DOT/FAA/AM-97/5

Office of Aviation Medicine
Washington, D.C. 20591

RECEIVED
MAR 25 1997
ERAU LIBRARY

A Laboratory Model of Readiness-to-Perform Testing. I: Learning Rates and Reliability Analyses for Candidate Testing Measures

Kirby Gilliland
Robert E. Schlegel
The University of Oklahoma
Norman, Oklahoma 73019

February 1997

Final Report

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

Doc
FAA
AM
97
05

U.S. Department of Transportation
Federal Aviation Administration

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-97/5		

4. Title and Subtitle	5. Report Date
A Laboratory Model of Readiness-To-Perform Testing. I: Learning Rates and Reliability Analyses for Candidate Testing Measures	February 1997
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
K. Gilliland, Ph.D. and R.E. Schlegel, Ph.D.	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
The University of Oklahoma 1000 Asp, Room 314 Norman, OK 73019	
	11. Contract or Grant No. DTFA-02-93-D-93088
12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	
	14. Sponsoring Agency Code

15. Supplemental Notes

This work was performed under contract from the Human Resources Research Division of the FAA's Civil Aeromedical Institute (CAMI) as part of task AAM-B-95-HRR-193. Thomas E. Nesthus, Ph.D., served as CAMI's research technical representative.

16. Abstract

The concept of Readiness to Perform (RTP) refers to that state in which a person is prepared and capable of performing a job for which that person is willingly disposed and is free of any transient risk factors, such as drugs, alcohol, fatigue, or illness. A large study involving thirty-two subjects, five RTP tests, two work sample tasks, and three self-report measures was conducted at the University of Oklahoma to provide the FAA with a basic laboratory investigation of the "Readiness-to-Perform" (RTP) testing concept. This report (Volume I) presents the objectives and methodology for the overall study, along with the analysis of data from the initial training phase of the study. Learning rate information for more than seventy task measures used in the study, as well as information on the reliability of those measures, is presented. Based on the learning curve analysis, it was determined that the customary initial high rate of learning was complete for most of the candidate RTP tasks by the tenth training session. A few tasks required additional sessions, but major learning effects for nearly all of these task measures did not exist beyond the middle of the second week of training. Even following five full weeks of experience, most task measures showed some continued improvement, with the rate of improvement determined by the complexity of the task. Test-retest correlation coefficients computed for various intervals throughout the study confirmed that the RTP tasks provided multiple, highly repeatable measures that were both reliable and stable over the latter four weeks of the study. Interpretations of several interesting findings with respect to performance assessment are provided within the framework of job performance and RTP testing.

17. Key Words		18. Distribution Statement
Performance-Based Testing Readiness-To-Perform Reliability Analyses	Fitness-For-Duty Learning Rates Validity Testing	Document is available to the public through the National Technical Information Service Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	70	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PREFACE

This report documents one portion of a larger project completed at the University of Oklahoma under Task Order 93T65153 of contract DTFA-02-93-D-93088 for the Federal Aviation Administration (FAA). Funding for the effort was provided by the FAA Human Resources Research Division, Human Factors Research Laboratory, at the Civil Aeromedical Institute (CAMI). This work is related to previous efforts addressing Readiness-to-Perform Testing under contract DTFA-02-92P23499, which is reported in FAA Office of Aviation Medicine technical report DOT/FAA/AM-93/13.

As outlined in the Statement of Work, a large study was conducted to provide the FAA with a basic laboratory investigation of the "Readiness-to-Perform" (RTP) testing concept. This report presents an analysis of the initial phase of that larger effort. Specifically, this report presents learning rate information for the various candidate measures used in the larger RTP study, as well as information on the reliability of those measures.

Several individuals deserve recognition for their contributions to the project. Randa L. Shehab, Luz-Eugenia Cox-Fuenzalida, Ioannis Vasmatzidis, and Rhonda Swickert served to coordinate the numerous facets of the study. Their contributions to the study design, subject recruitment and retention, data collection, reduction and analysis, and report writing were invaluable. The authors gratefully acknowledge the hardware and software support provided by Scott Mills and the contributions of Arasendran Sellakannu in the collection, reduction, summarization, and analysis of the vast amounts of data. The graduate assistants and the undergraduate support team who worked on this project devoted long hours, often at unusual times of the day, to the collection of the data reported here. Much appreciation is due these CAMI contributors: Dr. Thomas E. Nesthus, who served as the Contracting Officer's Technical Representative (COTR) for the contract and provided his full technical and personal support for the research; Mr. Howard Harris for his contributions to the alcohol testing phase of the study; and Dr. Robert E. Blanchard and Dr. David J. Schroeder for their comments on drafts of this report. The authors also express their appreciation to Dr. James C. Miller for his helpful comments on an earlier draft of this report.

TABLE OF CONTENTS

		Page
1.0	INTRODUCTION	1
2.0	BACKGROUND	1
3.0	OBJECTIVES	3
4.0	METHODOLOGY	3
4.1	Project Design	3
4.2	Subjects	4
4.3	Test Battery	5
4.4	Equipment	7
4.5	Test Facilities	7
4.6	Experimental Procedure	7
4.6.1	Training	8
4.6.2	Simulated Work	8
5.0	RESULTS	11
5.1	Data Reduction	11
5.2	Learning Data Presentation	11
5.3	General Descriptive Statistics	12
5.4	General Performance Improvement	12
5.5	Subjective (Self-Report) Measures	33
5.6	Intertrial Correlations (Test-Retest Reliability and Differential Stability)	35
6.0	DISCUSSION	44
	REFERENCES	48
	APPENDICES	
A.	Performance Measures	A1
B.	Reliability and Differential Stability Coefficients	B1

LIST OF FIGURES

Figure		Page
1.	Laboratory Model of Readiness-to-Perform Testing	4
2.	Spatial Processing (RT Mean and Standard Deviation and Percent Correct)	14
3.	Critical Tracking (Maximum and Mean Lambda)	14
4.	Critical Tracking (Control Losses and RMS Error)	15
5.	Dual Tracking-Group Lambda (Control Losses and RMS Error)	16
6.	Dual Memory Search-Group Lambda (Mean RT and Percent Correct)	17
7.	Dual Memory Search-Group Lambda (Speed and Throughput)	17
8.	Dual Tracking-Individual Lambda (Control Losses and RMS Error)	18
9.	Dual Memory Search-Individual Lambda (Mean RT and Percent Correct)	19
10.	Dual Memory Search-Individual Lambda (Speed and Throughput)	19
11.	Switching-Manikin Task (Mean RT and Percent Correct)	20
12.	Switching-Mathematical Processing (Mean RT and Percent Correct)	20
13.	Switching (Throughput)	21
14.	NovaScan™-Visual Search and Vector Projection Task (Mean RT and % Correct)	22
15.	NovaScan™-Continuous Spatial Memory Task (Mean RT and Percent Correct)	22
16.	NovaScan™-Attention Task (Number Completed and Percent Correct)	23
17.	Air Traffic Scenarios Test (Crashes and Separations)	24
18.	Air Traffic Scenarios Test (Errors)	24
19.	Air Traffic Scenarios Test (Number at Destination, Percent at Destination, Delay)	25
20.	Air Traffic Scenarios Test (Changes in Direction, Altitude, Speed)	26
21.	MATB-Monitoring Task (RT Mean and Standard Deviation)	26
22.	MATB-Monitoring Task (Time-Outs)	27
23.	MATB-Monitoring Task (False Alarms)	28
24.	MATB-Monitoring Task (Key Repeats)	28
25.	MATB-Monitoring Task (Errors)	29
26.	MATB-Communications Task (RT Mean and Standard Deviation)	30
27.	MATB-Communications Task (Error Counts)	30
28.	MATB-Communications Task ("ENTER" Repeats)	31
29.	MATB-Tracking (RMS Error)	32
30.	MATB-Resource Management Task (Mean Tank Level and Deviation)	32
31.	MATB-Resource Management Task (Pump Activity)	33
32.	Physical State Questionnaire (Mean Score)	34
33.	Mood Scale II (Mean Response)	35
34.	Mood Scale II (Mean Response Time)	36

LIST OF TABLES

Table		Page
1.	Summary of Task Codes	7
2.	Task Orders During Training	8
3.	Individualized Lambda Values for Dual Task	10
4.	ATC Scenario and MTB Script Characteristics	10
5.	Means and Standard Deviations for Key Dependent Measures	13
6.	Test-Retest Correlations Over 24-Hour Periods	38
7.	Test-Retest Correlations Over 48-Hour, One-Week, and Two-Week Periods	40
8.	Average Intertrial Correlations for Differential Stability Analysis	41

A Laboratory Model of Readiness-to-Perform Testing

I: Learning Rates and Reliability Analyses for Candidate Testing Measures

1.0 INTRODUCTION

The concept of Readiness to Perform (RTP) [1] defines the "*state* in which a person is prepared and capable of performing a job for which the person is willingly disposed and is free of any transient risk factors, such as drugs, alcohol, fatigue, or illness" (Gilliland and Schlegel, 1993). In general, it is assumed that exposure to risk factors typically results in degraded performance, but it is also possible that performance might be enhanced, at least temporarily, after exposure to some selected risk factors. Readiness-to-Perform *testing* assesses performance capability, typically prior to initiating work activities. When performance capability deviates from some established baseline level, then it is assumed that some risk factor or combination of risk factors are influencing that capability. In this manner, RTP measures have been applied as simple screening devices for risk factor assessment.

In 1993, Gilliland and Schlegel reviewed the concept of RTP and found that the use of RTP measures has rarely been reported beyond a few proprietary studies. Further, the validity for the use of many RTP measures rests primarily on pre-existing literature demonstrating the effects of drugs and stress on human task performance. Noticeably absent are studies investigating the actual reliability and validity of specific RTP tests. As a result, the Federal Aviation Administration (FAA) sponsored a large-scale investigation of selected RTP tests conducted by researchers at the University of Oklahoma. The research approach was to develop a laboratory model of RTP testing. This laboratory model approach provided an opportunity to explore, within a highly controlled environment, some of the fundamental problems associated with RTP testing (see Gilliland and Schlegel, 1993; 1995). The laboratory model approach also provided the ability to explore a number of possible risk factors that could not be introduced in workplace-based research. Likewise, the approach was sufficiently flexible to accommodate the simultaneous evaluation of multiple RTP tests, including a comparison of proprietary RTP tests and laboratory "benchmark" tests. Research of this type is essential to support the validity of RTP testing in a general sense, and to address numerous questions related to the implementation of RTP testing in the workplace. The results of this laboratory model, including the influence of risk factors on RTP tests, are reported in Volume II of this report.

The purpose of this report (Volume I) is to present the analysis of data from the initial training phase of the larger RTP laboratory model study. Specifically, this report presents learning rate information for the various candidate measures used in the larger RTP study, as well as information on the reliability of these measures. Because this report focuses on the initial training phase of the RTP laboratory model study, only information relevant to that portion of the study is reported. More explicit information about the overall design and methods used in later phases of the RTP laboratory model study is presented in Volume II. However, because both volumes of this report bear on the same general issues related to the RTP concept and the RTP laboratory model study, some general information is included in both reports. While the duplication of information was kept to a minimum, certain amounts of background and methodological information are included in both volumes to allow them to stand alone as complete, coherent, and rational presentations of their respective portions of the overall research effort.

2.0 BACKGROUND

The use of RTP testing as a screen for risk factors is based on the assumption that the RTP test used will detect when people fail to perform at their normal (or

[1] The authors have adopted the term "readiness-to-perform" testing for purposes of broader accuracy and clarity. However, it should be noted that readiness-to-perform testing has also been occasionally referred to as "fitness-for-duty" testing, more often in a military context.

usual) performance level. This is typically accomplished by having workers practice extensively on the RTP test. This leads to the establishment of baseline performance on the RTP test for each worker. Then, for each worker, future (often daily) performance assessments are compared to the established performance baseline. When test performance deviates significantly from the established baseline level, it is *assumed* that some risk factor, such as drugs, alcohol, stress, illness, or fatigue, is influencing performance capability. Thus, RTP testing does not identify the specific risk factor that may be present, but rather assesses performance capability at a specific point in time. It is in this manner that RTP measures are often used in business and industry as simple screening devices for risk factor assessment.[2]

The use of RTP testing has advantages over biochemical drug screening techniques. In a past survey of Fortune 1000 companies, 79% of the responding CEOs claimed that substance abuse was a significant problem in their company (Freudenheim, 1988). The use of blood or urine screening for many of these companies is simply too expensive. Behavior-based RTP screening, however, can be considerably less costly. In addition, RTP testing provides immediate results, does not invade the employee's privacy or compromise dignity (see Hamilton, 1991; Maltby, 1990), and seems to be more readily accepted by employers and employees (see Gilliland and Schlegel, 1993, for a more extended review of the advantages and disadvantages of RTP testing).

Unfortunately, very little research has been published on RTP testing and its usefulness. There is, however, some reason to believe that RTP testing would be effective. For many decades, the field of psychology has developed and reported on the use of numerous tasks for assessing an enormous range of human capabilities. With the advent of modern microcomputers, many of these tasks have been programmed for automatic presentation and data collection. Within the last several years, large batteries of human performance tasks have been developed. Some of the more notable of these batteries are the U.S. Air Force Criterion Task Set (Shingledecker, 1984; Schlegel and Gilliland, 1990), the Unified Tri-Services Cognitive Performance Assessment Battery (Hegge, Reeves, Poole, and Thorne, 1985; Perez, Masline, Ramsey, and Urban, 1987; Schlegel and Gilliland, 1992), the AGARD STRES Battery (Santucci, Farmer, Grissett, Wetherell, Boer, Gotters, Schwartz, and Wilson, 1989; Schlegel and Gilliland, 1992), the Walter Reed Performance Assessment Battery (Thorne, Genser, Sing, and Hegge, 1985; Schlegel and Gilliland, 1992), and the U.S. Navy PETER Battery (Bittner, Carter, Kennedy, Harbeson, and Krause, 1986). These batteries have made it possible to test a wide range of abilities fairly rapidly, with considerable accuracy, and often with vast data storage capability. These tasks have been used to measure performance, to screen personnel, and as metrics for assessing the influence of such factors as drugs, stress, and fatigue on performance.

In a critique of RTP assessment, Gilliland and Schlegel (1993) reviewed many of these batteries and noted their role as precursors to many of the RTP tests now available. In fact, nearly every behavioral RTP measure appears to have had its origin in prior computer-based tasks. At the same time, there is a large amount of research literature exploring the effects of such risk factors as drugs, alcohol, stress, and fatigue on human abilities. This literature was also briefly reviewed by Gilliland and Schlegel (1993) and is important for two reasons. First, this literature establishes a relationship between task performance and the influence of risk factors. It appears that most risk factors have fairly pronounced effects on a wide range of abilities and certainly on the performance of a broad range of tasks, many of which are included in the task batteries mentioned above. Second, this literature provides important insights into which tasks will be more or less sensitive to the influence of specific risk factors.

It is the combination of advances in task battery development and the increasing knowledge that risk factors do affect performance on such tasks that has provided the impetus for RTP development. However, the application of Readiness-to-Perform testing in the workplace to aid in the process of risk factor screening has developed so rapidly that many of the critical linkages between laboratory task performance and RTP measures, as used in the field, have not been established. These critical links, often the basis for substantiating validity, would only reside in well-constructed research on the nature of the RTP concept and the function of specific RTP measures. In addition, there are numerous unanswered questions regarding basic issues in implementing RTP testing.

In their theoretical analysis, Gilliland and Schlegel (1993) outlined several basic problems related to RTP testing for which there is little or no research. For example, very little data have been presented with regard to reliability estimates for specific RTP tests.

[2] It should be noted that RTP testing is theoretically capable of detecting both decreases and increases in performance, as compared to baseline performance. Some risk factors, such as stimulant drugs, are known to increase performance ability.

Very little proprietary research has been released regarding the predictive validity of specific RTP measures, and even less archival literature on RTP exists. Thus, it is not clear how well RTP tests actually perform in risk factor detection in the workplace. There are also numerous questions regarding the selection, implementation, and use of RTP tests. For instance, can an RTP test that has predictive validity only for risk factors, and not for job performance, be an effective RTP test? This question is fundamental to the entire RTP approach and sets the stage for addressing the following questions: Is a personalized baseline, or some combination of personalized baseline and group performance, the most effective standard for assessing an individual's RTP performance? Can a single RTP test detect more than one risk factor? Is daily RTP testing required, or can testing be performed intermittently? Should RTP testing be performed only once per shift, or more often?

The list of unanswered questions is extensive, and the void of knowledge is even more overwhelming if one considers that each of these questions ought to be answered for each RTP test selected for use. Yet, there are ways to address a number of these questions within a more general framework that will provide relevant and important information for RTP testing in general. The current RTP laboratory model study was designed to address RTP testing using that general framework to more efficiently explore fundamental questions of the RTP concept and testing approaches. The following section briefly describes the objectives of the overall study, including the objectives of the initial phase of the study reported in this volume.

3.0 OBJECTIVES

The main objective of the overall project was to provide the FAA with a large-scale, highly controlled, laboratory investigation addressing the use of "Readiness-to-Perform" (RTP) testing. Two major issues were addressed. First, the basic integrity of the model was assessed. It was essential to examine the effectiveness of the model because the quality of the research rests on the soundness of the model. Included in this basic model assessment were:

(1) investigations of the number of sessions needed to bring subjects to an asymptotic performance level on the RTP and simulated work tasks,
(2) examinations of RTP test and job task reliability,
(3) examination of the relationships among RTP test performance and job performance, and
(4) examination of the relationships among variations in RTP test performance and variations in job performance.

Because the establishment of validity is so central to the integrity of RTP testing, the second major research issue was validity. This issue involved:

(1) determination of the ability of each of the RTP tests to identify the presence of risk factors (sleep loss, alcohol, and antihistamine), and
(2) in a broader sense, investigation of the relationships between common risk factors and both RTP test performance and job performance.

The specific risk factors investigated were sleep loss (30 hours), low doses of alcohol (.03% to .05% breath alcohol level), and common antihistamine (4 mg. chlorpheniramine maleate). The risk factor validation component of the study was not an attempt to extend or duplicate other in-depth efforts of the FAA to validate RTP testing with regard to the influence of alcohol. Rather, this study attempted validation (in a more basic manner) concentrating on a range of risk factors and demonstrating at a very fundamental level whether these risk factors influenced the selected RTP tests. Without such confirmation, the model would lack both integrity and validity. Another feature that distinguishes this research effort is a concentration both on identifying valid RTP <u>tests</u> and valid RTP <u>measures</u>. Any RTP test may have several dependent measures, any one of which might be a potentially valid RTP measure. This study explored not only the typical response time and accuracy measures but additional dependent measures that were reasonably derived from the RTP tests.

This volume of the overall report focuses on the first two topics (above) that examine the effectiveness of the model, specifically the number of sessions needed to bring subjects to an asymptotic performance level on the RTP tests and simulated work tasks (or learning rate information), and examinations of both candidate RTP test and job task reliability.

4.0 METHODOLOGY

4.1 Project Design

As mentioned above, the approach adopted in this research was the construction of a laboratory model of RTP testing. The basic model for RTP testing is presented in its simplest form in Figure 1. The overall study consisted of four main stages. The first stage involved subject screening, pre-testing, and selection. During the second stage, subjects underwent orientation and training on RTP tests and simulated work (job) tasks. In this stage, the individualized comparative baselines for later RTP test comparisons were constructed.

The third stage provided a stable Simulated Work Period. Subjects were impressed with the view that they were being hired to "work at a job" each day in much the same manner as any typical worker. They arrived at the lab, were administered the RTP tests, and performed their "job." The fourth stage of the study provided Specialized Investigation Periods. This stage was actually conducted on weekends during the Simulated Work Period. Testing periods varied according to the requirements of the specific research protocol. Testing activity during this stage included examinations of the risk factors of sleep loss, alcohol, and antihistamines. All testing during this stage was designed to minimize any influence on daily testing sessions during the week. For example, sleep loss test sessions were conducted on Friday night following daily testing sessions and were completed by Saturday afternoon, allowing subjects a day and a half (two nights sleep) to recover before daily testing sessions resumed.

This report focuses primarily on the establishment of stable performance levels during the initial orientation and training of the subjects. For that reason, much of the cited data are related to Stage Two and the early part of Stage Three. However, in many cases complete data for all sessions throughout Stages Two and Three are reported.

4.2 Subjects

Thirty-two subjects participated in this study. Subjects were recruited from University of Oklahoma psychology and engineering classes, the general student body, and the Norman, Oklahoma regional community. Because of the possible adverse effects on pregnancy of risk factors such as alcohol and antihistamines, all subjects were male. They ranged in age from 21 to 43 years with a mean of 25.2 and a standard deviation of 5.5 years. All subjects signed an Informed Consent Form approved by the University of Oklahoma Institutional Review Board—Norman Campus.

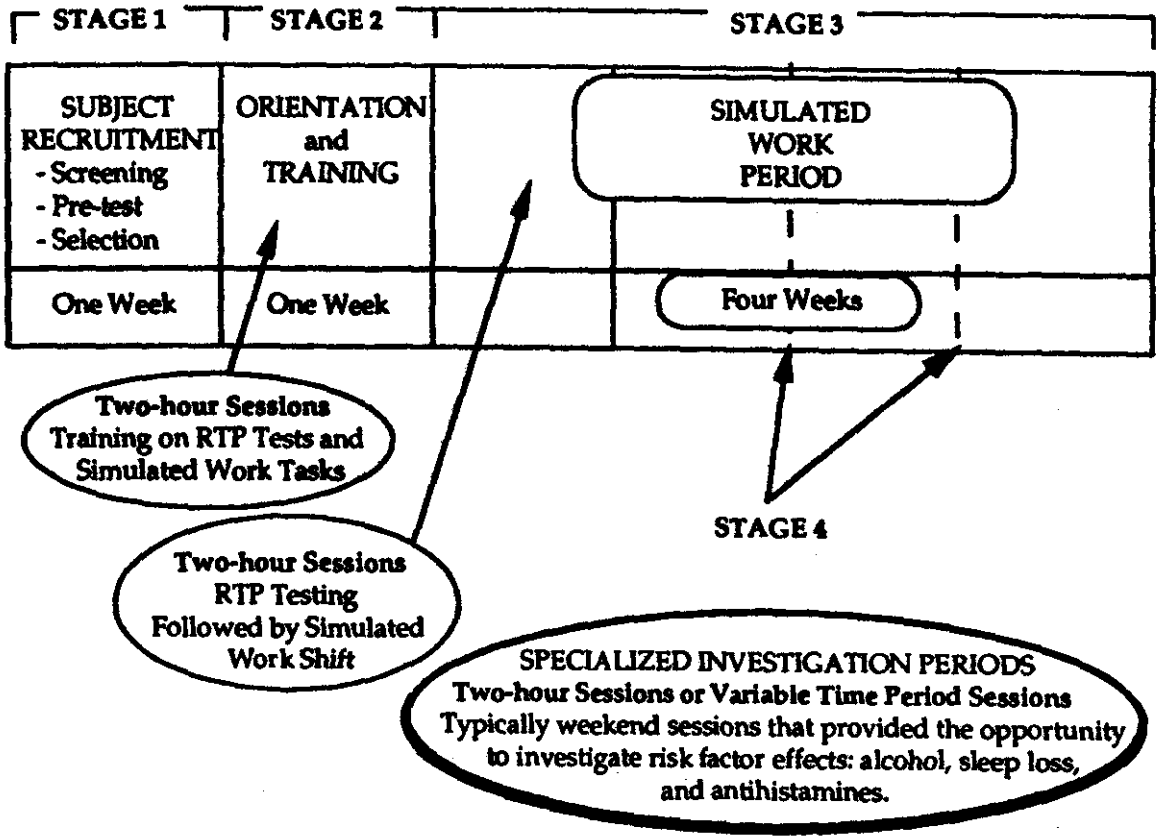

Figure 1. Laboratory Model of Readiness-to-Perform Testing.

Because data collection extended across five weeks and required participation on two weekends, a bonus payment system was used to increase motivation and study completion rate. Subjects were paid a base rate for approximately 64 hours of testing. Upon completion of the study, subjects were given an additional bonus for every hour of participation.

Two subjects were dropped during the first week of training for lack of schedule compliance. However, they were immediately replaced with alternate subjects. Three other subjects did not complete the experiment as scheduled. One subject dropped after three weeks (Session 20) due to personal problems. The other two subjects were dropped for lack of schedule compliance with one subject completing through Session 19, while the other completed through Session 20.

All subjects were surveyed for self-reported normal (or corrected-to-normal) vision, normal hearing, and the absence of any central nervous system stimulant or depressant medications. Due to the nature of the risk factors, additional relevant information about alcohol, caffeine, medication, and possible drug use was obtained. On average, subjects consumed 3.5 alcoholic beverages per week, although individual consumption ranged from 0 to 18 beverages per week. Subjects reported that, on average, they drank two or three times a month, with beer as the primary alcoholic beverage. Caffeine consumption was relatively low. Average coffee consumption was 1.7 cups per day. The average coffee consumption was skewed by one subject who reportedly drank two to three 12-cup pots of coffee daily. Average cola consumption was 1.3 cans per day. Only two subjects reported regular use of medication. Well into the testing regimen, one subject did indicate that he was a problem drinker and occasionally used drugs with alcohol.

4.3 Test Battery

Prudent selection of the candidate RTP tests was viewed as critical to the success of the project. The candidate RTP tests used in this study were selected on a rational basis, as outlined below. First, an extensive effort was made to selectively review the literature on task batteries and the influence of risk factors on human performance. This activity was essential because most of the readily available evidence of task sensitivity to risk factors is found in these two areas. Much of the basic review had been conducted (Gilliland and Schlegel, 1992). However, additional selective reviews focused specifically on identifying tasks that appeared sensitive to risk factor effects and would therefore serve as promising RTP test candidates.

Several factors were then considered in selecting the RTP tests for this study. One of the most important involved the offerings and logic provided by various RTP vendors based on their determination of effective tests. Another important factor was whether the test reflected the specific information processing skills used in typical safety-sensitive jobs—that is, jobs such as aircraft piloting or air traffic control in which the safety of the worker, co-workers, and others relies on prompt and correct decisions and actions. Identification of these skills affects task selection in terms of the information provided by a specific task regarding the cognitive processes or information processing stages affected by one or more risk factors.

