM behavior was observed. The results from an analysis of variance (ANOVA) showed that both mental resource requirements (in combination with flight path complexity) and the number of concurrent tasks created significant effects on task prioritization. This study confirmed that an increased workload can have negative effects on task initiation and task prioritization performance, which increases the likelihood of CTM errors.

2.1.2.7 Cockpit Task Prioritization Factors

Colvin (2000) developed a study to identify which factors pilots use to determine task priority and which tasks they will allocate their attention to. In his study, pilots flew arrival procedures in a part-task simulator. Two knowledge elicitation techniques (intrusive and retrospective) were used to probe the subjects for factors that influenced their attention prioritization scheme while performing multiple concurrent flight tasks.

Colvin concluded that 12 factors affected task prioritization in his study. Two major factors were task status, with a total of 51 instances (30%) reported, and task procedure, with 48 instances (28%) reported. Other important factors that were reported frequently included verifying information, reported 13 times (8%) and task importance, reported 12 times (7%). Other factors that were reported less frequently included: rate of

change, needed information, urgency, time/effort required, salience of stimulus, consequences, resist forgetting, and expectancy.

Colvin's study provides a better understanding of the task prioritization processes in the cockpit multi-tasking environment.

2.1.2.8 CTM Aids

Funk and Chou (1991) set out to develop CTM aids to help pilots to manage tasks and reduce CTM errors in the cockpit. Three general approaches to this are CTM aiding, training, and simulation. Funk and Chou (1991) and Chou, Madhaven, and Funk (1996) made several recommendations for a computer based system to help pilots manage tasks better. The first version of a CTM aid was the Cockpit Task Management System (CTMS). The CTMS, which operated in a part-task flight simulator environment, consisted of software and a display that computed and presented to the pilot information on task state (present / future), status (satisfactory / unsatisfactory), and importance to flight safety. It displayed a list of up-coming tasks and in-progress tasks, and when appropriate, recommended which tasks to attend to. In an experimental comparison between the CTMS with an unaided condition, an analysis of variance (ANOVA) showed that the CTMS reduced task prioritization error by 41% at an α=0.1 significance level and it reduced the number of incomplete tasks by 82% at an α=0.05 significance level.

Funk and Braune (1999) developed another CTM aid called the AgendaManager (AMgr). It was meant to help pilots prioritize tasks more effectively based on task importance with respect to flight safety. The AMgr used a speech recognition system to determine pilot goals by decoding clearance read-backs (e.g., "Roger, climb and maintain 15,000 feet"). Then software modules, called *function agents*, monitored the progress of each task. A display informed the pilot of the status of each task. In the display, higher priority tasks were given greater prominence (Figure 2.2).

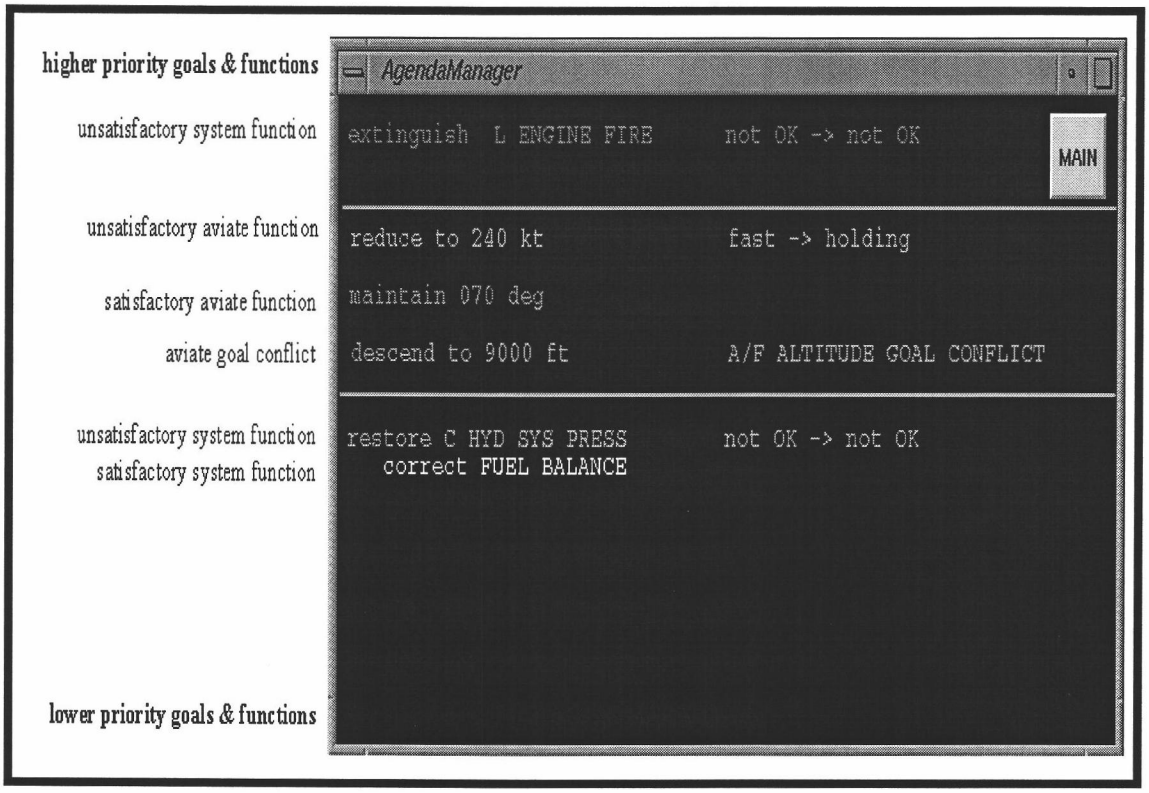


Figure 2.2 The AgendaManager (AMgr) Two Part-Task Simulator Interface.

An experimental comparison of the AMgr with a model of a conventional warning and alerting system showed that the AMgr improved CTM performance. Ten airline pilots flew two challenging flight scenarios (one with the AMgr and one without). The pilots correctly prioritized 72% of the time with the AMgr but only 46% of the time with the conventional system at an α=0.05 significance level in an ANOVA. Also, with the AMgr, participants managed to keep concurrent tasks at a satisfactory level for 65% of the time versus 52% in the conventional system at an α=0.05 significance level in an ANOVA.

Both aiding systems, CTMS and AMgr, were effective in part-task simulator evaluations and therefore suggest that aiding may be an effective way of improving CTM performance in an operational environment. However, the speech recognition technology used in the AMgr would not have been reliable in a demanding cockpit acoustic environment.

2.1.2.9 Training

Since the CTMS and AMgr aids seem to have some challenges (high workload demand and validation), improving CTM performance through effective training was investigated. Bishara and Funk (2002) undertook the development, delivery, and evaluation of CTM training in two studies. They set out to develop methods to train pilots to properly prioritize tasks to avoid interruptions and distractions, and to test the effectiveness of the training in a part-task simulator. Implementing a successful CTM training program would have many advantages and it would not incur the cost of new equipment and software.

In the first study, 12 instrument rated pilots were randomly assigned to 3 training groups: control, descriptive CTM training, and prescriptive CTM training. Each participant flew a difficult IFR pre-training flight scenario in Microsoft Flight Simulator 2000 in which 19 challenge points occurred. *Challenge points* were used to identify task prioritization errors committed by the pilot at specific instances in a flight simulation scenario. A prioritization error rate was calculated based on the percent of challenge points successfully accomplished. Following the first scenario, the control group received no training while the other two groups received CTM training.

The descriptive group received verbal training that included CTM definitions, theories, and suggestions on how to be aware of potential CTM errors. The prescriptive group received a mnemonic technique to reinforce CTM awareness. This mnemonic was called *APE*: *A* for assess, *P* for prioritize, *E* for execute.

Both CTM training groups showed an improvement in CTM performance from pre-training to the post-training flight (significant at the α=0.05 level in ANOVA) and the control group showed no change in performance. The results suggest that CTM training was effective (Figure 2.3). The prescriptive and descriptive groups made significant reduction in prioritization error rates.

In conclusion, the findings by Bishara and Funk suggested that CTM can be trained and that CTM performance can be improved by training.

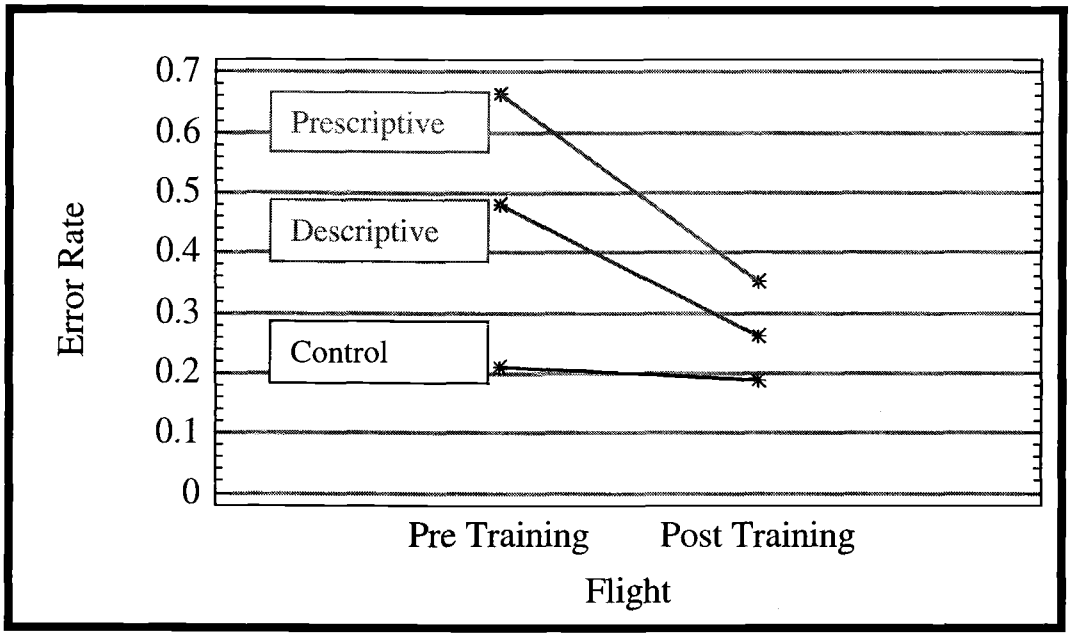


Figure 2.3 Mean Task Prioritization Error Rates for Pre-Training and Post-Training.

2.1.2.10 Studies of the Nature of CTM

In an effort to better understand the nature of CTM and how to improve it by training, Colvin and Funk (2000) and Bishara and Funk (2002) used part-task simulator studies. Much like these studies, Shakeri and Funk (2003), Nicolalde, Uttl, and Funk (2003), and Chen and Funk (2003), used a similar approach to investigate CTM performance by using a much simpler software program which serves as a multi-tasking environment in which the behavior of multiple subsystems were modeled.

Shakeri and Funk (2003) developed *Tardast* (Persian for juggler) as part of a National Aeronautics and Space Administration (NASA) grant to analyze and measure human multi-tasking performance. It was developed and used for studying the nature of CTM due to the challenges of using a flight simulator. The name *Tardast* was later changed to the name Task Management Environment (TME).

A study by Shakeri and Funk, using the TME, concluded that none of the participants could beat the near-optimal score of a heuristic search algorithm in any of the scenarios. It was observed that participants overreacted to poor task performance penalties by attempting to handle too many tasks. It was also found that participants' strategic task management was more significant in getting a good score than their tactical task management.

To learn more about cognitive abilities and CTM, Nicolalde, Uttl, and Funk (2003) compared cognitive abilities with task management performance and developed an augmented version (adapted from Wickens and Hollands, 1999) of the *Stage Model of Human Information Processing* to include several other cognitive processes (Figure 2.4).

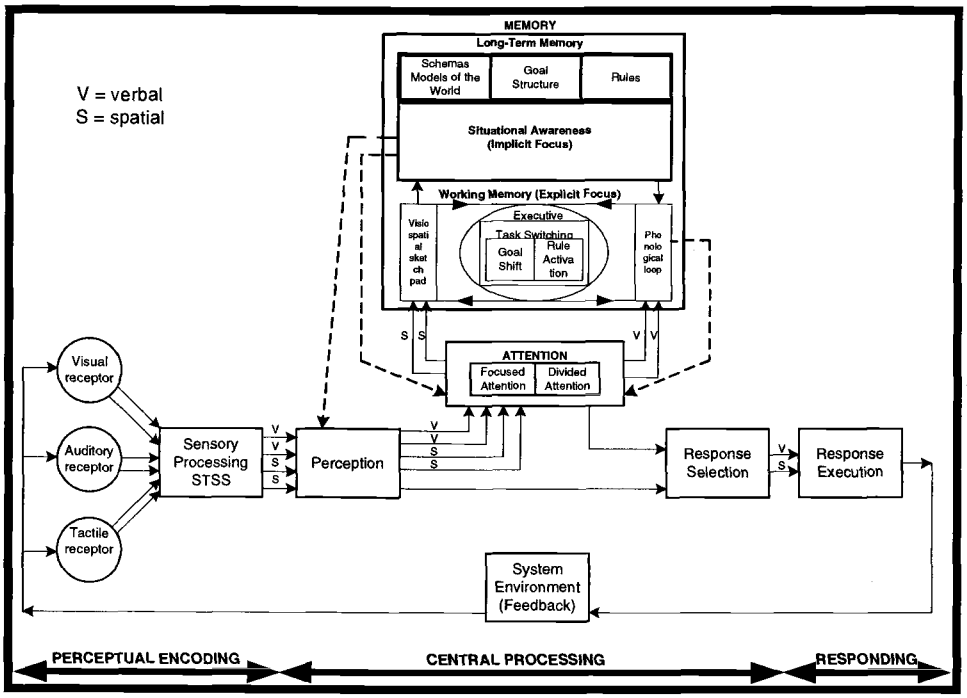


Figure 2.4 The Augmented Stage Model of Human-Information Processing.

To assess cognitive ability, one part of the study that was used for comparison included cognition tests (e.g. Simple Reaction Time and Card Sorting Test) and neuropsychological tests (e.g. Path Finding Test, Synonym Test, and Melbourne Decision Making Questionnaire).

In the second part of the study, cognitive performance was compared to task management performance using the TME to determine if cognitive performance can help predict CTM performance. The fundamental objective of this study was to determine if participants who scored well in certain cognitive tests, also performed well in the TME. The results showed that there were low correlations between cognitive measures and TME measures (Figure 2.5).

	AGE	Gender	Verbal IQ	Working Memory	Reaction Time	Decision Time	TME-EZ	TME-DF
AGE								
Gender	**-0.14**							
Verbal Intelligence	0.32	**0.20**						
Working Memory	0.08	0.06	**-0.22**					
Reaction Time	0.14	0.21	0.02	**0.23**				
Decision Time	0.14	0.06	-0.06	0.55	**0.51**			
TME-EZ	-0.10	-0.02	0.05	-0.02	-0.12	**-0.05**		
TME-DF	-0.07	0.01	0.08	-0.15	-0.17	-0.11	**0.49**	

Figure 2.5 Correlations of Cognitive and TME Performance.

Nicolalde, Uttl, and Funk concluded that this suggested that CTM cannot be explained in terms of a few simple cognitive processes. While these processes might be components of CTM, it is more likely that CTM is a complex combination of them, drawing on working memory and other mental resources.

Chen and Funk (2003) used data from Nicolalde and Funk's study, in which participants played the Task Management Environment (TME), to assess their task management performance. Chen and Funk used five fuzzy models that incorporated combinations of factors which were developed to model human task management performance in a simulated task management environment.

Chen and Funk concluded that a comparison of model predictions with human participant data suggested that human task management strategies consider task importance, task status, and task urgency in choosing which task to attend to next. Also, they concluded that the ability to shed lower priority tasks (based on task importance) is a very important determining factor of task management performance.

2.1.2.11 Need for Tools to Predict CTM Performance

Studies by Shakeri and Funk (2003), Nicolalde, Uttl, and Funk (2003), and Chen and Funk (2003) have made significant contributions to CTM research. However, little has been proven about the validity of their studies since the tool that was used for measuring task management performance was the TME. With respect to measuring CTM performance, there is a need for valid tools that can predict CTM performance accurately, precisely, and reliably.

2.2 Simulation Tools for Studying Concurrent Task Management

2.2.1 CTM Simulation Tools

Previous studies of Concurrent Task Management (CTM) have used part-task flight simulators to create challenging multitasking environments in which to study CTM behavior. Such an environment has the advantages of being complex enough to be interesting, while still having face validity with comparison to a real airplane's cockpit displays, controls, and tasks. In order to understand CTM fully, we must conduct research using tools that yield findings generalizable to the "real-world."

This study compared CTM performance measured using the TME with that obtained in a more realistic flight simulator to determine the TME's external validity. The remainder of this chapter continues the literature review with material relevant to the study.

2.2.1.1 The Task Management Environment

To overcome the disadvantages of using a "real-world" simulator, the Task Management Environment (TME) was developed. The TME is a software program that simulates an abstract system composed of up to 15 simple, dynamic subsystems. The TME was developed in the Microsoft Visual Basic 6.0 programming language. The program is compatible with a Microsoft Windows operating system.

The TME serves as a multi-tasking environment in which the behavior of the multiple subsystems can be specified by the experimenter. In this study, this TME program was investigated to determine if it can be used as a tool to assess a participant's performance in attending to and managing multiple concurrent tasks.

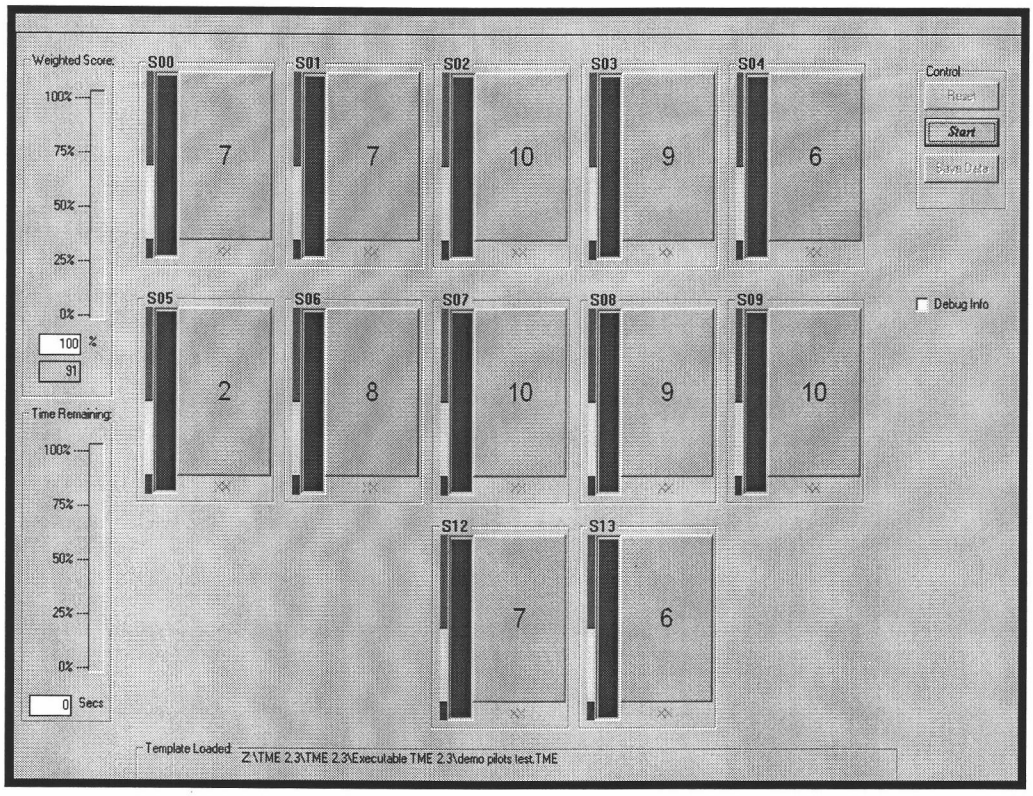


Figure 2.6 A Computer Model of the Task Management Environment.

In the TME (Figure 2.6), each subsystem has parameters that define how fast the status bar will increase or decrease indicating an improvement or deterioration of the subsystem's status. Two of the main parameters are the *Correction Rates* (CR) and *Deterioration Rates* (DR). These parameters can be adjusted to account for the dynamic nature of each subsystem. Mancayo and Funk (2003) and Shakeri, and Funk (2003) defined these parameters as follows:

- **Correction Rates (CR):** The rate at which a subsystem will correct or increase the accomplishment of a subtask.
- **Deterioration Rates (DR):** The rate at which a subsystem will deviate or decrease the accomplishment of a subtask.

There are two types of subsystems within the TME software program, *Continuous Subsystems* and *Discrete Subsystems*. In a continuous subsystem, if unattended to by an operator, the normal behavior of the individual continuous subsystem's status decreases at a constant rate of DR, until it reaches 0%. This type of subsystem is managed by clicking the mouse button to the right of the particular subsystem's status bar on the grey button. Here, the status will increase at a constant rate of CR until it reaches 100%.

When the operator, releases the mouse button the status begins to decrease at a rate of DR again (Funk et al, 2003).

The behavior of a discrete subsystem is very similar, except that its normal status remains at 100%, even without operator attention until a random "subsystem failure" event occurs. At that time, its status decreases at a rate of DR until the operator clicks the subsystem's button. The decreased status level will hold for a few seconds (controlled by the experimenter) during which the subsystem's button disappears. Then, the button reappears and the status continues to decrease at a rate of DR until the operator clicks the button again and the decreasing rate again pauses. This continues until a pre-determined number of clicks restore the subsystem status fully to 100%. Discrete subsystems will remain at a 100% status until another random "subsystem failure" event occurs (Funk et al, 2003).

Each subsystem presents the participant with a simple control task similar to that of a "real-world" system. A continuous control task in the TME simulates a continuous control task in a "real-world" system (e.g. a pilot trying to control the aircraft at a precise altitude or direction or speed). If the continuous control task is performed successfully, the corresponding TME task status would be 100% for that continuous subsystem. A discrete control task in the TME simulates a discrete control task in a "real-world" system (e.g. a pilot trying to lower the landing gear or restart the engine during flight). If performed successfully, the corresponding TME task status would be 100% for that discrete subsystem. As in the "real-world," an operator can often perform just one task at a time due to human limitations and capabilities, the TME allows the participant to perform just one control task (operate just one subsystem) at a time (Funk et al, 2003).