A number of other important factors were also considered in selecting the candidate RTP tests. For example, each of the selected tasks had some evidence of being sensitive to risk factor effects (see the above literature review description). The tasks varied in the cognitive resource or ability needed to perform them (e.g., psychomotor ability, spatial ability, memory, or attention allocation). The tasks also varied in complexity. Some were simple human performance tasks common to the psychological literature (e.g., tracking tasks and spatial processing task). Other tasks were of intermediate difficulty (e.g., dual task) And, some tasks were quite complex and challenging (e.g., Switching task and NovaScan™ task). Some tasks were selected because they appeared to assess basic abilities and resources, while others were selected because they appeared to be related to common job requirements (e.g., memory plus divided attention). Thus, certain candidate RTP tests appeared to have *a priori* predictive validity for risk factors alone, and others had *a priori* predictive validity for both risk factors and job performance.

As noted previously, commercial RTP tests are either directly or indirectly related to tasks in existing human performance task batteries. For that reason, some of the candidate RTP measures were selected from existing performance assessment batteries. In addition, the commercial RTP test recently acquired by the FAA was included among the candidate RTP tests in this study, as was another commercially-related RTP task. Inclusion of these measures provided the opportunity to further explore their validity and to compare their effectiveness to the "benchmark" candidate RTP tests drawn from the various human performance assessment batteries. Five performance tasks were selected as candidate RTP tests.

The simulated work tasks selected for this study had to be simple enough to be learned within a few training sessions, yet complex enough to provide a challenge to the subjects and some degree of intrinsic

motivation. The degree to which these tasks provided some degree of similarity to tasks of importance in the aeronautical environment was also given consideration. The simulated work tasks were a low-fidelity air traffic control simulation, named the Air Traffic Scenarios Test (Aerospace Sciences, 1991; Broach and Brecht-Clark, 1994; Gilliland and Schlegel, 1992; Weltin, Broach, Goldbach, and O'Donnell, 1992), and the Multi-Attribute Task Battery (MATB) developed at the NASA Langley Research Center (Comstock and Arnegard, 1992). The MATB includes (1) a monitoring task that consists of both a set of response time stimuli and a set of probability monitoring dials, (2) a communications task, (3) a compensatory tracking task, and (4) a resource management task that simulates a complex fuel tank management task. Because this multi-faceted task was designed to approximate the aircrew operations environment, this task brings an added degree of ecological validity to the study.

Brief descriptions of the RTP tests, work samples, and subjective rating scales are provided below.

RTP Tests

Spatial Processing (SPA) - involves indicating whether a rotated pattern of histograms is the same as one previously presented. The test lasts three minutes.

Critical Tracking (TRK) - involves tracking an unstable object along a single axis on the display using a trackball for two minutes.

Dual Task (DUL) - involves performing the Sternberg Memory Search while Tracking. The Sternberg Memory Search involves indicating whether a letter is the same as one of those in a previously memorized set. The test lasts three minutes.

IML Switching Task (NTI) - involves responding to one of two tasks presented simultaneously on each screen display. In the Manikin task, the subject presses a key to indicate which hand of a manikin holds a matching symbol. In Mathematical Processing, the subject presses a key to indicate whether a sum of three numbers is greater or less than five. The test lasts four minutes.

NovaScan™ FAA Task (NSF) - involves integrated responses to three tasks. For two of the tasks, stimulus screens are presented in directed attention fashion with a series of stimuli from one task alternating with a series of stimuli from the other task. In addition, a vigilance/attention task is performed for every stimulus screen. In the Visual Search and Vector Projection task, the subject searches for two labeled vectors, makes mental rotations of the vectors based on verbal on-screen instructions, and responds as to whether the rotated vectors would ever intersect after mentally projecting to infinity. In the Spatial Memory task, the subject memorizes the position and shape of a missing symbol for later comparison with the next spatial memory stimulus screen. For the Attention task, subjects look for the presence of small symbols in the corners of each screen. The test is based on a fixed number of stimuli, and test time is thus a function of subject proficiency.

Work Sample Tasks

Air Traffic Scenarios Test (ATC) - an approximation of the air traffic control environment which involves the directing of planes to their destinations using altitude, speed, and heading changes. The work version of the task lasts 25 minutes.

Multi-Attribute Task Battery (MTB) - an approximation of the air crew operations environment, which includes a monitoring task (a set of lights and a set of dials), an auditory communications task, a compensatory tracking task, and a resource management task involving the monitoring and control of fuel tank levels. The work version of this task lasts 40 minutes.

Subjective (Self-Report) Measures

Activity State Questionnaire (ASK) - an expanded form of the Pennebaker Physical Symptoms Checklist to assess the current state of physical health. The questionnaire consists of responding to 25 items scored on a seven-point scale. Subjects also responded to two questions regarding their level of preparedness for task performance. The test takes approximately two minutes.

Mood Scale II (MOO) - involves pressing a numbered key to indicate the level of agreement with each of 36 descriptive adjectives to assess the current mood in the categories of activity, happiness, depression, anger, fatigue and fear. The test takes approximately two minutes.

NASA Task Load Index (TLX) - ratings of task workload using the categories of mental, physical, temporal, performance, effort, and frustration. This collection of ratings was obtained following each work sample task. Providing the ratings takes approximately one minute. (Note: While the NASA TLX ratings were among the subjective ratings administered, they were completed by the subjects after each of the simulated work tasks and, thus, were linked logically to the workload generated by these tasks. The use of these TLX ratings cannot be used as a reflection of overall workload experienced by the subject during the entire testing session. Therefore, TLX ratings are included in Volume II of this report, which addresses, among other issues, relationships among the various measures.)

Table 1. Summary of Task Codes.

Task	Code
Spatial Processing	SPA
Critical Tracking	TRK
Dual Task (Group Lambda)	DULG
Dual Task (Individual Lambda)	DULI
IML Attention Switching	NTI
NovaScan™ FAA	NSF
Air Traffic Scenarios Test	ATC
Multi-Attribute Task Battery	MTB
Activity State Questionnaire	ASK
Mood Scale II	MOO
NASA Task Load Index (ATC)	TLX

Table 1 presents a summary of the task codes used throughout the remainder of the report when referring to the various tasks.

4.4 Equipment

All tasks were presented on eight microcomputer workstations. Each workstation consisted of a Gateway 486-33 MHz processor with the necessary input devices ("Anykey" keyboard, Microsoft mouse, Kensington Expert Mouse 4.0 trackball, CH Products Flightstick, and NovaScan™ interface box). All data were recorded on these machines and on subject diskettes and then downloaded to a central data management system (Gateway 486-66 MHz) for data reduction and analysis using the Statistical Analysis System (SAS) and Microsoft Excel. In cases of emergency, this machine also served as a backup workstation. A Macintosh Quadra 950 was used almost exclusively for graphics and desktop publishing using Microsoft Excel and Microsoft Word. The tremendous volume of data generated in large-scale, multi-day studies such as this places large demands on data reduction and graphing capabilities. In addition, report preparation required the full-time dedication of this system. Testing was automated to allow a subject to perform the tests independently and in the minimal amount of time. Of course, multiple experimenters were present at all times to monitor the subject's safety and performance, and provide assistance, if needed. The software automatically performed all housekeeping functions, such as subject identification, file naming, test sequencing, and data backup.

4.5 Test Facilities

All testing was conducted in a quiet laboratory space located in the basement of Dale Hall at the University of Oklahoma. The testing workstations were approximately 3 ft. wide and 3 ft. deep and were located in one room (approximately 13 ft. by 20 ft.). The stations were divided by 3-in. thick acoustic panels to minimize distractions. The computers and response devices were placed on tables at the testing stations positioned at a height of approximately 28 in.

Another room of approximately the same size served as the data reduction and project management office. A third room served as an auxiliary room for interviewing, orientation, and miscellaneous activities. All of these rooms represent modern laboratory space with centrally controlled heating and air conditioning. Temperature in the testing room was maintained at approximately 68° F throughout the sessions.

4.6 Experimental Procedure

Data were collected from subjects over a five-week period. Subjects met for an initial one-hour orientation, during which they completed consent forms and questionnaires and were provided with a study review packet. Task demonstrations were also provided during the orientation to familiarize subjects with the tasks prior to training. All subjects were scheduled for one two-hour session per day, five days each week. In addition, subjects were asked to reserve two specific weekends for the risk factor studies.

Table 2. Task Orders During Training.

Session									
Monday		Tuesday		Wednesday		Thursday		Friday	
1	2	3	4	5	6	7	8	9	10
SPA	ATC	SPA	SPA	SPA	SPA	SPA	SPA	SPA	SPA
SPA	MTB	TRK	TRK	TRK	TRK	TRK	TRK	TRK	TRK
NTI		NSF	NSF	NSF	NSF	NSF	NSF	NSF	NSF
NTI		DUL	DUL	DUL	DUL	DUL	DUL	DUL	DUL
TRK		NTI	NTI	NTI	NTI	NTI	NTI	NTI	NTI
TRK			Break		Break		Break		Break
NSF-TRAIN			ATC		ATC		ATC		ATC
DUL			ATC		ATC		MTB		MTB
DUL			MTB		MTB		MTB		MTB
NSF			MTB		MTB		MTB		MTB
					MTB				

4.6.1 Training

Training began the Monday following orientation and continued throughout the first week. Session numbers 1 through 10 were designated for training, with two sessions completed each day. Because some tasks were complex in nature and more difficult to learn, each task required a different training schedule. Therefore, to ensure optimal training on each task, the tasks presented in the different sessions varied. Table 2 summarizes the task order during the first week. Session 1 required subjects to perform two trials each of Spatial Processing (SPA), Attention Switching (NTI), and Critical Tracking (TRK), followed by the training version of NovaScan™ FAA (NSF), two trials of Dual Task (DUL), and finally the testing version of NSF. For Session 1 only, all tasks were presented with instructions. Session 2 introduced subjects to abbreviated versions of the Air Traffic Scenarios Test (ATC) and Multi-Attribute Task Battery (MTB). On each of the remaining days of training, the first sessions (i.e., Sessions 3, 5, 7, and 9) were identical and contained one trial each of SPA, TRK, NSF, DUL, and NTI. Session 4 duplicated Session 3 but added two abbreviated trials each of ATC and MTB. On the third day of training, subjects completed Sessions 5 and 6. Session 6 duplicated Session 5 but added two shortened trials of ATC and three shortened trials of MTB. On the last session of the final two days of training (Sessions 8 and 10, respectively), subjects completed a standard length trial of ATC, along with three abbreviated trials of MTB.

The training scheme described above was sufficient for most subjects to achieve acceptable levels of performance. However, two of the tasks were problematic for some subjects. Five subjects required additional explanations regarding the Manikin portion of the NTI task. These subjects were allowed to perform an additional trial of the NTI task. This typically occurred following Session 1. The NovaScan™ FAA task was problematic for seven subjects. Two of the seven were provided additional training trials on the vector component of the task after Session 1. In addition, these seven subjects were provided some level of additional training after Session 4. Five of the subjects were given additional training on both portions of the NovaScan™ FAA task, while another was given training on only the vector task and the other on only the spatial memory task. With the additional training trials, these subjects were able to provide acceptable performance levels.

The interval between tests was subject-determined, that is, the tests did not start automatically. Subjects were required to press a key to start the next task. This allowed an opportunity for the subjects to ask questions and receive feedback. Summary feedback was provided at the end of each task during all sessions. A minimum break of three minutes was enforced between the set of RTP tests and the work sample tasks.

4.6.2 Simulated Work

After the initial week of training, subjects provided four additional weeks of working data (Sessions 11

through 30). One two-hour session was performed each day. At the start of each session, each of the RTP tests was performed, followed by complete trials of ATC and MTB. In addition, subjective scales were added to the battery starting with Session 11. The Mood Scale II (MOO) and the Physical Symptoms (Activity) State Questionnaire (ASK) were both performed prior to the RTP tasks. Subjective workload assessment using the NASA Task Load Index (TLX) was conducted after ATC and MTB.

Risk factor investigations were conducted on the weekends following Sessions 20 and 25. On each of these weekends, all subjects participated and were divided into two test groups (sleep loss or alcohol) based on subject availability. On the second weekend, as counterbalanced, the subjects were tested on the remaining risk factor. Session numbers 31 through 33 were designated for sleep-loss testing, and session number 34 was designated for alcohol testing, regardless of the counterbalanced order.

All RTP test parameters remained fixed after Session 11, with the exception of DUL. The difficulty parameter (lambda) of the tracking portion of the DUL task was initially set to 2.0. This value represented a relatively low level of difficulty at which almost all subjects were able to attain perfect performance with regard to control losses (a score of 0). To make the task more sensitive to variations in subject ability, the lambda parameter was increased following Session 12, and two variations of the task were performed by all subjects in subsequent sessions. One variation (DULI) used an individualized lambda value set to 70% of the average of the subject's maximum lambda values for Sessions 7 through 10 of the TRK task. The other variation (DULG) used a group lambda value that was established as the average of all subjects' individualized lambda values. The group lambda value was set at 3.7. Table 3 presents the individualized lambda values implemented in Session 13 for each subject.

From Session 11 on, the RTP test order varied but was balanced across subjects and across days within subjects, as was the order of ATC and MTB. In addition, there were six alternate scenarios for the ATC task and five alternate scripts for the MTB task. These were also balanced across subjects and across days within subjects. Characteristics of the ATC scenarios and MTB scripts for training and work are summarized in Table 4.

The various RTP test orders were developed to minimize interference between consecutive tasks (e.g., hand fatigue from consecutive TRK and DUL trials). Sessions 11 and 12 used eight different orders. For each subject, the RTP test orders were randomly assigned such that a balance of orders was obtained across all 32 subjects (4 subjects for each of 8 orders). For a given subject, different test orders were assigned for Sessions 11 and 12. Sessions 13 through 30 used six test orders. These orders were randomly used within each subject in three blocks of six (18 sessions) such that each block contained a complete set of the six orders. Sessions associated with risk factors (Sessions 31 through 34) used four unique orders randomly assigned within each subject, using each of the four orders once (four sessions). For a given subject, the same order was never presented on consecutive sessions.

The order of the ATC and MTB tasks was alternated, with either ATC performed before MTB (order 1) or vice versa (order 2). For Session 11, the order was randomly assigned but balanced across subjects. For Session 12, the opposite order was used for each subject. For Sessions 13 through 32, the orders were blocked in sets of four sessions. The order for the first session was randomly selected and this specified the order of the remaining three sessions in the block. For example, if the first session was order 1, then the order of the four sessions in the block was 1-2-2-1. On the other hand, if the first session was order 2, then the order of the four sessions in the block was 2-1-1-2. Sessions 33 and 34 used randomly assigned orders balanced across subjects and across sessions within subjects.

For the ATC task, two different scenarios were used by all subjects for Sessions 11 and 12. For Sessions 13 through 30, six different scenarios were used. Within each block of six consecutive sessions, a random ordering of the six scenarios was developed for each subject. Thus, each block of six sessions contained a complete set of the six scenarios. Within the randomization, there was a restriction prohibiting the assignment of the same scenario to two consecutive sessions for the same subject. Also, for a given a subject, Sessions 18, 23, 33, and 34 all used the same scenario to enable a baseline comparison with the risk factors. The assignment of the six available scenarios for use by a particular subject as that subject's baseline scenario was balanced across subjects. Sessions 31 and 32 used randomly selected scenarios that were different from the baseline scenario for that subject.

There were five unique scripts for the MTB task for Sessions 11 through 30. As with ATC, the order of the scripts varied randomly for a given subject, such that each block of five sessions contained the complete set of five scripts with the restriction against consecutive sessions using the same script. Once again, a given subject used the same script for Sessions 18, 23, 33, and 34 to enable baseline-treatment comparisons. The scripts used in Sessions 31 and 32 were randomly selected from the group of five and differed from the assigned baseline script.

Table 3. Individualized Lambda Values for Dual Task.

ID	Lambda	ID	Lambda
201	4.0	219	3.3
202	3.7	220	3.7
204	3.6	221	3.8
205	3.7	222	3.7
206	3.9	223	3.4
207	2.9	224	3.3
208	3.3	225	3.7
209	3.8	226	4.2
210	3.8	227	3.4
211	3.9	228	3.6
212	4.2	229	3.2
213	4.1	230	3.6
215	3.7	231	3.4
216	4.7	232	3.3
217	3.9	233	3.7
218	4.0	234	3.6

Table 4. ATC Scenario and MTB Script Characteristics.

	ATC			MTB	
Session	Scenario	Planes (#)	Length (sec)	Script	Length (min)
1	S1	3	300		
2	S2	5	333	GRAPE (A)	10
3	S3	12	460		
4	S4	16	750	GRAPE (B and A)	10
5	S5	16	750		
6	S4	16	750	RTPXX (E, A, and D)	10
7	S5	16	750		
8	S6	43	1500	RTPXX (A, B, and E)	10
10	S7	45	1500	RTPXX (D, B, and C)	10
11	RTPA	40	1500	RTP40 (A, B, C, D, or E)	40
12	RTPB	40	1560	RTP40 (A, B, C, D, or E)	40
13-34	RTP1-6	45	1500	RTP40 (A, B, C, D, or E)	40

5.0 RESULTS

This section of the report presents a discussion of the training and baseline data from the study. While the reported analyses emphasize initial training and baseline sessions (especially Sessions 1 through 10), many of the discussions and figures present data through Session 30. These additional data are included simply because it could be argued that the Work Simulation Stage represents nothing more than an extended set of baseline sessions, that is, sessions conducted under standard laboratory conditions without the introduction of any risk factors. In addition, it may be important for some researchers to examine the longer-term stability of such task variables.

5.1 Data Reduction

This project involved the collection of a massive database. Only a portion of those data are summarized within this report. Appendix A presents a list of more than 150 performance measures and the codes used to represent them in the SAS databases and analyses. Approximately 13,275 data observations (Subjects x Sessions x Tasks), each containing numerous dependent measures, were collected over the course of more than 1300 subject sessions. It is noteworthy that of the 13,275 observations, fewer than 50 were lost due to equipment or procedural errors, and the majority of the losses occurred during the first week of training. Very few outlier data points were removed prior to the summaries and analyses. The deleted observations resulted from identifiable hardware, software, or subject errors. In instances where subjects inadvertently reversed response keys for an entire trial, the raw data files were rescored to provide correct summary information.

The procedure for data reduction involved several phases. Raw and summary data files from the individual subject PC diskettes and workstation hard drives were transferred to the Gateway 486/66 MHz data management computer. SAS DATA step input programs were used to extract the data from the summary files and to create individual SAS databases for each task. The SAS UNIVARIATE procedure was used to provide extensive descriptive statistics for each dependent variable. These analyses were reviewed for questionable data points that could be the result of procedural errors or data outliers. Data points in question were corrected where possible and removed when necessary (see paragraph above). The next step in data reduction involved computing summary statistics across all subjects to aid in evaluating the average performance pattern across sessions for each task measure.

The next stage in the data analysis involved editing and reduction of the data base. Each time the tasks in this study were collectively administered they generated over 150 measures. Many of these measures were highly correlated with one another; others were of minor value in assessing performance on the given task. After reviewing many of these dependent variables, it was determined that the major analyses for this study would focus on a reduced subset of principal performance measures for each task. This reduced subset of variables contained the major performance measures that have traditionally been used to assess performance for each of the tasks (e.g., reaction time, percent correct, RMS error). The reduced subset of dependent measures is listed in Table 5 under Section 5.3, General Descriptive Statistics.

5.2 Learning Data Presentation

Selecting a method for presenting learning rate data is a daunting process at best. In fact, the learning curve has been a topic of interest for decades in experimental psychology (e.g., Barlow, 1928; Gulliksen, 1934; Thurstone, 1919). Learning rate presentations typically provide figures of the trial-by-trial data for visual inspection, along with accompanying tables of standard descriptive statistics. The identification of such characteristics as asymptotic performance level and stable baseline periods is often based on visual analysis or, in some cases, by the application of curve-fitting procedures.

Certainly, sophisticated methods for deriving learning curve parameters have been suggested (see Mazur and Hastie, 1978; Restle and Greeno, 1970; Spears, 1985). Many excellent examples exist of mathematically sophisticated curve-fitting comparisons that provide examinations of the degree to which various exponential equations fit various data sets (e.g., Gulliksen, 1934; Mazur and Hastie, 1978). However, there is considerable controversy over which method of curve parameter estimation is best. Some have even suggested that many psychologists have simply ceased to be concerned with learning curve shapes (Mazur and Hastie, 1978).

What seems to be ignored in this debate is that no method is probably adequate for all or even most cases, and that the method for deriving such parameters is probably best determined on a case-by-case basis, given a number of scientific and pragmatic considerations. In this regard, the "level of analysis" seems to be an important issue. Researchers vary in their needs for precision. Researchers in highly specialized areas of psychophysics or in areas of learning model comparisons often work with tasks that have well-defined, highly stable learning curve characteristics. In

such cases, these researchers need to apply highly refined exponential equations to detect very small differences in learning curve parameters. In contrast, researchers in areas of computerized task assessment and applied human factors are more often concerned with identifying in general terms when subjects have completed rudimentary learning processes, recognizing that more refined learning processes for performance tasks may continue for some time. This need for less refined estimates of curve parameters, as well as general disagreement among more sophisticated curve-fitting techniques, has supported the more frequent use of simple descriptive techniques based on visual analysis.

Another issue that mediates the decisions of how to derive curve parameters is the recognition that more sophisticated curve-fitting procedures are typically best accomplished when knowledge of the data structure (that is, some theory of the learning process underlying the data) is linked to the selection of the specific exponential equation that is being applied to the data. In simpler terms, learning curves vary from smooth positively accelerating to negatively decelerating to S-shaped forms. No simple mathematical curve-fitting procedure can be applied to all these cases with equal effectiveness. For these reasons, this report will concentrate on an extensive descriptive statistical analysis and will include a fairly standard approach based on visual analyses.

5.3 General Descriptive Statistics

An extensive set of descriptive statistics was generated for each dependent variable across all subjects for each session. Over 1400 tables of descriptive statistics and box plots (by session, for over 50 dependent variables) were then reviewed to examine the integrity of the dataset. Of particular concern were sessions where individual subjects were shown to perform significantly out-of-range in comparison with other subjects. Such sessions usually suggested equipment or procedural problems. For example, on a few occasions a trackball would suffer an intermittent failure due to loose connectors, software problems would lead to intermittent failure of joysticks, and subjects would occasionally fail to hit the appropriate response keys. When such cases were discovered, daily logs of equipment and procedural problems were consulted. When it could be confirmed that such aberrant data were due to equipment or procedural problems, those data were eliminated from the analysis (which included fewer than 50 individual subject's trials out of over 13,000 as noted above). Of course, there were cases where subjects performed out-of-range with no apparent explanation. These constituted a fairly small percentage of the sessions, and there was no reason to believe that these cases were distributed in a non-random fashion. When there was no explanation for such variation, the cases were left in the analyses and were assumed to represent normal variation among subjects.

Table 5 presents means and standard deviations for a sample of dependent measures for each task, across a sample of testing sessions (Sessions 1, 10, 20, and 30). These dependent variables represent those most commonly used by researchers but are not exhaustive.

Table 5 presents the performance of subjects at the beginning (Session 1) and, in general terms, the end of the training period (Session 10), and at two points in the Work Simulation period where stable performance would be expected (Sessions 20 and 30). It is presumed that much, if not all, of the typical learning curve effect would be absent from these latter data. Thus, these data would provide very good estimates of stable baseline data on these measures for subjects such as those used in this study. Tables that provide means and standard deviations across all sessions for these variables are available from the authors, along with a representative sample of the SAS UNIVARIATE descriptive statistics tables and box plots for a select group of dependent variables.

5.4 General Performance Improvement

To examine the pattern of learning or skill acquisition for the various tasks, data for all sessions were summarized graphically. Task learning is indicated by faster response times, higher accuracy and throughput, and fewer control losses over time. In general, performance improved rapidly over the first three to five sessions. The rate of improvement leveled off by the eighth to tenth session for those tasks without parameter changes during that period. Detailed differences, as a function of task, are presented in the discussions that follow. Note that the graphs of most performance measures begin with Session 1, but other measures begin with sessions ranging from 2 through 13. As explained in Section 4.6 in more detail, this is because not all tasks or questionnaires were introduced at the first training session (Session 1) and because the final implementation of some tasks depended on establishing baseline performance criteria on similar tasks (e.g., tracking based on group and individual lambda values).

Table 5. Means and Standard Deviations for Key Dependent Measures.

Task	Measure	Session 1 Mean	Session 1 Std. Dev.	Session 10 Mean	Session 10 Std. Dev.	Session 20 Mean	Session 20 Std. Dev.	Session 30 Mean	Session 30 Std. Dev.
Spatial Processing	MNCORRT	1580	(349)	1176	(344)	1015	(270)	919	(272)
	SDCORRT	632	(261)	402	(218)	404	(276)	320	(168)
	PC	82	(14)	87	(8)	95	(6)	90	(5)
Critical Tracking	MAXL	4.3	(90.0)	5.4	(0.6)	5.7	(0.7)	6.1	(0.5)
	MEANL	3.1	(0.7)	4.5	(0.6)	4.8	(0.7)	5.2	(0.6)
	CTLOSS	17.0	(4.6)	11.1	(1.7)	10.3	(1.9)	9.3	(1.3)
	RMS	58.0	(6.1)	54.7	(5.6)	54.5	(9.1)	54.4	(7.7)
Dual Task (Group)	CTLOSS	11.7	(17.7)	0.4	(1.0)	11.5	(14.3)	8.1	(14.3)
	RMS	44.5	(21.7)	22.4	(15.2)	53.4	(15.3)	46.8	(16.2)
	MNCORRT	989	(373)	695	(174)	676	(142)	638	(173)
	PC	91	(20)	99	(2)	96	(6)	98	(2)
	SPEED	65.3	(21.3)	90.8	(19.3)	92.4	(19.0)	100.0	(24.6)
	THRUPUT	59.9	(25.6)	89.5	(19.2)	88.9	(20.2)	98.4	(24.1)
Dual Task (Individual)	CTLOSS	-	-	-	-	10.0	(11.7)	5.3	(8.7)
	RMS	-	-	-	-	54.8	(13.1)	47.8	(16.2)
	MNCORRT	-	-	-	-	699	(210)	629	(141)
	PC	-	-	-	-	98	(2)	98	(2)
	SPEED	-	-	-	-	91.1	(20.6)	100.3	(23.5)
	THRUPUT	-	-	-	-	89.4	(19.7)	98.2	(22.3)
Switching Task (Manikin)	MANCORRT	3803	(1181)	2212	(856)	1873	(552)	1562	(507)
	MANPC	77	(19)	97	(5)	98	(2)	99	(2)
	MANTP	13.8	(7.0)	30.1	(11.9)	34.2	(10.3)	41.5	(12.5)
	MANCORTX	4158	(1495)	2323	(892)	2011	(588)	1750	(601)
	MANPCX	77	(19)	96	(6)	98	(2)	99	(3)
Switching Task (Math)	MTHCORRT	4238	(1205)	2454	(630)	2565	(687)	2097	(548)
	MTHPC	87	(16)	97	(3)	98	(3)	97	(2)
	MTHTP	13.6	(5.1)	25.5	(7.5)	27.8	(8.3)	29.9	(8.5)
	MTHCORTX	4663	(1241)	2610	(711)	2676	(741)	2332	(613)
	MTHPCX	84	(17)	96	(5)	98	(4)	97	(4)
NovaScan™ FAA Task	VECCRT	12891	(3599)	9860	(2265)	9226	(2666)	8008	(2923)
	VECPC	89	(11)	93	(10)	92	(13)	90	(15)
	VATNPC	98	-	100	-	99	-	100	-
	MEMCRT	4385	(1263)	3243	(888)	3032	(882)	2598	(1030)
	MEMPC	88	(11)	93	(6)	93	(8)	94	(7)
	MATNPC	96	-	100	-	99	-	99	-
Air Traffic Scenarios Test	PCDEST	46	(22)	72	(18)	93	(9)	98	(3)
	DELAY	1.6	(0.9)	50.8	(16.2)	30.6	(12.4)	18.1	(8.4)
	CRSHAC	0.0	(0.0)	2.9	(4.1)	1.1	(1.9)	0.1	(0.6)
	CRSHBD	0.0	(0.0)	1.2	(2.1)	0.1	(0.3)	0.0	(0.2)
	CRSHAP	0.1	(0.2)	0.2	(0.5)	0.9	(3.7)	0.2	(0.5)
	SEPAC	0.0	(0.0)	17.3	(29.4)	5.9	(15.6)	1.4	(2.0)
	SEPBD	0.1	(0.2)	9.4	(13.0)	1.1	(2.8)	0.1	(0.3)
	ERRDEST	0.0	(0.0)	1.4	(2.9)	0.2	(0.5)	0.1	(0.4)
	ERRGTALT	0.0	(0.0)	0.6	(0.8)	0.2	(0.4)	0.0	(0.2)
	ERRAPALT	0.0	(0.2)	1.3	(1.3)	0.4	(1.1)	0.2	(0.5)
	ERRGTSPD	0.0	(0.0)	0.1	(0.3)	0.0	(0.0)	0.0	(0.0)
	ERRAPSPD	0.1	(0.4)	4.4	(3.8)	3.5	(3.3)	3.4	(5.2)
	NDIR	11.3	(2.9)	127.0	(32.6)	131.3	(17.9)	132.4	(17.2)
	NALT	1.8	(1.4)	44.5	(10.2)	55.7	(10.4)	55.2	(6.5)
	NSPD	6.3	(2.5)	62.0	(14.5)	72.9	(13.5)	68.6	(9.3)
Multi-Attribute Task Battery (MATB)	LTSRT	2.8	(1.5)	2.0	(0.7)	1.7	(0.4)	1.7	(0.4)
	DLSRT	7.0	(2.1)	5.4	(1.8)	4.4	(1.1)	4.2	(1.4)
	MONRT	4.8	(1.2)	3.7	(1.1)	3.0	(0.6)	2.9	(0.8)
	LTSFA	0.0	(0.2)	0.1	(0.3)	0.4	(0.7)	0.5	(0.7)
	DLSFA	0.5	(1.5)	0.2	(0.5)	2.2	(4.2)	3.8	(12.2)
	MONFA	0.6	(1.5)	0.3	(0.5)	2.6	(4.1)	4.3	(12.1)
	MONER	3.4	(2.4)	0.8	(0.9)	4.0	(4.5)	5.9	(12.0)
	COMCRT	6.3	(2.4)	4.7	(1.8)	3.9	(1.5)	3.8	(1.4)
	COMER	1.1	(1.9)	0.4	(1.1)	2.1	(1.6)	2.0	(2.2)
	TRKRMS	45.4	(12.7)	45.0	(12.4)	48.5	(17.7)	45.8	(16.4)
	TNKMAD	345	(298)	232	(130)	211	(179)	191	(126)
	TNKACT	48	(18)	72	(27)	344	(146)	367	(185)

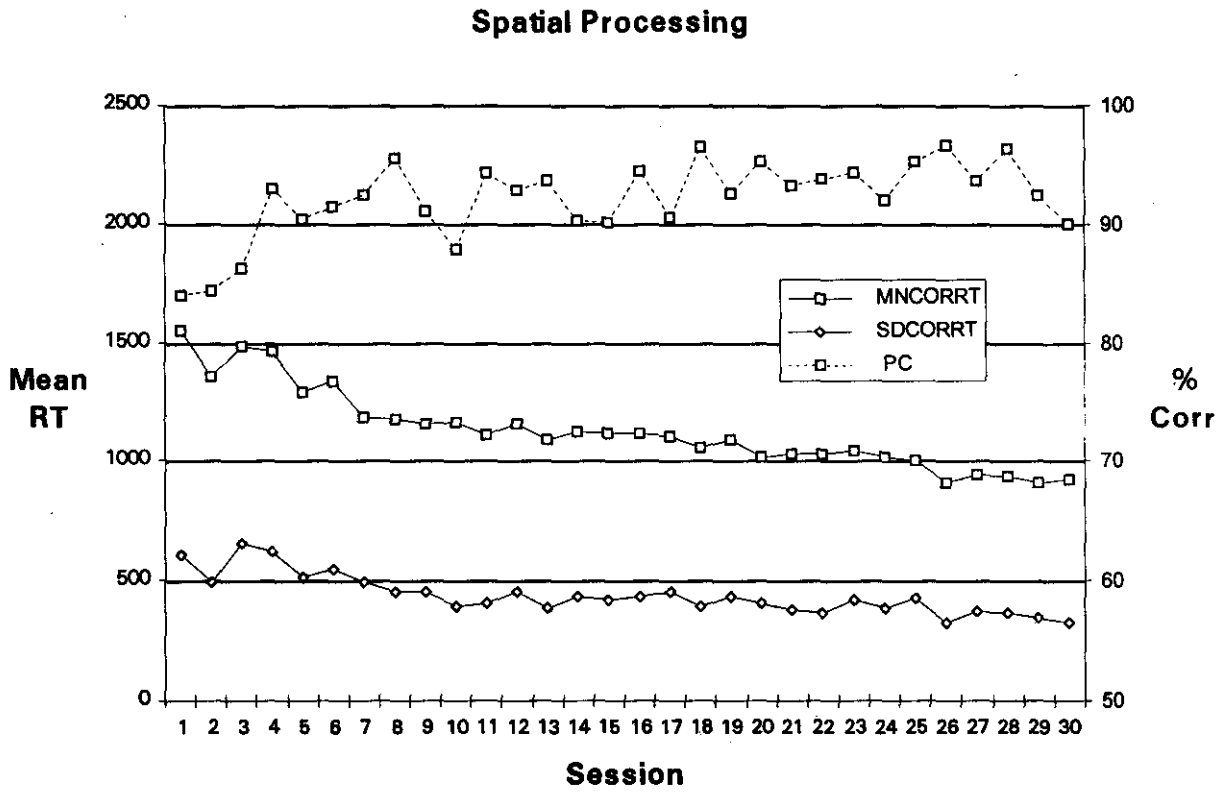

Figure 2. Spatial Processing (RT Mean and Standard Deviation and Percent Correct).

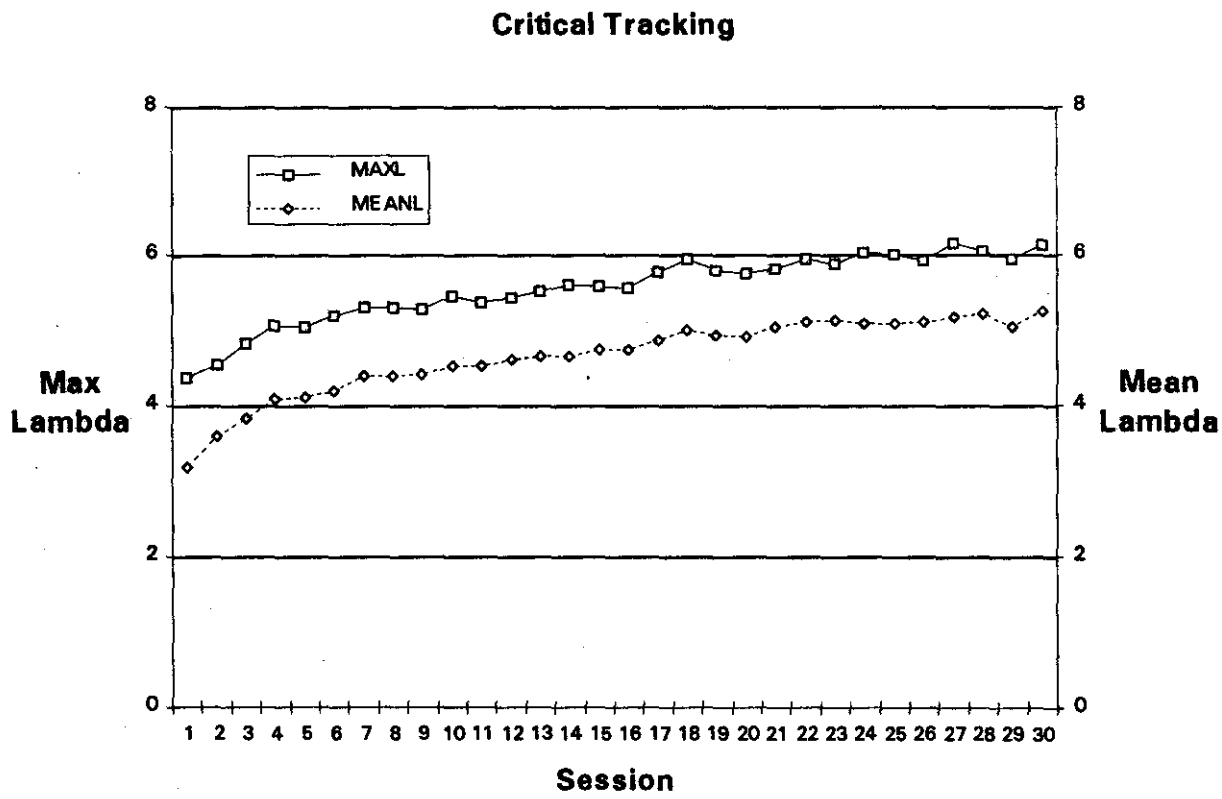

Figure 3. Critical Tracking (Maximum and Mean Lambda).

Spatial Processing Task

Three variables were used to assess spatial processing performance: mean correct response time (MNCORRT), standard deviation of correct response times (SDCORRT) and percent correct (PC). These are presented in Figure 2. A considerable amount of learning occurred during the first seven sessions of the spatial processing task, as indicated by the MNCORRT variable. However, it is also apparent that after this initial period, more modest gains were made throughout the thirty sessions. The MNCORRT learning curve never reached a clearly defined plateau, although by the end of the training period (Session 10) considerable stability in the response had been achieved. This view is supported by the fact that the SDCORRT and PC curves also show a greater degree of stability at this same time (i.e., variability and accuracy begin to show less fluctuation than that demonstrated in the earlier sessions). In fact, PC increased rapidly and remained above 90% after Session 3, with the exception of Session 10. Considerable learning of the task process was complete by Session 4.

Critical Tracking Task

Figure 3 reveals that both performance measures for this task (maximum lambda during the trial and the mean of the lambda's at control losses) increased from Session 1 through Session 30, again indicating some degree of continued performance improvement throughout the study. Much like the spatial processing task, tracking performance seems to have improved most rapidly during the first few sessions, providing a fairly distinct inflection point after the fourth session. While learning continued after this point, the rate slowed considerably. This trend was also noted in the control losses, a function of performance proficiency, which also improved throughout the sessions, as illustrated in Figure 4. However, after Session 10, the change is very slight. In contrast, RMS error remained approximately constant from Session 1 through Session 30 and was therefore not a very sensitive performance measure for this variation of the task.

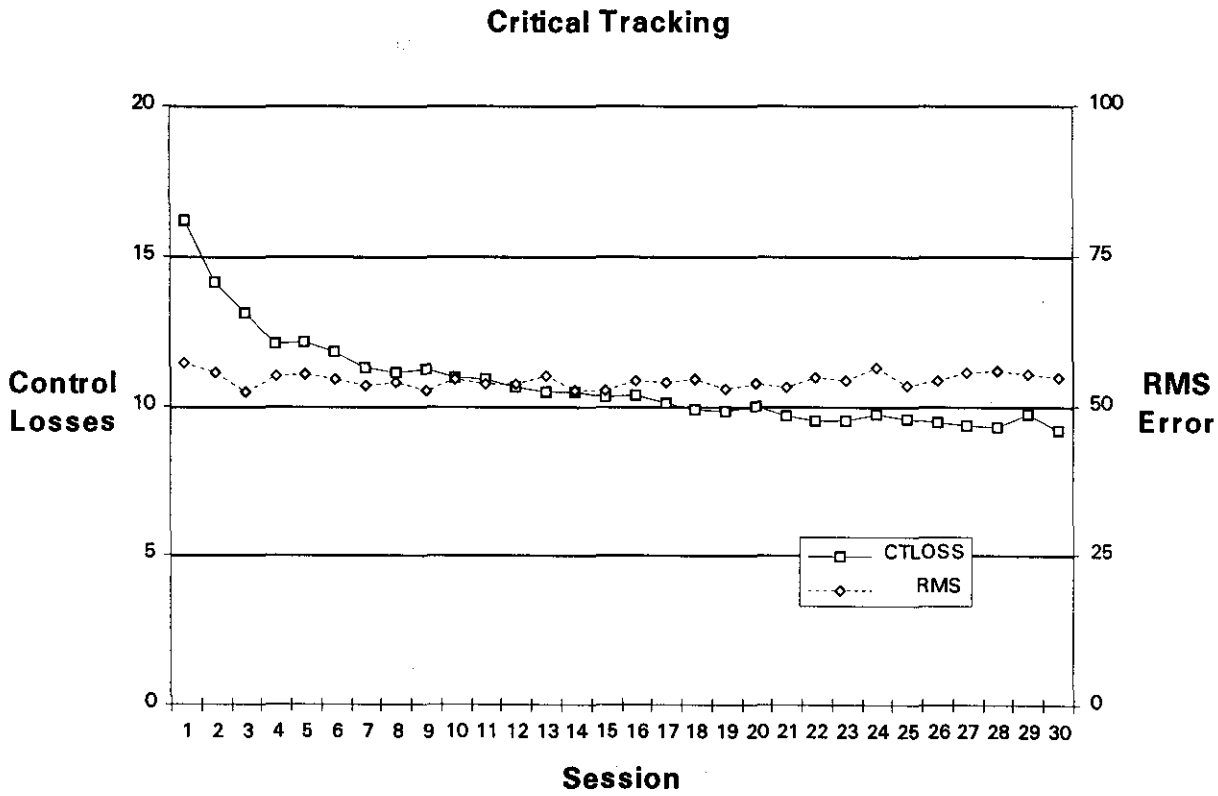

Figure 4. Critical Tracking (Control Losses and RMS Error).

Dual Task - Group Lambda

Tracking: Figure 5 presents Dual Task-Group Lambda tracking performance. Prior to the change in lambda from 2.0 to 3.7, the number of control losses appeared to stabilize close to zero by Session 8, and RMS error exhibited considerable improvement, although questionable stability. Following the lambda change, both measures increased dramatically and then showed a second phase of improvement through Session 30. The majority of improvement for control losses occurred during the first five sessions at the higher lambda. At Session 9, data were eliminated for two subjects (S213 and S219) who had malfunctioning trackballs during that time.

Memory Search: Mean overall response time (MNALLRT) and mean correct response time (MNCORRT), shown in Figure 6, suggest that learning continued until the last session prior to the lambda change (Session 12), although most learning was complete by Session 5. Two to three sessions were required for Memory Search performance to recover following the change in lambda value. After this, only modest amounts of improvement can be observed through Session 30. With respect to mean incorrect response time (MNINCRT), a definite downward trend can be observed for the sessions prior to the lambda change. After the lambda change, incorrect response time means were very erratic, probably due to the few number of incorrect responses. With the exception of the first session, the measures of percent correct of all stimuli (PC) and percent correct of all responses, excluding time-outs (PCRESP), were essentially identical. Following a substantial improvement from Session 1 to Session 2, percent correct showed little change over the course of the study. The lambda change appeared to have a slight effect on the percent correct variables for the first two sessions following the change. As seen in Figure 7, speed and throughput followed the same pattern as the response time measures. The influence of the lambda change on both response time and percent correct is easily observed in the composite measure of throughput.

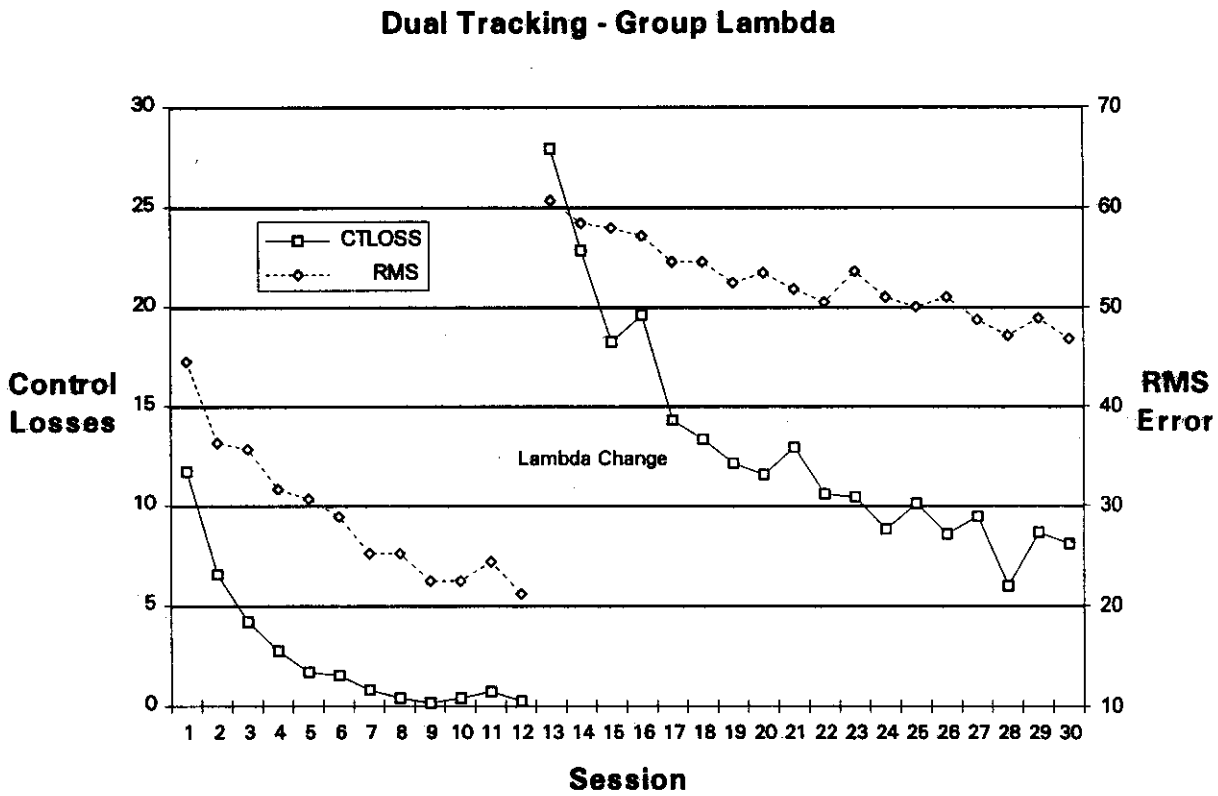

Figure 5. Dual Tracking-Group Lambda (Control Losses and RMS Error).

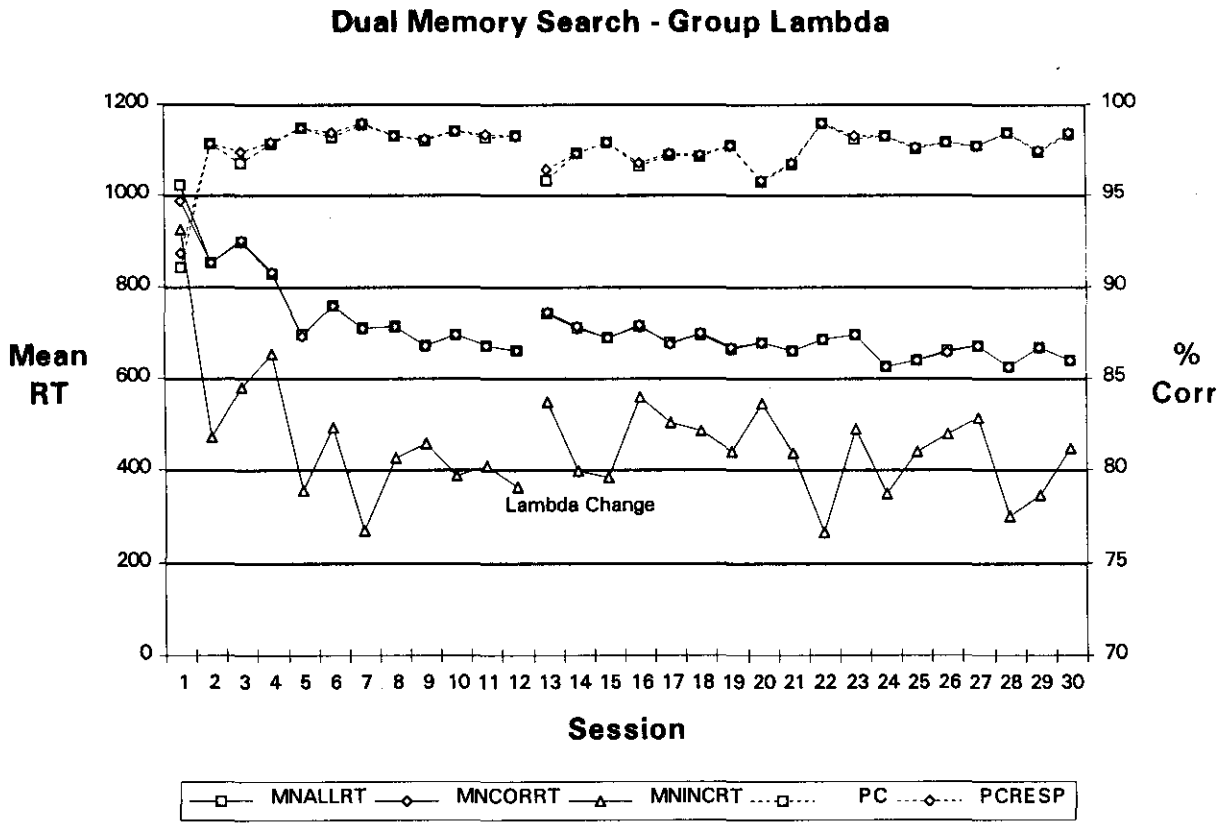

Figure 6. Dual Memory Search-Group Lambda (Mean RT and Percent Correct).

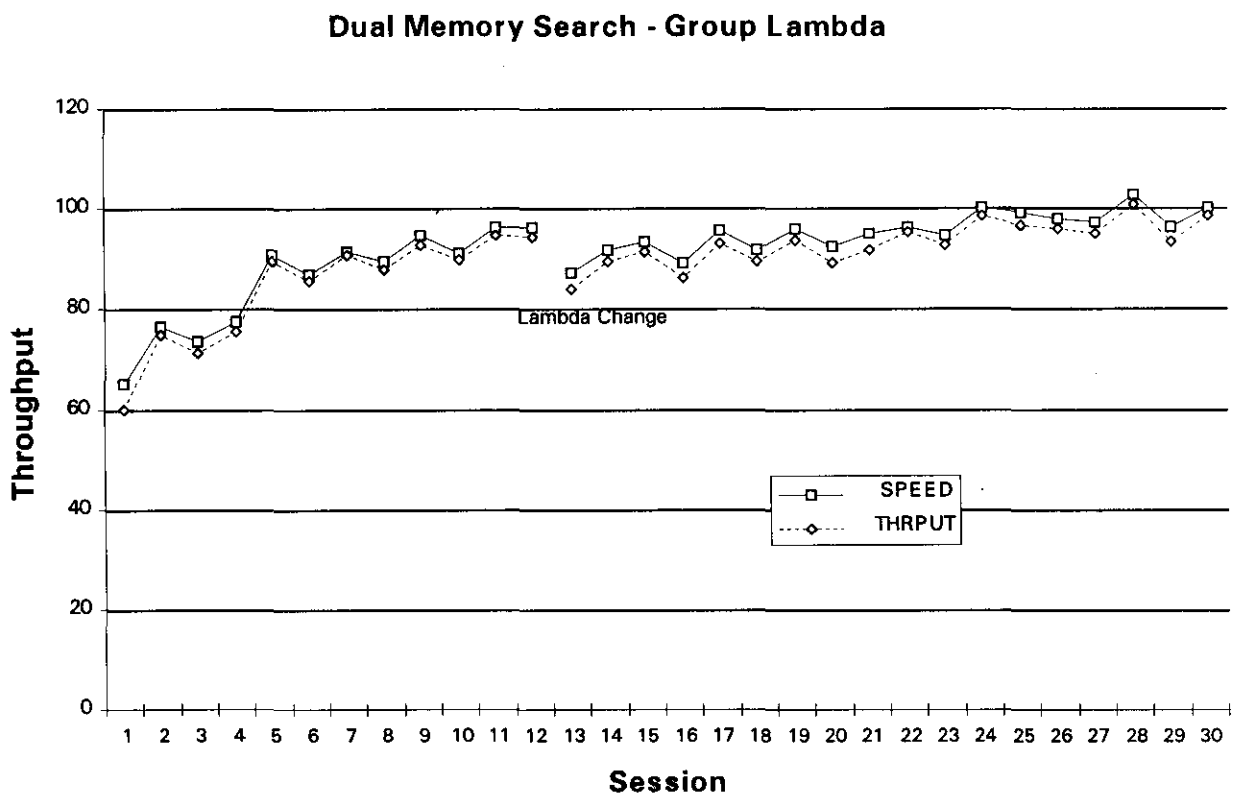

Figure 7. Dual Memory Search-Group Lambda (Speed and Throughput).