The goal of the TME participant is to maintain or keep each subsystem at a satisfactory status level. To do so, the subsystem must be in the green (50-100%) range. The participant must try to avoid entering the yellow (10-50%) range or the red (0-10%) range. The TME automatically calculates a CTM performance score based on this objective. The TME calculates a *Total Weighted Score* (TWS), which is the summation of all subsystem cumulative scores and reflects overall task management performance. The TWS for a subsystem at any time is qi, where the variable q is a qualitative transform of the subsystem's current status level as follows:

- **q=+1,** if its status level is **satisfactory**
- **q= 0,** if its status level is **unsatisfactory**
- **q= -1,** if its status level is **very unsatisfactory**

The variable *i* is the subsystem's importance, the number appearing directly to the right of the subsystem's blue status bar in the interface. For example, for a subsystem whose value is five (i=5), a participant will earn a value of five points (qi= 1x5) for a subsystem if maintained in the green status, zero points (qi= 0x5) if maintained in the yellow status, and negative five points (qi= -1x5) if maintained in the red status.

2.2.1.1.1 *How to Play*

The object of the game is for the participant to keep the blue bar in the green zone as best as he or she can to achieve a high score. The higher the value of importance *i*, the more a subsystem's status will affect the participant's final score. If the blue bar is in the green zone, the participant will score points. If the blue bar drops to the yellow zone, the participant will not gain any points. If, the blue bar drops to the red zone, the participant will lose points negatively.

To get the blue bar to rise; a participant must "Left-Click" (using the computer mouse) on the large numbered button located to the right of the blue bar until he or she thinks that the blue bar is at a satisfactory position for the time being. The participant will have to either: "Click and Hold" or "Click and Release," depending on how the individual task functions (continuously or discretely). Then, the participant should try to attend to the other tasks and raise their blue bars to satisfactory positions.

To begin play, a participant must "Left-Click" on the "Start" button located in the upper right hand of the screen using the mouse. The blue bar for each subsystem (depending on how many are included in the set up) will begin to drop at different speeds.

2.2.1.2 Flight Training Device

A Flight Training Device (FTD) is a simulation tool that can be used for pilot training to practice every aspect of flying an airplane including take-off, climb, cruise, descent, and landing. Knowing that the practice of flying an airplane would be very

dangerous and require high monetary costs, an FTD is used to prevent dangerous incidents/accidents and reduce the costs needed to acquire the amount of acceptable training required for various levels of aviation proficiency. There are various FTD simulators that are used for the different levels of aviation ranging from recreational to commercial to military.

At Oregon State University, research on aviation and CTM has been limited to the use of an "off-the shelf" video game called Microsoft Flight Simulator (MFS) including a toy yoke and toy pedals). Although the instrumentation has some degree of realism for training practice, it is not recognized by the Federal Aviation Administration (FAA) as a certifiable training tool due to its lack of realism. An FAA certification means that a particular FTD is recognized as "close-to the real-thing" and it can be used as a training tool to practice flying an airplane. MFS has many disadvantages. It is very static, it has no force-feedback, and any non-pilot can operate it and perform at a high level.

However, some FTD's are certified by the FAA because they can simulate flying an airplane to a very high degree of realism. These simulators include actual mockup instrumentation that looks like a real airplane (having face validity), they have the design of a cockpit, force-feedback within the controls and require a qualified level of pilot experience.

2.2.1.3 The Frasca 141

In this research, data that was collected from another experiment (Hoover, 2005) was used for comparison. In Hoover's experiment, she used the Frasca 141 FTD to simulate flying an airplane. The Frasca 141 is intended to simulate light general aviation aircraft (Figure 2.7).

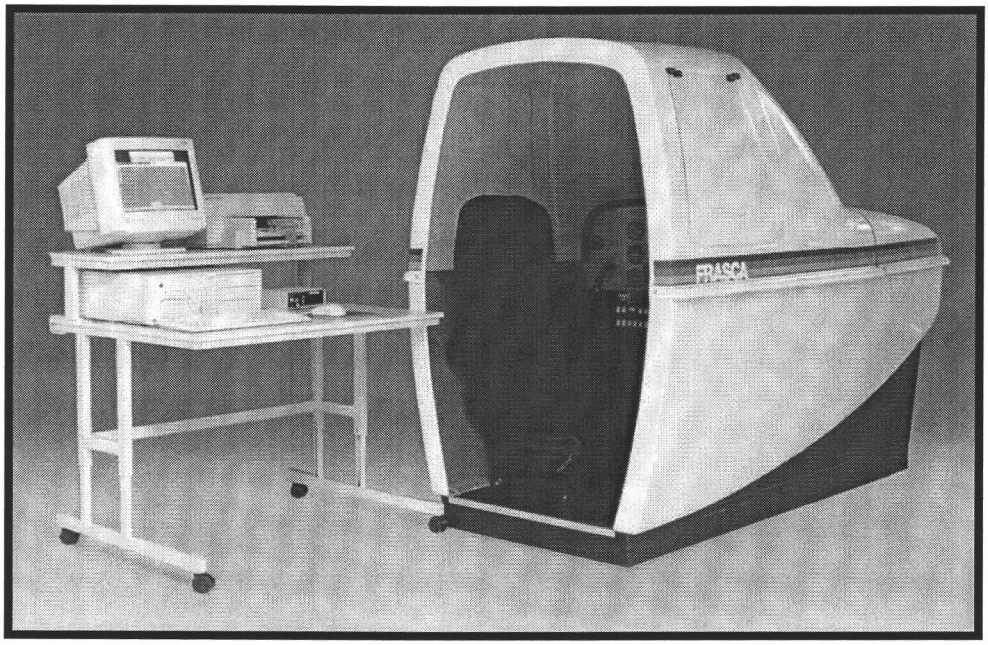


Figure 2.7 The Frasca 141 Simulator and Instructor's Station.

The Frasca 141 was used because it has the capabilities of accommodating recurrent flight training, incurs low simulation costs, and has a high degree of realism. The cockpit equipment and systems' controls and indicators were modeled after actual aircraft parts. Hoover used the latest Frasca 141 technology to provide reliable and realistic simulation (Frasca, 2005).

Human performance in the Frasca 141 can be measured without basing performance on predicted flight data alone. It represents high fidelity in flight simulation from starting the engine, runway taxi, takeoff, climb, cruise, slow-flight, stalls, descents, and landing.

The Frasca 141 has numerous features. These include: FAA approval under 14 CFR parts 61 and 141, guaranteed FAA Level 2 or 3 qualification, Jeppesen Navigational Database for the entire continental United States, graphical instructor station, true vision, and multiple performance configurations.

The Frasca 141 FTD includes a control yoke, foot pedals, Global Positioning System (GPS), operable circuit breakers, custom panel layout, Bendix/King flight control systems, Bendix/King avionics, Electronic Flight Instrument Systems (EFIS), and a wide-cockpit for a second pilot (Figure 2.8).

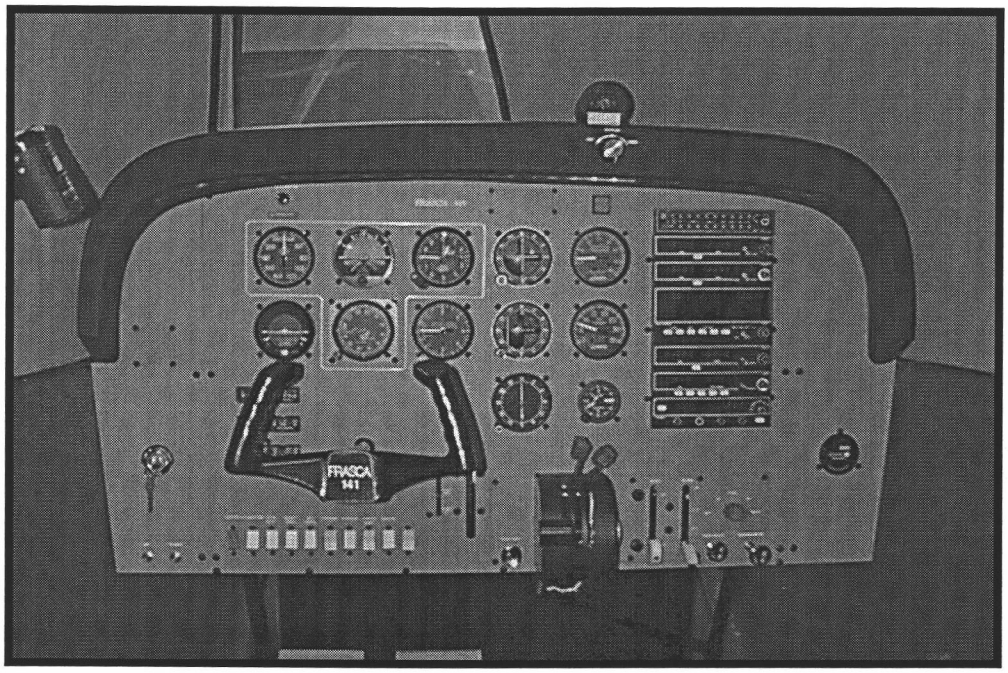


Figure 2.8 The Frasca 141 Instrument Displays and Controls.

2.3 Measuring Human Performance

2.3.1 Overview

This section will introduce a review of definitions and literature relevant to the measurement of human performance. The motivation of this section is to define the types of measurements recommended to measure human performance, to define validity, to understand the quality of test predictors, and to understand how to make inferences on a predictor test for various types of validity.

2.3.2 Introduction

Human performance is commonly measured using standards of accuracy, speed, training time, and satisfaction. *Accuracy* is defined as freedom from errors. *Speed* often means increasing the speed of human performance to a maximum. *Training* time is the total time required to bring system users to a desired level of performance. It answers the question: How much time will it take for a user to learn how to operate the system? One of the main goals of a designer is to find ways of designing activities so that training time is reduced to the minimum. *User Satisfaction* may be the most critical measurement of

all. It concerns whether or not the human performing an activity in a particular context receives satisfaction. It answers the question: Does the activity provide a reward for that particular situation (Bailey, 1982)?

2.3.2.1 Human Measuring: Objective vs. Subjective

System performance is frequently concerned with monitoring the functioning of an entire system that includes the human as a component. For example, we might be interested in monitoring the "safety" of a nuclear power plant; however there are no really clear measures that reflect safety of that overall system. Olson (1988, as cited in Wickens, 1998) attempted to identify a set of consistent and objective safety indicators for a nuclear power plant and found that "safety" was related to a number of different issues including radiation, maintenance, training, experience, etc. This was evaluated by quality, availability, and relationship. They found that no single measure was an adequate indicator of nuclear power plant safety, but there was a need for multiple indicators to properly measure the indicators.

In the application of science, traditional beliefs have focused on mostly objective measures rather than subjective measures (Wickens, 1998). Kosso (1989, as cited by Wickens, 1998) described subjective measures as those that rely on human experience, judgment, perception, or cognition. These factors are assumed to distort or unfavorably alter the reality that is actually being measured. For this reason alone, many researchers agree that measuring human performance based upon objective measures (e.g. the number of correct keys a piano player can play in 5 minutes) is more desirable than subjective measures. For example, many researchers would be more accepting of measuring mental workload from an objective measure of heart rate compared to a subjective rating system.

As Wickens points out, with computer technology becoming more advanced and widely used, it becomes easier for researchers to program a computer to measure something. For example, a number of objective measures could be obtained from a flight simulator including speed, altitude, rates of ascent/descent, etc). However, do these objective measures tell us anything about performance? If so, how do we interpret it? If we take an in-depth look at literature in regards to objective and subjective measure, it is

pretty clear that each has its own positive uses but the debate over which one to use has some issues. In 1987, Mikulincer and Hobfoll (Wickens, 1998) found that objective and subjective measures indicated that both were predictive of combat stress reaction and post traumatic disorders while subjective parameters were stronger predictors of the two.

Meanwhile, other researchers have found that some cases conclude subjective measuring is superior to objective measuring. Hennessy (1990, as cited by Wickens, 1998) notes that human performance testing and evaluation is usually conducted in "real-world" environments where tasks are performed under different conditions that may vary widely. In such circumstances, subjective measures have several advantages over objective measures including easy summarization and a condensed process of score rating. Whereas, objective data require large amounts of data reduction before performance can be analyzed.

2.3.2.2 Human Performance vs. System Performance

Many times, system designers seek to measure human performance but actually measure system performance. Too often, people associate poor system performance with poor human performance without taking into account other system factors.

In the past, many airline accidents have been attributed to human error or pilot error. However, after more accident investigation, other systems factor components are found to be equally at fault. For example, in many airline incidents/accidents, investigators attribute the mishap to poor or degraded human performance prematurely. The frequency of this leads the investigator to ignore pertinent information regarding the adequacy of the airplane including weathering, serviceability, misleading gauges, or work overload in the cockpit. Thus, human performance may be affected by systems factors or even poor design.

Taylor (1957) provides a good example of problems associated with measuring human performance rather than system performance by comparing a boy-on-a-bicycle with that of a boy-on-a-pogo-stick. He stated:

> "In this case, the main performance measurement is the speed at which they travel a quarter of a mile. After several trials we find that the boy on the bicycle consistently

> traveled the distance in a shorter time. What can we conclude about human performance in this situation? The answer is: very little. The performance measure we selected is not a measure of human performance but a measure of system performance: boy-on-a-bicycle vs. boy-on-a-pogo-stick. It is apparent that one system is better than the other but this does not mean that human performance in one system is better than human performance in the other. The boy-on-a-pogo-stick actually may have been doing a better job of pogo-stick-jumping rather than bicycle-riding. As long as we are dealing with a system-level performance measure, human performance can only be inferred, and the inference in this case could easily be misleading."

Simulators (including TME and Frasca 141) are used to assess human performance and, more significantly, to predict how the human will perform in the "real-world" environment. For the TME, the predictor is used to see how well the participant manages multiple concurrent tasks. For the Frasca 141, the predictor is used to determine if the pilot is qualified to fly a real airplane. Even though FAA certification does not equal scientific validity, the Frasca 141 can be viewed as equivalent to the "real-world."

Still, although the FAA does not define multi-tasking in an airplane cockpit as CTM, the FAA does recognize flying as a domain where the management of tasks is implied to be very important. With respect to that, the Frasca 141 can be used as an environment to measure CTM performance. Since the Frasca 141 has been judged by the FAA as being realistic, a favorable comparison of CTM performance in the TME, with that in the Frasca 141, would establish some realism for the TME as well.

2.3.3 Assessing the Quality of Predictors

A *predictor* is any variable used to forecast a criterion. For example, in the medical field, body temperature can be used as a predictor of illness (illness is the criterion). There is no limit to the variables that can be used for this purpose. History has explored a multitude of tools that can be used as potential predictors of performance. In CTM research, recent studies have used the TME as a potential predictor of CTM performance.

All predictor variables, like other measuring tools, can be assessed in terms of their quality or goodness including consistency, accuracy, and precision. In psychology, the quality of measurement is judged by two psychometric criteria: reliability and validity. If a predictor is not both reliable and valid, it is useless (Muchinsky, 2003). In industrial and organizational psychology, the quality of a study is usually evaluated in terms of three categories: construct validity, internal validity, and external validity. These characteristics are important because without validity, the results of a study cannot be used to draw any conclusions (Wickens, 1998).

The review of literature for this study combined psychological measurement and engineering quality control concepts to cover the most intricate aspects of validity pertinent to the methodology for this study.

2.3.3.1 Validity

Validity is used as a means for making inference upon a test to decide if it is measuring exactly what it was intended to measure. A *test* is defined as a procedure intended to establish the quality, performance, or reliability of something, especially before it is taken into widespread use (New Oxford, 2001). In this research study, a TME trial run and an evaluated simulator session are individual tests. Different tests manifest different degrees of validity.

Muchinsky (2003) defines *validity* as a standard for evaluating tests that refer to the accuracy or appropriateness of making inferences from test scores. Muchinsky further explains that valid measures yield "correct" estimates of what is being assessed. Validity is separate from reliability in that it is dependent upon the use of a test. It refers to the test's appropriateness for predicting or making inferences about a certain criteria. For example, a given behavioral test might be highly valid for predicting employee productivity but totally invalid for predicting employee absenteeism.

Uses of the word "valid" will vary from situation to situation. A computer software programmer might think of the word valid in a different way than a person holding a valid driver's license or a group of musicians who have been considered valid since they sold 1 million records. With the wide degree and different interpretations of the word valid, Muchinsky uses the following metaphor:

> "There is the tendency to think of test validity as being equivalent to an on/off light switch-either a test is valid or it isn't. It is probably more accurate to think of test validity as a dimmer light switch. Tests manifest various degrees of validity, ranging from none-at-all to a-great-deal. At some point along the continuum of validity, practical decisions have to be made about whether a test manifests "enough" validity to warrant its use. To carry the light switch analogy further, a highly valid test sheds light on the object (construct) we seek to understand. Thus, the test validation process is the ongoing act of determining the amount of "illumination" the test projects on the construct."

In terms of the TME, stating that it is a valid test for predicting CTM performance would suggest that the procedures of the TME are a "correct" estimate of a participants' CTM performance in a "real-world" domain such as aviation. For example, if the TME is a valid test, then it could be used to predict how well a participant would manage tasks in an airplane cockpit.

There are several types of validity and they all involve determining the appropriateness of a test for making inferences. These types of validity include: construct validity, criterion related validity, content validity, face validity, internal validity, and external validity.

2.3.3.2 Construct Validity

Construct validity is useful in establishing operational measures for the concepts being studied by way of case study methods. This includes using multiple resources of evidence, a chain of evidence, and subjective data collection (Yin, 1989). *Construct validity* refers to the degree to which the researchers manipulated what they wanted to and the degree with which they measured what they wanted to. For example, if a researcher was measuring the independent variable "fatigue," he or she would want to expose the participants to different levels of fatigue (Wickens, 1989).

Muchinsky (2003) defines construct validity as the degree to which a test is accurate and represents a faithful measure of the construct that it aims to measure. A *construct* is a theoretical concept that has been proposed by psychologists to explain aspects of behavior such as intelligence, motivation, mechanical comprehension, etc.

The quest is to find the linkage between what is being measured by a test and its theoretical construct. Construct validity is arguably the most critical factor for determining the worth of applied or basic research.

For example, the Frasca 141 flight simulator aims to measure the construct of pilot performance. Since the Frasca 141 has been certified by the FAA, this means that the Frasca 141 represents a faithful measure of how a pilot will perform if he or she were to fly an actual airplane. Since this measurement has been certified by the FAA, it is considered to be a known measure of predicting actual pilot performance.

To establish construct validity of the predictor, this study sought to compare scores on the experimental test (TME) with known measures of CTM (these are obtained in real flight or in an FAA approved simulator such as the Frasca 141). If the test is a valid predictor, then the scores are closely related and exhibit the same patterns or similarities to the known measures of CTM. In statistical terms, a high correlation must exist between our test and the known measure of CTM to be valid. The correlation is determined by the correlation coefficient value. This value is either convergent or divergent.

If the correlation is high, then the coefficient is referred to as *convergent validity coefficient.* If a low correlation exists, the coefficient is referred to as *divergent validity coefficient.* A divergent validity coefficient in this study would suggest that the TME scores are not related to the Frasca 141 scores. For example, variables in this study such as gender, physical strength, or eye color might have divergent validity coefficients (hypothetically) because they have no relation to CTM performance since CTM is a manifestation of cognitive abilities, perception, judgment, and training. For the results (of the CTM studies that have used the TME) to have value, CTM as measured by the TME would have to have a convergent validity coefficient.

2.3.3.3 Criterion Related Validity

Muchinsky (2003) defines *criterion related validity* as the degree to which a test forecasts or is statistically related to a criterion. It is an important manifestation of validity in that it answers "how much" a predictor is related to a certain criterion. There are two kinds of criterion related validity: concurrent and predictive.

In measuring *concurrent validity*, an experimenter is concerned with how well a test (also known as the predictor) can predict a criterion concurrently. For example, if we wanted to predict graduate student grade point averages, we might base their future performance on a predictor. If the predictor is a valid measure of graduate student performance (grade point average), there will be a high correlation between test scores and grade point average. The purpose of assessing concurrent criterion related validity is so the test can be used at a later time with the knowledge that it is predictive of the criterion.

In measuring *predictive validity*, an experimenter collects predictor information and uses it to "forecast" future performance. For example, a university might use a test called the Graduate Records Examination (GRE) to predict the criterion of how well one might do in a graduate program. When predictor scores are correlated with statistical data, the resulting correlation is called the validity coefficient.

The *validity coefficient* reveals the degree of association between two variables. For significance, a moderately acceptable validity coefficient is in the .50-.70 range, strongly acceptable validity coefficient is in the .70-.90 range, and a very strongly acceptable validity coefficient is in the .90-1.0 range. If the correlation between the predictor and the criterion are greater, we will know more about the criterion based on the predictor. Muchinsky states,

> "A correlation of 1.0 indicates a perfect prediction. However, Lubinsky and Dawis (1992) noted, tests with moderate validity coefficients are not necessarily flawed or inadequate. The results attest to the complexity of human behavior."

Human behavior is very complex and can be influenced by many factors such as motivation, fatigue, skill, and even luck. Certain predictors are valid for predicting only certain criteria but they can be influenced by other factors. Although a predictor may have an insignificant validity coefficient, this may have been due to the effects associated with sources of unwanted variability. To be successful in this study, a validity coefficient at least 0.5, which is moderately acceptable, would have to be attained.

2.3.3.4 Content Validity

Muchinsky (2003) defines *content validity* as the degree to which Subject Matter Experts (SME's) agree that the items in a test are a representative sample of the domain of knowledge the test aims to measure. It involves the degree to which a predictor covers a representative sample of the behavior being assessed. This type of validity is widely used for validation of flight simulators and driving simulators (Frasca, 2003 and NHTSA, 2005).

Content validity is a manifestation of construct validity. It has been mostly used in tests to indicate how well a person has mastered a specific skill or area of content. For example, in order to be considered content valid, an achievement test on flying an airplane must contain a representative sample or mix of test items covering the domain of aviation, such as ascent/descent, direction, altitude, speed, and ATC communication. If all of the questions were about ascent/descent, the test would not be a balanced representation of the content of flight simulation. If a person scores high on a content valid test of flight simulation, we can infer that he or she is very knowledgeable about flying an airplane.

Assessing "content validity" is different from assessing "criterion related validity." It is not based on correlation coefficients. It is assessed by SME's in that particular field. For example, pilots would define the domain of aviation and then help with the design and set up of the flight simulation devices and test procedures. The SME's would determine the "content validity" of the test ranging from "not-at-all" to "highly-valid."

In this study, although research in regards to CTM SME's has not been identified, it is logical to state that the TME could benefit from the experience of CTM SME's to agree that the items used in the test are representative of the domain of CTM.