Dual Task - Individual Lambda

Tracking: Figure 8 reveals that both control losses and RMS error decreased continuously from Session 13 through Session 30. As with the group lambda version of the task, most of the improvement occurred during the first five sessions at the individualized lambda (control losses from 27 to 10 and RMS error from 63 to 57). The remaining improvement was less rapid but constant (control losses from 10 to 5 and RMS error from 57 to 48). Compared with performance using the group lambda of 3.7, average performance with the individualized lambda values was better. That is, absolute levels of these dependent variables reached lower levels more rapidly under the individual lambda condition, as compared to the group lambda condition (cf. Figure 5 and Figure 8).

Memory Search: Figure 9 presents performance measures for the Dual Memory Search-Individual Lambda task. Mean overall response time (MNALLRT) and mean correct response time (MNCORRT) for this task showed minimal improvement from Session 13 to Session 30. It might be argued that slight improvement is noticeable after Session 23, but this is not evidence for a remarkable learning effect during the last seven sessions. Mean incorrect response time (MNINCRT) presents a fairly erratic trend. Again, this instability may result because the number of responses comprising these points is sharply reduced following the early stages of learning, thereby creating a more volatile measure. Both percent correct measures (PC and PCRESP) remained stable throughout all sessions. Finally, speed (responses per min) and throughput (correct responses per min) show a fair degree of stability from Session 13 through Session 30 (Figure 10). However, there is slight but noticeable improvement across the sessions.

Switching

Manikin Task: As revealed in Figure 11, the two response time variables, MANCORRT and MANCORTX, show continued improvement throughout the study. The learning curves for both variables never reached a plateau. On the other hand, the percent correct measures (MANPC and MANPCX) appear to reach a plateau of 98% at Session 17. Mean response time for transition stimuli was slightly longer than for all stimuli combined. The measure of throughput (MANTP) increased from Session 1 through Session 30 (see Figure 13).

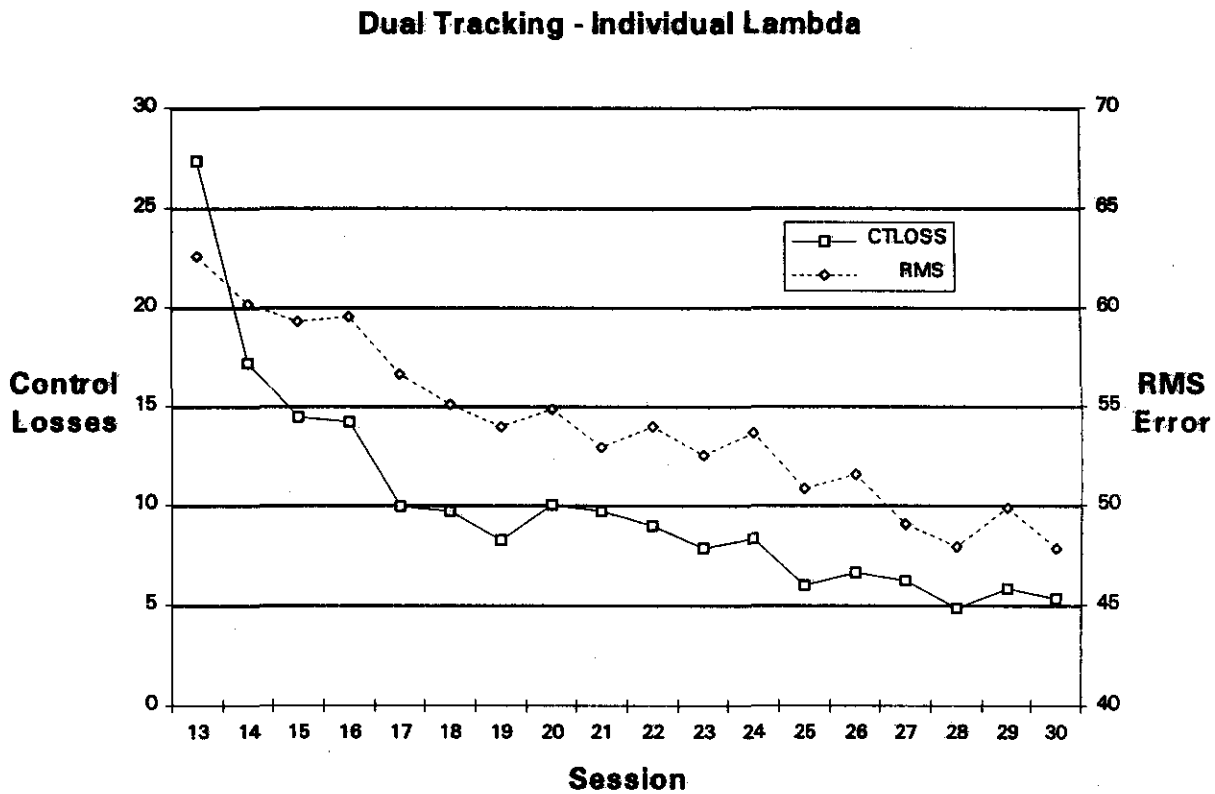

Figure 8. Dual Tracking-Individual Lambda (Control Losses and RMS Error).

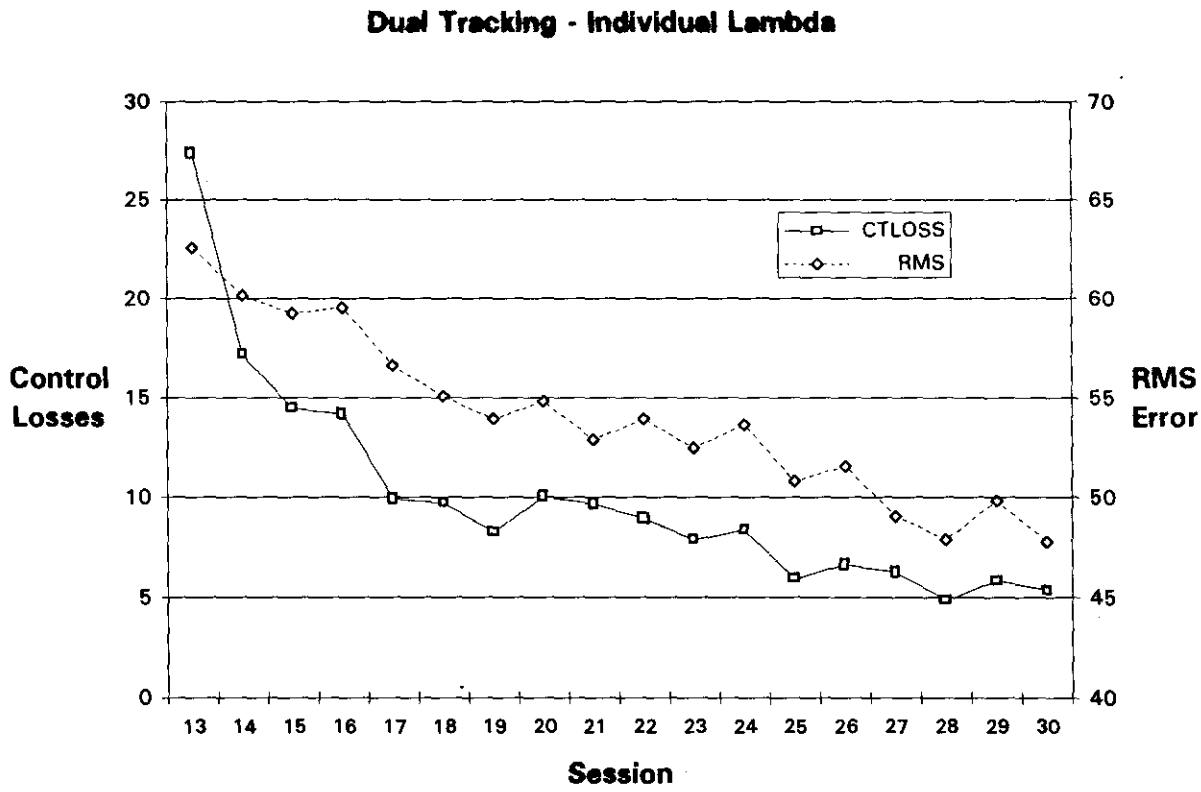

Figure 9. Dual Memory Search-Individual Lambda (Mean RT and Percent Correct).

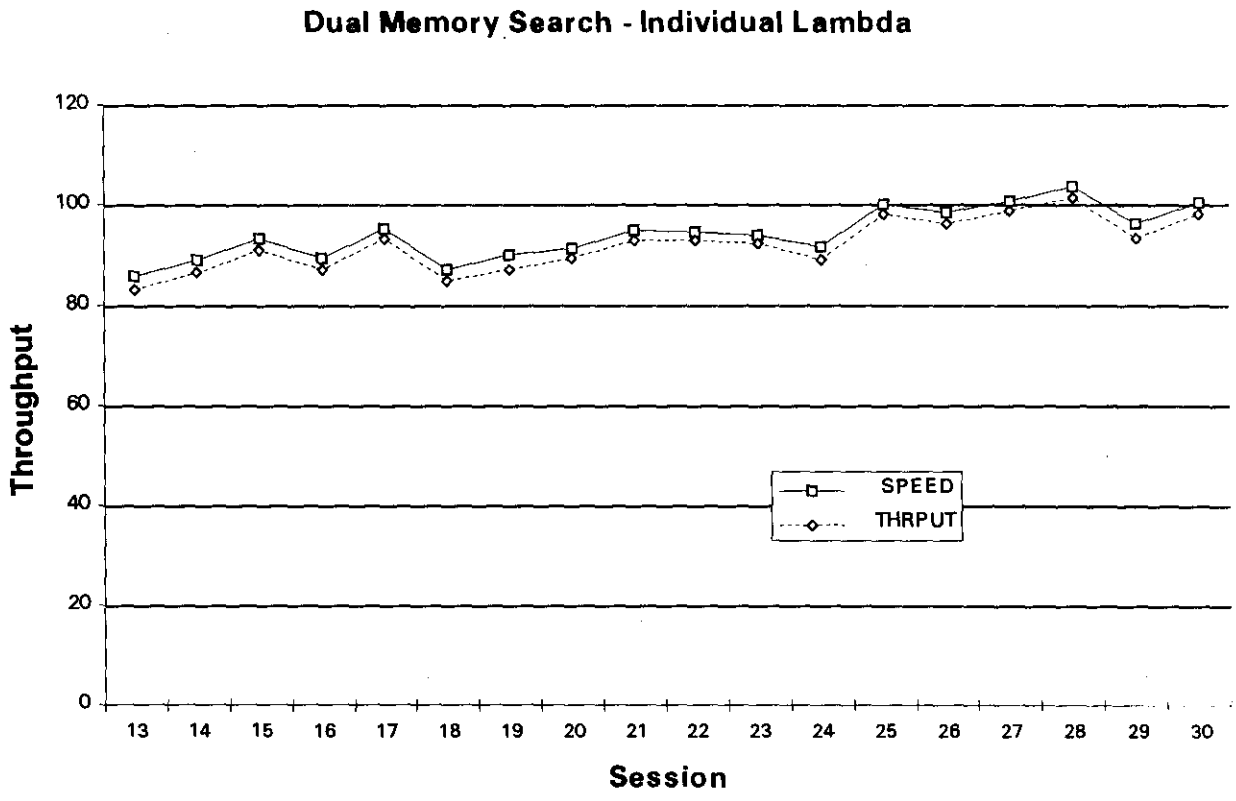

Figure 10. Dual Memory Search-Individual Lambda (Speed and Throughput).

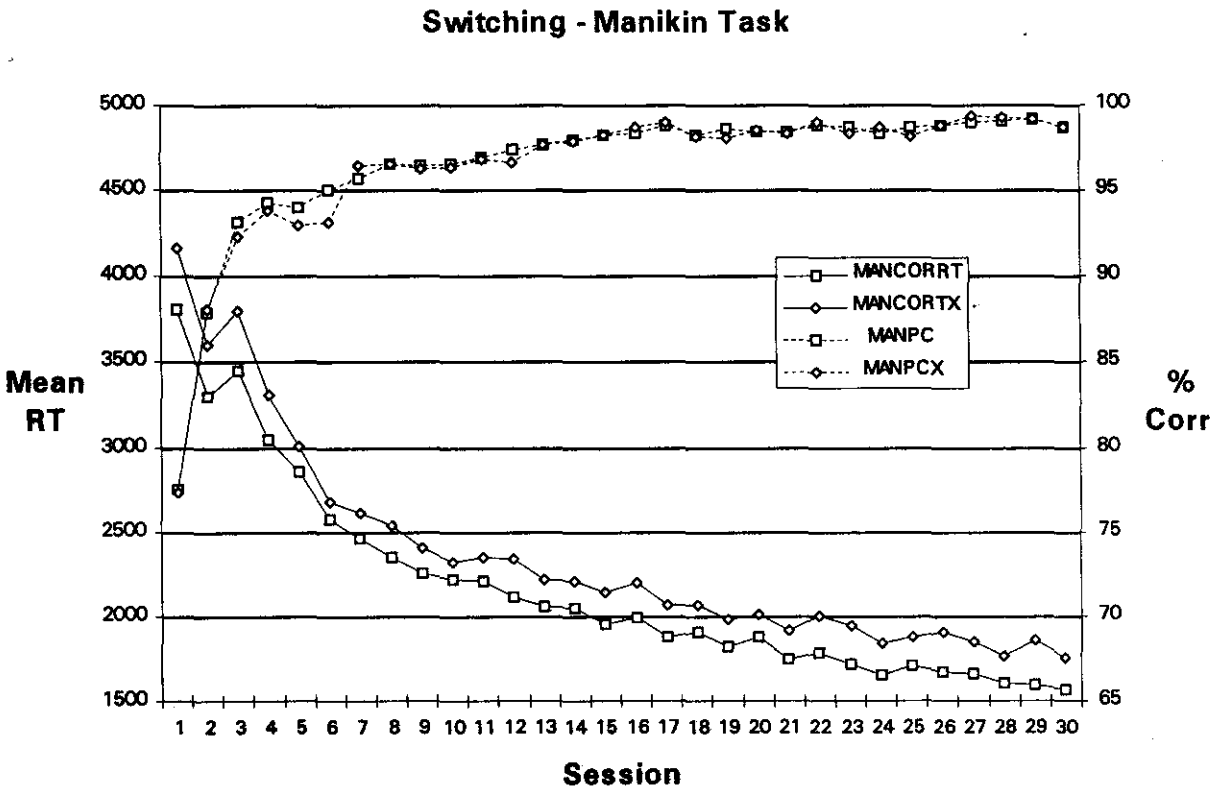

Figure 11. Switching-Manikin Task (Mean RT and Percent Correct).

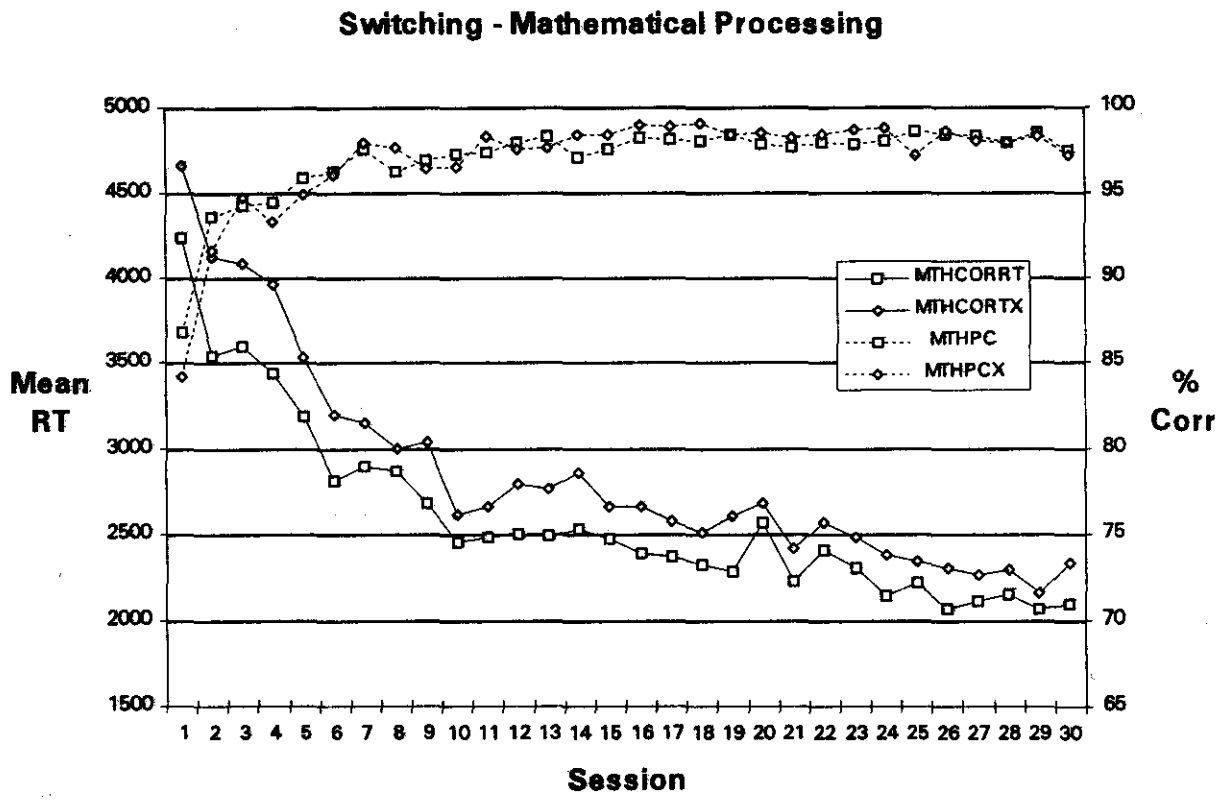

Figure 12. Switching-Mathematical Processing (Mean RT and Percent Correct).

Mathematical Processing Task: As with the Manikin Task, the response time variables shown in Figure 12 for Mathematical Processing (MTHCORRT and MTHCORTX) continued to show improvement throughout the study, although there is evidence of stability from Session 26 on. Again, the mean response time for transition stimuli was slightly longer than for all stimuli combined. The percent correct measures (MTHPC and MTHPCX) were relatively constant after the first 10 sessions. Mirroring the pattern for response time, Mathematical Processing throughput (MTHTP) presented continuous improvement over the entire study, but with a lower apparent asymptote than manikin throughput (see Figure 13).

NovaScan™

Vector Projection Task: Vector Projection correct response times (VECCRT) are presented in Figure 14. A downward trend is visible from the first to the last session, indicating that the subjects continuously improved, although much of that improvement was obtained in the first seven sessions. The standard deviation of the correct response times (VECCSD), however, showed minor variability and slight improvement in the first few sessions and was then fairly stable throughout the remaining sessions. The percent correct measure (VECPC) also improved rapidly and remained between 90% and 94% throughout testing. In general, much improvement was seen in this task during the first several sessions, and while improvement was seen following these sessions, the rate of improvement was sharply reduced as was the trial-to-trial variation.

Continuous Spatial Memory Task: The measure of correct response time for this task (MEMCRT) decreased from Session 2 through Session 30, as shown in Figure 15. A similar, although much less pronounced, pattern existed for the standard deviation of the correct response times (MEMCSD). On the contrary, percent correct (MEMPC) stabilized by Session 7 and varied between 92% and 96%.

Attention Task: On Vector Projection screens, the number of attention acknowledgments (VATNACK) coincided with the number of attention requests (VATNREQ; as seen in Figure 16) for almost every session, indicating that the accuracy for this task was close to 100% from the first session (Session 2). False alarms (VATNFA) were essentially zero for all sessions (with the exception of Session 29). On Spatial Memory screens, the number of attention acknowledgments (MATNACK) was essentially the same as the number of attention requests (MATNREQ; as seen in Figure 16) for all sessions, again indicating that accuracy for this task was close to 100% from the

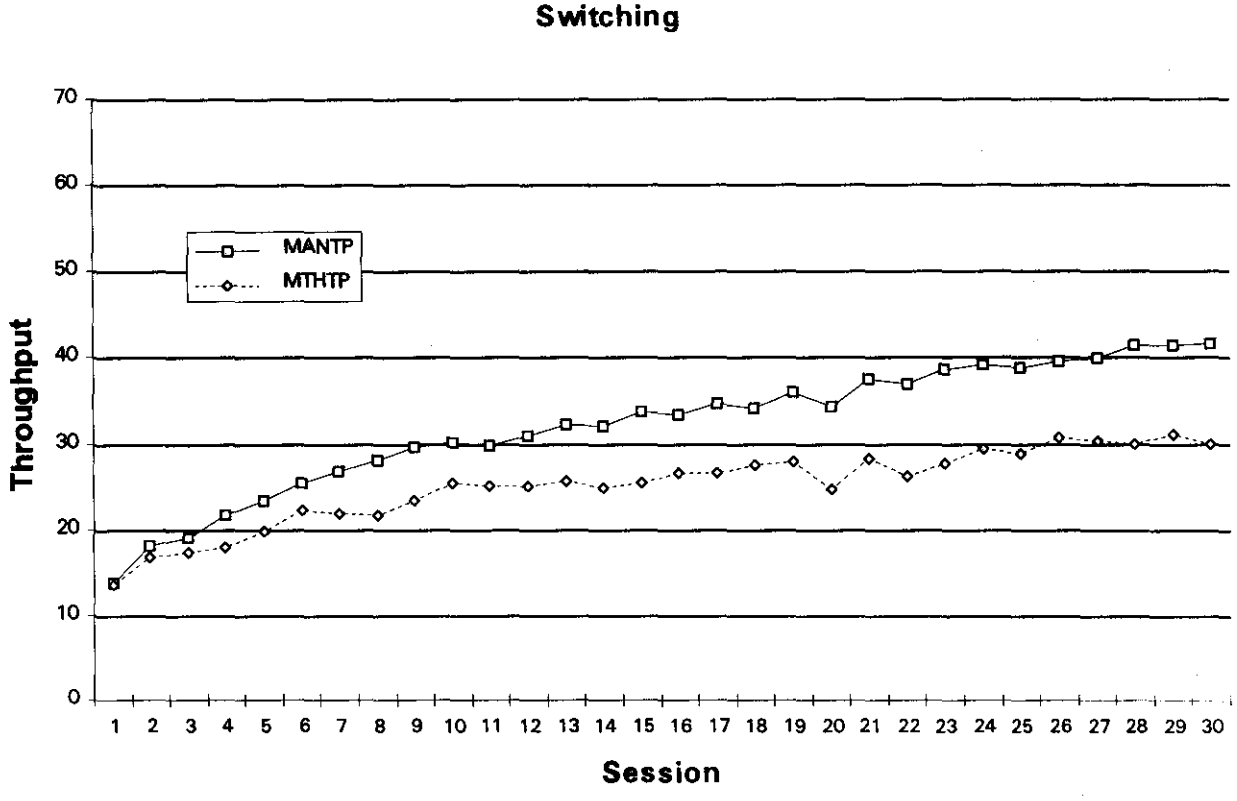

Figure 13. Switching (Throughput).

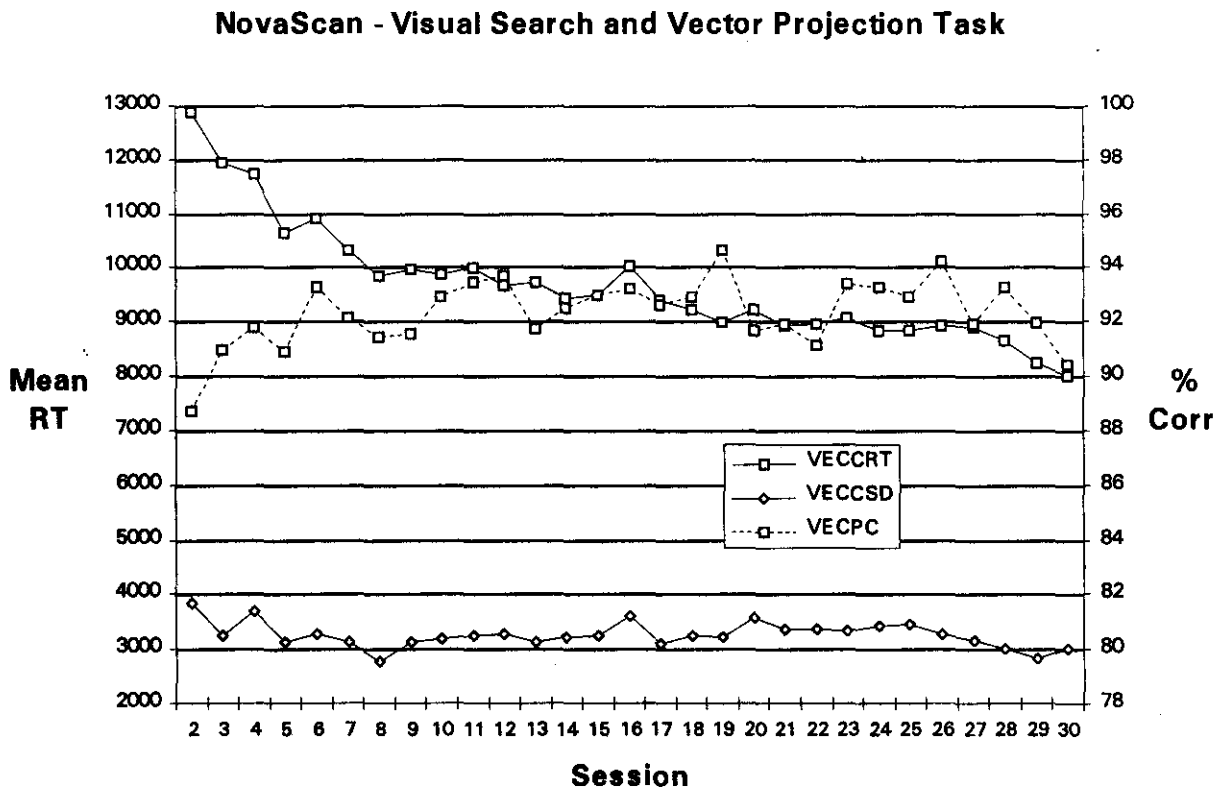

Figure 14. NovaScan™-Visual Search and Vector Projection Task (Mean RT and Percent Correct).