2.3.3.5 Face Validity

Muchinsky (2003) defines *face validity* as the appearance that items in a test are appropriate for the intended use of the test by the individuals who take the test. It is based on people's judgment. This is concerned with the appearance of the test items. It

answers the question "does it look like it should test what is intended?" Estimates of content validity are made by SME's. However, estimates of face validity are made by participants (or test subjects). Keep in mind, it is possible for a test to be content valid while not being face valid and vice versa. In some instances, the face validity could have a profound effect on whether the participants will perceive the test to be an appropriate and legitimate means of prediction.

With respect to the TME, Funk et al (2003) have implied that the TME has some face validity. That is, it looks like or appears to be an appropriate test for CTM by individuals who operate the TME (Shakeri and Funk, 2003).

2.3.3.6 Internal Validity

Wickens (1998) defines *internal validity* as a condition in an experiment where the causal or independent variables and no other extraneous variables caused a change in the effects being measured. Campbell and Stanley (1963) state:

> "It (internal validity) is used for explanatory and causal case studies only and not for descriptive or exploratory studies to establish a causal relationship where certain conditions are shown to lead to other conditions as distinguished from spurious relationships."

Here a researcher does pattern matching, explanation building, and time-series analysis of data. A control group study is usually set up to determine whether event x led to event y.

In some cases, sources of unwanted variability can possibly cause a change in effects. Since the study about the TME is an exploratory study, internal validity will be implied with respect to the use of data from another experiment by Hoover (2005). It is important to note that some of the data collected by Hoover was used for statistical analysis in this study. Hoover's experiment used a control group design to measures the effects of training on CTM performance.

2.3.3.7 External Validity

Wickens (1998) defines external validity as the degree to which we can generalize the results of a study to other people, tasks, and/or settings. In some respects, this might be the single most important form of validity (assuming that all the others are accounted for) because it gives yield to value in the professional community. Results that are determined to be too narrow may eliminate all applicability of the study, except of course, learning that this experiment is one that should not be repeated under the exact same conditions. Lack of external validity may be due to the usage of unrealistic, simple tasks or settings.

Assessment of this type of validity on the TME is crucial. Based on the widely used validation technique of correlation, a high correlation between the TME and the Frasca 141 would shed light on how well TME performance can be generalizable outside of the TME. Also, external validation of the TME would suggest that the TME could be used as a screening tool or research tool to predict how well a participant might perform in any domain such as an airplane pilot, a surgeon, or even just a cook.

On the other hand, if the criterion does not imply that the TME is externally valid, then we cannot give much weight to the previous TME experiments until further evidence has been collected. Also, the lack of external validity would imply that the TME could not be used for CTM research, at least without modifications to correct the lack of validity.

External validity has the most implication to this study. The level of significance in the correlation results will determine what we can imply from the relationship between the TME (predictor) on CTM performance (the criterion).

2.3.3.8 Predictor Construct Goals

Assessment is based on the notion of making an inference about something. In engineering, we make assessments based on "something" to determine whether someone is likely to perform well in a particular domain. That "something" is a construct that is used to predict satisfactory performance. That "something" could be the individual's CTM performance.

The literature reviewed above on CTM and validity has been bound together to pose four related questions: How do we assess the characteristics needed for an individual that will be successful in a particular domain like aviation? What measures should be used to predict performance? What predictor can we use for CTM? Is the TME a valid CTM research tool?

3 Research Objectives

Concurrent Task Management (CTM) is the process by which human operators (such as pilots, drivers, and surgeons) of complex systems perform to allocate their attention among multiple concurrent tasks.

Complex systems (such as aircraft cockpits, automobiles, and operating rooms) impose a high resource demand on the limited capacity of the human operator. Adams, Tenney, and Pew (1991) have predicted that these complex systems have matured to the point that they "support, complement, and extend the capabilities of their human operators." Knowing this, a proliferation of disasters has been associated with the use of these complex systems.

Thorough understanding of human operator's knowledge, skills, and abilities associated with CTM might help us to prevent these disasters. But to develop that understanding, we must develop valid tools to measure human operators' CTM performance. Since the TME exhibits some face validity, it has been hypothesized that it may have the potential to be useful for predicting human CTM performance in a "real-world" environment such as an airplane cockpit. However, it raises an important question as to whether or not it is a good enough tool. Would an exploratory study support the idea that TME accurately measures CTM performance?

Since the Frasca 141 flight simulator has been recognized and certified by the Federal Aviation Administration (FAA) as a valid human performance measurement tool for aviation, it was used to determine if the TME is externally valid through an exploratory study.

The research hypothesis for this study is stated as follows:

H_0: CTM scores in the TME do NOT correlate to CTM scores in a more realistic task management environment.

H_1: CTM scores in the TME correlate to CTM scores in a more realistic task management environment.

To test the null hypothesis, TME tests were performed to measure pilot CTM performance. CTM performance in the TME was compared to CTM performance in the Frasca 141.

Significant correlations between the TME scores and the Frasca 141 scores would suggest that there was a relationship between CTM performance in the TME and CTM performance in the Frasca 141. Therefore, significantly positive correlations would suggest that the TME can be used to as a tool to predict pilot CTM performance and that it can be generalizable to the "real-world."

The next chapter describes the experimental methodology that was used for this study including experiments, participants, facilities, testing apparatus, experimental procedures, measures, and data analysis procedures.

4 Experimental Methodology

4.1 Introduction

To briefly summarize previous Concurrent Task Management (CTM) research, Shakeri and Funk (2003) found that strategy is an important factor in CTM performance and that humans are not optimal. Funk and Chen (2003) developed a fuzzy model where status, urgency, and other factors affected people's task management performance. Nicolalde and Funk (2003) found that CTM performance cannot be predicted based on cognitive abilities alone, but rather CTM is a complex cognitive process that must be studied as a unit.

Although these studies have made significant contributions to CTM research, it is not clear how much validity can be attributed to their findings since they relied on the Task Management Environment (TME) as a measurement tool. This raises the fundamental question: Can the findings from these studies be generalized to the "real-world?"

In order to seek evidence towards answering this fundamental question, this study used an experimental approach to compare the TME with a certified, "real-world" simulator (the Frasca 141), to determine the external validity of TME.

4.2 Overview

It can be very difficult to find a source of pilots for research. The major challenges include time and the high costs incurred from using a Flight Training Device (FTD) such as the Frasca 141. In order to accommodate these challenges, this study combined two experiments that were composed of the same homogeneous participant population. The results of the second experiment by Stebel (the focus of this study) relied on data collected from the first experiment by Hoover in order to compare the relationship between the two data sets. The model in Figure 4.1 represents the two experiments, where P = pilot participant and r = relationship of the data.

The first experiment was performed by Hoover to determine the effects that training would have on CTM performance using a control group design (Hoover and Funk, 2005). In Hoover's experiment, "true" CTM performance data from the Frasca 141 was recorded. *True* CTM performance can be defined as data collected from the Frasca 141 including a pilot's task prioritization error rate. Simply put, Hoover's study conveniently provided a set of "true" CTM performance data against which CTM data collected using the TME could be compared. Although Hoover's experiment identified some interesting results with respect to training pilots, the details of that experiment are beyond the scope of this study.

The second experiment (and focus of this study) was merely concerned with "piggy-backing" on Hoover's experiment in order to obtain true CTM performance data. The second experiment was performed by Stebel to determine if the TME has external validity by using a correlation analysis to compare CTM performance data from the TME with that obtained in the higher fidelity Frasca 141. In Stebel's experiment, CTM performance data from the TME was recorded to investigate a relationship with "true" CTM performance data found in the Frasca 141. These two experiments are explained more thoroughly in this chapter.

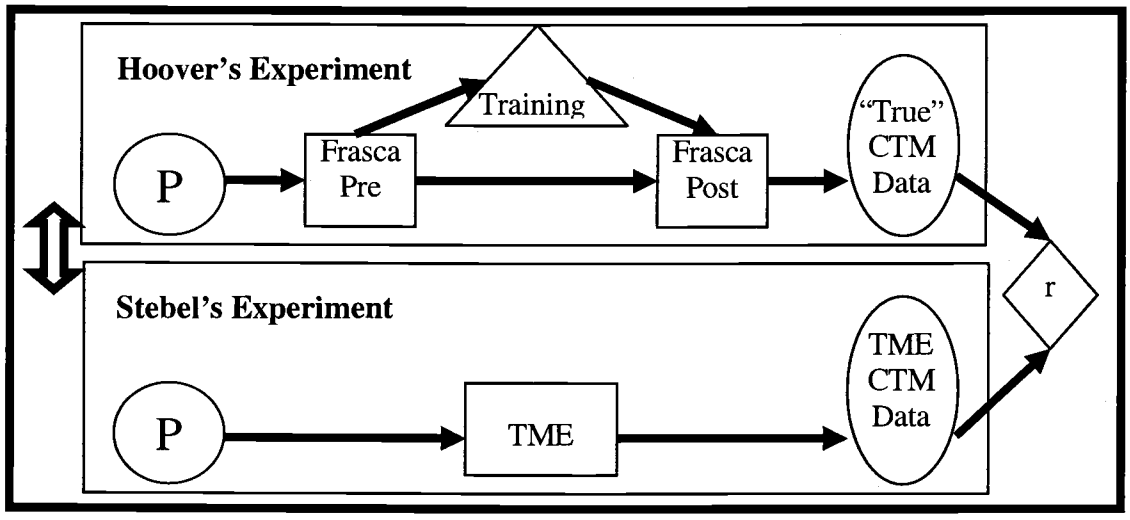


Figure 4.1 Model of Data Compared in Hoover's Experiment to Stebel's Experiment.

4.3 Hoover's Experiment

4.3.1 Participants

Twenty-five student pilots enrolled in the Flight Technology Program at Central Washington University (CWU) participated in Hoover's experiment. The participants that took part in Hoover's experiment received credit towards their academic curriculum. All participants were required to meet the following requirements: enrollment in at least the second year of the CWU Flight Technology Program, the same minimum amount of FTD experience, a private pilot's license, and an instrument rating of Stage II or check schedule for Stage II. Each participant was asked to give informed consent, perform two flight simulations using the Frasca 141, and fill out a pre-experiment questionnaire. Each participant received course credit for his or her participation.

4.3.2 Testing Facility

All participant data in Hoover's experiment was collected at the CWU Flight Technology Building located near Bowers Field at Midstate Aviation, Inc. Both FTD's used in her experiment were located in the main simulator lab. The testing environment was controlled to eliminate the possibility of extraneous variables that could affect participant performance.

4.3.3 Testing Apparatus

4.3.3.1 The Frasca 141

In Hoover's experiment, two identical Frasca 141 flight simulators (Figure 2.7) were used. Both simulators were fully approved for instrument training under 14 CFR Part 91 and Part 141. The 141 set up was chosen because a relatively large number of the participant population had met the requisite number of hours and experienced flight hours logged on the Frasca 141. By using this simulator, data could be collected in a timely and simultaneous manner. The Instructor Station was used to collect data regarding system failures, environmental systems, instrument failures, and performance measures. An FAA certified flight instructor was paid to initiate, assess, debrief, and monitor test results by using the nearby Instructor Workstation (Figure 2.7).

4.3.3.1.1 Frasca 141 Configuration

The Frasca 141 was configured electronically and mechanically to resemble a Piper PA28R-200 aircraft. This particular aircraft simulation included a normally aspirated single engine retractable-gear low-wing aircraft with a single-boost pump and a vacuum pump. The avionics package included a Bendix/King stack including dual KY196 coms, dual KN53 NAV radios, KDI572 DME, KR87 ADF, KT76A transponder, KMA 24 audio panel with marker beacons, and a certified GNS430 IFR en route and approach GPS (Hoover et al, 2005).

4.3.3.1.2 Frasca 141 Scenario

The Frasca 141 simulations were conducted in a Line Oriented Flight Training (LOFT) format using standard instrument charts and procedures. Simulated flights began at Snohomish County Airport at Paine Field in Everett, WA. Pilots flew the Paine 2 departure and contacted Seattle Center for vectors to the Instrument Landing System (ILS) approach to Runway 13R at Boeing Field in Seattle, WA. After flying the missed approach procedure (MAP), pilots held at BLAKO intersection until given directions by Air Traffic Control (ATC) to navigate to the ANVIL intersection for the ILS Runway 16R at Seattle-Tacoma (SEATAC) International Airport in Seattle, WA. Pilots switched to radar vectors at 15 miles southeast of ANVIL for the localizer course intercept for the ILS Runway 16R. This particular LOFT was chosen because it involved many complicated procedures in a short time frame. This gave the pilot a high workload environment during a majority of the one hour flight simulation (Hoover, 2005).

4.3.4 Experimental Procedure

Participants were tested individually using a Pre-Test and Post-Test design (Campbell and Stanley, 1963). Each Pre-Test and Post-Test lasted approximately one hour. Due to scheduling constraints and FAA instructor costs, some instances called for two pilots to be tested simultaneously on different Frasca 141 FTD's. After signing an informed consent form to meet human subject testing guidelines, participants completed a pre-experimental questionnaire providing demographic data. Then, participants were

randomly split into two groups (a control group and an experimental group). The control group was comprised of nine participants and the experimental group was comprised of ten participants. Each participant performed a Pre-Test flight simulation (Frasca Pre-Test) and Post-Test flight simulation (Frasca Post-Test) using a Frasca 141 that followed standard operating procedures designated by the FAA and CWU. The Frasca Post-Test was performed approximately two weeks after the Frasca Pre-Test. To avoid between-group interaction, participants were asked not to discuss the experiment with anyone (Hoover, 2005). This is essential to the present study because new information learned by the experimental group could have had an effect on the control group performance in the Frasca Post-Test.

During a two week period, participants in the control group did not receive any additional training but the participants in the experimental group attended two ninety-minute learning sessions that included reading, self-study, group learning activities, and a guided discussion. The first ninety-minute session consisted of a class discussion related to aviation human factors, aeronautical decision making, situational awareness, workload management, and cockpit task management. The second ninety-minute session included individual project reports regarding incidents and accidents attributed to CTM error where participants were asked to role play (Hoover, 2005).

Directly following the Post-Test flight simulation, participants were voluntarily recruited by a flyer (Appendix 1) and word of mouth to participate in the second experiment by Stebel.

4.3.5 Measures

4.3.5.1 Task Prioritization Error Rate

The objective performance measures in the first experiment were collected from a Frasca Pre-Test and Post-Test. In Hoover's experiment, performance measures from both the Frasca Pre-Test and Frasca Post-Test were used to compare the differences between the control group and the experimental group scores. During Hoover's experiment (Frasca Pre-Test and Frasca Post-Test) there are 14 challenge points that include 20 total tasks. Hoover defined a *Task Prioritization Error* as when a participant

diverts his or her attention from a more important or more urgent task to a less important or less urgent task. Hoover based this on the CTM research by Funk et al (2003) and by the FAA Practical Test Standards stated by FAA-S-8081-4C with respect to altitude, airspeed, heading, intercepting and tracking course, use of checklists, procedures, and ATC communications. A task prioritization error score of 1 was recorded every time the participant made an error during a challenge point. The error rate was the number of errors made divided by the number of possible challenge points. The pilot's goal was to make as few errors as possible in the flight simulation, especially at the challenge points.

While Hoover's experiment collected data on several variables, this study was mostly interested in CTM performance in the Frasca Post-Test.

4.4 Stebel's Experiment

4.4.1 Participants

A subset of nineteen participants from the first experiment (Hoover's Experiment) volunteered their time to be involved in the second experiment (P-value). The participants were recruited through the use of flyers (see Appendix 1) posted in the Flight Technology Building near Bowers Field and by word of mouth. All participants were required to meet the following requirements: current enrollment in the Flight Technology Program at Central Washington University (CWU), completion of Hoover's experiment using the FRASCA 141, at least a second year status in the CWU Flight Technology Program, the same minimum amount of FTD experience, a private pilot's license, and an instrument rating of Stage II or check schedule for Stage II.

Each participant was asked to give informed consent. Then, participants were given verbal training of the TME, performed six TME trials, and answered a post-experiment questionnaire. After completing the post-test questionnaire, each participant was compensated $20 for his or her time.

4.4.2 Testing Facility

All participant data was collected at the CWU Flight Technology Building located near Bowers Field at Midstate Aviation, Inc. The TME data was collected in two

separate classrooms to control for simultaneous testing and to eliminate extraneous variables that could have an affected participant performance.

After collecting all of the participant data at CWU, the data was analyzed at Oregon State University in Corvallis, OR.

4.4.3 Testing Equipment

To gather participant data using the TME, two different computers were used in this experiment. One computer was located in each control room.

The first computer used in this experiment was a Dell Inspiron 7000 laptop running the Microsoft Windows NT operating system located in room 1. The Inspiron 7000 had a processor speed of 366 MHz and 256 MB of RAM. This laptop had a 15 inch monitor and standard mouse. Fifteen of the participants were tested using this computer. All participant data was stored on Iomega PC/Mac 100 MB zip disks.

The second computer was an Intel 2003 Laptop running the Microsoft Windows XP Professional operating system located in room 2. The Intel 2003 laptop had a processor speed of 500 MHz and 128 MB of RAM. This laptop had a 13 inch monitor and optical mouse. This laptop was plugged into a separate 15 inch Dell flat-screen monitor in order to see the entire TME interface. All participant data was stored on Iomega PC/Mac 100 MB zip disks.

4.4.4 Testing Apparatus

4.4.4.1 The Task Management Environment

The Task Management Environment (TME) was used for the second experiment by Stebel. In this experiment, the TME was modified to fit the needs of measuring human performance in the aviation domain. It was modeled after aircraft tasks and subsystems to measure the CTM behavior of a pilot participant.

4.4.4.1.1 TME Configurations

In the second experiment, the TME was designed to model 12 subsystems. The number of subsystems, used in this experiment, was determined by the major tasks and

subsystems involved with flying an airplane. These 12 tasks and subsystems were based on the Aviate, Navigate, and Communicate Theory (ANC) including ten Aviate Tasks, one Navigate Task, and one Communicate Task. *A*=Aviate Tasks are involved with keeping the airplane upright and flying. *N*=Navigate Tasks are involved with the pilot determining where he or she is located and where he or she is going. *C*=Communicate Tasks are involved with the pilot communicating with Air Traffic Control (ATC).

The TME subsystems were configured to model the corresponding aircraft tasks and subsystems. Two private pilots were used as Subject Matter Experts (SMEs) to assist in determining importance (*i*) values with relation to a particular task or subsystem in an airplane. The *i* values ranged from 2 to 10 (Table 4.1).

Aircraft Subsystem Parameters		**TME Subsystem Parameters**				
Aircraft Task and Subsystem	**ANC Task**	**TME Subsystem**	***i***	**DR**	**CR**	**Behavior**
Manage physical control of airplane: yoke & pedals	A	S00	7	9	108	discrete
Manage speed of wind on plane by climb/descent: airspeed indicator	A	S01	7	10	90	discrete
Manage straight/level flight by up/down pitch: attitude indicator	A	S02	10	18	76	discrete
Manage altitude height (ft): altimeter indicator	A	S03	9	14	86	continuous
Manage VOR, ADF, & GPS: radio control station	N	S04	6	5	98	discrete
Manage communication w/ATC: communication link	C	S05	2	4	110	continuous
Manage coordination of turning airplane using rudders: turn coordinator	A	S06	8	11	90	discrete
Manage heading/direction N,E,S,W: directional gyroscope	A	S07	10	15	92	discrete
Manage rate of climb: vertical speed indicator (VSI)	A	S08	9	12	90	discrete
Manage landing capabilities: landing gear switch	A	S09	10	10	108	continuous
Manage oil/press. Levels for instruments: vacuum indicator	A	S12	7	10	94	continuous
Manage airplane hydraulic braking: braking control	A	S13	6	6	109	discrete

Table 4.1 The TME Configurations.

4.4.4.1.2 TME Mixed Scenario

As discussed previously, there are two types of TME subsystems: continuous and discrete. Previous CTM research using the TME (Vasquez, 2004) determined the relationships of mixed task scenarios based on performance. This research showed that there are small correlations between continuous and mixed scenarios and discrete and mixed scenarios, which suggested that the optimal scenario is a mixed scenario. Thus, for this experiment, the TME was configured to with a mix of both continuous and discrete subsystems much like that of an airplane.

In this experiment, the TME was configured to include four discrete subsystems (S03, S05, S09, and S12) and eight continuous subsystems (S00, S01, S02, S04, S06, S07, S08, and S13) based on how fast, moderate, or slow a particular subsystem changes. Each subsystem had a deviation rate (DR) and correction rate (CR) that was modeled with respect to the rate of failure associated with a particular aircraft task or subsystem.

4.4.5 Experimental Procedure

Participants were tested individually, each in a single session lasting approximately one hour in a classroom at Central Washington University separate from the Frasca 141 classroom. Due to scheduling constraints, there were four instances were two pilots were tested simultaneously in different classrooms. Participants were asked to sign an informed consent form (Appendix 2). Each participant received the same TME training from a verbal script (Appendix 3). At this time, participants could ask questions before beginning the experiment.

After the verbal training, each participant performed six TME trials using a mixed scenario. Past TME experiments showed that it takes ten minutes of practice time for a participant to learn to mange the TME (Nicolalde and Funk, 2003). All six TME trials were recorded for analysis.

Each trial was set to run for five minutes. When the five minutes timed out, a CTM score, computed by the TME, was recorded by the experimenter. After each individual trial, the participants were given up to five minutes of resting time before beginning the next trial test.

After completing the six TME testing trials, each participant filled out a post-experiment questionnaire (Appendix 4) that included demographic and subjective questions.

4.4.6 The Measures

Objective performance measures were gathered from the six TME testing trials. The TME software recorded three types of files for each trial test. The first file was a text document that displayed all of the information with respect to scenario type. It can be opened with any word processor or spreadsheet.

The second file recorded the characteristics of each subsystem (i.e. subsystem number, correction rate, deterioration rate, behavior, test time and date, scenario length, and total weighted score).

The third file, recorded all of the raw data from each subsystem (i.e. subject ID, test time, time increments and corresponding status of the subsystem, and number of the subsystem that was attended in that specific time increment).

4.4.6.1 Total Weighted Score

In this experiment, the Total Weighted Score (TWS) was used to compare performance. The TWS was the score that was earned by the participant upon the completion of a TME test trial. The cumulative score for each subsystem was the mean instantaneous score since the beginning of the trial time. The TWS was the summation of all subsystems' cumulative scores and reflects an overall CTM performance measure, weighted according to subsystem importance (Funk et al, 2003).