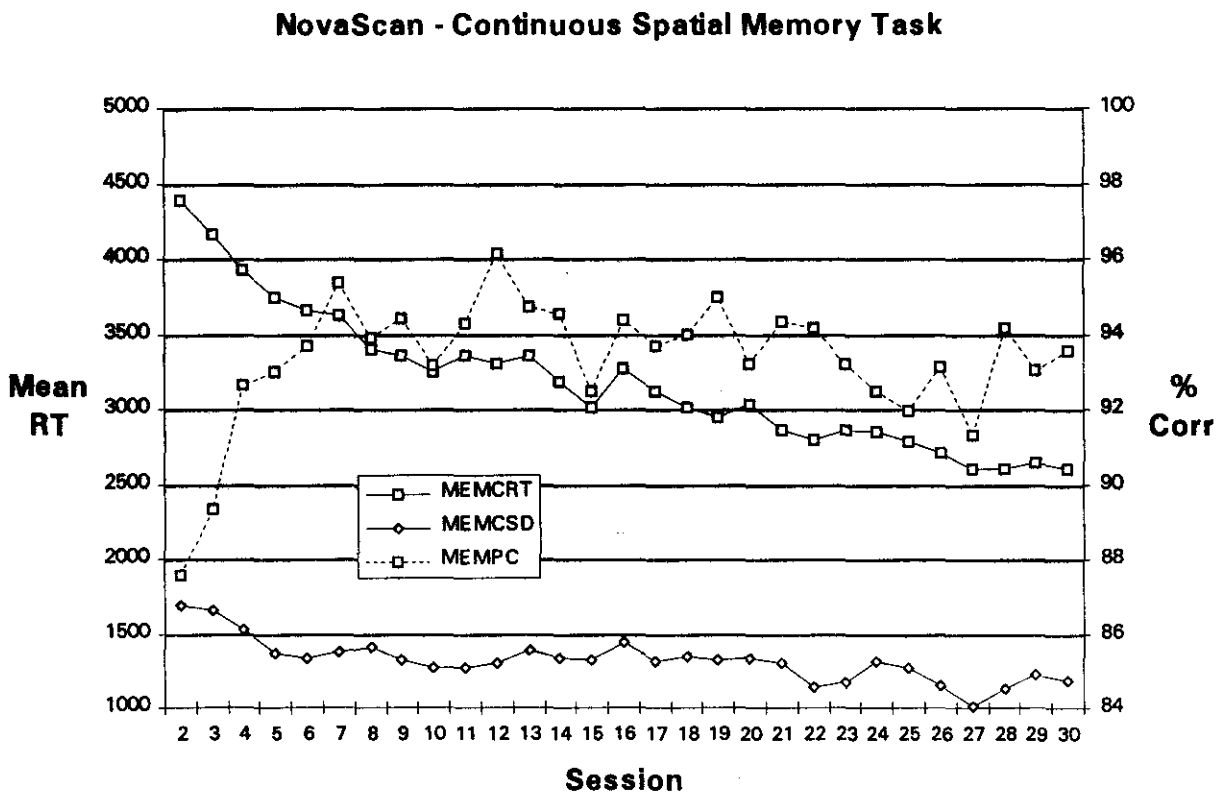

Figure 15. NovaScan™-Continuous Spatial Memory Task (Mean RT and Percent Correct).

first session. The average number of false alarms (MATNFA) was zero for all sessions except Sessions 8, 25, and 29.

In general, the two more central NovaScan™ tasks (Visual Search and Vector Projection, and Continuous Spatial Memory) yielded percent correct and variability measures that stabilized fairly quickly. However, the response time measures showed modest and continual change (improvement) across five weeks of testing.

Air Traffic Scenarios Test

For the easier (and shorter) ATST scenarios used during Sessions 1 through 7, crashes involving aircraft, the airspace boundary, and airports (i.e., CRSHAC, CRSHBD and CRSHAP) were few from the first session (Session 1), as is evident in Figure 17. Separation errors, both SEPAC (number of separation violations with other aircraft) and SEPBD (number of separation violations with the air space boundary) were also low. Both crashes and separation violations showed evidence of a learning curve from Sessions 4 through 7. These sessions all involved a more difficult scenario (an increase to 16 planes). Following another increase in scenario difficulty and length at Session 8, aircraft crashes and separation errors increased dramatically. In particular, Session 10, involving scenario S7 (45 planes, 1500 seconds), had an excessive number of errors (CRSHAC = 3, SEPAC = 17). Performance recovery was somewhat erratic through Session 22, followed by a more orderly decrease to a respectable 0 to 1 errors in each category.

Low error rates for airport speed and altitude (ERRAPSPD and ERRAPALT), gate speed and altitude (ERRGTSPD and ERRGTALT), and destination (ERRDEST), presented in Figure 18, suggest that the basic rules of the task were learned in the earliest test sessions (Session 1). The variable that seemed most affected by the scenario difficulty change was ERRAPSPD (number of speed errors when landing at the airport). This variable increased considerably for Session 10 and remained high for the remaining sessions. In fact, from Session 13 through Session 30, ERRAPSPD increased at a slight but constant rate. Contrary to intuition, this may have been due to increased proficiency and a motivation to further reduce delay times by attempting to change landing speed at the last instant. This change in strategy

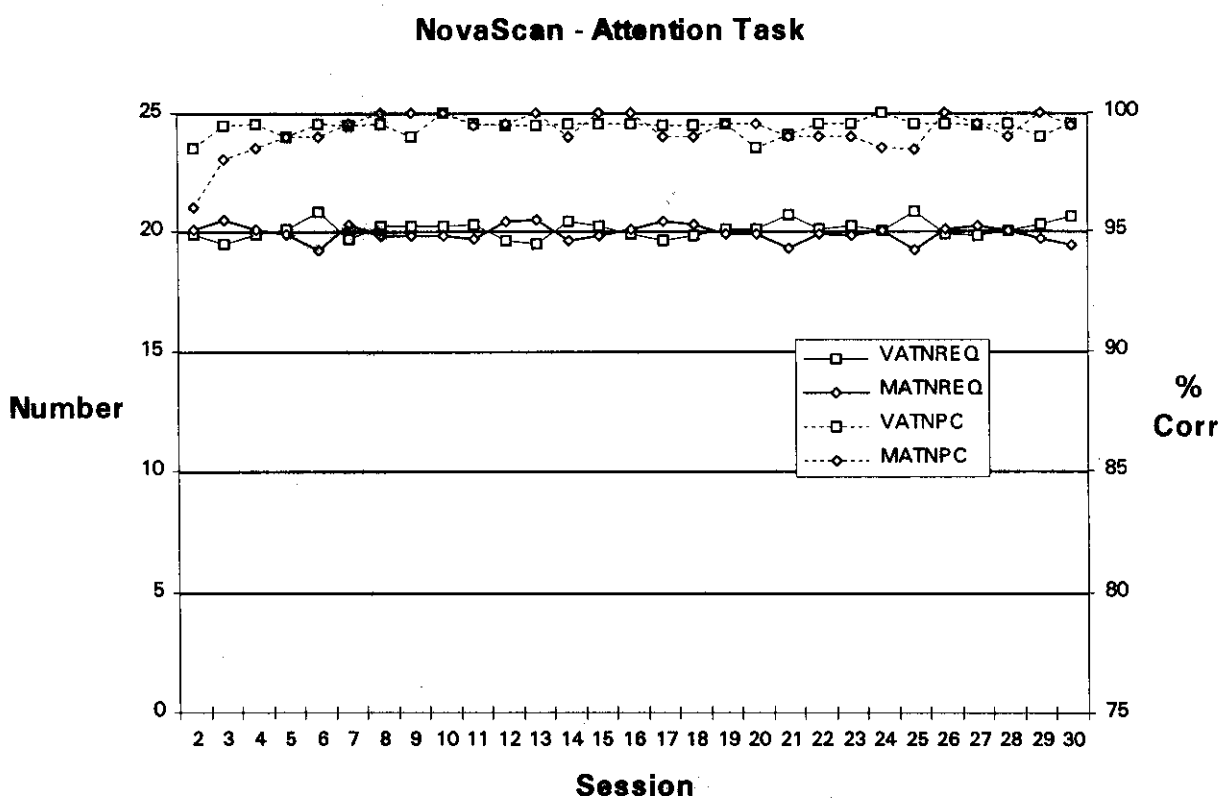

Figure 16. NovaScan™-Attention Task (Number Completed and Percent Correct).

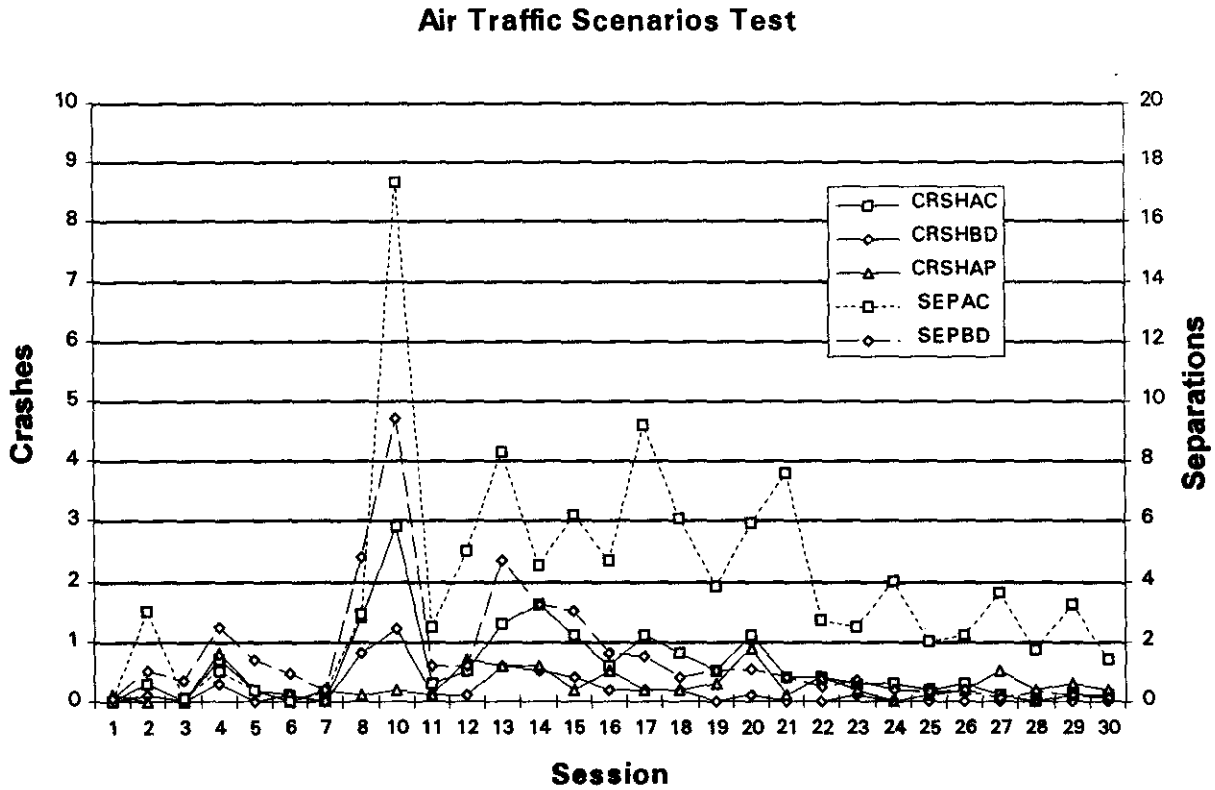

Figure 17. Air Traffic Scenarios Test (Crashes and Separations).

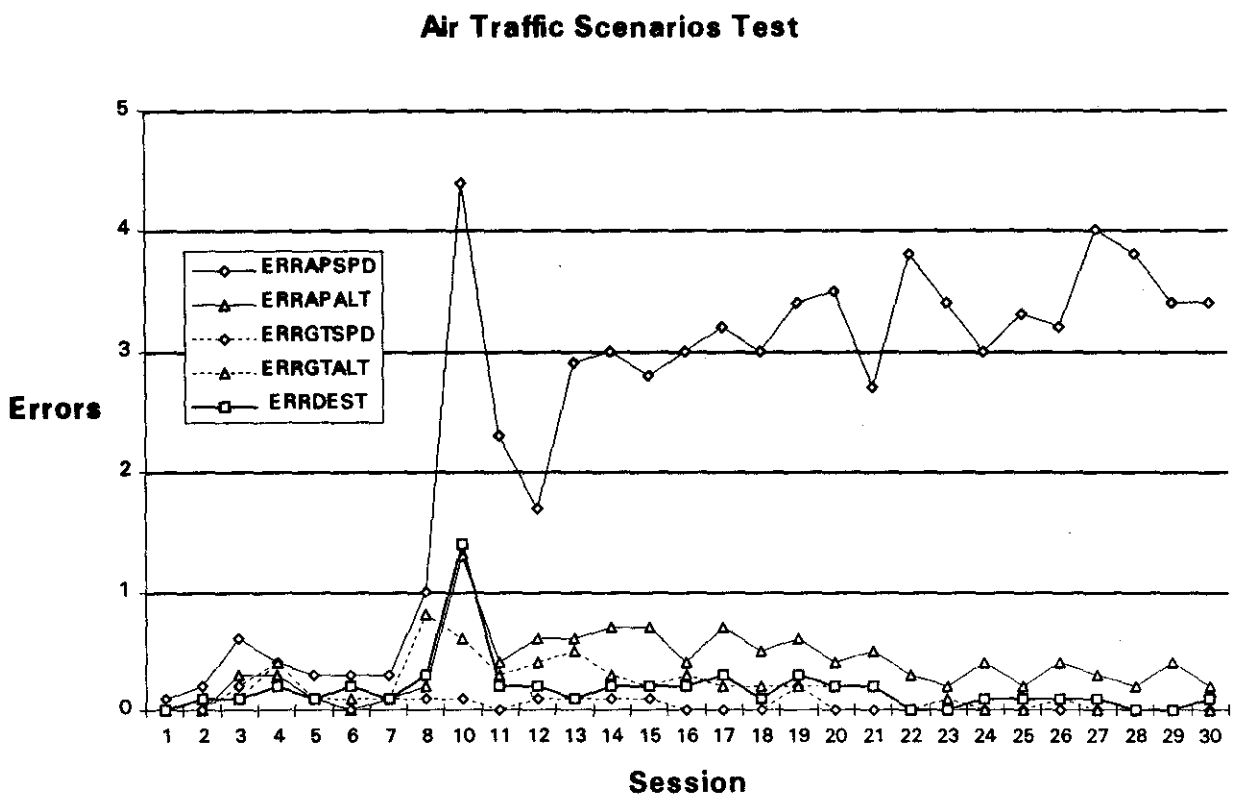

Figure 18. Air Traffic Scenarios Test (Errors).

undoubtedly led to a great number of incidents in which speed change was forgotten or not caught at the last instant prior to landing, thereby resulting in higher error rates. This explanation is supported by the substantial continued improvement in the delay score for planes arriving at their destination (DELAY) from Session 13 through Session 30, as seen in Figure 19.

Also from Figure 19, it can be seen that the number of airplanes arriving at the correct destination (NDEST) increased as a function of the number of planes in the scenario (i.e., before Session 8, compared to after Session 8) and the proficiency of the controllers (i.e., evidence of learning curves after each increase in difficulty level). The percentage of planes successfully arriving at the correct destination (PCDEST) improved dramatically between Session 1 (46%) and Session 6 (97%). Following the change to more difficult scenarios, a second stage of improvement occurred, with PCDEST reaching 98% at Session 30.

Data from Figure 20 suggest that the indicators of control activity, that is, the number of direction, altitude and speed commands issued (NDIR, NALT, NSPD, respectively), varied primarily as a function of the number of planes, but also to a much lesser degree as a function of controller proficiency. From Session 13 on, these measures were fairly stable. The number of direction changes (NDIR) appeared to remain constant, the number of altitude changes (NALT) increased slightly, and the number of speed changes (NSPD) decreased slightly.

In more general terms, the ATST seems to be a complex task, the rules of which are learned quite quickly. Many of the performance requirements of the task are also learned quickly; however, there is clear evidence that complex trade-offs may be occurring that affect performance measures well beyond initial training sessions.

Multi-Attribute Task Battery

Monitoring Task: Figure 21 reveals that mean response time for lights (LTSRT) appears to stabilize by Session 12. Response time for dials (DLSRT) demonstrated rapid improvement over the first six sessions and then continued slight improvement through the end of the study. The average of these two monitoring response time variables (MONRT) suggests that again subjects improved most in early sessions (probably the first six) and then showed gradual improvement throughout later sessions. The standard deviations of these variables also indicate continued

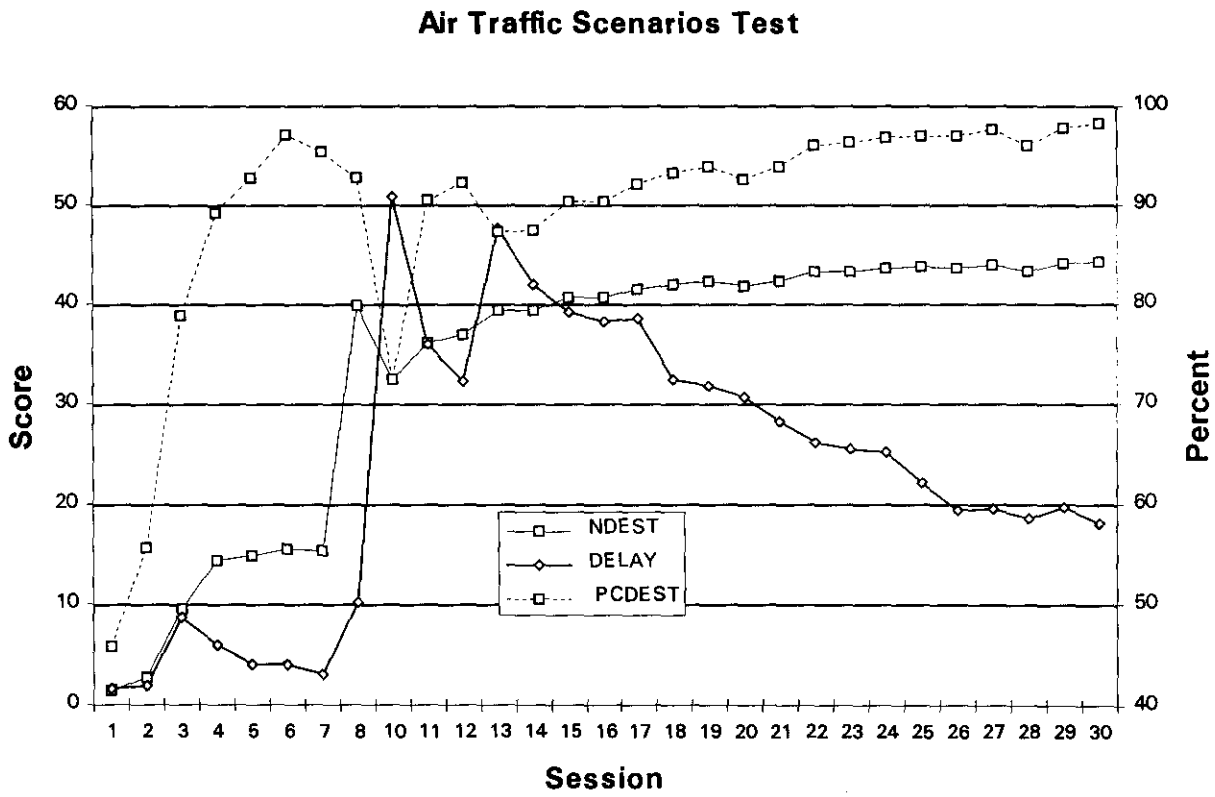

Figure 19. Air Traffic Scenarios Test (Number at Destination, Percent at Destination, Delay).

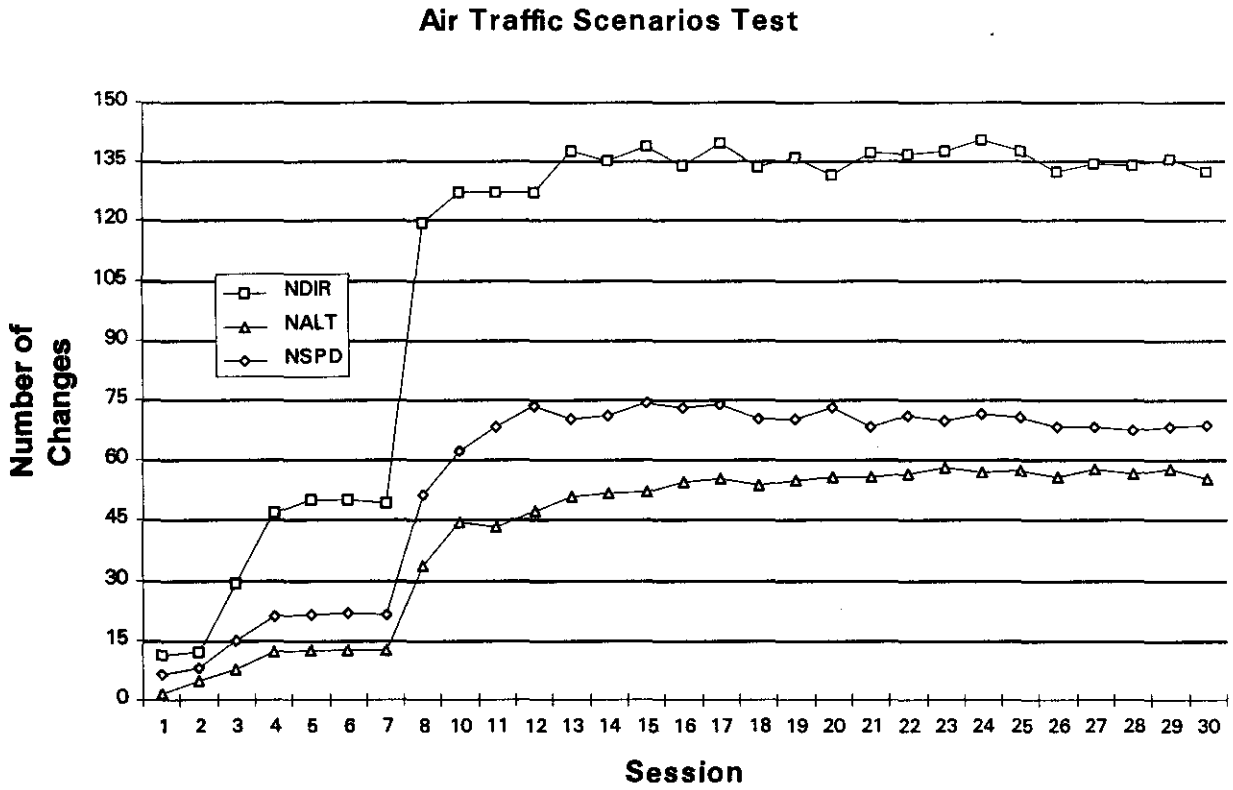

Figure 20. Air Traffic Scenarios Test (Changes in Direction, Altitude, Speed).

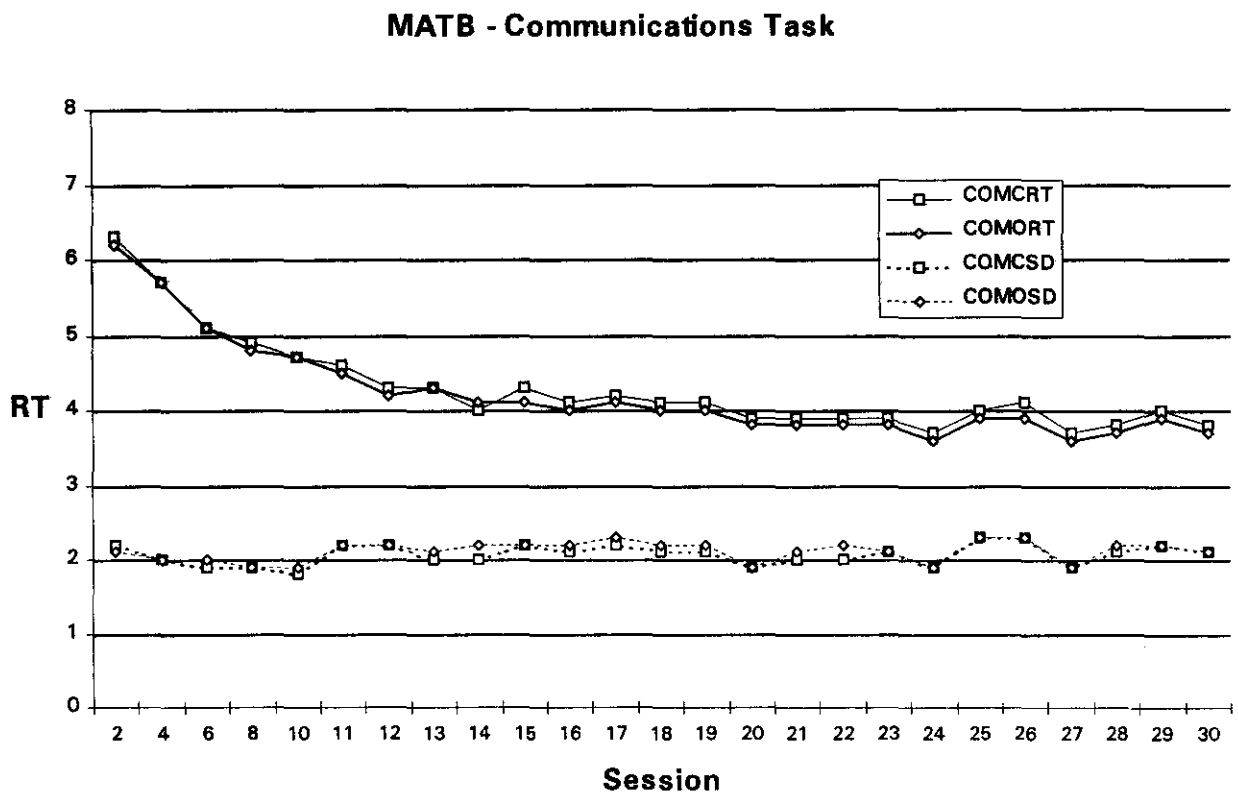

Figure 21. MATB-Monitoring Task (RT Mean and Standard Deviation).

improvement over the first few sessions. In particular, the standard deviation for lights (LTSSD) reached a plateau at Session 20, the standard deviation for dials (DLSSD) reached a plateau at Session 24, and the standard deviation for lights and dials combined (MONSD) reached a plateau at Session 22. In all cases, the majority of the improvement was completed during the first six sessions.

Figure 22 presents data for time-out errors for lights (LTSTO), time-out errors for dials (DLSTO), and time-out errors for lights and dials combined (MONTO). The number of time-out errors for lights (LTSTO) decreased rapidly after the first session and approached zero as early as Session 10. Time-out errors for dials (DLSTO) and, therefore, time-out errors for lights and dials combined (MONTO) decreased considerably from Session 2 to Session 10 and then increased (by a factor of four) over the next two sessions as a result of increasing the length of the task from 10 minutes to 40 minutes. Following this change, both variables presented a slight downward trend through Session 25.