4.4.6.2 Demographic Measures

The demographic variables were the measures recorded in the post-experiment questionnaire (Appendix 4) and included age and gender and strategies and comments. All of the variables were set to an objective scale except for the strategies. Strategy variables were subjective in nature since they related to the participants' opinions.

Additional demographic data was collected in the post-experiment questionnaire (Appendix 8) including: computer experience, video game experience, flying experience,

physical impairments, and fatigue levels. This data was not used for analysis in this study but was collected for the possibility of performing a multivariate analysis in future research.

4.4.6.2.1 Age and Gender

The participants were asked to answer a question regarding their age to determine the average sample age (years) and age range (youngest to oldest) of the sample population. Also, the participants were asked to answer a question regarding their gender to determine the gender balance of the sample population.

4.4.6.2.2 Strategies and Comments

The participants were asked to answer two subjective questions to determine if the participants used any task management strategies when playing the TME and if they noticed any task management similarities between flying an airplane and playing the TME.

4.5 Data Analysis Procedure

StatGraphics and Microsoft Excel were used to perform data analysis such as normalization of the data, tests of normality, descriptive statistics and outlier treatment, frequency distributions, histograms and scatter plots, t-tests, parametric correlations, and non-parametric correlations.

The main focus of this study was the correlation analysis (parametric and non-parametric). It was performed to compare CTM performance in the Frasca 141 (Task Prioritization Error Rate) with CTM performance in the TME (TWS). This was instrumental in determining the external validity of the TME.

4.5.1 Data Normalization

When a measurement of time is expressed in minutes and another in seconds, and operations between the two are going to be performed, data normalization is needed (StatSoft, 2005).

> "The technique of data normalization is used to adjust a series of values using a transformation function in order to make them comparable to some specific point of reference. Data normalization is required when there is an incompatibility of the measurement units across variables that might affect the result of operation between variables."

Simply put, data normalization transforms variables into compatible measuring units. In this study, data normalization was necessary since the performance scores in the TME and the performance scores in the Frasca 141 had different measuring scales. By normalizing this data, the scores were transformed into one measurement scale percentages ranging from 0% (minimum, low) to 100% (maximum, high).

In the Frasca 141, the performance measure was Task Prioritization Error Rate. The best possible score, based on the Task Prioritization Error Rate scale, was 0. This was converted into a percentage. For example, if a participant scored 5, out of a possible 20, based on the Task Prioritization Error Rate, his or her score was normalized into a score of 75% (15 divided by 20).

In the TME, the performance measure was TWS. The best possible score, based on the TWS scale, was 91. This was converted to a percentage. For example, if a participant scored 90, out of a possible 91, based on the TWS scale, his or her score was normalized into a score of 99% (90 divided by 91).

4.5.2 Normality Tests

A common application for distribution fitting, before parametric testing, is testing for data normality. If the data is not normally distributed, then non-parametric tests are recommended (Lapin, 1998 and Montgomery, 2001).

To test for normality, several tests were run to determine whether the variables could model a normal distribution. The Chi-Square Test (represented as Chi^2-stat) was used to divide the range of variables into 12 equally probable classes and compared the number of observations in each class to the number expected. The Shapiro-Wilks Test (represented as W-stat) was used to divide the quantiles of the fitted normal distribution to the quantiles of the data. The standardized Z-Score Test for Skewness (represented as Z-score) was used to look for a lack of symmetry in the data.

To test for goodness-of-fit, several tests were run to determine whether specific data set variables could model a normal distribution. The Chi-Square Test (represented as Chi^2-stat) was used to divide the range of variables into non-overlapping intervals and compared the number of observations in each class to the number expected based on the fitted distribution. The Kolmogorov-Smirnov Test (represented as D-stat) was used to compute the maximum distance between the cumulative distribution of variables and the central distribution frequency of the fitted normal distribution.

4.5.2.1 Outlier Treatment

After testing for normality, additional tests were performed to determine if there were any outliers that needed treatment in the data set variables. Two inter-quartile values (25% and 75%) where calculated in order to determine the quartile ranges. Inter-quartile ranges (minimum extreme value and maximum extreme value) were used to find outliers. Any outliers found would be transformed to the nearest value inside the mean plus or minus two inter-quartile ranges. For example, in a variable where the 25% inter-quartile equals 68.5, the 75% inter-quartile equals 88.5, and the mean equals 77.3, if 68.5 is subtracted from 88.5, the inter-quartile value is 20. If the inter-quartile value is multiplied by 2 and then added that to the mean of 77.3, the maximum extreme value is 46.3. If the inter-quartile value is multiplied by 2 and then subtracted from the mean of 77.3, the minimum extreme value is 108.3.

4.5.3 Descriptive Statistics

Univariate descriptive statistics were calculated in this experiment to determine the population size (n), mean (mean), standard deviation (σ), minimum score (min), and maximum score (max). These were calculated after the normality tests were treated for potential outliers.

4.5.4 Frequency Distributions

The frequency distributions were computed for each of the descriptive statistic measures to determine the percentile ranges of the percentiles (10%, 25%, 33.3%, 66.6% and 75%). The percentile ranges, with respect to the frequency distributions, were

calculated to see if any additional inferences could be made about the data set variables based on the frequency of participant performance.

4.5.4.1 Histograms

Histograms were generated for each of the normality tests to graphically represent the frequency distribution. The shape of the distribution portrays the frequency of values from different ranges of the variables. The histograms allow one to evaluate the normality of the empirical distribution with the normal curve. The horizontal (x) axis represented the uniform ranges of possible scores for each of the variables. The vertical (y) axis represented the number of participants that scored within each range.

4.5.5 Learning Curve

The learning curve was generated by plotting the mean performance values for each TME trial for the entire sample population. Two standard deviation bars were plotted for the mean of each trial.

4.5.6 T-tests

The T-test was used to evaluate the differences in means between two groups. For example, the T-test can be used to test for a difference in test scores between a group of patients who were given a drug and a control group who received a placebo. According to StatSoft (2005), the T-test can be used even if the sample sizes are very small as long as the variables are normally distributed within each group and the variation of scores in the two groups is not reliably different. If normality is not met, then one can evaluate the differences in means between two groups using one of the non-parametric alternatives to the T- test.

In this experiment, a T-test was used to determine if significant improvement occurred between successive TME trials.

4.5.7 Parametric Correlation

In this study, Pearson correlations were calculated to measure the strength of the linear relationship between two variables, x and y. The x-variable (independent) was the TME measure and the y-variable (dependent) was the Frasca 141 measure. The correlation between the two variables reflected the degree to which the variables were related. This analysis was used assuming that both the x and y variables were approximately normally distributed.

A *correlation coefficient* (also known as the Pearson's product-moment coefficient or validity coefficient) was calculated to determine if there was a high or low correlation between the x and y variables. It was signified by *r* ranging from -1.0 to 1.0, where 1.0 is a perfect positively skewed correlation and -1.0 is a perfect negatively skewed correlation. If r is close to 0.0, the dispersion was large and the variables were uncorrelated. If the results illustrate that there was a low P-value calculation (less than 0.05), then this proves that there was evidence to reject a null hypothesis. This would suggest that there was a statistically significant relationship between the two variables x and y (Winks Stat Software, 2005).

The R^2 is called the *coefficient of determination* and was used for interpretation of the proportion of variance in y that is contained in x.

The *P-value* is defined as the smallest level of significance that would lead to a rejection of the null hypothesis (Montgomery, 2001). In this case, the null hypothesis stated that there was no statistically significant relationship between the two variables x and y.

4.5.7.1 Scatter Plots

The scatter plots were generated as a visual representation to assess the variable relationships and trends, between the TME measures and Frasca 141 measures, with a regression line.

4.5.8 Non-Parametric Correlation

In this study, Spearman's rank correlations were calculated to measure the strength of the linear relationship between two variables, x and y.

Wikipedia (2005) states that this measures how well an arbitrary monotonic function could describe a relationship between two variables without making any assumptions about the frequency distribution of the variables.

Unlike the Pearson's correlation, the Spearman's rank correlation does not require the variables to be normally distributed, nor does it require the variables to be measured on interval scales. It assumes that the variables under consideration are measured on at least an ordinal (rank order) scale (Winks Stat Software, 2005)

In this study, the x-variable (independent) was the TME measure and the y-variable (dependent) was the Frasca 141 measure. For each correlation analysis, the x and y data sets were ranked into two ordered series, where N is the number of pairs of values, the raw scores were converted to ranks, and the differences (D) between the ranks of each observation on the two variables were calculated using the equation:

$$\mathbf{r}=\frac{1-6(\Sigma D^2)}{N(N^2-1)}$$

A *correlation coefficient* (also known as Spearman's rank coefficient or validity coefficient) was calculated to determine if there was a high or low correlation between the x and y variables. It was signified by *r* ranging from -1.0 to 1.0, where 1.0 is a perfect positively skewed correlation and -1.0 is a perfect negatively skewed correlation. If r is close to 0, the dispersion was large and the variables were uncorrelated. If the results illustrate that there was a low P-value (less than 0.05), then this proves that there was evidence to reject a null hypothesis. This would suggest that there was a statistically significant relationship between the two variables x and y (Winks Stat Software, 2005).

To test whether an observed value of r is significantly different from zero, the observed value can be compared with tables (Montgomery, 2001) for various levels of significance. For sample sizes above about twenty, the variable t_0 is used to test whether an observed r is significantly different from the theoretical value. This was calculated using the equation:

$$t_0 = \frac{r}{[\sqrt{(1-r^2)}/(n-2]}$$

For example, if t_0=0.2155>1.3333, this means there was not a significant relationship between x and y because 0.2155 is NOT greater than 1.3333 (where 0.2155 was the calculated t_0, 1.3333 was the value found using the t-distribution tables at a 90% confidence level where α=.10). In this scenario, if it were true that the t_0 value was greater than 1.3333, then there would be a significant relationship between x and y, where, H_0= no relationship between variables.

R^2, the *coefficient of determination*, was used for interpretation of the proportion of variance in y that is contained in x.

5 Results

5.1 Overview

This chapter presents the results of the methodology and analyses described in chapter 4. It describes data normalization, descriptive statistics, normality tests, frequency distributions (including outlier treatment), plots, parametric correlations (Pearson), and non-parametric correlations (Spearman rank) based on objective findings. It also includes summaries from the post-experiment questionnaire including demographic information and subjective information.

5.1.1 Normality Tests

The variables in this experiment, which were defined as CTM performance measures, were analyzed for comparison. The normality tests showed how well each of the data set variables fit a normal distribution. These variables were defined as follows:

- **TME Mean** grand mean of TME mean scores (trials 2-6) for all participants.
- **TME 1** mean of TME trial one scores for all participants
- **TME 2** mean of TME trial two scores for all participants.
- **TME 3** mean of TME trial three scores for all participants.
- **TME 4** mean of TME trial four scores for all participants.
- **TME 5** mean of TME trial five scores for all participants.
- **TME 6** mean of TME trial six scores for all participants.
- **Frasca Pre** mean of Frasca Pre-Test scores for all participants.
- **Frasca Post** mean of Frasca Post-Test scores for all participants.

StatGraphics was used to calculate tests of normality. StatGraphics based its decision of whether the data was normally distributed by the lowest P-value. In the normality tests, it chose the normality test that proved to have the lowest P-value amongst the other normality tests and then based a decision upon that value. StatGraphics chose the lowest confidence interval possible based upon the alpha significance level. The *null hypothesis* (H_0) was defined as: the TME performance measures are normally distributed. Table 5.1 shows the values for the Chi^2-stat, W-stat, Z-score, D-stat, and their respective P-values.

Variables	Chi-Square Test		Shapiro-Wilks Test		Z-Score Test		Kolmogorov-Smirnov Test	
	Chi^2-stat	P-value	W-stat	P-value	Z-score	P-value	D-stat	P-value
TME Mean	5.63	0.78	0.96	0.57*	0.51	0.61	0.12	0.96
TME 1	11.95	0.22*	0.97	0.79	0.40	0.69	0.16	0.75
TME 2	10.68	0.30*	0.98	0.92	0.43	0.67	0.13	0.91
TME 3	9.42	0.40	0.95	0.34*	0.75	0.45	0.17	0.65
TME 4	17.00	0.05**	0.94	0.23	0.94	0.35	0.13	0.91
TME 5	11.95	0.22*	0.96	0.57	0.47	0.64	0.12	0.95
TME 6	4.37	0.89	0.96	0.62	0.72	0.47*	0.11	0.98
Frasca Pre	18.2	0.03***	0.93	0.19	0.76	0.44	0.20	0.45
Frasca Post	1.7	0.30*	0.95	0.43	0.99	0.32	0.13	0.92

Table 5.1 Normality Test Results.

*Lowest selected P-value chosen to determine normality where $\alpha = 0.10$ @ 90% confidence. Don't reject H_0.

**Lowest selected P-value chosen to determine normality where $\alpha = 0.05$ @ 95% confidence. Don't reject H_0.

***Lowest selected P-value chosen to determine normality where $\alpha = 0.05$ @ 95% confidence. Reject H_0.

The results of the normality tests indicated that all data set variables passed for normality (except for Frasca Pre) since the lowest P-value amongst the tests performed equaled a value that was greater than or equal $\alpha = 0.10$. This means the idea that the variables came from a normal distribution with 90% or higher confidence can NOT be rejected.

The data set variable TME 4 passed the normality tests at a 95% confidence since the lowest P-value amongst the tests performed equaled 0.05, which is greater than or equal to the $\alpha = 0.05$. The data set variable Frasca Pre did NOT pass the normality tests at a 95% confidence since the lowest P-value amongst the tests performed equaled 0.03, which is less than the $\alpha = 0.05$.

If the experiment was based on the Shapiro-Wilks Test, Z-Score Test, or Kolmogorov-Smirnov Test only, all of the data set variables would pass for normality since the lowest P-value amongst the tests performed equaled a value that was greater than or equal to at least $\alpha = 0.10$. This means the idea that the variables came from a normal distribution with 90% or higher confidence can NOT be rejected. Simply, most of the data set variables in these tests are normally distributed at a 90% or higher confidence with $\alpha = 0.10$

5.1.1.1 Outlier Treatment

Table 5.2 shows the inter-quartile calculations, necessary for outlier treatment, for data set variables including mean score, 25% inter-quartile value, 75% inter-quartile value, minimum extreme value, maximum extreme value, and the number of outliers found.

The outlier treatment calculations found that the data set variables did not have any outliers. Thus, none of the participant scores were removed from the data set variables. All of the participant scores were included for further analysis.

Data Set Variables	Mean	25 % Inter-Quartile	75% Inter-Quartile	Minimum Extreme	Maximum Extreme	Outliers
TME Mean	77.3	68.5	88.5	46.3	108.3	0
TME 1	57.2	50.0	65.0	26.2	88.2	0
TME 2	72.1	62.5	80.5	36.1	108.1	0
TME 3	72.7	72.7	65.5	36.7	108.5	0
TME 4	76.3	68.5	88.0	37.3	115.3	0
TME 5	77.4	69.5	88.5	39.4	115.4	0
TME 6	78.2	71.0	89.0	42.2	114.2	0
Frasca Pre	65.0	65.0	85.0	35.5	115.5	0
Frasca Post	75.0	75.0	90.0	51.8	111.8	0

Table 5.2 Outlier Treatment Values.

5.1.2 Descriptive Statistics

After determining if any outliers needed to be treated, descriptive statistics were calculated for all of the data set variables. The descriptive statistics included: population size (n), mean (mean), standard deviation (σ), minimum score (min), and maximum score (max) (Table 5.3).

The descriptive statistics indicated that the lowest TME mean was 57.2 (TME 1) and the highest TME mean was 78.2 (TME 6).

Also, each data set variable had a similar σ. With respect to a previous CTM study (Nicolalde & Funk, 2003), the variability in participant performance is much lower in this study. While the number of participants (specific aviation training) in this study was only nineteen, the σ was approximately fifteen or less, compared to Nicolalde and Funk's study of 94 participants (random students) where the σ was approximately 21 to 24.

Data Set Variable	n	mean	σ	min	max
TME Mean	19	77.3	13.2	46	99
TME 1	19	57.2	12.4	33	80
TME 2	19	72.1	12.9	45	95
TME 3	19	72.7	14.2	45	94
TME 4	19	76.3	14.1	46	95
TME 5	19	77.4	13.7	46	98
TME 6	19	78.2	14.1	49	99
Frasca Pre	19	75.5	14.7	45	95
Frasca Post	19	81.8	12.6	50	100

Table 5.3 Descriptive Statistic Values.

5.1.3 Frequency Distributions

The percentile ranges were calculated as follows: 10.0, 25.0, 33.3, 66.6, and 75.0. This means that 10.0%, 25.0%, 33.3%, 66.6%, and 75.0% of the population scored below the calculated value for each of the values, respectively. For example, (in the data set variable TME 6) 10.0% of the population scored below 60.200% or less, 25.0% of the population scored below 71.000% or less, 33.3% of the population scored below 72.994% or less, 66.6% of the population scored below 85.988% or less, and 75.0% of the population scored below 89.000% or below (Table 5.4).

After calculating upper limit percentile ranges, the results indicate that by definition one third of the participants scored above the 66.6 percentile range, two thirds of the participants scored below the 66.6 percentile range, one fourth of the participants scored above the 75.0 percentile range, and three fourths of the participants scored below the 75.0 percentile range.

Data Set Variable	Percentiles Ranges				
	10.0	25.0	33.3	66.6	75.0
TME Mean	65.067	68.500	70.986	85.944	88.500
TME 1	41.800	50.000	52.982	64.000	65.500
TME 2	58.600	62.500	64.988	78.976	80.500
TME 3	53.600	65.500	68.988	80.988	83.500
TME 4	58.000	68.500	69.994	85.976	88.000
TME 5	65.200	69.500	70.000	82.940	88.500
TME 6	60.200	71.000	72.994	85.988	89.000
Frasca Pre	55.000	65.000	69.970	85.000	85.000
Frasca Post	69.000	75.000	79.970	85.000	90.000

Table 5.4 Frequency Distribution: Percentile Ranges.

5.1.3.1 Histograms

Figure 5.1 shows the frequency distribution of the data set variable TME Mean. It is easy to see that six participants scored a mean score between 64 and 74. The graphical representation (as shown by the curve) portrays a normal distribution.

The kurtosis (peakedness) of the performance measures, appears to be bimodal (having two peaks). This evidence might suggest that the sample was not homogeneous but possibly its elements came from two different participant populations, each more or less normally distributed.

Additional histograms were generated for each of the data set variables. These histograms are included in Appendix 6 for reference.

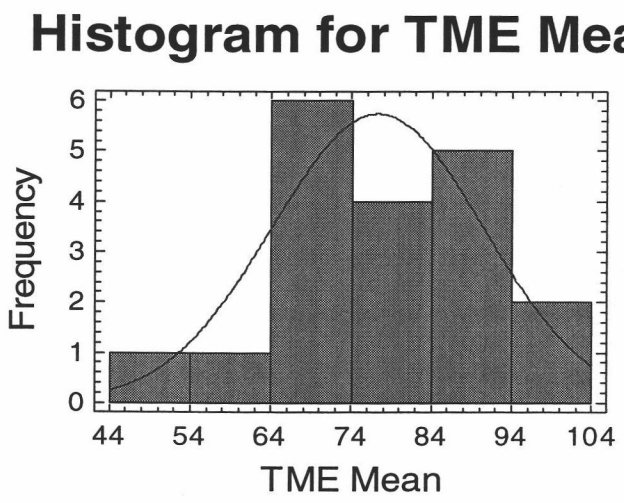


Figure 5.1 A Histogram Representing the Frequency Distribution of the TME Mean.

5.1.4 Learning Curve

The resulting learning curve of participant performance by the mean score of each TME trial was generated (Figure 5.2). An asymptotic mode was reached in the last five trials since the learning curve shows that trials two through six seemed to have stopped learning significantly. However, there is a small increase between trials two and six, but not much.

The graphical representation of the learning curve indicates that participants were still learning how to play the TME in trial one. However, beginning in trial two,

participants ceased to improve substantially in learning. This suggests that the data from trials two, three, four, five, and six can be used in the correlation analysis.

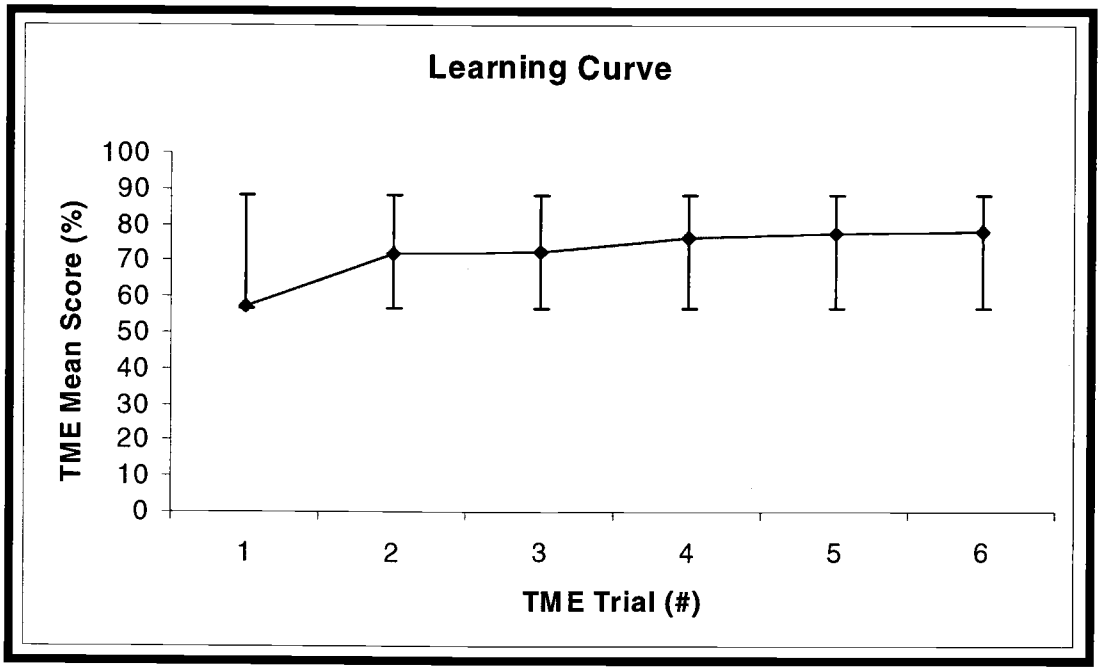


Figure 5.2 The Learning Curve for the Mean TME Score per Trial Number.

5.1.5 T-tests

The T-test paired sampling strategy was used to determine if and when the participants learned to operate the TME based on a more objective comparison. This was done to shed more light on the learning curve (Figure 5.2). If a matched pair was found to be the same or equal to zero, then that was interpreted as a way of stating that a participant has already reached his or her best performance on the earlier trial. This was used to decide which TME scores would be the "best" predictors to use for correlation analysis.

The T-tests calculated a mean difference between paired samples. For example, the mean difference between TME 6 and TME 1 is 21.1. If the mean difference varies widely from zero, then the two paired samples are not equal or are considered to be the same. If the mean difference is close to zero, then the trials are considered to be equal or the same (Table 5.5).

In the first T-test, the null hypothesis was defined where the comparison of the difference between means TME 6 and TME 1 was equal to 0.0. The alternative

hypothesis was defined where the comparison of the difference between means TME 6 and TME 1 was not equal to 0.0. Since the P-value ($2.1483x10^{-5}$) for this test was less than 0.05, the null hypothesis was rejected at the 95.0% confidence level. As a result, the difference in means between TME 6 and TME 1 was not equal to 0.0. Also, this evidence suggested that the participants were still learning in TME trial 1 and that TME trial 1 was not a "best" predictor of participant performance.

In the remaining T-tests, the null hypothesis was defined where the comparison of the difference between means TME 6 and TME 2, TME 6 and TME 3, TME 6 and TME 4, and TME 6 and TME 5 were all equal to 0.0. The alternative hypothesis was defined where the comparison of the difference between means TME 6 and TME 2, TME 6 and TME 3, TME 6 and TME 4, and TME 6 and TME 5 was not equal to 0.0. Since the P-value for these tests were not less than 0.05, the null hypothesis was NOT rejected at the 95.0% confidence level.

As a result of these T-tests, the difference between means in trials TME 2, TME 3, TME 4, and TME 5 while matched with the mean in trial TME 6 are all equal to 0.0. This evidence suggested that participants were not still learning after trial 1 and that the mean scores from TME trials 2 through 6 are the best predictors of participant performance (Table 5.5).

Test	Paired Test	Mean Difference	T-stat	P-value	Hypothesis Testing	Conclusion
1	TME 6 and TME 1	21.0526	4.8828	$2.1483x10^{-5}$	Reject	NOT same
2	TME 6 and TME 2	6.1052	1.3948	0.1716	Don't reject	same
3	TME 6 and TME 3	5.4737	1.1926	0.2408	Don't reject	same
4	TME 6 and TME 4	1.9473	0.4621	0.6726	Don't reject	same
5	TME 6 and TME 5	0.8421	0.1868	0.8529	Don't reject	same

Table 5.5 T-test Paired Sampling Results.

5.1.6 Parametric Correlations

Since the normality tests proved that most of the data fit a symmetrical normal distribution, parametric correlations were performed using parametric correlation. Each individual TME trial (including TME Mean, the average of TME trials 2 through 6) was initially compared with Frasca Post. For additional analyses, they were also compared

with Frasca Pre. To determine external validity, this study was mostly concerned with the resulting correlation between the "best" predictor of CTM performance (TME Mean) and "true" CTM performance (Frasca Post).

Table 5.6 shows the statistical results from the parametric correlations including: correlation coefficient (r), the coefficient of determination (R^2), and the statistical significance (P-value), where, H_0=there is no relationship between variables.

Fourteen parametric correlations were performed. The results of the 1st test revealed that since the P-value was greater or equal to α=0.10, there is not a statistically significant relationship between Frasca Post and TME Mean at the 90% or higher confidence level (fail to reject H_0). The R^2 value indicated that the model as fitted explains 0.32% of the variability in Frasca Post. The r value equals -0.0569, which indicated a relatively weak negative relationship between the variables.

None of the 14 tests showed a significant correlation r=0.5 (moderate: r=0.5, strong: r=0.7, very strong: r=0.9). The highest significant correlation coefficient was r=0.1152 between Frasca Pre and TME 3. None of the parametric correlations were significant enough to say that there is a relationship between TME performance and Frasca 141 performance. In fact the resulting parametric correlations were very insignificant since they were all very close to zero.

	Frasca Pre			Frasca Post		
Variables	**r**	**R^2**	**P-value**	**r**	**R^2**	**P-value**
TME Mean	-0.0115	0.0001	0.9627	-0.0569	0.0032	0.8169
TME 1	0.0467	0.0022	0.8495	0.0797	0.0064	0.7457
TME 2	-0.0209	0.0004	0.9323	-0.1351	0.0182	0.5814
TME 3	0.1152	0.0133	0.6386	-0.0546	0.0030	0.8243
TME 4	-0.0719	0.0052	0.7700	-0.1904	0.0120	0.6557
TME 5	-0.0203	0.0004	0.9341	-0.0541	0.0029	0.8260
TME 6	0.0583	0.0034	0.8125	0.0024	5.68×10^{-6}	0.9923

Table 5.6 Parametric Correlation Results.

5.1.6.1 Scatter Plots

Scatter plots were generated to compare each dependent variable *x* (TME CTM performance) with the independent variable *y* (Frasca CTM performance). Figure 5.3 shows a scatter plot of all of the participants' TME Mean scores with comparison to their respective Frasca Post scores. This plot shows that many of the participant performances

were not closely related to the regression line. The regression line represents the "best fit" correlation equation (y=-0.0541x+86.022) based on the variables. In this scatter plot, the regression line is sloped in a negative direction. This scatter plot gives inferences to conclude that there is a flat and small correlation since the regression line is mostly negative and sloped down to the right. In this case, x clearly is not a good predictor of y.

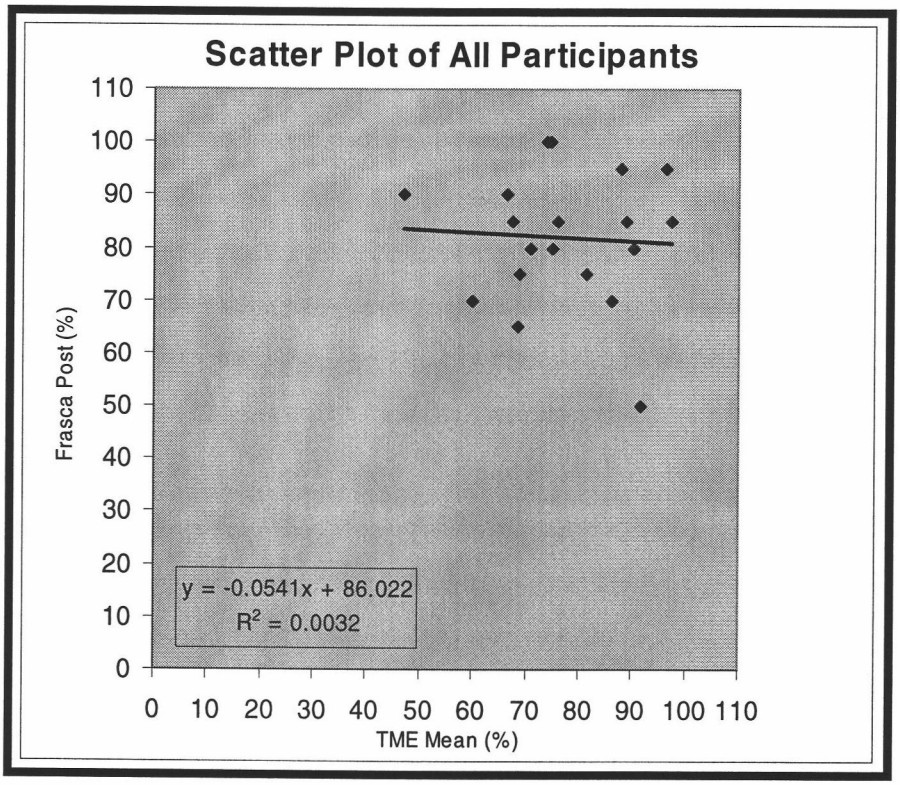


Figure 5.3 A Scatter Plot of TME Mean and Frasca Post Using Parametrics.

See Appendix 7 for additional scatter plots with respect to each of the resulting correlations between TME trials and Frasca Post.

5.1.7 Non-Parametric Correlations

Although the normality tests proved that most of the data followed a normal distribution, further analyses using non-parametric correlations were performed using non-parametric correlation. This analysis was not technically necessary but since the study had a small sample size of nineteen (thirty to one hundred recommended for parametric analysis), the normality tests and T-tests might not be very representative of

the experiment itself but perhaps, instead, it is only representative of this particular sample group.

The results of the 1st test (Table 5.7) reveal that since the calculated t_0=0.0812 is NOT greater than the t-distribution table value of 1.3333, there is not a statistically significant relationship between Frasca Post and TME Mean at the 90% or higher confidence level (accept H_0). The R^2 value indicates that the model as fitted explains 0.04% of the variability in Frasca Post. The r value equals 0.0197, indicating a relatively weak positive relationship between the variables.

None of the 14 non-parametric correlations reached an acceptable level of correlation of at least r=0.5, where moderate: r=0.5, strong: r=0.7, or very strong: r=0.9. None of the correlations were strong enough to say that there was a relationship between TME performance and Frasca 141 performance.

	Frasca Pre			Frasca Post		
Variables	**r**	**R^2**	**t_0**	**r**	**R^2**	**t_0**
TME Mean	0.0522	0.0027	0.2155>1.3333	0.0197	0.0004	0.0812>1.3333
TME 1	0.0233	0.0005	0.0961>1.3333	0.0504	0.0025	0.2081>1.3333
TME 2	-0.1162	0.0135	-2.0035>1.3333	-0.0566	0.0032	-0.2337>1.3333
TME 3	0.0882	0.0078	0.3651>1.3333	-0.0522	0.0027	-0.2155>1.3333
TME 4	-0.0031	9.61×10^{-6}	-0.1366>1.3333	0.0145	0.0002	0.0598>1.3333
TME 5	0.0715	0.0051	0.2956>1.3333	0.1364	0.0186	0.5676>1.3333
TME 6	0.1241	0.0154	0.5156>1.3333	0.1004	0.0101	0.4166>1.3333

Table 5.7 Non-Parametric Correlation Results

5.1.8 Demographics

The post-experiment questionnaire (Appendix 5) provided information about age, gender, TME strategy and additional information (Appendix 8).

5.1.8.1 Age and Gender

The sample population that was tested in this study was composed of 19 student pilots from Central Washington University (CWU). Only 2 of the participants were female. The participant ages ranged from 19 to 26. The average participant age was about 21 years (Table 5.8).

Sample Population	
Participants	19
Male	17
Female	2
Age Range	19 to 26
Avg. Age	21.32
Age σ	1.77

Table 5.8 Age and Gender Demographics.

5.1.8.2 Strategies and Comments

Additional subjective questions were asked in the post-experiment questionnaire to determine if the participants used any strategies for the TME testing and if they noticed any similarities to flying any airplane with the TME software (Appendix 5).

Participants were asked if they used any strategy to perform well in the TME game. The results showed that eleven participants focused mainly on the importance value of each subsystem, four participants managed all discrete subsystems at once when the blue bar was close to the yellow zone and on the second click, one person tried to keep everything as level as possible without shedding tasks, one person gave up totally on subsystem S05 (importance value of 2), one participant tried to keep all of the bars in at least the yellow zone, one participant worked sequentially, and two participants learned how to click and hold on a subsystem while scanning with the mouse cursor (meanwhile the current subsystem would improve) until another subsystem was found to be in an unsatisfactory state. One participant chose not to answer the question.

Also, participants were asked if they noticed any similarities in playing the TME game to flying an airplane. The results provided multiple opinions. Six participants noticed a similarity in scanning. Four participants noticed a similarity in multi-tasking. Three participants noticed a similarity in task prioritization importance between flying an airplane and playing the TME game. Two participants noticed no similarity in playing the TME and flying an airplane. One participant noticed a similarity in task management in general. One participant noticed a similarity in attention demand (resource allocation). Two participants noticed little or no similarity between playing the TME and flying an airplane. One participant chose not to answer.

Finally, participants were asked if they had any comments or suggestions concerning the TME game. The results showed that five participants found the TME game to be challenging, four participants found the TME game to be fun, two participants found the TME game to look good for measuring task performance, one participant found the TME game to be good for practicing instrument scanning, one participant wanted more practice, one participant noticed that he or she did not have much room to improve after the first few trials, one participant questioned if the scan and hold strategy was good for measuring, and one participants chose not to answer.

6 Discussion

6.1 Overview

This chapter discusses the meaning and significance of the findings in this study. The main objective was to investigate the external validity of the TME with respect to measuring human CTM performance.

Chapter 1 introduced a "real-world" example of how poor CTM can result in human error, damage, and loss of life. Chapter 2 summarized previous research in regards to CTM theory, CTM experimentation, and validity. Chapter 3 summarized the research objectives and hypothesis generation that was used for this exploratory study, based on external validation practices. Chapter 4 summarized the research methodology that was used for comparing human CTM performance using the TME and Frasca 141. Chapter 5 presented the results of experimentation for a sample population of 19 pilot participants. Correlation analysis was performed between the Frasca 141 performance measure, Task Prioritization Error Rate, and the TME performance measure, Total Weighted Score (TWS).

This chapter discusses the findings (especially from the correlation analyses), discusses why the results were surprising, gives implication for CTM and TME research, discusses limitations of the study, introduces recommendations for future use of the TME, and gives a brief summary of conclusions.

6.2 Summary of Findings

6.2.1 The Correlations

The statistical findings indicate that TME performance does NOT correlate significantly with Frasca 141 performance. Due to the small sample size of 19 participants, both parametric and non-parametric correlation analyses were performed to make inferences on the data gathered (although most of the data was shown to have a normal distribution). Refer to tables 5.6 and 5.7 in chapter 5 for significant values.

Based on T-tests, the best predictors for comparison were determined to be the TME Mean measure (trials 2 through 6) and the Frasca Post measure. The results from the parametric analysis showed that r was -0.0569 at a significance level of α=0.10. The results of the non-parametric analysis showed that the r was 0.0197 at a significance level of α=0.10. Both of these results indicate that the correlation coefficients are very close to 0, thus they are not accepted at a minimal level of significance where r=+/-0.5. With these findings, scientific inference can be made that there were no statistically significant correlations between CTM performance in the TME and CTM performance in the Frasca 141.

Additional correlation analyses were performed to shed more light on the research hypotheses. In the parametric analyses, the highest correlation coefficient was r=-.1904 between TME 4 and Frasca Post. However, this suggests that there is a negative correlation between the two measures. In the non-parametric analyses, the highest correlation coefficient was r=0.1152 between TME 3 and Frasca Pre. These findings suggest that there are no significantly positive relationships between TME performance and Frasca 141 performance. Thus, it can be concluded from this study that TME performance cannot be used to predict CTM performance in a more realistic environment.

6.3 Why Were the Test Results Surprising?

The results were especially surprising since the TME had shown some face validity in previous CTM research. With respect to the correlation findings, it was disappointing to uncover scientific evidence that none of the correlations between TME performance and Frasca 141 performance showed any significantly positive relationships. Not only were there no significantly positive correlations, almost all of the correlations were very close to zero.

Also, it was disappointing to find that 9 out 14 (64%) parametric correlations and 4 out of 14 (29%) non-parametric correlations and 13 out of 28 (46%) combined correlations showed a slightly negative correlation. This suggests that the parametric tests provided little or no evidence of a positive correlation and that the relationship

between the two measures was negative. However, it is unclear as to how much credence can be placed on these findings due to the limitations of this study.

6.4 Implications for CTM and TME Research

It has already been established by Muchinsky (2003) that there is a wide degree of different interpretations of the word valid. However, external validity was defined by Wickens (1998) as the degree to which we can generalize the results of a study to other people, tasks, and/or settings. External validity can be thought of as a means for proving that something can be applied to the "real-world." To prove for external validation, correlation analysis is the scientific standard.

Based on evidence from this study, a key conclusion was made about the external validity of TME, which was the basis of this study. Due to the fact that the TME was compared to a certified "real-world" simulation tool, the Frasca 141, the lack of a significantly positive correlation would suggest that the TME does not have external validity. This implies that the TME has no predictive validity as a research tool or screening tool to predict CTM performance. Also, these findings implied that previous research by Shakeri and Funk (2003), Nicolalde and Funk (2003), and Nicolalde, Uttl, and Funk (2003) have been invalidated to some degree. This is due to the fact that they used the TME as a tool to measure CTM performance. However, it does not completely invalidate their findings and their contributions should not be discarded for two reasons: more research is needed to determine the validity of the TME and they made additional contributions (e.g. augmenting the stage model of human information processing etc.).

Content validity is the degree to which Subject Matter Experts (SME's) agree that the items in a test are a representative sample of the domain of knowledge the test aims to measure (CTM performance). It is the degree to which a predictor covers a representative sample of the behavior being assessed. Based on the definition of content validity, CTM SME's agreed that the TME covers a range of knowledge in the domain for the test (e.g. task status, task importance, task prioritization, strategy, attention, and multi-tasking). Since CTM SME's helped design, configure, and model the TME, it could be concluded that the TME has some content validity. However, it is unclear

whether the TME is actually measuring CTM performance or something else. More research is needed to make a conclusive statement on content validity.

Still, evidence suggests that the TME does have some face validity. That is, it "looks-like" it would measure pilot CTM performance. Pilot SME's were used to help model the TME after aircraft subsystems by deviation rates, correction rates, and usability. By definition, face validity would imply that the TME appears to be an appropriate measure of CTM performance based on subjective results (Muchinsky, 2003). Funk et al (2003) claimed that the TME has some face validity. The results of the post-experiment questionnaire (see Appendix 5) in this study support Funk et al's claims since 18 out of 19 participants (94.7%) stated that they noticed a similarity between task management in the TME and flying an airplane. However, the high percentage might only be representative of this small sample population.

6.5 Limitations of this Study

The limitations of this study may have had an effect on the results. Several limitations have to be considered. First, the sample size of 19 is very small. An ideal sample population would have been around 100 participants (Hayter, 2002). However, testing this specific sample population was limited to one weekend where the participants were tested following their Frasca Post test. Also, data collection was challenging due to a lack of funding and difficulty in locating a source of pilot participants.

Given the time constraints to run the experiment, participants were not given any practice in using the TME prior to the experiment. Thus, participants improved during the testing by learning how to play the game. T-tests were used to determine what data to use for analysis. The T-tests concluded that TME trials 2 through 6 were considered to be the same since they were proved to be statistically equivalent. Thus, the TME Mean for trials 2 through 6 was used for correlation analysis.

Although it might be unrealistic, it would have been ideal to measure pilot CTM performance while flying an airplane and compare that to the pilot CTM performance in the TME. However, this would possibly endanger lives, encounter limited resources, demand more time, and incur significant costs.

In this study, the experimental construct was CTM performance. Since construct validity is defined as the degree to which a test accurately represents a faithful measure of the construct that it aims to measure, findings from this experiment suggest that the study might have lacked construct validity to some degree. The statistical findings show that there are divergent correlation coefficients (existence of a low correlation). It is a possibility that this study was not actually measuring CTM performance in both the Frasca 141 (known measures) and the TME (experimental measures).

Although certification by the FAA means that the Frasca 141 can be viewed as equivalent to flying an airplane in the "real-world" for part of a pilot's training, it has not been validated scientifically as a CTM measurement tool. Given that the FAA acknowledges task management to be important in flying an airplane, Hoover's study can only imply that the Frasca 141 can be considered to be a realistic task management environment.

Also, it is possible that the measurement scales from the Frasca 141 and the TME may not have been measuring CTM performance by definition. For comparison, the performance measures were normalized into a percentage. Performance in the Frasca 141, in the Hoover's Experiment, was normalized into a percentage of successful prioritization decisions out of 14 discrete challenge points (e.g. 1 error out of 20 potential task prioritization errors in the Frasca equals 95%). Task Prioritization Error Rate was defined by Hoover (2005) as the opportunity, during a challenge point, for a participant to divert his or her attention from a more important or more urgent task to a less urgent or less important task. However, performance in the TME, in Stebel's Experiment, was based on a continuous measure called the Total Weighted Score (TWS). The TWS is calculated every 0.1 second for each subsystem. This implies that the TWS records 1 error for every 0.1 second for each subsystem, when a subsystem is unsatisfactorily managed (status level = yellow or red) until that subsystem has been improved to a level of satisfaction (status level = green). Thus, there might be a better way of measuring CTM performance with respect to the TME.

With respect to internal validity, findings from this study suggest that it does not meet the requirements of internal validity. Internal validity is defined by Wickens (1998) as a condition in an experiment where the causal or independent variables and no other

extraneous variables caused a change in the effects being measured. Campbell and Stanley (1963) stated that internal validity is not usually used for exploratory studies but it is used for causal studies where a control group experimental design is used to determine whether event x led to event y. Since this study relied upon data from Hoover's Experiment, it is possible that there were unwanted sources of variability that might have affected the Frasca Post measures that were used for comparison in this study. As described above, Hoover's Experiment included a control group and experimental group design. Both groups performed two flight scenario simulations two weeks apart. However, the experimental group was debriefed before their second attempt. The debriefing helped clear up any misunderstandings about CTM. The results of the "Hoover's Experiment" showed that the experimental group made a 54% decrease in their respective Task Prioritization Error Rate. There is a possibility that the debriefing could have caused an unwanted source of variability that may have had an effect on the Frasca Post measures, which may have had an effect on the correlations in the "Stebel's Experiment." In Figure 5.1, the histogram for TME Mean shows that the distribution of data appears to be bimodal (having 2 peaks). This might suggest that the sample is not homogeneous but it is possible that this could indicate that there are two different populations, each more or less normally distributed. The existence of two separate populations could contaminate control of the experiment and contaminate the data.

Previous CTM research has concluded that participants who use a prioritization strategy perform better than those who do not (Colvin and Funk, 2000). A post-experiment questionnaire in Stebel's Experiment (Appendix 5) found that most participants used a strategy of prioritizing based upon task importance and urgency. Two participants found an "Improve and Move" strategy where they could improve on one TME subsystem by clicking and holding the mouse button while moving the mouse cursor to an urgent task simultaneously. Two participants learned how to do this in the TME and they performed significantly better than the rest of the population. They averaged a normalized score of 87% and 89% but reached close to 99% at times.

6.6 Recommendations for Future Use of the TME

Based on the findings and experience gained conducting this study, there are several recommendations to make with respect to the future use of the TME for CTM research.

6.6.1 Increase Sample Population

A larger sample population would have cut down on variance in the correlation analysis. In this study, a larger sample size would have allowed the grouping of data among better performers and worse performers, in which case it could have decreased the variability in the scores and resulted in possibly higher correlations among variables. A sample population of approximately 100 participants would be desirable.