Figure 23 reveals that the number of false alarms for lights (LTSFA) remained close to zero for all sessions. On the contrary, the number of false alarms for dials (DLSFA) was low for the first five sessions, and then increased in a somewhat erratic manner following the task length change on Session 11. This pattern suggests that following Session 10, performance deteriorated over time. After data collection was completed, it was determined that at least five subjects generated numerous false alarms throughout some sessions by periodically and repeatedly pressing the dials response keys (see Figure 24). It is assumed that the subjects did this as a "preemptive strategy" to allow more time for the other MATB tasks without having to constantly monitor the dials. Removal of the data for these subjects eliminated peaks on Sessions 16 through 18 and smoothed the plot considerably. However, it did little to influence the increases seen in Sessions 26-30. It may have been that this performance strategy was adopted by more subjects at this point. The pattern of false alarms for lights and dials combined (MONFA) was very similar to that of DLSFA.

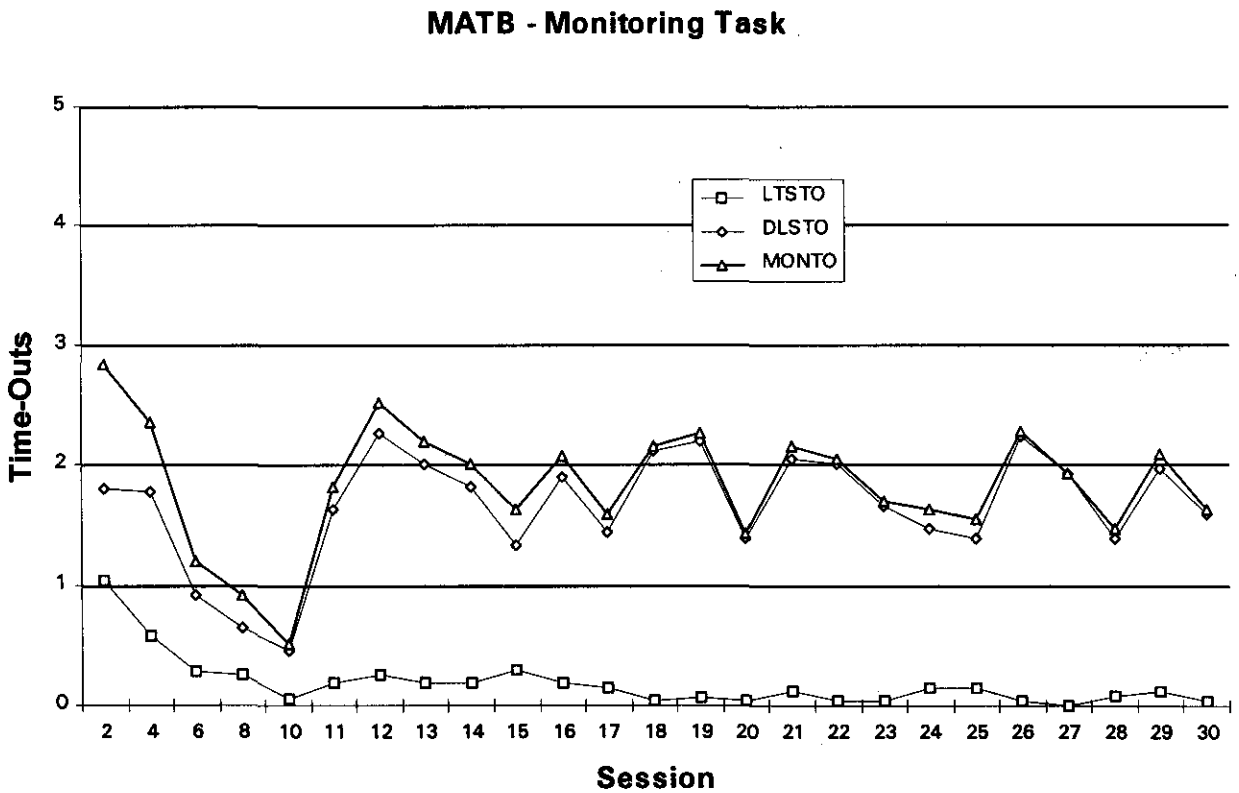

Figure 22. MATB-Monitoring Task (Time-Outs).

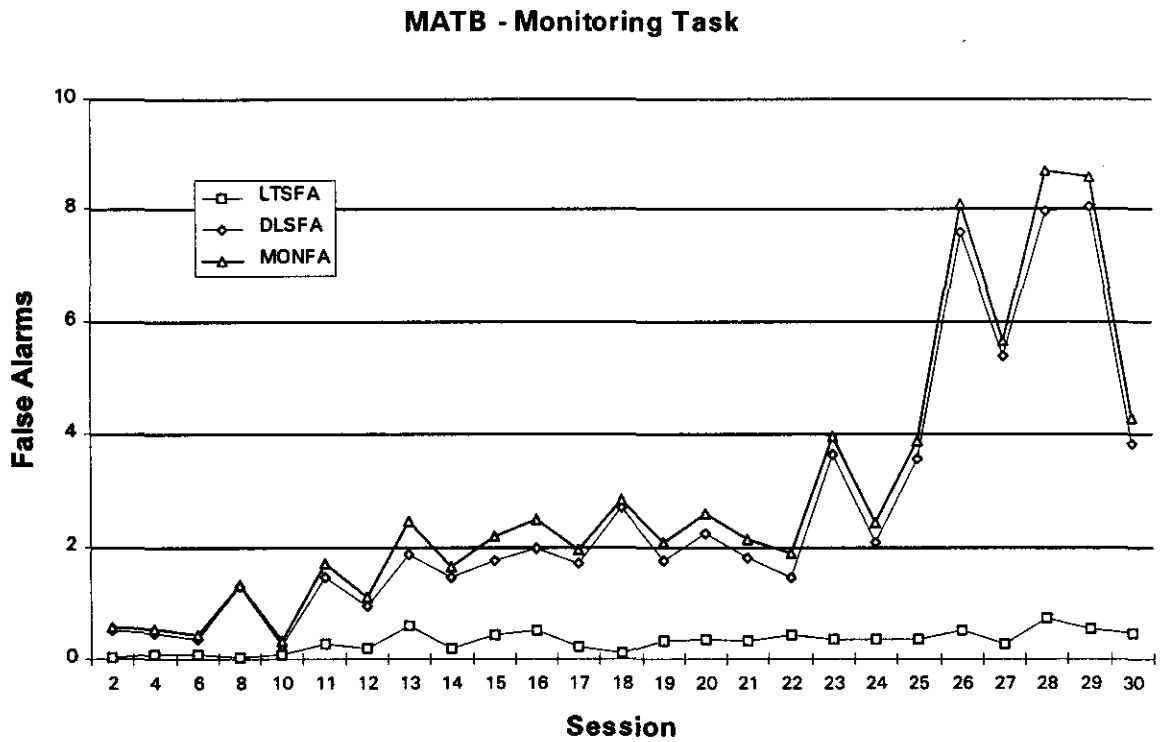

Figure 23. MATB-Monitoring Task (False Alarms).

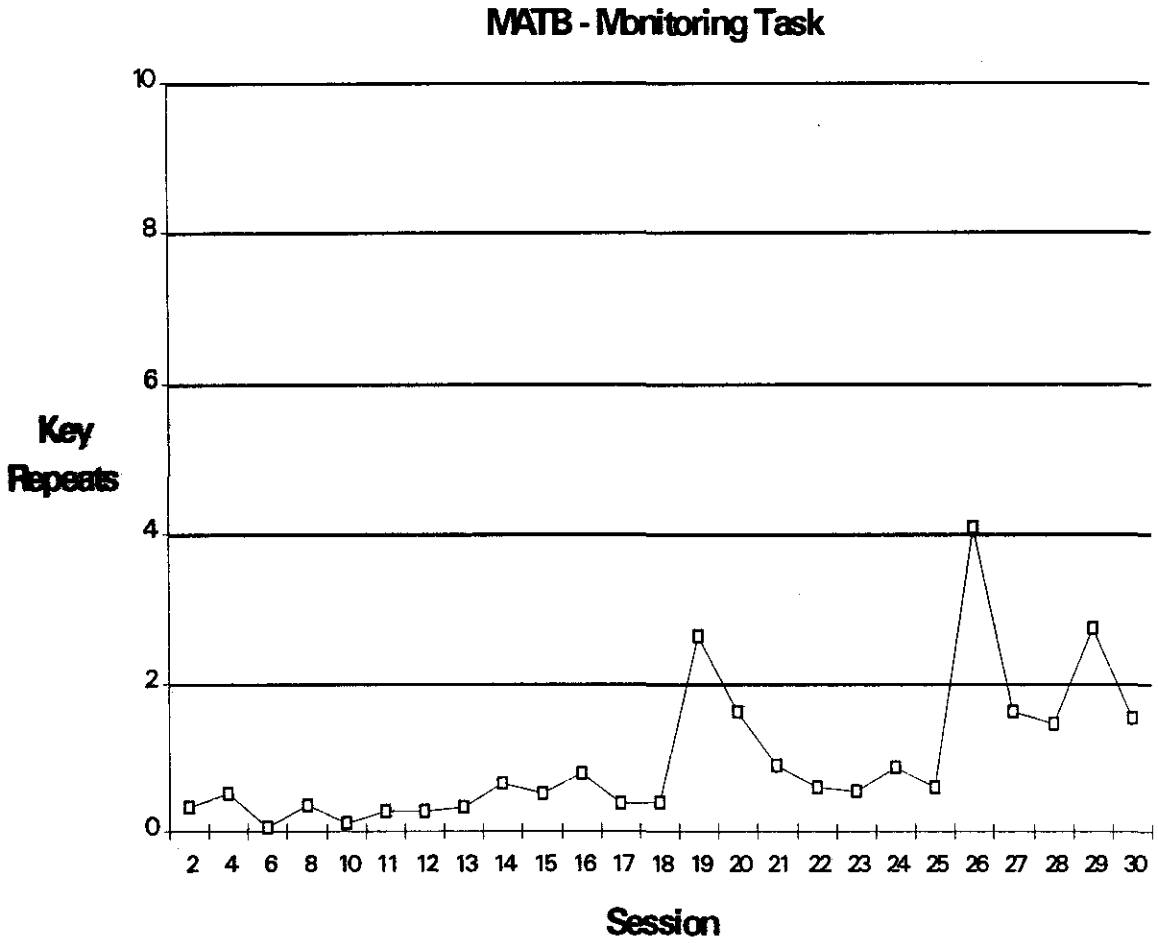

Figure 24. MATB-Monitoring Task (Key Repeats).

As can be seen in Figure 25, the combination of time-out and false alarm errors for lights (LTSER) was very low, as expected, and never exceeded the value 1 beyond the first session. This trend held even after Session 11 when the task length was increased. The time-out and false alarm errors for dials (DLSER) indicates continued improvement from Session 2 to Session 10. After this session, DLSER followed the pattern of DLSFA, as discussed above. Time-out and false alarm errors for lights and dials combined (MONER) was almost identical to DLSER.

Communications Task: Figure 26 illustrates that both mean response time for correct responses (COMCRT) and mean overall response time (COMORT) decreased from 6.4 seconds (Session 2) to approximately 4.0 seconds by Session 14, and thereafter showed only slight improvement. The measures of standard deviation of response time for correct responses (COMCSD) and standard deviation of response time for overall responses (COMOSD), on the other hand, remained constant at approximately 2 seconds throughout the entire study.

All error variables stabilized quickly and remained remarkably stable, as seen in Figure 27. Time-out errors (COMTO) decreased slightly from Session 2 to Session 4, and then remained constant up to Session 10. Following the task length increase at Session 11, COMTO averaged 2.7 errors per session. Five subjects were identified as consistently forgetting to press the "Enter" key after setting the communication frequency, thus generating a time-out (see Figure 28). Removal of these subjects' data resulted in a reduction in COMTO to 0.6 errors per session. Of the other error variables, othership false alarms (COMYFA) and accuracy errors (COMYAC) were essentially zero for all sessions, whereas unexplained errors (COMUNER) were zero for the sessions between 2 and 10, but increased very slightly for the remaining sessions. Ownship accuracy errors (COMAC) stabilized at approximately 1.2 per session. The total number of errors (COMER) followed a pattern similar to COMTO, increasing at Session 11, and stabilizing at a value slightly above four thereafter. Removal of the outlier data reduced the average COMER to 2.3

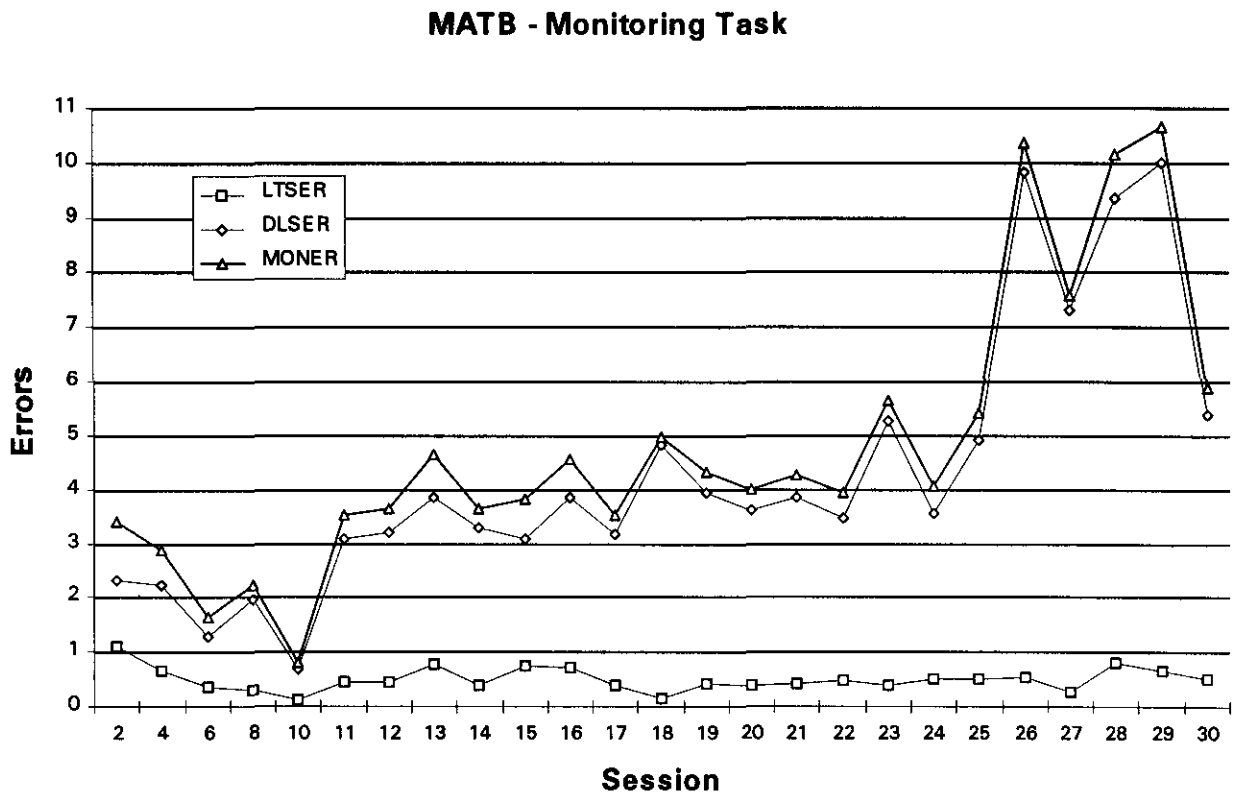

Figure 25. MATB-Monitoring Task (Errors).

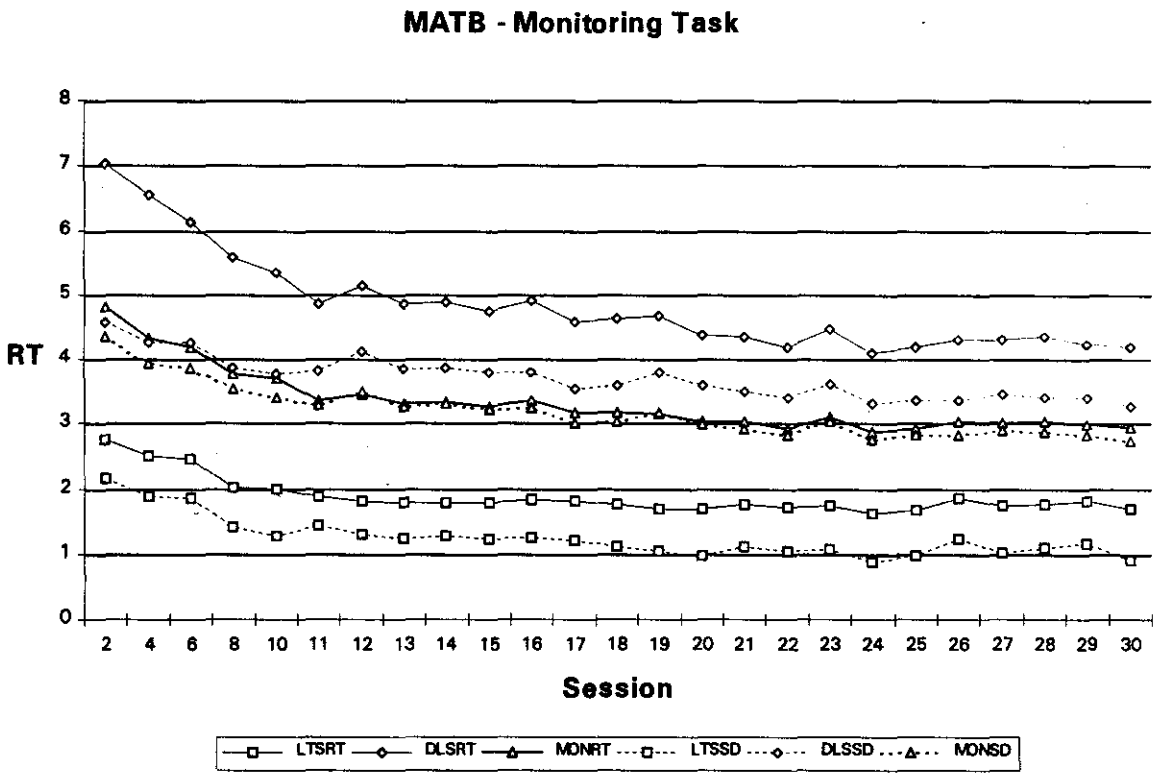

Figure 26. MATB-Communications Task (RT Mean and Standard Deviation).

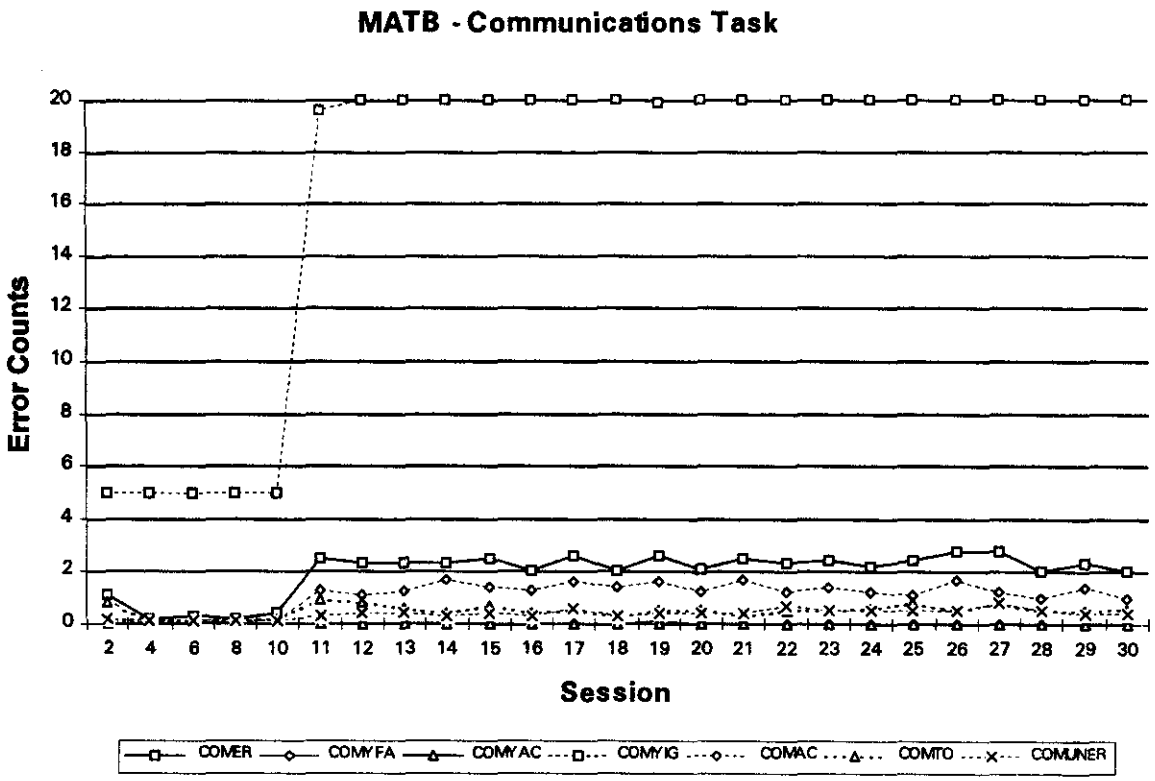

Figure 27. MATB-Communications Task (Error Counts).

errors per session. Finally, the number of othership messages correctly ignored (COMYIG) stabilized at the maximum possible level at each session (i.e., 5 prior to Session 11 and 20 after Session 11), indicating that subjects performed at the best possible level with respect to this variable. In general, error rates for this aspect of the task were quite stable across all sessions.

Tracking Task: As seen in Figure 29, Root Mean Square error (TRKRMS) showed considerable variability across sessions. In fact, there is little evidence of marked improvement within difficulty levels. Following Session 15, RMS error fluctuated erratically, with a peak value of 53.5 at Session 19. A number of variables may account for such erratic performance on this task. First, there appears to be a software problem in the MATB program that occasionally locked out the joystick in all or part of one axis rendering the subject unable to track throughout the two-dimensional tracking array. As a result, a number of subjects were removed from this dataset because of this software/mechanical problem. However, short-term intermittent errors of this type may have gone undetected and may have led to the higher levels of variability seen in later sessions, although this is only speculation. Second, the later sessions were more difficult mainly because they were longer. The longer the test session, the greater the likelihood that other tasks, especially the Resource Management Task, create compounding problems that draw on greater amounts of resources. Finally, even though the subject is required to perform all the tasks within the MATB simultaneously, the Tracking task is the only task that is highly continuous in nature; that is, it requires constant high levels of attention and action. The Resource Management task is continuous, but fairly slow moving, so one can time share easily. The remaining tasks have continual, but relatively low frequencies of events, which make them easy to time share also. Thus, as increased resources are needed to maintain or rescue a failing task, the task most likely affected is the one that demands the most constant level of attention, i.e., the Tracking task. This is especially true in a multi-task environment, such as the MATB, where all tasks are viewed as approximately equal in importance.

Resource Management Task: The mean level of resources for Tank A (TNKAMN) fluctuated considerably for the first three sessions, and then started to increase gradually from a low value of 2410 (Session 6) stabilizing at a level of about 2470 units (see Figure 30). While this was still below the desired 2500, it

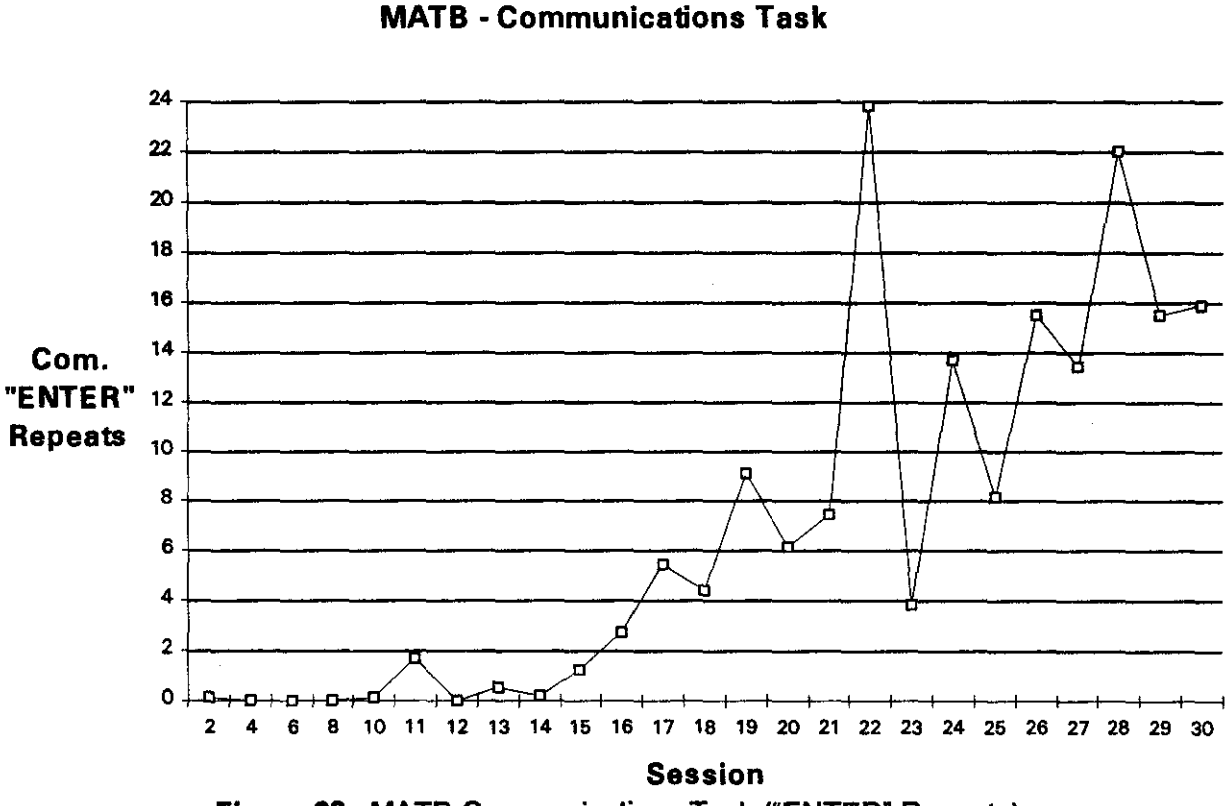

Figure 28. MATB-Communications Task ("ENTER" Repeats).