6.6.2 Compare Discrete Error Rates

Based on the correlation results between CTM performance in the Frasca 141 and CTM performance in the TME, it would be desirable to the investigate and evaluate the Frasca 141 Task Prioritization Error Rates and discrete error rates, not yet defined, in the TME. This may have been a major reason why no correlation was found in this study. The experiment may have been set to measure two different things.

This future research could be performed using the data from this study. Investigation of the data files in Stebel's Experiment could be used to identify errors (possibly based on challenge points). These errors could be initially termed **D**iscrete **E**rror **R**ates of the **T**ask Management Environment (DERT). A discrete error can be defined as when a participant attends to (clicks on) a subsystem when a subsystem with a higher importance is in an unsatisfactory (yellow or red zone) state. A *Discrete Error Rate* is the rate of occurrence, either the number of discrete errors divided by the number of discrete error opportunities or the number of discrete errors divided by time. It is possible that this approach may yield a significantly different outcome with respect to determining the external validity of the TME. If that were the case, and the result was that TME showed scientific evidence of external validity, it would be appropriate to perform a reliability study with a new set of subjects to verify that the predictor test is

actually the best beyond this sample, perhaps by a different pilot population. Different domain studies (e.g. driving an automobile) would help us understand if the TME is a cross-domain multi-tasking environment.

6.6.3 TME Modification

With respect to the TME, there is still reason for exploring its validity further, especially since it has face validity and some content validity. Modifying the TME might improve its external validity.

6.6.3.1 Incorporate Fast Rates of Change.

The TME does not allow the user to make quick and sudden changes at a fast rate. For example, if a pilot is operating an airplane in an emergency situation, the pilot can pull the yoke hard and fast to compensate quickly. However, the TME does not have that capability beyond the system parameters. If it were possible to incorporate this type of fidelity, that might increase the external validity of TME. Possibly, the "Click and Hold" method in TME could be replaced with a "Click and Drag" method where participants could click on the blue bar and drag it up at faster rates.

6.6.3.2 Incorporate Multiple Management Controls.

The TME does not allow the user to change the status of two subsystems at once and this may be unrealistic. For example, a pilot can change altitude, attitude, and direction by moving the yoke and other controls at once. One way to incorporate this might be to add a small button to each subsystem. Clicking on that button (at least two subsystems) could be used to control several subsystems simultaneously.

6.6.3.3 Incorporate Distractions

The TME does not account for additional alarms (in some situations) much like flying an airplane. These could be thought of as distractions or interactions as the CTM research identifies. A failed discrete subsystem might be a distracter where its display is more salient. For example, when a pilot is flying an airplane and a landing gear light malfunctions, the pilot is distracted from flying the plane. Therefore, the TME could be

set up where some subsystems rely upon the success of others for their individual performance. The incorporation of audio cues could be used to alarm participants that an important subsystem is in a critical state.

6.6.3.4 Incorporate Voice Control

The TME does not include a verbal control. Incorporation of a verbal control would increase the opportunity for realistic task management by including communication. For example, when a pilot is flying an airplane and the air traffic controller is speaking to him or her, the pilot has to concurrently manage aviate and navigate tasks as well as communication tasks. Driving research has recognized this as a task management problem with respect to cell phone usage while driving. A standard telecommunication device could be used to account for this type of task.

6.7 Summary and Conclusions

Although the Task Management Environment (TME) has some face validity, the results of this study showed no significant correlation between CTM performance in the TME and CTM performance in a certified "real-world" simulator, the Frasca 141. Based on standard validation practices, this study suggests that the TME does not have external validity and that the TME cannot be used as a research tool or a screening tool to make generalizations about a participant's CTM performance to the "real-world."

Due to time constraints, the main objective of this study was fulfilled and no other analyses were performed. Still, several complementary analyses of the data set could have shown different relationships among the experimental and control groups. Future research in comparing the Frasca 141 Task Prioritization Error Rates with the Discrete Error Rates of the TME (DERT) might lead to new conclusions about the external validity of the TME.

With these implications and recommendations in mind, it is clear that more research is needed to determine the validity of the TME. The TME should not be abandoned because it has some face validity and because the limitations in this study may have led to the weak correlations. Further analysis and/or modification of the TME might lead to new conclusions about its validity as a research tool.

References

Adams, M., Tenney, Y., and Pew, R. (1991). State-of-the-art report: Strategic workload and the cognitive management of advanced multi-task systems. SOAR CSERIAC 91-6: Crew System Ergonomics Information Analysis Center Wright-Patterson Air Force Base, OH.

Bailey, R. (1982). Human Performance Engineering: A Guide for Systems Designers. Englewood Cliffs, NJ: Prentice Hall, Inc.

Bishara, S. and Funk, K. (2002). The effectiveness of cockpit task management training on task prioritization performance in simulated flight. Unpublished master's thesis, Industrial Engineering, Oregon State University, Corvallis, OR.

Boeing Company. (1998). Statistical summary of commercial jet aircraft accidents: worldwide operations 1959-1997. Seattle: Boeing Commercial Airplane Group. www.boeing.com.

Campbell, D. and Stanley, J. (1963). Experimental and Quasi-Experimental Designs for Research. Rand McNally and Company.

Chapanis, A. (1996). Human Factors in Systems Engineering (1st ed.). New York: John Wiley and Sons, Inc.

Chen, J. and Funk, K. (2003). A fuzzy model of human task management performance. Unpublished master's thesis, Industrial Engineering, Oregon State University, Corvallis, OR.

Chou, C. (1991). Cockpit task management errors: A design issue for intelligent pilot-vehicle interfaces. Unpublished doctoral dissertation, Industrial Engineering, Oregon State University, Corvallis, OR.

Chou, C., Madhaven, D., and Funk, K. (1996). Studies of cockpit task management errors. The International Journal of Aviation Psychology. 6 (4), 307-320.

Clause, C., Mullins, M., Nee, M., Pulakos, E., and Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example. Journal of Personnel Psychology. 51, 193-208.

Colvin, K. and Funk. (2000). Factors that affect task prioritization on the flight deck. Unpublished doctoral dissertation, Industrial Engineering, Oregon State University, Corvallis, OR.

Colvin, K., Dodhia, R., and Dismukes, K. (2005). Is pilots visual scanning adequate to avoid mid-air collisions? Symposium conducted at the proceedings of International Symposium of Aviation Psychology, Oklahoma City, OK.

Damos, D. (1991). Multiple Task Performance. London, England: Taylor and Frances.

Frasca Simulators, Inc. (2005). Frasca Model 141: Single Engine Training Device. www.frasca.com/web pages/brochures/141bro.htm.

Funk, K. (1991). Cockpit task management: Preliminary definitions, normative theory, error taxonomy, and design recommendation. The International Journal of Aviation Psychology, 1, (4), 271-285.

Funk, K. and Braune, R. (1999). The AgendaManager: A knowledge-based system to facilitate the management of flight deck activities. Symposium conducted at the proceedings of the World Aviation Congress, October 19-21, San Francisco, CA.

Funk, K., Colvin, K., Bishara, S., Nicolalde, J., Shakeri, S., and Chen, J.Y. (2003). Training Pilots to Prioritize Tasks: Theoretical Foundations and Preliminary Experiments (NASA Grant NAG 2-1287: Final Report). Industrial Engineering, Oregon State University, Corvallis, OR.

Funk, K. and Shakeri, S. (2003). A comparison between humans and mathematical search based solutions in managing multiple concurrent tasks. Unpublished master's thesis, Industrial Engineering, Oregon State University, Corvallis, OR.

Hoover, A. and Funk, K. (2005). Experimental analysis of task prioritization training for a group of university flight technology students. Unpublished doctoral dissertation, School of Education, Oregon State University, OR.

Kaplan, R. and Saccuzzo, D. (2001). Psychological Testing: Principles, Applications, and Issues (5th ed.). Belmont, CA: Wadsworth.

Lapin, L. (1998). Probability and Statistics for Modern Engineering (2nd ed.). Prospect Heights, IL: Waveland Press, Inc.

Lubinski, D. and Dawis, R. (1992). Aptitudes, skills, and proficiencies. Handbook of Industrial and Organizational Psychology (2nd ed.), 1-59. Palo Alto, CA: Consulting Psychologist Press.

Mancayo, J. and Funk, K. (2003). Development of TME Version 2.0. Unpublished master's project, Industrial Engineering, Oregon State University, Corvallis, OR.

McQueen, N. (1917). The Distribution of Attention. Cambridge, United Kingdom: Press Syndicate of the University of Cambridge.

Montgomery, D. (2001). Design and Analysis of Experiments (5th ed.). John Wiley and Sons, Inc.

Moray, N., Dessouky, M., Kijowski, B., and Adapathya, R. (1991). Strategic behavior, workload, and performance in task scheduling. Human Factors and Ergonomics Society, 33 (6), 607-629.

Muchinsky, P. (2003). Psychology Applied to Work: An Introduction to Industrial and Organizational Psychology (7th ed.). Wadsworth and Thomson Learning, Inc.

National Highway Traffic Safety Administration. (2005). The National Advanced Driving Simulator (NADS) Is the Most Sophisticated. www-nrd.nhtsa.dot.gov/departments/nrd-12/nads/NADS.htm.

National Transportation and Safety Board. (1979). Aircraft Accident Report. United Airlines, Inc., McDonnel-Douglas DC-8-61, December 28, 1978, Portland, OR, N8082U, (NTSB-AAR-79-7), Washington, D.C.

New Oxford American Dictionary. (2001). New York, New York: Oxford University Press, Inc.

Nicolalde, J., Uttl, B. and Funk, K. (2004). The augmented stage model of human information processing: How well do cognitive abilities drawn from the stages in the model predict concurrent task management performance? Symposium conducted at the Proceedings of IIE Annual Conference, Portland, OR.

Pattipati, K. and Kleinman, D. (1991). A review of engineering models of information processing and decision making in multi-task supervisory control. London, England: Taylor and Francis.

Pew, R. and Mavor, A. (1998). Modeling human and organizational behavior. National Research Council, Washington D.C: National Academy Press.

Raby, M. and Wickens, C. (1994). Strategic workload management and decision biases in aviation. The International Journal of Aviation Psychology, 4, (3), 211-240.

Ranter, H. (2005). Press release: Record low number of airliner accident fatalities in 2004, Aviation Safety Network. January 1, 2005: Washington, D.C.

Reason, J. (1990). Human Error. Cambridge, United Kingdom: Press Syndicate of the University of Cambridge.

Rogers, W. (1996). Flight deck management: A cognitive engineering analysis. Symposium conducted at the proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting, Philadelphia, PA.

Rubenstein, J., Meyer, D., and Evans, J. (2001). Human Perception and Performance: Executive control of cognitive processes in task switching. Journal of Experimental Psychology, 27, (4), 763-797.

Ruffel-Smith, H. (1979). A simulator study of the interaction of pilot workload with error, vigilance, and decisions (NASA Technical Memo 78472). NASA Ames Research Center, Moffett Field, CA.

Shakeri, S. and Funk, K. (2003) A comparison between humans and mathematical search-based solutions in managing multiple concurrent tasks. Symposium conducted at the proceedings of the 2003 Industrial Engineering Research Conference, May 18-20, 2003, Portland, OR.

StatSoft, Inc. (2005). Electronic Statistical Textbook. www.statsoft.com.

Taylor, F. (1957). Psychology and the design of machines. Journal of American Psychologist, 12, 249-258.

Tulga, M. and Sheridan, T. (1980). Transaction on Systems: Man and cybernetics, dynamic decisions and workload in multi-task supervisory control. Journal of IEEE, 10, (5), 217-232.

Vasquez, C. (2004). A preliminary study for task management environment external validation: Correlation between continuous, discrete and mixed scenarios. Unpublished master's thesis, Industrial Engineering, Oregon State University, Corvallis, OR.

Wickens, C. (2005). Aviation Human Factors Division, University of Illinois-Champaign. Attentional tunneling and task management. Symposium conducted at the proceedings of the International Symposium of Aviation Psychology, Oklahoma City, OK.

Wickens, C. and Damos, D. (1980). The acquisition and transfer of time-sharing skills. Journal of Act Psychologica, 6, 569-577.

Wickens, C., Gordon, S., and Liu, Y. (1998). An Introduction to Human Factors Engineering. New York: Addison-Wesley Longman, Inc.

Wickens, C. and Hollands, J. (1999). Engineering Psychology and Human Performance (3rd ed.). New York: Harper Collins Publishers.

Wikipedia. (2005). Spearman's Rank Correlation. www.answers.com/topic/spearman-s-rank-correlation-coefficient.

Winks Statistics Software. (2005). Statistics Tutorial: Pearson's Correlation Coefficient. www.texasoft.com/winkpear.html.

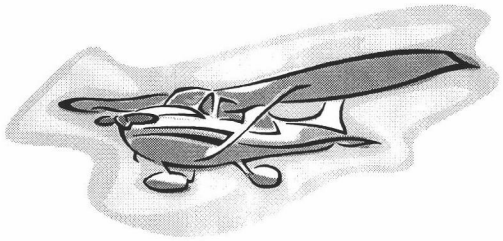
Yin, R. (1989). Case Study Research: Design and Methods (Revised ed.). Sage Publications.

Appendices

Appendix 1 – Recruitment Flyer

CENTRAL WASHINGTON UNIVERSITY
Your future is Central.

Pilot Volunteers Needed



Pilot Volunteers are needed for an experiment at Central Washington University to study how well people manage multiple tasks. Participants will play a simple (but relevant) computer game and fill out a questionnaire. The experiment will take no more than 1 hour.

Time Needed:

Only **1 Hour** of your time will be required directly following your second FRASCA 141 flight in Professor Hoover's study.

Requirements:

1. You must have completed Professor Hoover's experiment in the FRASCA 141
2. A Private Pilots License
3. Instrument Stage II completed or stage II check scheduled

Compensation:

$20 (subject to completion of playing the computer game and questionnaire.)

Contact Information:

For information or to volunteer, call or meet with Jeffrey Stebel, and ask for "Jeff "at (541) 619-4158 or send an e-mail to stebelj@engr.orst.edu .

research in cooperation with

Oregon State UNIVERSITY | College of Engineering

Appendix 2 – Informed Consent Document

Oregon State UNIVERSITY | **College of Engineering**

"Informed Consent Document"

Purpose. This project is being conducted to learn and measure how well people manage multiple concurrent tasks, like when flying an airplane. Also, it will help to determine the validity of the software tool called the Task Management Environment (TME Version 2.3).

Procedure. The experiment will be conducted as follows. First, each participant will fill out this informed consent form. Then, each participant will receive training instruction on a simple computer game. Additionally, each participant will play the computer game (TME) for a period of no longer than 1 hour and his/her performance data will be recorded. After playing the game, each participant will fill out a short questionnaire. Only a ID #, identifying a participant, will be recorded with study data. The total length of the experiment should NOT be more than approximately 1 hour. A list linking name and ID # will be kept only long enough to request each participant's pre- and post-test FRASCA 141 scores from participation in Professor Hoover's experiment, which will be used for statistical comparison in this study. After matching scores, Professor Hoover will delete the link to names so there is no possibility of identifying any individual in this study.

Risks. Each participant will understand that the probability and magnitude of harm, inconvenience, or discomfort anticipated in this study are no greater than those encountered in daily life.

Benefits. Each participant will understand that he/she will have satisfaction of knowing that he/she is participating in a study that will enable researchers to more fully understand human behavior and cognitive performance during Concurrent Task Management (CTM). Findings may increase the safety, efficiency, and effectiveness in aviation systems as well as other domains.

Compensation. Each participant will understand that he/she will be given $20 cash upon the total completion of participation in this experiment. Compensation will NOT be given, under any circumstances, to participants who do NOT fully complete: consent form, 3 practice trials, 3 test trials, and questionnaire.

Voluntary Participation. Each participant will understand that his/her participation in this study is voluntary and can be withdrawn at any time. However, if chosen to do so, he/she will NOT receive compensation.

Confidentiality. Each participant will understand that the data collected in this study will be stored in a secure computer file via an ID # (only accessible by Jeffrey M. Stebel). No personal identifiers will be included. It is understood that the anonymous data will be kept for 6 years by Oregon State University and may be used for new research. After 6 years, it will be destroyed. All results will be presented in reports and publications in a manner that will make it impossible to identify individuals.

Participant's Statement. I understand that I will have the opportunity to ask questions and receive satisfactory answers from Jeffrey M. Stebel who will be conducting this experiment at (541) 619-4158 or [stebelj@engr.orst.edu]. I understand that any further questions concerning this experiment should be directed to his Faculty Advisor, Dr. Kenneth H. Funk [funkk@engr.orst.edu] at (541) 737-2357.

If I have questions about my rights as a research participant, I may contact Central Washington University's Human Protections Administrator at the CWU HSRC Office at (509) 963-3115.

My signature below indicates that I have read and that I understand the process described above, give my informed and voluntary consent to participate in this experiment and give permission for Professor Hoover to release my FRASCA scores for use in this experiment.

Name of Participant (*please print*):_____ Date: _____

Signature of Participant: _____

Procedure for Obtaining FRASCA Scores

1. Stebel will assign an ID# to each name and record data only under ID#.

2. Stebel will provide Hoover with a list of names and his corresponding ID #s, along with signed consent forms for documentation of release of scores.

3. Hoover will enter pre and post test scores for each participant on the list then delete all names, rendering scores anonymous, before returning list (and consent forms) to Stebel.

Appendix 3 – Experimental Script

"Experimental Script"

Greetings:

Hello, I am Jeff Stebel and you have received a flyer in regards to this study. Thank you for your interest and consideration in participating in this follow-up study. In general, this study has been designed to learn how well people manage multiple tasks. Management of tasks is something we as humans do in any type of system that we operate (e.g. cooking, driving a car, performing surgery, or flying an airplane). In order for me to study your ability to manage tasks, I would like to ask you for 1 hour of your time. You will be asked to play a simple (but relevant) computer game and fill out a questionnaire. If you complete the experimental data collection process in full, you will be compensated $20 for your time.

Informed Consent:

At this time, please read the "Informed Consent Document." This document will explain the purpose of this experiment, its procedures, risks, benefits, compensation, voluntary participation, confidentiality, and additional information. If you feel that you would still like to participate in this experimental study; please print, sign, and date the document at the bottom.

How to Play the Game:

Now that you have given consent to be a participant in this study, I would like to explain how to play the game. This is a computer model of a simple system consisting of several very simple subsystems. We call it the "Task Management Environment," also known as TME Version 2.3. As you can see, each of these vertical bars represents the status of a subsystem. Each subsystem has a numeric value of importance represented by a number. The object of the game is to keep the bar in the green zone as best as you can and achieve a high score. The higher the value of importance (or subsystem number), the more this subsystem's status will affect your final score. If the bar is in the green zone you will score points. If the bar drops to the yellow zone you will not gain any points and if the bar drops to the red zone you will lose points negatively. If a very important subsystem's bar drops into and stays in the red, it will hurt your score a lot. If an unimportant bar drops into the red, it will hurt your score only a little. The relative degree of harm to your score depends on the relative importance of the subsystems. At the end of the 5 minute trial time period, the software will automatically calculate a weighted score based on how well you kept the bars in the green zones (which is how well you managed each task collectively). You will be asked to perform 6 trials.

To get the blue bars to rise; "Left-Click" (using the computer mouse) on the large numbered button to the right of the blue bar until you think the blue bar is at a satisfactory position for the time being. You will have to either: "Click and Hold" or "Click and Release," depending on how the individual task functions (*continuously* or *discretely*). Then, you should try to attend to the other tasks and raise their blue bars to a satisfactory position, in the green zone. You will do this until time expires. You have 5 minutes. The higher valued tasks are more important. Thus, they have more impact on your final score.

To Begin Play

To begin play, please "Left-Click" on the "Start" button located in the upper right hand of the screen using the mouse. The bars will begin to drop at different speeds. Please begin when you have an understanding and are ready to proceed. Do you have any questions before we begin?

Good Luck and Please Begin!

Appendix 4 – CTM Performance Measures

Oregon State UNIVERSITY | College of Engineering

"CTM Performance Measures"

Participant	FRASCA Score (%)		TME Score (%)					
	Pre-Test	Post-Test	1	2	3	4	5	6
1	85	100	55	62	71	69	75	80
2	90	80	64	82	88	87	91	93
3	70	85	73	86	69	95	98	99
4	80	75	43	72	82	84	75	85
5	85	95	64	72	86	90	86	88
6	70	75	67	79	80	75	69	62
7	55	50	50	82	88	89	94	92
8	45	65	55	60	48	68	67	70
9	90	90	43	45	45	46	46	49
10	80	85	66	72	74	74	78	76
11	85	80	50	53	64	70	70	73
12	55	95	63	89	94	94	98	97
13	65	70	33	60	56	60	67	53
14	95	85	53	79	80	86	91	90
15	65	100	57	77	69	79	77	66
16	80	80	68	77	81	76	77	72
17	85	90	37	65	55	50	70	79
18	95	70	80	95	85	89	83	86
19	60	85	65	63	67	68	58	76

Appendix 5 – Post-Experiment Questionnaire

Oregon State UNIVERSITY | **College of Engineering**

"Post-Experiment Questionnaire"

Participant ID #: _____

1. What is your age? _____ **years**

2. What is your gender?
(Please X the appropriate box below)

☐ **M** ☐ **F**

3. How long have you been using computers? _____ **years**

4. How many hours do you use computers? _____ **per day**

5. What do you consider your level of computer expertise?
(Please X the appropriate box below)

☐ **1** = beginner (least)
☐ **2** = some experience
☐ **3** = moderate experience
☐ **4** = much experience
☐ **5** = expert (highest)

6. Do you play video games?
(Please X the appropriate box below)

☐ **Yes** ☐ **No**

7. How many hours do you play video games? _____ **per day**

8. How many days do you play video games? _____ **per week**

9. How long have you been flying? _____ **years**

10. How many hours have you logged flying? _____ **total**

11. With respect to flying, do you have:
(Please X all that apply)

☐ **Private Pilots License**
☐ **Stage II Instrument Ratings**
☐ **Other** *(please indicate)* _____

12. Are there any unusual circumstances that have happened today that may affect your abilities performing in this study? *(Please X the appropriate box below)*

☐ **Yes** ☐ **No**

13. Have you consumed a high amount of caffeine today?
(Please X the appropriate box below)

☐ **Yes** ☐ **No**

14. Are you taking any medications?
(Please X the appropriate box below)

☐ **Yes** ☐ **No**

15. How rested do you feel today compared to a normal day?
(Please X the appropriate box below)

☐ **1** = exceptionally less rested than normal
☐ **2** = moderately less rested than normal
☐ **3** = normally rested
☐ **4** = moderately more rested than normal
☐ **5** = exceptionally more rested than normal

16. If you normally wear corrective lenses for computer use, were you wearing them during this experiment? *(please X the appropriate box below)*

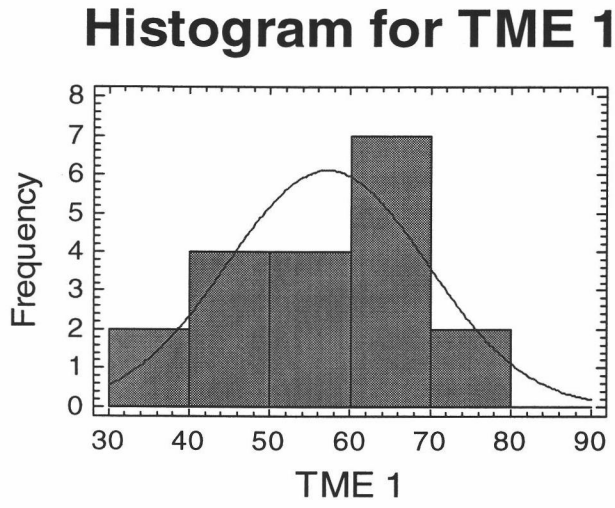
☐ **Yes** ☐ **No** ☐ **I don't normally use them**

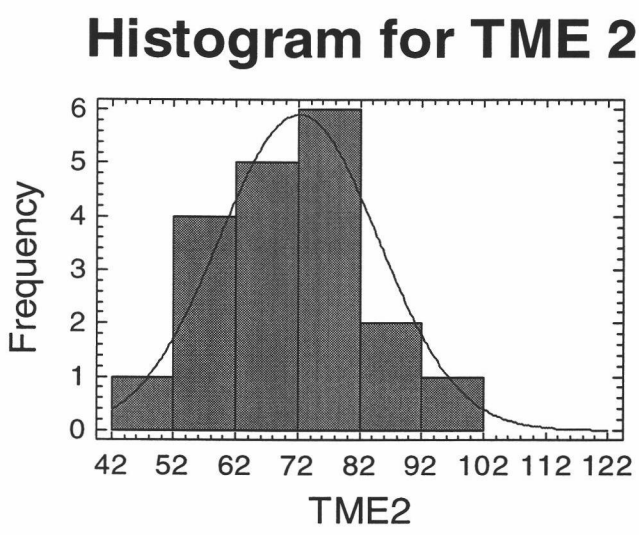
17. Did you use any kind of strategy to perform well in this game (i.e. – Did you shed tasks to achieve a higher valued score? In other words, did you give up on some subsystems in order to do well in others?)?

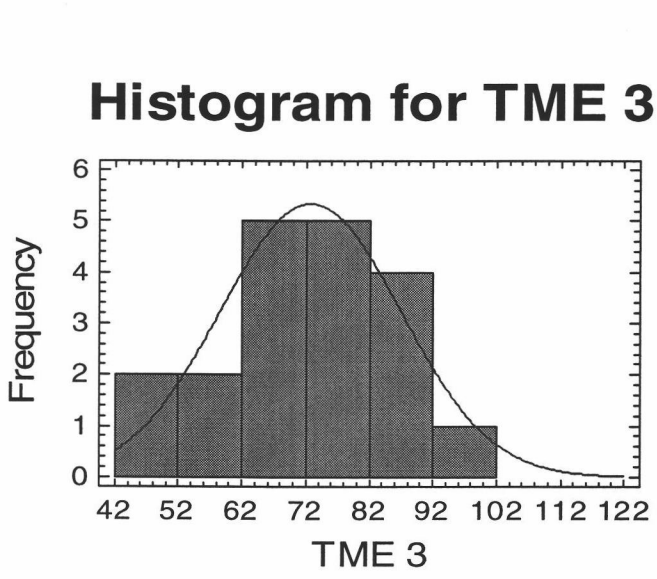
18. Did you notice any similarities in playing this game with respect to flying an airplane?

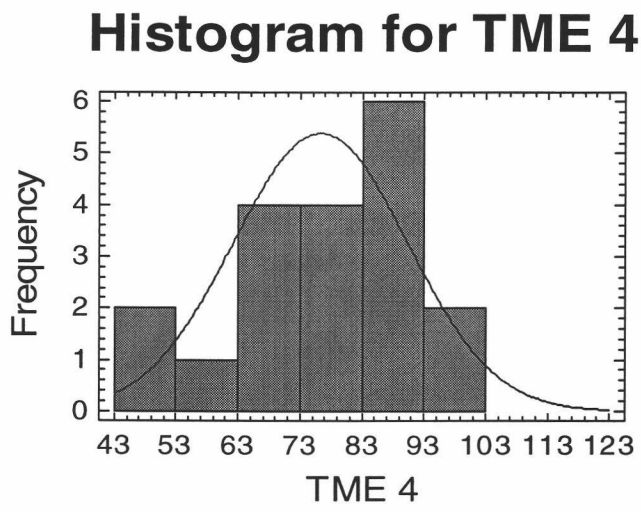
19. Write any comments or suggestions concerning this experiment with regards to: the experimental setup = *challenging* ?, the TME software game, your strategy in playing the game, what you think, etc.
(Thank you for your time!)

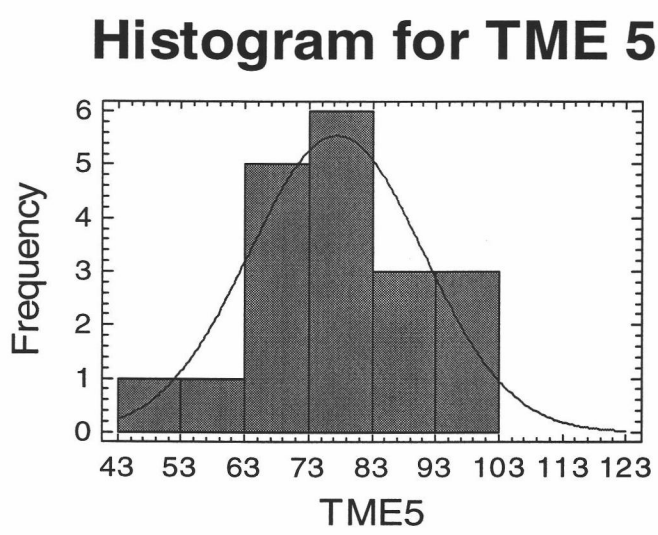
Appendix 6 – Histograms

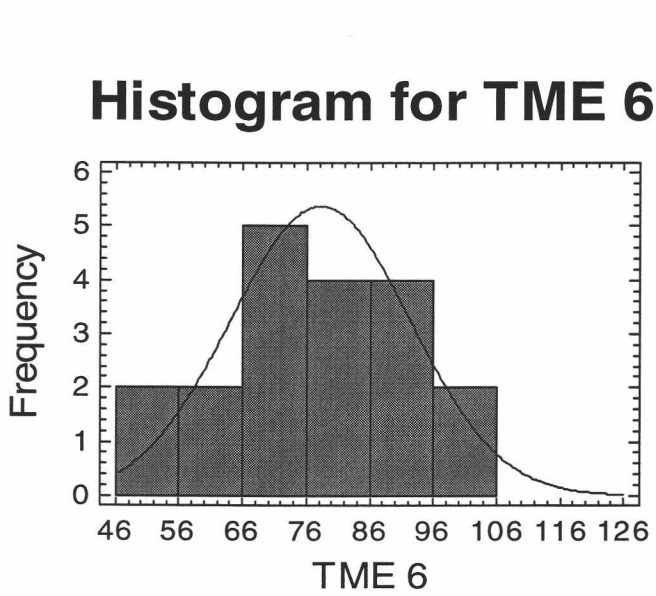

Histogram for TME 1
Frequency
0
1
2
3
4
5
6
7
8
30
40
50
60
70
80
90
TME 1


Histogram for TME 2
Frequency
0
1
2
3
4
5
6
42
52
62
72
82
92
102
112
122
TME2

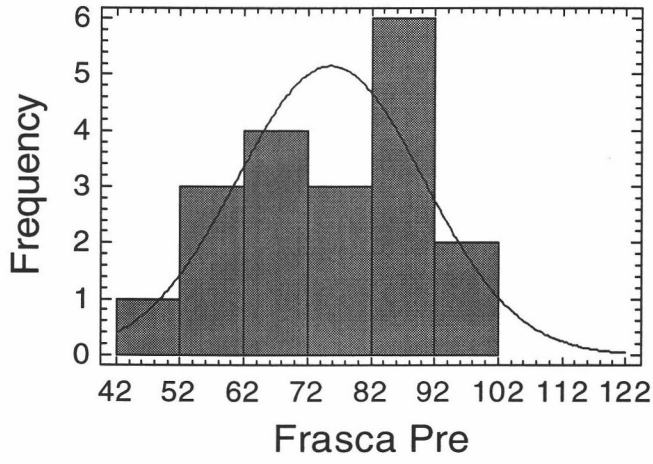

Histogram for TME 3
Frequency
0
1
2
3
4
5
6
42
52
62
72
82
92
102
112
122
TME 3


Histogram for TME 4
Frequency
0
1
2
3
4
5
6
43 53 63 73 83 93 103 113 123
TME 4

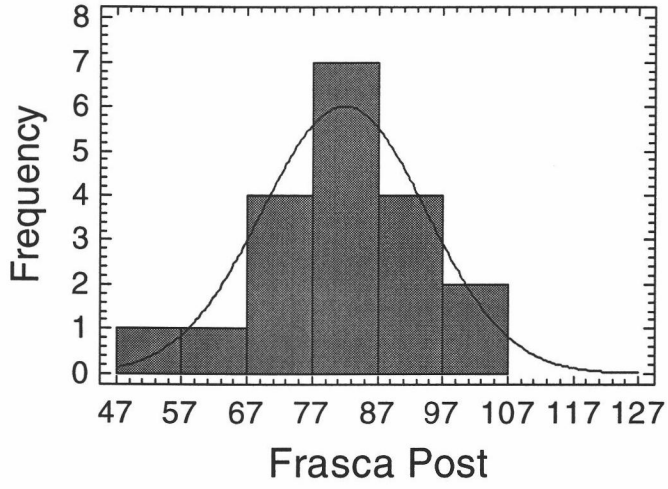

Histogram for TME 5
Frequency
0
1
2
3
4
5
6
43 53 63 73 83 93 103 113 123
TME5


Histogram for TME 6
Frequency
0
1
2
3
4
5
6
46 56 66 76 86 96 106 116 126
TME 6

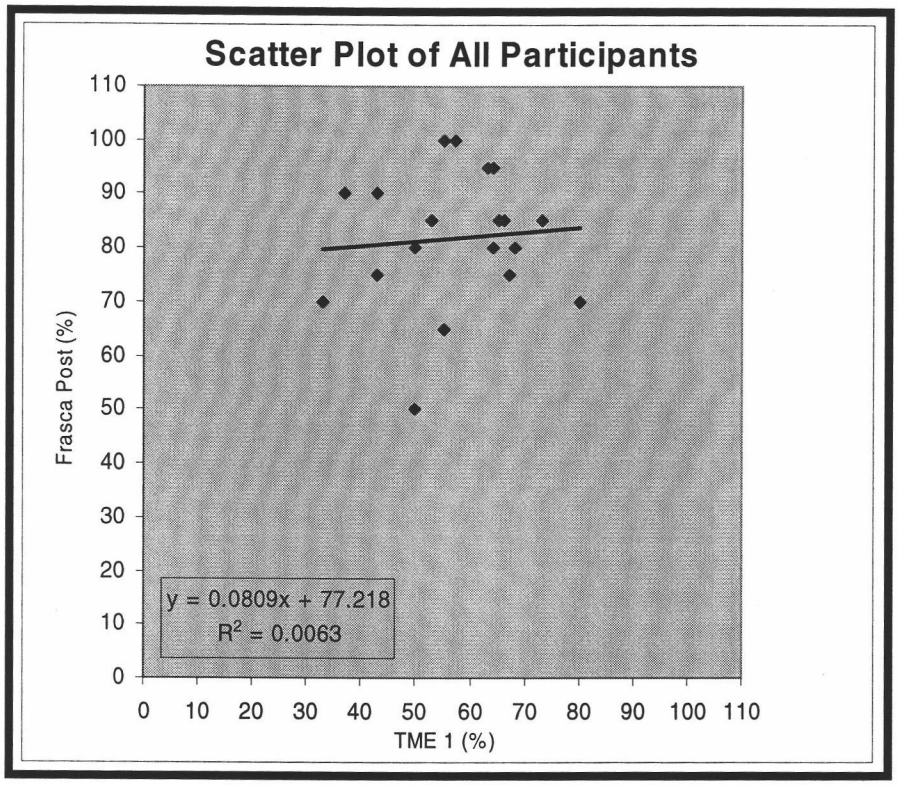
Histogram for Frasca Pre

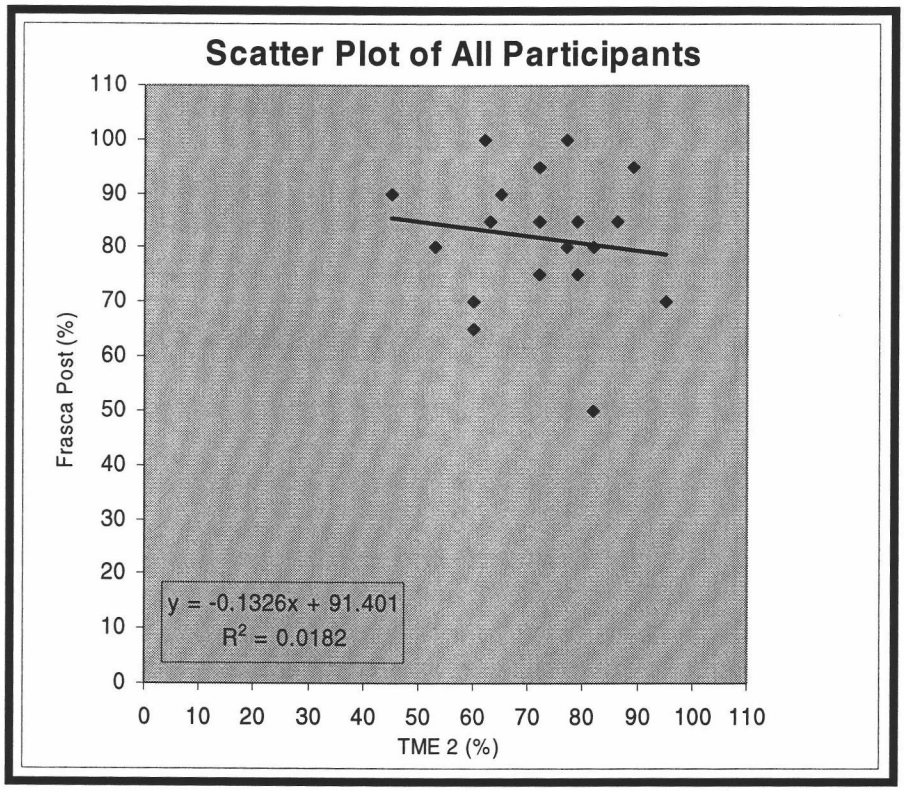


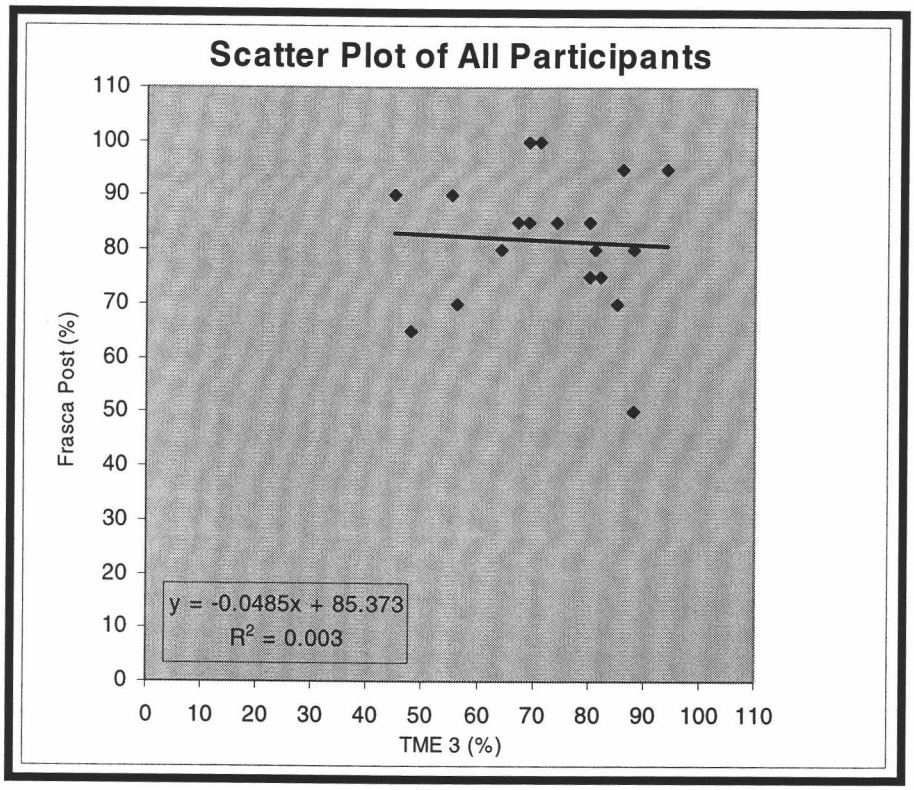
Histogram for Frasca Post

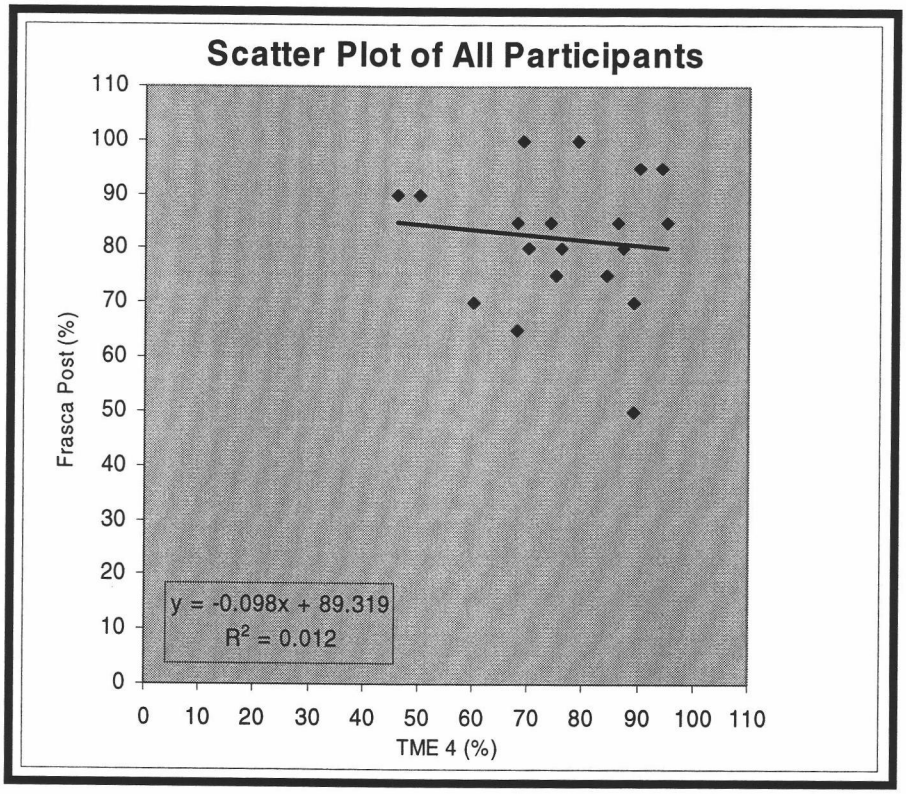


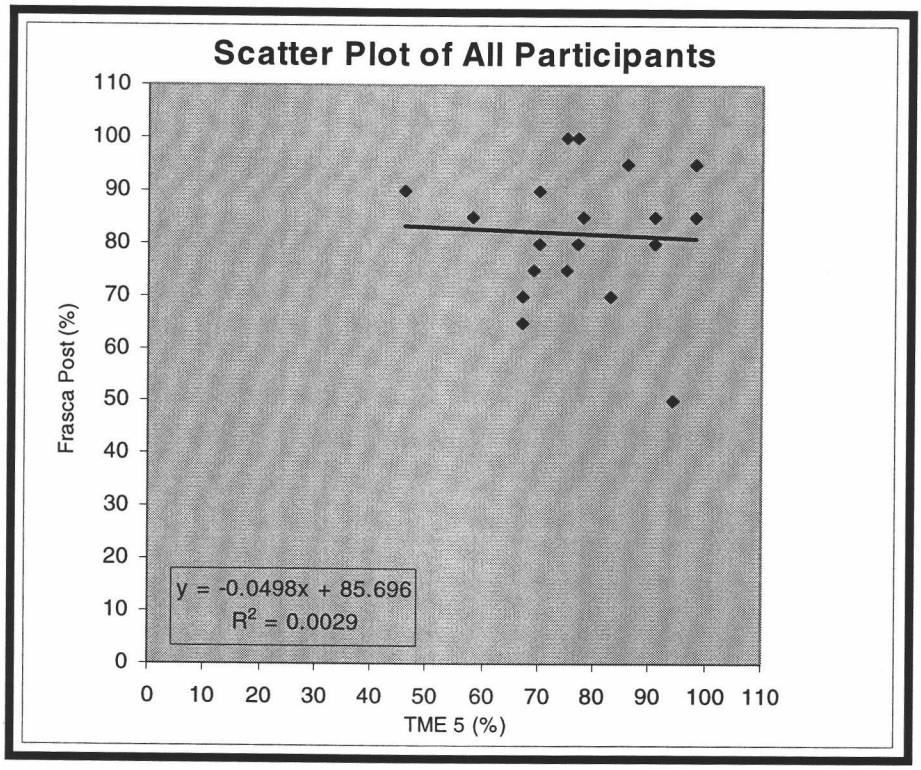
Appendix 7 – Scatter Plots

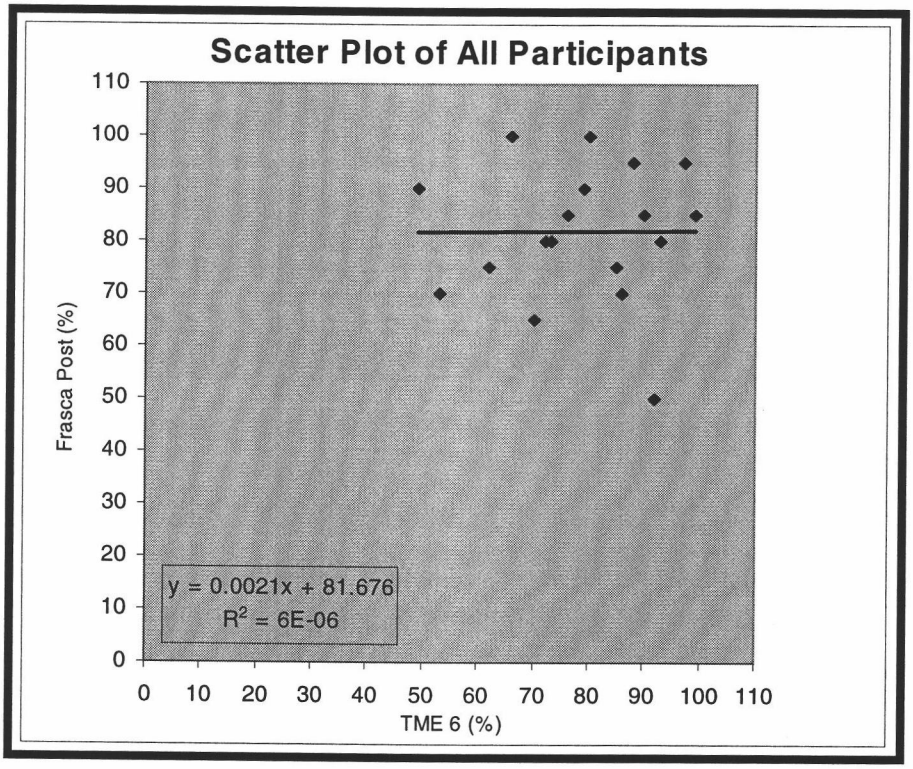

Scatter Plot of All Participants
Frasca Post (%)
TME 1 (%)
y = 0.0809x + 77.218
R2 = 0.0063
0
10
20
30
40
50
60
70
80
90
100
110