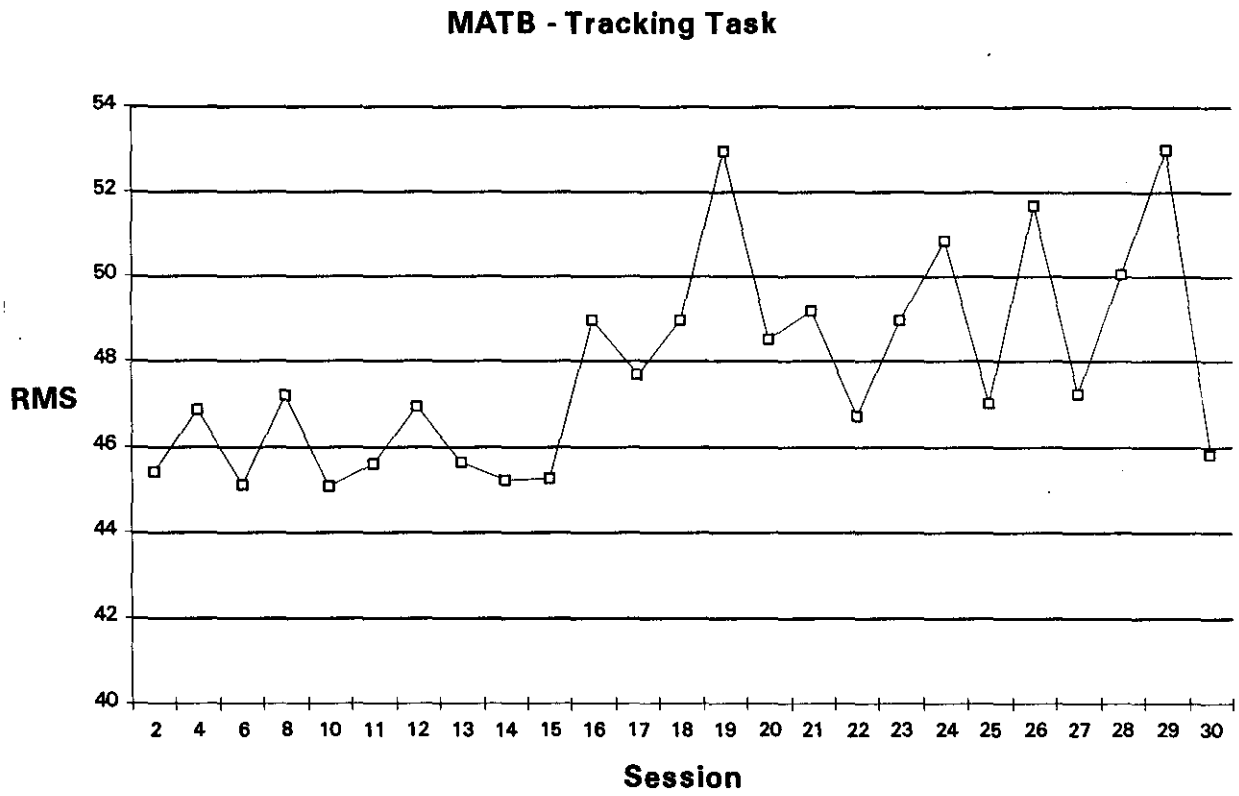

Figure 29. MATB-Tracking (RMS Error).

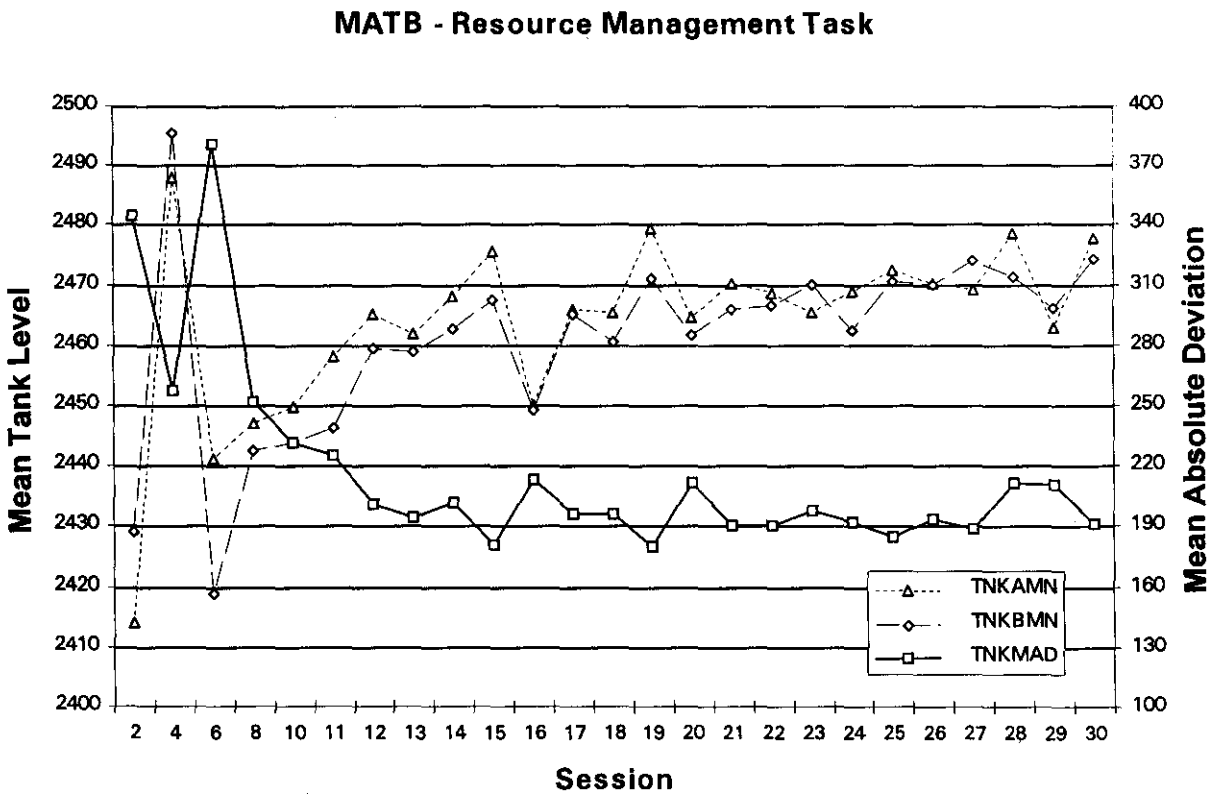

Figure 30. MATB-Resource Management Task (Mean Tank Level and Deviation).

may have been the result of a strategy employed by many subjects to bring the tank levels up to the 2500 level (or somewhat above that), and then let them drain down over an acceptable period of time. This freed the subject to attend to other tasks as opposed to keeping the tank level more close to the 2500 level, but nearly ensures an average score somewhat less than 2500. A similar trend was observed for the mean level of resources for Tank B (TNKBMN). The Mean Absolute Deviation from 2500 units for Tanks A and B (TNKMAD) dropped fairly quickly and showed only slight improvement after Session 13. The lack of any large increases in performance effectiveness after earlier, easier sessions suggests that subjects probably selected a reasonably effective Resource Management strategy fairly early and then made minor improvements in their application of it.

Finally, pump activity (TNKACT), seen in Figure 31, showed a continuous, although modest, increase from the first session (Session 2) through the easier sessions, and then a similar level of increase through the earlier sessions at the more difficult level. By Session 21 it appeared to level off. This performance looks much like the inverse of TNKMAD—that is, subtle, increased, and more effective use of the pumps (TNKACT) leads to slight but continual improvement in TNKMAD. Thus, subjects probably do not evidence a stable level of pump control actions until about Session 21, but the overall level of improvement from early sessions is not great.

5.5 Subjective (Self-Report) Measures

Beginning with the first Work Simulation Session (Session 11), subjects provided daily self-report measures of current physical symptoms (Activity State Questionnaire) and predominant emotional state (Mood Scale II) prior to initiating their test session. Also, during the test session, subjects rated the subjective workload associated with each of the two work simulation tasks (the ATST and MATB). Because these workload ratings are tied to task performance on a specific day, they are more relevant to the comparative task analyses included in Volume II, and will not be addressed here.

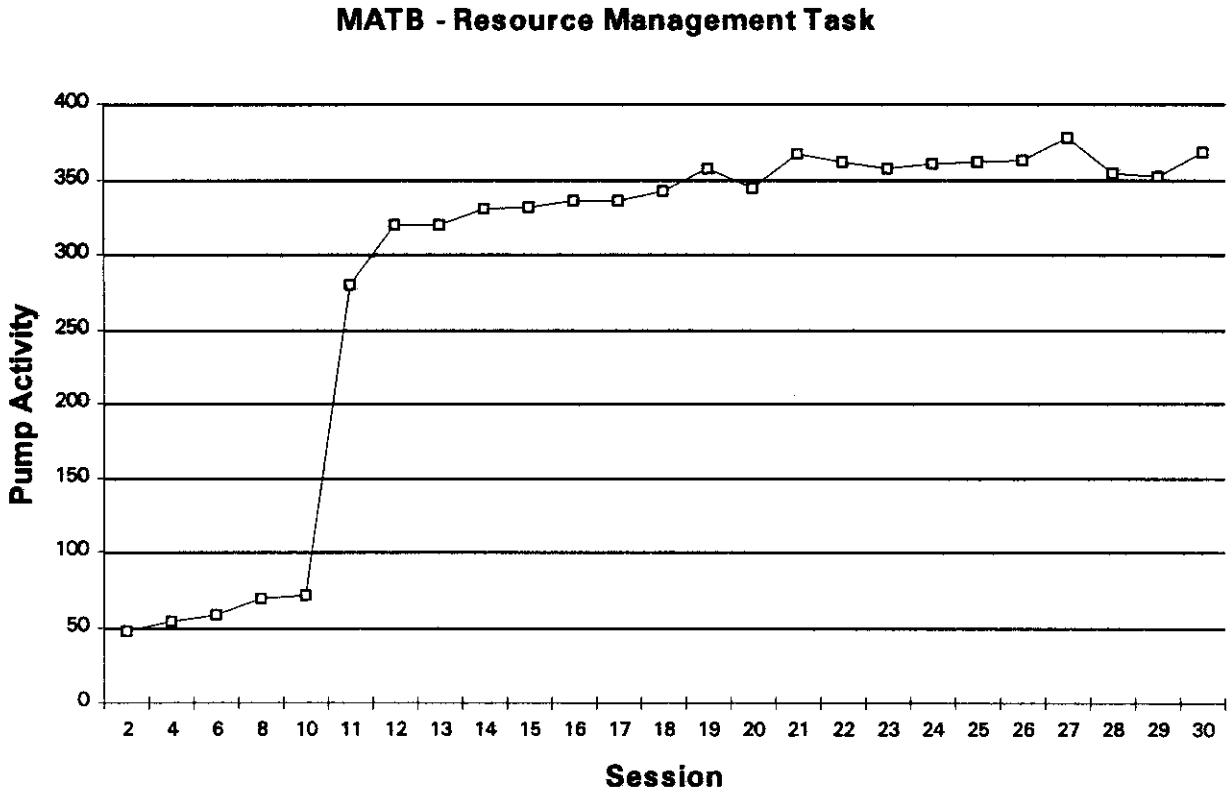

Figure 31. MATB-Resource Management Task (Pump Activity).

Activity State Questionnaire

Figure 32 presents the data obtained from the daily Activity State Questionnaire (ASK). This questionnaire was included as a general measure of physical symptoms experienced by the subjects. The questionnaire includes an expansion of the Pennebaker Symptom/Emotion questionnaire (Pennebaker, 1982). The variable PHYSICAL represents this general scale of physical symptoms. As is clear from the figure, subjects rated themselves relatively consistently across sessions on this scale. There was a slight elevation on the first day, probably related to anxiousness associated with new surrounds and new demands for performance on unfamiliar tasks. Aside from that deviation, there is little in the results that suggest the subjects varied significantly in their average physical condition. These data are also consistent with Pennebaker's (1982) normative data. The present scale was twice as long as the original Pennebaker scale and the subjects' ratings were approximately twice the magnitude of the published means—that is, adjusted for test length, the scores of the subjects in this study are very close to the published normative data for the Pennebaker scale.

In addition, the ASK provided subjects the opportunity to rate their general level of preparedness for performing that day. The PREP scale score (range 2 to 14) shows the subjects' responses (Figure 32) for this measure. The subjects rated themselves clearly above average (average = 8.0) in feeling prepared to perform, and this was fairly consistent across all sessions.

These ratings were important at the group level for two reasons. First, these results suggest that, as a group, the subjects in this study were feeling well and prepared to perform their tasks consistently across the testing sessions. Second, these data support the view that any historical artifacts, including risk factor assessment conducted on weekends, did not adversely affect day-to-day performance during the week. This belief is further supported by the lack of cyclical variation in the performance data that could be associated with weekend testing dates.

Mood Scale II

Subjects also reported their moods daily by responding to adjectives on the Mood Scale II using a 3-point scale before beginning the RTP tasks. A

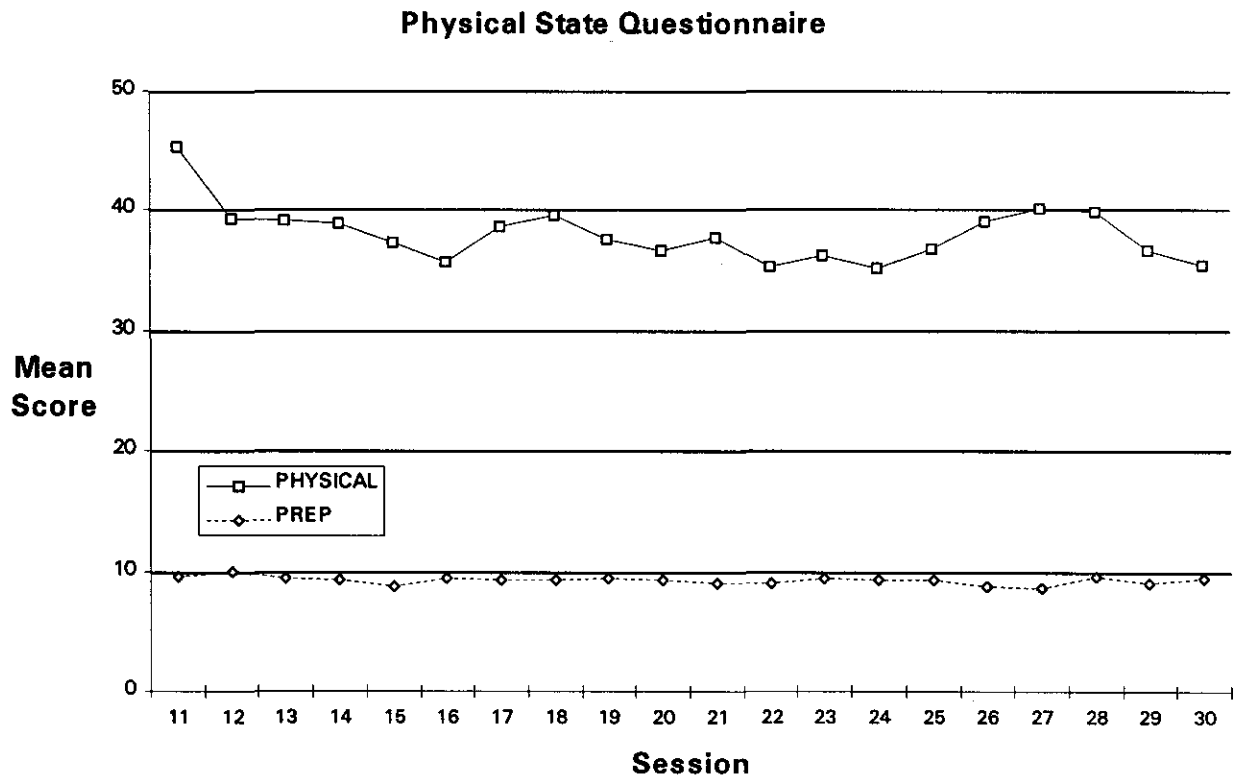

Figure 32. Physical State Questionnaire (Mean Score).

response of "1" indicated that the subject did not feel that the adjective described the current mood, while a response of "3" indicated that the adjective adequately described the subject's mood. The adjectives are divided into six categories (Activity, Happiness, Depression, Anger, Fatigue, and Fear). In general, subjects ranked average in Activity and Happiness. For Activity, the mean scores centered around 2.0 on the 3-point scale. For Happiness, the scores averaged 2.2. On the contrary, the scores for Depression, Anger, and Fear were close to 1 (the lowest possible score), indicating that the subjects were not depressed (1.1), not angry (1.2), and not fearful (1.1). The Fatigue category scores were slightly higher (1.4) and occasionally reached mean values of 1.5 across all subjects. As Figure 33 for Mood Scale II ratings shows, the mean values computed across subjects were fairly consistent throughout the study.

Figure 34 presents the response times of subjects to the mood items. Overall, response times decreased across the first nine sessions that the Mood Scale was administered (Sessions 11 through 19). Subjects apparently became more efficient in answering the mood questions, which was probably a function both of gaining familiarity with the test and of learning to use input keys for the computer more effectively. What is interesting in these data is that it took the subjects 200 to 400 msec longer to respond to the Activity, Happiness, and Fatigue adjectives. It is possible that people with positive attitudes, such as those in this study, can quickly decide that they are not depressed, angry, or fearful, and their responses to the corresponding adjectives are made quickly and automatically. These same subjects take longer to determine the extent of state variables more characteristic of them, such as activity, happiness, and fatigue levels. The longer response times for the Fatigue adjectives provide some evidence that more conscious thought was devoted to these stimuli.

5.6 Intertrial Correlations (Test-Retest Reliability and Differential Stability)

Because RTP testing is usually based on intertrial comparison of performance, using RTP tests with a high degree of reliability is essential. In this regard, RTP test reliability plays an important role in both the integrity and quality of RTP testing. In classical terms, reliability refers to the replicability of a measure, that is, whether a test can be applied over time and provide much the same result (see Lord and Novick, 1968; Guilford, 1954; Gulliksen, 1950; Guttman, 1955). Typical measures of reliability include test-retest techniques in which a test is administered twice, with the intervening time period commonly ranging from 24 hours to several weeks, although it can also be immediately following or as long as several years.

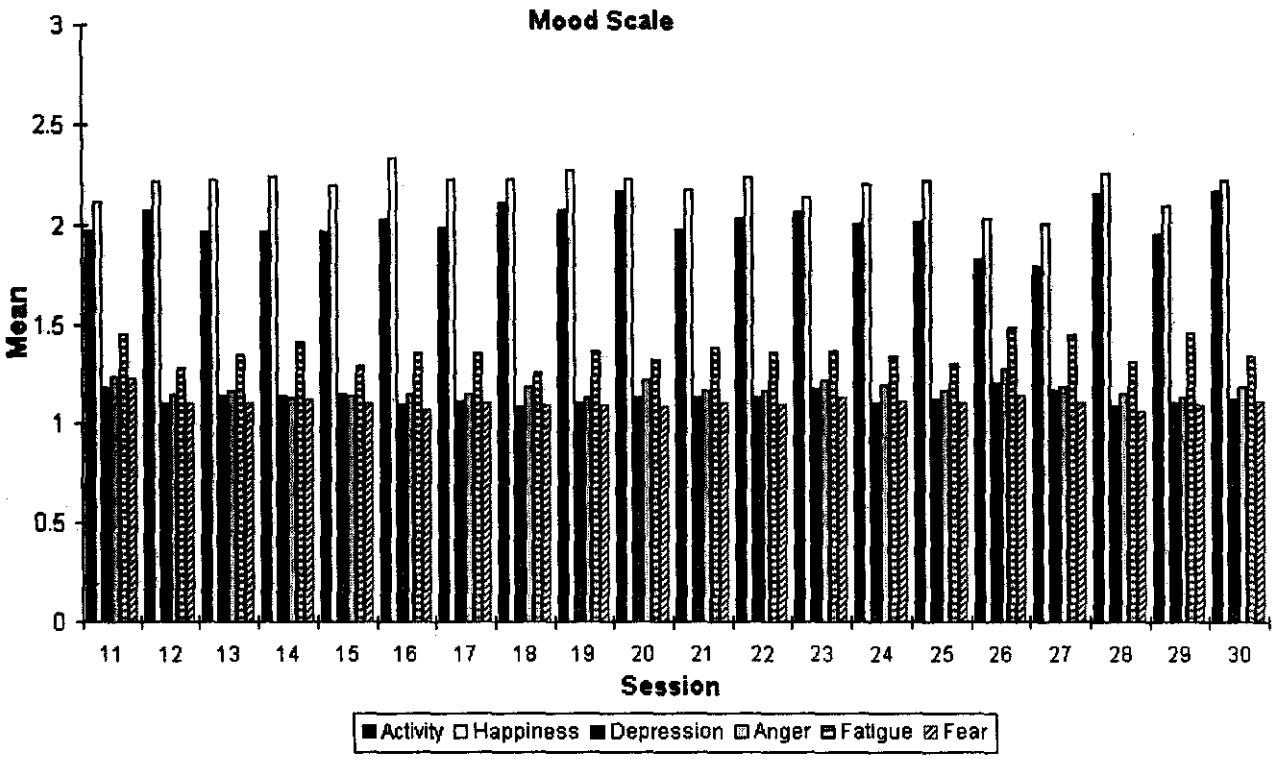

Figure 33. Mood Scale II (Mean Response).

One of the most common indices of reliability is the correlation between test administrations (see Lord and Novick, 1968; Guilford, 1954). This model for estimating reliability is based primarily on a psychometric (or pencil-and-paper) testing perspective in which there is often a general assumption that the test-retest interval is free of activities related to test content (that is, individuals do not practice or review test items between testing sessions). Performance testing presents a uniquely different situation because, during the test-retest interval, individuals often involve themselves in activities related directly or indirectly to those being tested. These additional test sessions or activities related to test performance may affect the correlation derived (Gulliksen, 1950; Guttman, 1955). Thus, it seems important to consider what occurs historically (see Campbell and Stanley, 1971) between test administrations. What is of central importance, of course, is the research question at hand. If one is interested in understanding the enduring nature of some ability or, more likely, some trait as measured by some test, then controlling activities related to the measurement technique may be important—as in personality testing. If one wants to know how reliable a measure is over time *in the presence of continued practice*, then intervening involvement in the test skills would be an important element to include.

More recently, the term "stability" or "differential stability" has entered the dialogue surrounding issues of reliability (see Jones, 1980; Jones, Kennedy, and Bittner, 1981; Kennedy, Carter, and Bittner, 1980). In many cases, the distinction between these concepts is nominal at best. *Reliability* can refer to either internal consistency or repeatability, that is, the stability of a measure. *Stability* is therefore nothing more than one form of reliability (Guilford, 1954). Stability seems to be best understood within the framework of relative comparability across testing sessions. In this sense, stability seems to be that special form (or conceptualization) of reliability that is more easily applied to performance data. *Differential Stability* refers to a more sophisticated approach to establishing the stability of performance data through the analysis of patterns within a correlation matrix of task trials (see Jones, 1980; Jones et al., 1981; Kennedy et al., 1980). Differential stability is achieved when the relative performance between subjects is constant, the day-to-day variability is minimized, and the group mean has overcome most of the learning effect—although some continued improvement may still be seen (see Jones, 1980). Summarized simply, evidence for differential stability is suggested by the presence of a "superdiagonal" form within the correlation matrix—that is, the correlations of early trials with later

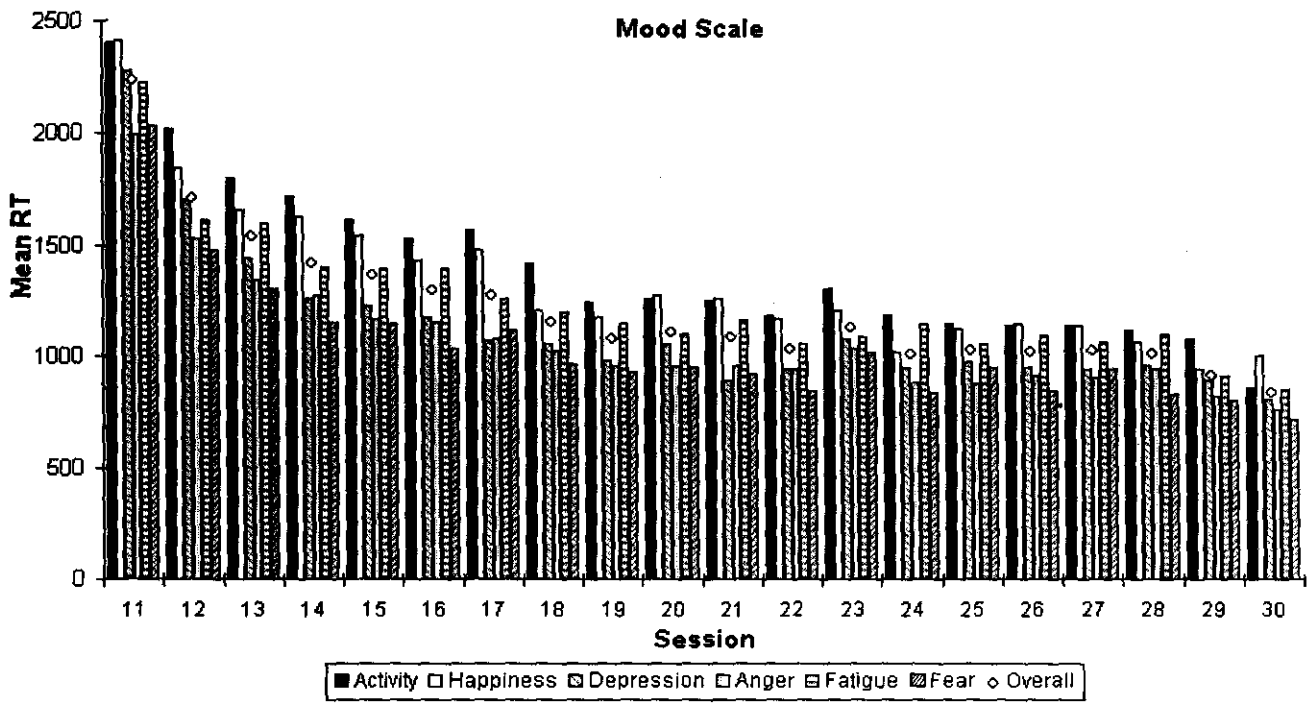

Figure 34. Mood Scale II (Mean Response Time).

trials are lower than the correlations among later trials. It is also the case that the correlations among early trials are lower than the correlation among later trials, and the trial-by-trial correlations across time increase.