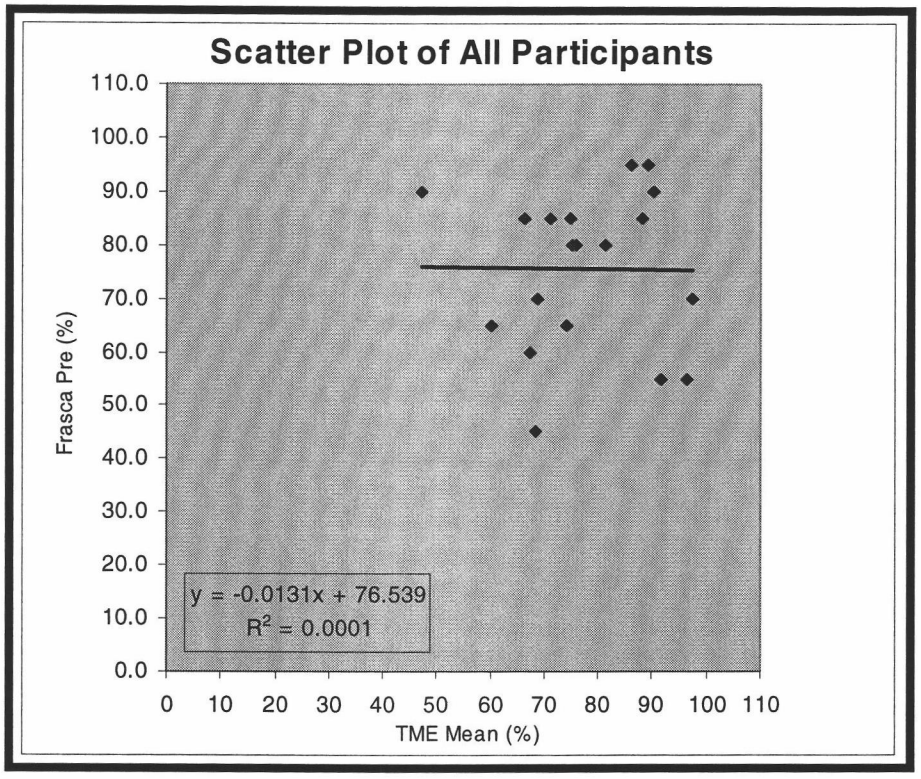

Scatter Plot of All Participants
Frasca Post (%)
TME 2 (%)
y = -0.1326x + 91.401
R2 = 0.0182
0
10
20
30
40
50
60
70
80
90
100
110

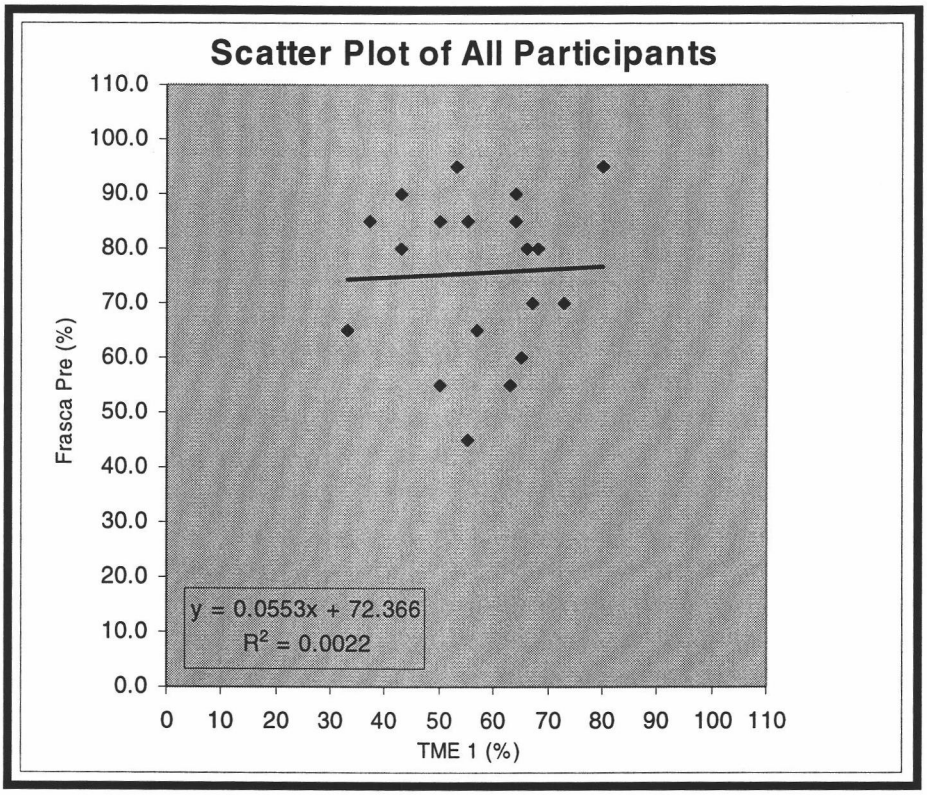

Scatter Plot of All Participants
Frasca Post (%)
TME 3 (%)
y = -0.0485x + 85.373
R2 = 0.003
0
10
20
30
40
50
60
70
80
90
100
110

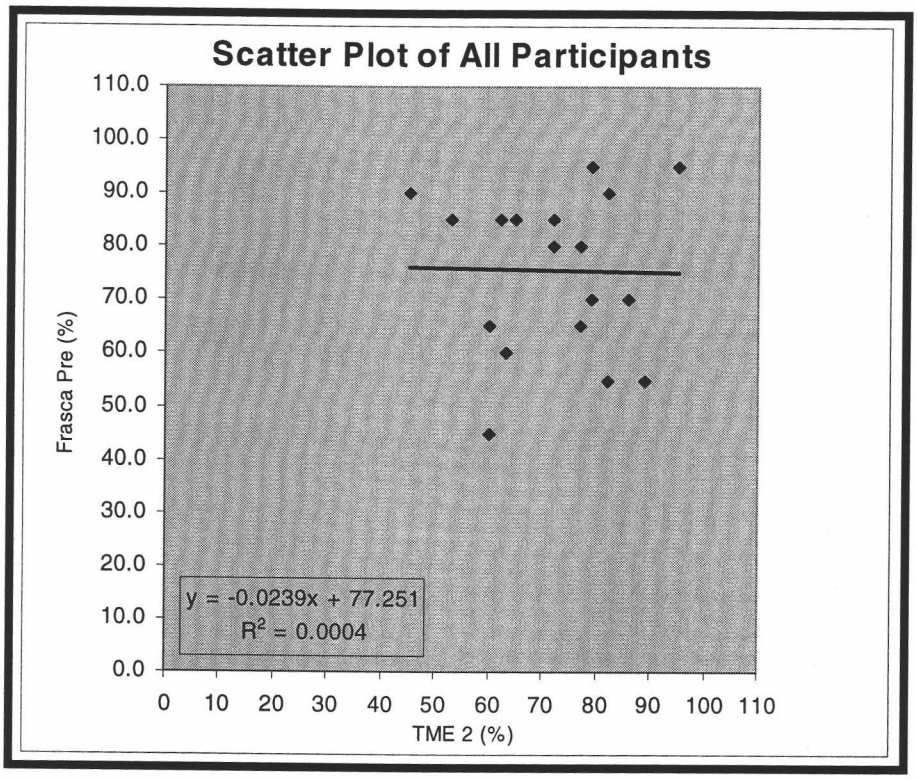

Scatter Plot of All Participants
Frasca Post (%)
TME 4 (%)
y = -0.098x + 89.319
R2 = 0.012
0
10
20
30
40
50
60
70
80
90
100
110

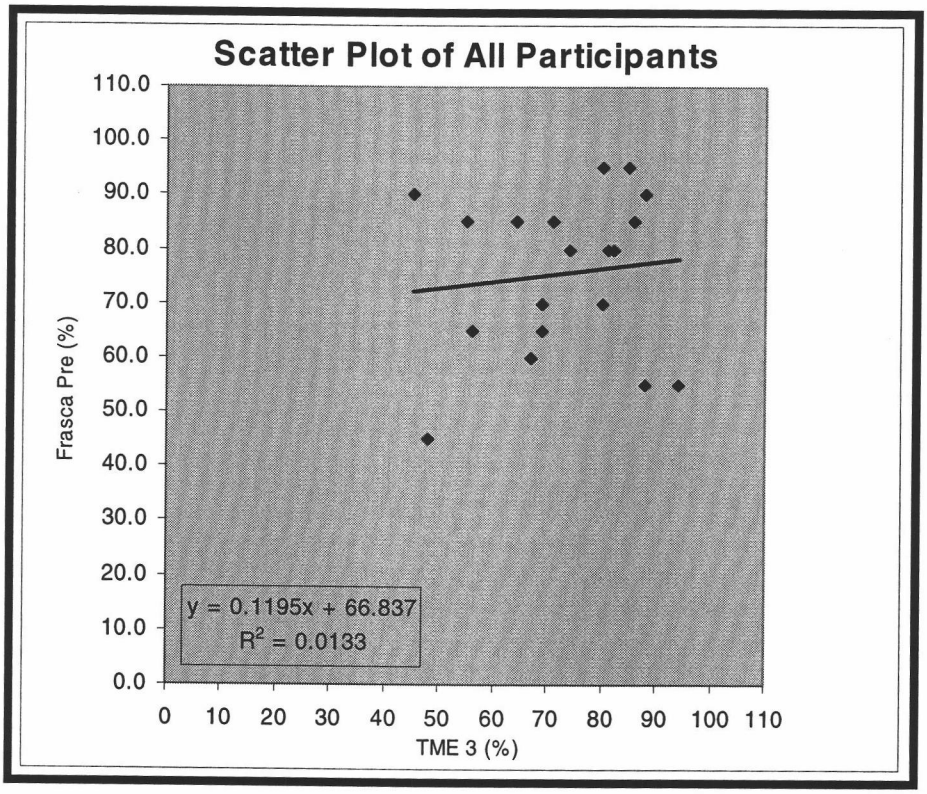

Scatter Plot of All Participants
Frasca Post (%)
TME 5 (%)
y = -0.0498x + 85.696
R2 = 0.0029
0
10
20
30
40
50
60
70
80
90
100
110

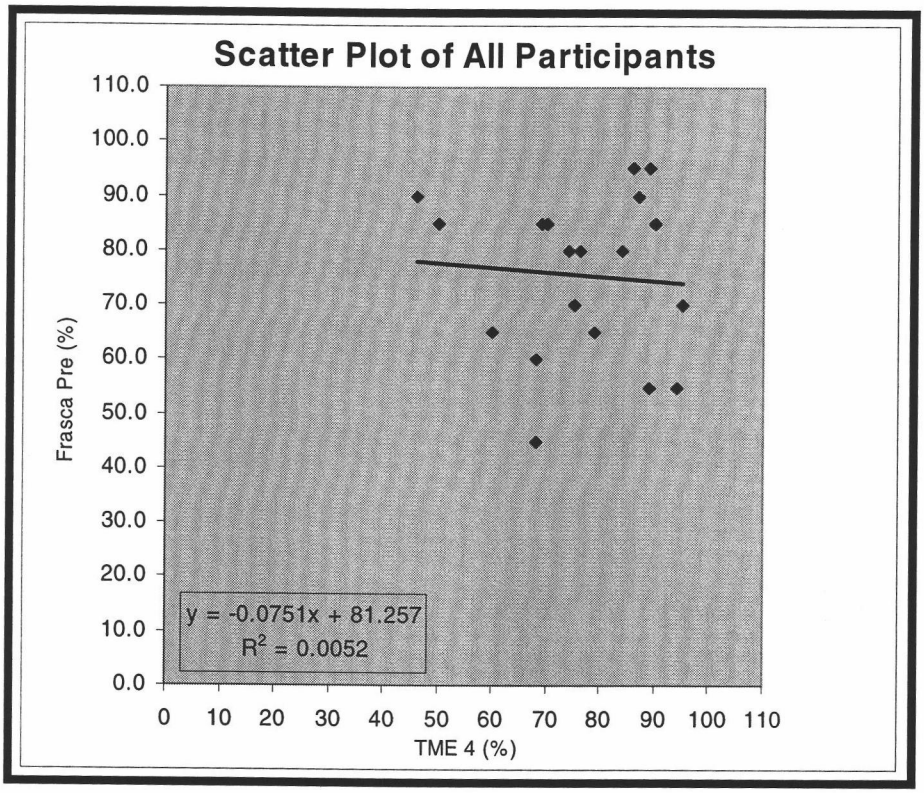

Scatter Plot of All Participants
Frasca Post (%)
TME 6 (%)
y = 0.0021x + 81.676
R2 = 6E-06
0
10
20
30
40
50
60
70
80
90
100
110

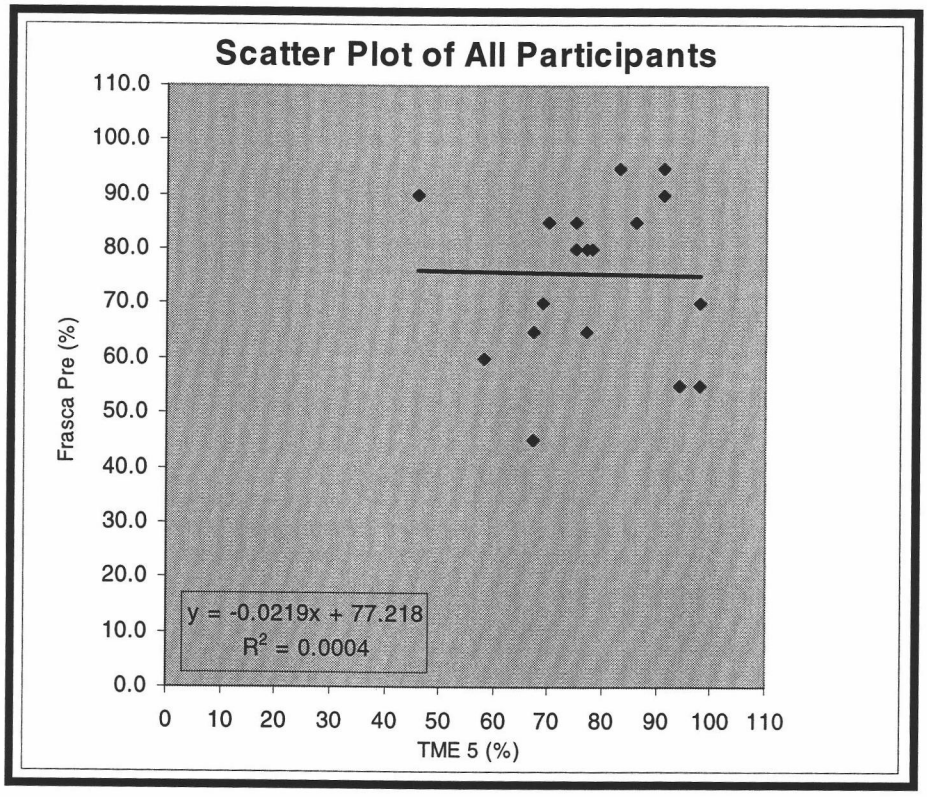

Scatter Plot of All Participants
Frasca Pre (%)
110.0
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0 10 20 30 40 50 60 70 80 90 100 110
TME Mean (%)
y = -0.0131x + 76.539
R2 = 0.0001

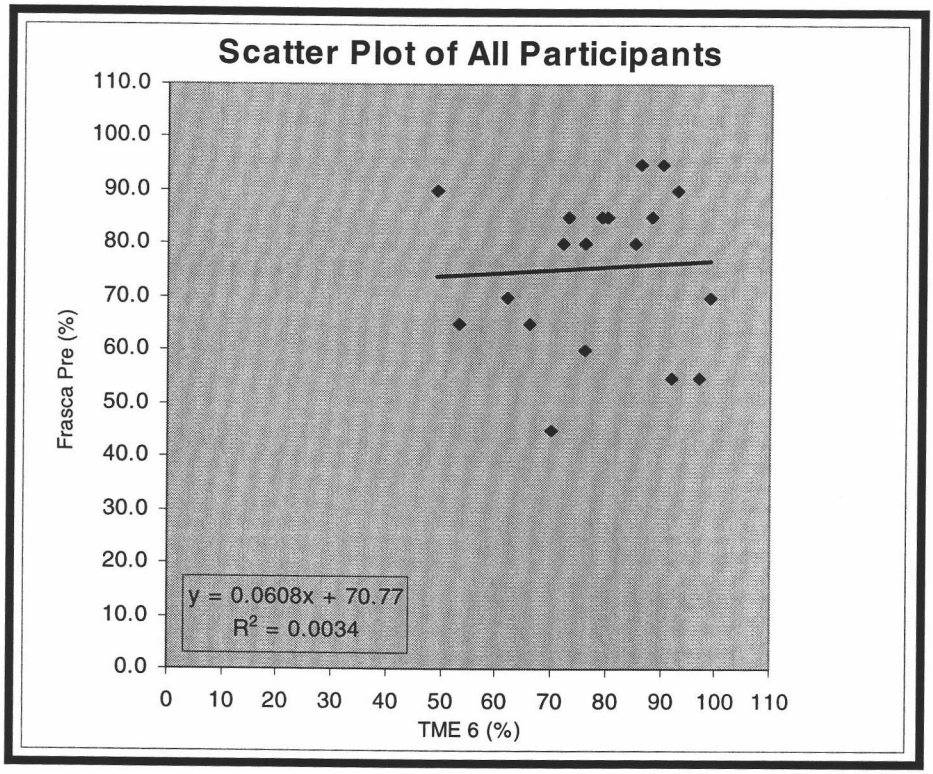

Scatter Plot of All Participants
Frasca Pre (%)
110.0
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0 10 20 30 40 50 60 70 80 90 100 110
TME 1 (%)
y = 0.0553x + 72.366
R2 = 0.0022


Scatter Plot of All Participants
Frasca Pre (%)
110.0
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0 10 20 30 40 50 60 70 80 90 100 110
TME 2 (%)
y = -0.0239x + 77.251
R2 = 0.0004


Scatter Plot of All Participants
Frasca Pre (%)
110.0
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0 10 20 30 40 50 60 70 80 90 100 110
TME 3 (%)
y = 0.1195x + 66.837
R2 = 0.0133


Scatter Plot of All Participants
Frasca Pre (%)
110.0
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
y = -0.0751x + 81.257
R2 = 0.0052
0 10 20 30 40 50 60 70 80 90 100 110
TME 4 (%)


Scatter Plot of All Participants
Frasca Pre (%)
110.0
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
y = -0.0219x + 77.218
R2 = 0.0004
0 10 20 30 40 50 60 70 80 90 100 110
TME 5 (%)


Scatter Plot of All Participants
Frasca Pre (%)
110.0
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0
10
20
30
40
50
60
70
80
90
100
110
TME 6 (%)
y = 0.0608x + 70.77
R2 = 0.0034

Appendix 8 – Post-Test Questionnaire Data

Computer Experience	
Range # of Years	7 to 15
Avg. # of Years	11.26
σ # of Years	2.7
Range # Hrs per Day	1 to 6
Avg. # of Hrs per Day	2.36
σ # Hrs per Day	1.42
Range Expertise Level	2 to 6
Avg. Expertise Level	3.26
σ Expertise Level	0.73
Video Game Experience	
Participants	19
Number Who Play	15
Avg. Hrs Per Day	0.75
Avg. Hrs Per Week	3.05
Flying Experience	
Flying Avg. (Yrs.)	3.05
Flying Range (Yrs.)	0 to 6
Flying σ (Yrs.)	1.4
Flight Log Avg.(Hrs.)	194.79
Flight Log Range (Hrs.)	80.91
Flight Log σ (Hrs.)	75 to 400
FTD Avg. (Hrs.)	18.79
FTD Range (Hrs.)	9.44
FTD σ (Hrs.)	8 to 42
Frasca Avg. (Hrs.)	15.53
Frasca Range (Hrs.)	7.91
Frasca σ (Hrs.)	8 to 42
Certification Level	
Private Pilots License	19
Stage II Instrument Ratings	16
Commercial License	7
Stage V Communication	3
CFI Stage I	1
Other	1

Fatigue Level	
Exceptionally < Normal Rested	1
Moderately < Normal Rested	4
Normally Rested	13
Moderately > Normal Rested	0
Exceptionally > Normal Rested	0
Physical Attributes	
Wearing Contacts	3
NOT Wearing Contact Lenses	4
Don't Require Corrective Sight	11
Unusual Amt of Caffeine Today	0
On Any Medications	0
Unusual Circumstance	1