What is perhaps overlooked in applications of differential stability analysis is that the nature of the task (and subsequently, the overall performance acquisition curve) may play a very important role in determining whether the pattern in the differential stability correlation matrix actually conforms to the superdiagonal form. Take, for example, the case of percent correct measures where ceiling effects are often found. In these cases, it is quite possible that a classic learning curve will be demonstrated with increasing correlations across trials. Then, due to the ceiling effect and associated lack of variability, repeated correlations over subsequent later trials are very low—perhaps near zero. While this case would violate the conceptual framework for establishing differential stability based on superdiagonal form, all of the subjects are doing well and their performance is very reliable. It is quite possible that this type of measure might be very good in detecting risk factors as well. Furthermore, researchers often know that there will be a learning phase in the acquisition of a task and often know approximately how many trials are needed to overcome this phase (see Schlegel and Gilliland, 1990, 1992). Therefore, in many cases, what is most important in demonstrating differential stability is *a period of time* in which pair-wise correlations among testing sessions are consistently high.

Both reliability and differential stability were assessed in this study. Reliability was assessed by examining trial-to-trial relationships at various testing intervals. Table 6 presents the 24-hour test-retest correlations for each of several dependent measures for each candidate RTP task and each job performance task. These comparisons represent sessions at the end of each week. Thus, Week 1 represents the test-retest reliability estimate at the end of the training period, in most cases, 10 training trials. The values for Week 1 are labeled Sessions 8-10 (instead of sequentially, i.e., 9-10) because during that first training week subjects were performing two testing sessions per day. Session 8 was on Thursday and was compared to Session 10, which was conducted 24 hours later on Friday. The values for other weeks represent reliability estimates for the last two successive days of that week. These values probably represent the best estimates of relatively immediate (24-hour) test-retest reliability.

Table 7 presents test-retest estimates for 48-hour, 1-week, and 2-week time intervals. *It is especially important to note* that, as discussed previously, these reliability estimates are based on test-retest periods throughout which the subjects continued task performance. One should *not* assume that these test-retest correlations represent good estimates of reliabilities between test sessions *without* intervening task involvement or practice. For example, it would be erroneous to assume that these estimates reflect the relationship between two test sessions between which the subject had no practice on the task. It has been clearly shown that task performance declines without continued practice (Schlegel, Shehab, and Gilliland, 1994), which would result in lower reliability coefficients, as compared to time intervals during which subjects had continued practice.

Differential stability was assessed by examining the pattern of grouped correlations. Table 8 presents *averaged pair-wise correlations* across those testing sessions where stable performance was expected—that is, from the end of the training phase through the end of the testing sessions. These average correlations were calculated by taking the mean of all possible pair-wise correlations involving the last three test sessions for each week. By examining these values, one is able to assess an important characteristic of the superdiagonal-correlational-matrix form supporting differential stability (Jones et al., 1981). Specifically, these values represent groups of correlations during a period in which the tasks were believed to be well-learned and performance ought to have been stable. Therefore, according to the concept of differential stability (Jones et al., 1981), these average correlations ought to be relatively high, and they should remain high and relatively consistent across the weeks of the study.

The following sections provide brief summaries for the reliability and differential stability estimates for key performance measures for each candidate RTP measure and for each job task. Interpretations of sporadic fluctuations in these variables may be quite ill advised at this level of analysis. Therefore, these summaries concentrate on the more general trends in these variables. Master tables that include estimates of reliability and differential stability coefficients for all task dependent variables can be found in Appendix B.

Spatial Processing Task

Test-retest reliabilities over 24 hours for the mean correct response time (MNCORRT) measure were quite high across all weeks (see Table 6). The same was

Table 6. Test-Retest Correlations Over 24-Hour Periods.

Task	Measure	Week 1 8-10	Week 2 14-15	Week 3 19-20	Week 4 24-25	Week 5 29-30
Spatial Processing	MNCORRT	0.85	0.92	0.89	0.88	0.93
	SDCORRT	0.58	0.77	0.70	0.64	0.66
	PC	0.34	-0.13	0.48	0.45	0.75
Critical Tracking	MAXL	0.78	0.74	0.71	0.72	0.77
	MEANL	0.84	0.65	0.88	0.85	0.81
	CTLOSS	0.76	0.68	0.89	0.81	0.74
	RMS	0.56	0.46	0.69	0.80	0.78
Dual Task (Group)	CTLOSS	0.48	0.66	0.75	0.75	0.69
	RMS	0.64	0.86	0.81	0.89	0.81
	MNCORRT	0.95	0.96	0.81	0.91	0.87
	PC	0.16	0.35	0.38	0.34	-0.08
	SPEED	0.94	0.96	0.86	0.92	0.89
	THRUPUT	0.92	0.95	0.82	0.90	0.85
Dual Task (Individual)	CTLOSS		0.46	0.71	0.74	0.53
	RMS		0.65	0.75	0.73	0.90
	MNCORRT		0.93	0.76	0.90	0.84
	PC		0.13	0.43	0.32	-0.03
	SPEED		0.90	0.87	0.90	0.89
	THRUPUT		0.87	0.84	0.90	0.84
Switching Task (Manikin)	MANCORRT	0.96	0.97	0.96	0.91	0.96
	MANPC	0.46	0.28	0.30	0.33	0.33
	MANTP	0.97	0.97	0.97	0.96	0.96
	MANCORTX	0.90	0.95	0.89	0.93	0.92
	MANPCX	0.31	-0.25	-0.14	0.46	0.76
Switching Task (Math)	MTHCORRT	0.90	0.96	0.93	0.90	0.95
	MTHPC	0.46	0.13	0.46	0.48	0.11
	MTHTP	0.95	0.97	0.94	0.96	0.97
	MTHCORTX	0.87	0.93	0.92	0.94	0.91
	MTHPCX	0.15	0.26	-0.03	0.57	0.20
NovaScan™ FAA Task	VECCRT	0.73	0.86	0.78	0.91	0.90
	VECPC	0.46	0.66	0.64	0.92	0.91
	VATNPC		-0.10	-0.03	0.61	0.12
	MEMCRT	0.75	0.65	0.88	0.92	0.91
	MEMPC	0.20	0.14	0.41	0.49	0.69
	MATNPC	0.40	-0.11	0.29	0.62	-0.07
Air Traffic Scenarios Test	PCDEST	0.72	0.82	0.65	0.63	0.18
	DELAY	0.38	0.71	0.59	0.83	0.90
	CRSHAC	0.57	0.76	0.67	0.15	-0.05
	CRSHBD	0.44	0.73	0.56	.	.
	CRSHAP	0.33	-0.05	0.14	-0.08	-0.18
	SEPAC	0.77	0.70	0.92	0.63	0.53
	SEPBD	0.33	0.75	0.75	0.28	-0.13
	ERRDEST	-0.11	0.04	0.06	-0.11	-0.08
	ERRGTALT	0.10	-0.08	0.01	-0.04	1.00
	ERRAPALT	0.39	0.40	0.69	0.49	0.18
	ERRGTSPD	-0.10	-0.09	.	.	.
	ERRAPSPD	0.42	0.42	0.72	0.56	0.91
	NDIR	0.61	0.57	0.77	0.83	0.71
	NALT	0.49	0.66	0.86	0.61	0.40
	NSPD	0.60	0.74	0.79	0.65	0.36
Multi-Attribute Task Battery (MATB)	LTSRT	0.35	0.71	0.57	0.90	0.61
	DLSRT	0.41	0.76	0.73	0.91	0.69
	MONRT	0.47	0.76	0.77	0.95	0.67
	LTSFA	-0.03	0.32	-0.13	0.16	0.02
	DLSFA	0.51	0.86	0.89	0.93	0.99
	MONFA	0.41	0.85	0.89	0.92	0.99
	MONER	0.59	0.83	0.78	0.84	0.97
	COMCRT	0.20	0.14	0.41	0.49	0.69
	COMER	0.93	0.80	0.95	0.90	0.90
	TRKRMS	0.95	0.91	0.70	0.84	0.54
	TNKMAD	0.82	0.70	0.85	0.99	0.83
	TNKACT	0.84	0.80	0.93	0.95	0.94

generally true for longer test-retest intervals (see Table 7). In fact, even at the longest interval (two weeks), the test-retest reliability was 0.86, which is fairly remarkable for a performance measure. The test-retest reliabilities for the standard deviation of correct response time measure (SDCORRT) for the Spatial Processing Task generally ranged in the 0.60 - 0.70 level across all time intervals. While not as high as the test-retest reliabilities for MNCORRT, these values represent at least marginally acceptable levels of reliability. The percent correct (PC) measure reliabilities were generally poor and highly variable. This was undoubtedly due to the ceiling effect that was common not only on this measure but also many of the other PC measures for other tasks. When subjects become skilled at tasks that require some component of accuracy they often emphasize this aspect of the tasks and therefore score at or near 100% in accuracy, especially if the task is not complex. The lack of variability in the scores leads to severe reductions in the correlations across test sessions or, as in this case, even negative correlations.

Evidence for differential stability can clearly be seen in the average correlations across weeks of the study for MNCORRT (see Table 8). Across the five-week period, the stability value ranged near the 0.90 level. The pattern of stability values for the SDCORRT variable was slightly more consistent than that seen for the test-retest reliabilities, and would probably be considered marginally acceptable for differential stability as well. The pattern of stability values for the PC measure was somewhat more consistent than the test-retest reliabilities for this measure, but they were too low and too variable to conclude differential stability in traditional terms. This is clear from a visual inspection of the data (see Figure 2).

Critical Tracking Task

The Critical Tracking Task had 24-hour test-retest reliabilities that were reasonably acceptable for the maximum lambda (MAXL), mean lambda (MEANL), and control losses (CTLOSS) measures (see Table 6). With only a few exceptions, these reliability coefficients remained in an acceptable to good range (0.71 to 0.89). The correlation coefficients for mean lambda (MEANL) were somewhat better than those for MAXL. This result was expected because MAXL represents only one data point from each trial, and thus exhibits greater random variation, whereas MEANL is the average of the lambda values for several control losses during the trial. The correlation coefficient for control losses (CTLOSS) was lowest for week two (0.68), reflecting the fact that almost all subjects stabilized between 9 and 13 control losses per trial following the first week (see Table 6). The correlation coefficients for root mean square error (RMS) were reasonably good for weeks four and five (0.77 and 0.73), but relatively low for the first three weeks (0.38 to 0.66).

The general trend of the reliability coefficients for extended time periods (see Table 7) was similar to the 24-hour reliabilities. However, individually, these coefficients were not as high overall. In most cases, these reliabilities would fall in a marginally acceptable category.

Average correlations across the five-week period (see Table 8) suggest that the MEANL measure evidenced reasonable differential stability. Both the MAXL and CTLOSS measures could probably be viewed as marginally acceptable with regard to their differential stability. The RMS measure failed to provide convincing levels of differential stability.

Dual Task - Group Lambda

Tracking: The 24-hour, test-retest correlation for tracking control losses (CTLOSS) was low at the end of week one (see Table 6), but somewhat higher for the other weeks (0.66 to 0.75). These latter reliability coefficients were marginally acceptable. The low value for week one reflects the zero, or near-zero, number of control losses scored by almost all subjects (91%) by the end of this week ("floor effect"). During the second week, the increase in lambda value at Session 13 helped differentiate subject performance, and this was reflected in the higher correlations across the remaining weeks. The RMS error reliability coefficients were generally quite good during the last four weeks (see Table 6). Again, the lower coefficient during week one was probably due to the lack of variability associated with lower task difficulty. Test-retest reliabilities for longer intervals (see Table 7) were quite erratic for the CTLOSS measure but were very good for the RMS measure.

Differential stability for RMS was also very good (highest average correlation coefficient = 0.87 for week four), indicating that RMS error was a more stable measure than CTLOSS, which suffered from the floor effect and had only marginally acceptable reliability coefficients (see Table 8).

Memory Search: High 24-hour, test-retest reliability coefficients were generally obtained for memory search MNCORRT for all weeks (see Table 6). Curiously, the longer-interval test-retest reliabilities were quite variable (see Table 7). It was observed that any correlation involving Session 23 was unusually low in comparison to the correlations not involving that session. The low correlations were traced to the unusually slow performance of Subject 231 on Session 23 (1761 msec vs. 717 msec). Removal of this subject's

Table 7. Test-Retest Correlations Over 48-Hour, One-Week, and Two-Week Periods.

		48-Hour			1-Week		2-Week
		Week 3	Week 4	Week 5	Week 3-4	Week 4-5	Week 3-5
Task	Measure	18-20	23-25	28-30	18-23	23-28	18-28
Spatial Processing	MNCORRT	0.84	0.85	0.93	0.84	0.89	0.86
	SDCORRT	0.62	0.56	0.76	0.68	0.76	0.77
	PC	0.52	0.32	0.51	0.06	0.37	0.32
Critical Tracking	MAXL	0.58	0.73	0.73	0.64	0.65	0.62
	MEANL	0.81	0.78	0.72	0.70	0.61	0.68
	CTLOSS	0.77	0.68	0.60	0.66	0.42	0.65
	RMS	0.61	0.74	0.55	0.56	0.56	0.33
Dual Task (Group)	CTLOSS	0.66	0.82	0.48	0.75	0.64	0.54
	RMS	0.87	0.85	0.87	0.82	0.89	0.81
	MNCORRT	0.86	0.59	0.78	0.57	0.58	0.73
	PC	0.51	0.07	0.63	0.20	0.36	0.46
	SPEED	0.91	0.85	0.87	0.83	0.79	0.80
	THRUPUT	0.89	0.85	0.87	0.84	0.77	0.78
Dual Task (Individual)	CTLOSS	0.35	0.39	0.51	0.32	0.36	0.39
	RMS	0.71	0.50	0.81	0.69	0.58	0.64
	MNCORRT	0.75	0.96	0.76	0.30	0.41	0.86
	PC	0.59	0.21	0.08	0.15	0.51	0.49
	SPEED	0.78	0.88	0.83	0.68	0.72	0.83
	THRUPUT	0.76	0.85	0.83	0.64	0.71	0.82
Switching Task (Manikin)	MANCORRT	0.94	0.97	0.96	0.89	0.85	0.90
	MANPC	0.34	0.67	-0.15	0.24	0.31	0.59
	MANTP	0.96	0.96	0.96	0.96	0.90	0.91
	MANCORTX	0.90	0.96	0.93	0.92	0.84	0.81
	MANPCX	-0.08	0.60	-0.22	0.21	-0.10	-0.03
Switching Task (Math)	MTHCORRT	0.88	0.90	0.95	0.88	0.88	0.89
	MTHPC	0.39	0.09	0.32	0.25	0.25	0.43
	MTHTP	0.91	0.97	0.97	0.92	0.95	0.92
	MTHCORTX	0.83	0.87	0.91	0.81	0.86	0.77
	MTHPCX	0.06	0.19	-0.17	0.39	0.39	0.67
NovaScan™ FAA Task	VECCRT	0.84	0.87	0.81	0.77	0.83	0.81
	VECPC	0.75	0.93	0.66	0.84	0.90	0.78
	VATNPC	0.08	-0.08	-0.11	-0.07	0.23	0.28
	MEMCRT	0.74	0.68	0.94	0.74	0.71	0.87
	MEMPC	0.60	0.38	0.43	0.61	0.64	0.51
	MATNPC	-0.12	0.07	-0.10	0.13	0.32	0.16
Air Traffic Scenarios Test	PCDEST	0.73	0.50	0.08	0.86	0.42	0.38
	DELAY	0.70	0.79	0.91	0.90	0.56	0.56
	CRSHAC	0.49	0.01		-0.07		
	CRSHBD	0.61			0.36		
	CRSHAP	0.28	0.05	-0.02	-0.11	-0.11	0.10
	SEPAC	0.97	0.57	0.03	0.92	0.24	0.14
	SEPBD	0.87	0.34	-0.13	0.65	0.02	0.11
	ERRDEST	-0.11	-0.05	-0.08	-0.05	-0.04	-0.05
	ERRGTALT	-0.16	-0.06	1.00	0.23	0.81	-0.07
	ERRAPALT	0.71	0.61	0.42	0.45	0.64	0.54
	ERRGTSPD				-0.04		
	ERRAPSPD	0.74	0.26	0.92	0.45	0.19	0.34
	NDIR	0.81	0.74	0.72	0.77	0.54	0.48
	NALT	0.75	0.81	0.66	0.68	0.76	0.63
	NSPD	0.65	0.44	0.67	0.72	0.27	0.45
Multi-Attribute Task Battery (MATB)	LTSRT	0.69	0.87	0.73	0.45	0.45	0.53
	DLSRT	0.73	0.82	0.78	0.71	0.59	0.65
	MONRT	0.77	0.86	0.75	0.69	0.52	0.66
	LTSFA	0.08	0.32	0.26	-0.14	0.14	-0.06
	DLSFA	0.95	0.99	0.99	0.68	0.95	0.49
	MONFA	0.95	0.99	0.99	0.67	0.95	0.48
	MONER	0.94	0.96	0.98	0.65	0.93	0.46
	COMCRT	0.60	0.38	0.43	0.85	0.86	0.85
	COMER	0.93	0.92	0.90	0.97	0.92	0.96
	TRKRMS	0.79	0.87	0.75	0.88	0.83	0.77
	TNKMAD	0.95	0.93	0.87	0.80	0.87	0.93
	TNKACT	0.95	0.96	0.96	0.88	0.94	0.86

Table 8. Average Intertrial Correlations for Differential Stability Analysis.

Task	Measure	Week 1 Ave 8-10	Week 2 Ave 13-15	Week 3 Ave 18-20	Week 4 Ave 23-25	Week 5 Ave 28-30
Spatial Processing	MNCORRT	0.88	0.90	0.87	0.88	0.91
	SDCORRT	0.59	0.69	0.69	0.64	0.71
	PC	0.38	-0.03	0.38	0.38	0.57
Critical Tracking	MAXL	0.75	0.63	0.69	0.70	0.70
	MEANL	0.87	0.69	0.83	0.80	0.73
	CTLOSS	0.80	0.68	0.80	0.72	0.60
	RMS	0.66	0.38	0.61	0.77	0.73
Dual Task (Group)	CTLOSS	0.14	0.74	0.69	0.78	0.68
	RMS	0.65	0.85	0.81	0.87	0.84
	MNCORRT	0.94	0.87	0.82	0.67	0.82
	PC	0.10	0.36	0.44	0.22	0.22
	SPEED	0.94	0.90	0.88	0.86	0.88
	THRUPUT	0.92	0.89	0.86	0.85	0.85
Dual Task (Individual)	CTLOSS		0.38	0.52	0.53	0.54
	RMS		0.68	0.71	0.60	0.84
	MNCORRT		0.90	0.76	0.92	0.80
	PC		0.33	0.40	0.20	0.06
	SPEED		0.86	0.84	0.88	0.86
	THRUPUT		0.85	0.80	0.87	0.84
Switching Task (Manikin)	MANCORRT	0.94	0.97	0.95	0.92	0.95
	MANPC	0.50	0.26	0.33	0.46	0.04
	MANTP	0.97	0.97	0.96	0.95	0.96
	MANCORTX	0.92	0.95	0.91	0.93	0.92
	MANPCX	0.40	-0.03	-0.05	0.50	0.12
Switching Task (Math)	MTHCORRT	0.91	0.95	0.91	0.90	0.95
	MTHPC	0.57	0.12	0.46	0.19	0.22
	MTHTP	0.96	0.96	0.94	0.96	0.96
	MTHCORTX	0.83	0.89	0.87	0.89	0.90
	MTHPCX	0.39	0.06	0.14	0.29	0.11
NovaScan™ FAA Task	VECCRT	0.77	0.87	0.85	0.90	0.80
	VECPC	0.63	0.78	0.73	0.92	0.80
	VATNPC	-0.09	-0.07	0.09	0.15	0.16
	MEMCRT	0.83	0.72	0.84	0.79	0.93
	MEMPC	0.14	0.28	0.52	0.51	0.51
	MATNPC	0.19	0.12	0.26	0.17	0.09
Air Traffic Scenarios Test	PCDEST	0.72	0.82	0.69	0.61	0.14
	DELAY	0.38	0.60	0.64	0.82	0.90
	CRSHAC	0.57	0.54	0.60	0.04	-0.05
	CRSHBD	0.44	0.62	0.64		
	CRSHAP	0.33	0.01	0.12	-0.03	-0.03
	SEPAC	0.77	0.68	0.93	0.70	0.27
	SEPBD	0.33	0.53	0.80	0.30	-0.12
	ERRDEST	-0.11	0.13	0.03	-0.08	0.28
	ERRGTALT	0.10	0.29	0.05	0.24	1.00
	ERRAPALT	0.39	0.38	0.66	0.56	0.29
	ERRGTSPD	-0.10	0.03	0.42		
	ERRAPSPD	0.42	0.50	0.75	0.49	0.92
	NDIR	0.61	0.68	0.78	0.74	0.74
	NALT	0.49	0.61	0.79	0.72	0.59
	NSPD	0.60	0.74	0.71	0.57	0.55
Multi-Attribute Task Battery (MATB)	LTSRT	0.39	0.74	0.62	0.90	0.61
	DLSRT	0.28	0.73	0.76	0.86	0.70
	MONRT	0.39	0.78	0.79	0.90	0.68
	LTSFA	-0.06	0.14	0.05	0.21	0.07
	DLSFA	0.59	0.85	0.92	0.95	0.99
	MONFA	0.48	0.85	0.92	0.95	0.99
	MONER	0.63	0.79	0.84	0.90	0.98
	COMCRT	0.14	0.28	0.52	0.51	0.51
	COMER	0.95	0.84	0.94	0.92	0.92
	TRKRMS	0.94	0.92	0.65	0.83	0.62
	TNKMAD	0.80	0.78	0.88	0.95	0.82
	TNKACT	0.87	0.85	0.94	0.95	0.96

data drastically increased the correlations involving Session 23 (ranging from 0.80 to 0.88) without changing any of the correlations for the other sessions. The average correlations across weeks in Table 8 confirm that MNCORRT was a reliable and stable measure across the five-week period, and that it was the data of one subject in Session 23 that appeared to distort the findings. The reliability coefficients for all levels of analysis for the total number of responses per minute (SPEED) and for throughput (THRUPUT) were uniformly high, suggesting both high reliability and differential stability across the five-week period.

The PC measure demonstrated very low and variable 24-hour test-retest correlations (lowest = -0.08 for week five, highest = 0.38 for week three). Once again, this result was due to the PC ceiling effect discussed for spatial processing. Low correlations of the same magnitude were also obtained for longer test-retest intervals (see Table 7) and averaged correlations (see Table 8).

Dual Task - Individual Lambda

Tracking: In general, the reliability coefficients for the individual lambda version of the Dual-Task were somewhat lower in comparison to the group lambda version. The individual lambda version is adjusted to equalize the relative difficulty of the task across subjects. As a result, this equalizes the general task difficulty for all subjects, thereby producing lower performance differentiation among subjects and a lower correlation in any intertrial correlation. As with the group lambda version of the dual task, fairly low and erratic 24-hour test-retest correlations were obtained for CTLOSS (see Table 6). These values dropped substantially for the longer test-retest intervals (see Table 7) and the averaged correlations (see Table 8). For RMS error, reasonably good reliability coefficients were derived for the 24-hour test-retest intervals (see Table 6), but the longer test-retest intervals and the averaged correlations were quite low (see Tables 7 and 8, respectively).

Memory Search: In general, high 24-hour test-retest correlation coefficients were obtained for MNCORRT. For this measure, the lowest coefficient was observed for week three (0.76) and the highest coefficient for week four (0.93). On the contrary, the one-week test-retest correlations between weeks three and four (0.30), and between weeks four and five (0.41) were not good. However, a high two-week correlation was demonstrated between weeks three and five (0.86). As with the group lambda version, the low correlations were traced to poor performance of Subject 231 on Sessions 22 through 25. The generally high level of reliabilities seen across the averaged correlations (see Table 8) suggests that the MNCORRT measure was both reliable and differentially stable.

The PC measure exhibited very low 24-hour test-retest correlations (highest = 0.43 for week three) due to the ceiling effect (see Table 6). The same was true for the longer interval test-retest correlations and averaged correlations (see Tables 7 and 8, respectively).

High 24-hour test-retest correlations were obtained for both SPEED and THRUPUT (see Table 6). Correlation coefficients for SPEED ranged from 0.87 (week three) to 0.90 (week two and four), and for THRUPUT from 0.84 (weeks three and five) to 0.90 (week four). Somewhat lower, but generally quite acceptable, longer interval test-retest correlations were obtained for both SPEED and THRUPUT. In general, the magnitudes of correlations for the Memory Search measures for the individual lambda version were comparable to those for the group lambda version.

Switching

Very high 24-hour test-retest correlation coefficients were obtained for a number of performance measures for the Switching Task. These include: the mean response time for correct responses for both the Manikin (MANCORRT) and the Mathematical Processing (MTHCORRT) tasks, the Manikin throughput (MANTP) and Mathematical Processing throughput (MTHTP), and the transition response time for correct responses for the Manikin (MANCORTX) and the Mathematical Processing (MTHCORTX) tasks (see Table 6). These high levels of reliability were also reflected in longer interval test-retest correlations and averaged correlations across weeks (see Tables 7 and 8, respectively).

Neither the percent correct nor transition percent correct measures for the Manikin (MANPC and MANPCX) or Mathematical Processing (MTHPC and MTHPCX) tasks demonstrated high reliability coefficients of any kind. These values were also quite inconsistent across all test-retest intervals and across the averaged sessions.

NovaScan™

In general, good 24-hour reliability coefficients (see Table 6) were obtained for the response time measures for the two NovaScan™ subtasks. In particular, mean correct response time for the Visual Search and Vector Projection task (VECCRT) exhibited strong correlations above 0.85, except for week one (0.73), which was still acceptable. Similarly, mean response time for the Continuous Spatial Memory task (MEMCRT) presented high correlations for weeks three through five. Somewhat lower reliability values

were found for weeks one (0.75) and two (0.65). The reliability coefficients for longer test-retest intervals (see Table 7) were quite good for the VECCRT measure, but were far less consistent, although probably still acceptable, for the MEMCRT measure. Both of these measures provided acceptable patterns of averaged correlations, suggesting acceptable levels of differential stability (see Table 8).

The 24-hour reliability coefficients for the visual search and vector projection percent correct (VECPC) reached very satisfactory levels only in weeks four and five (see Table 6). Longer interval test-retest reliabilities, which were generally based on these latter sessions, suggest fairly good reliability levels (see Table 7). The trend across weeks for the averaged correlations also suggests that differential stability seems to be developed at acceptable levels by week two and at much better levels by week four (see Table 8). The continuous spatial memory task percent correct measure (MEMPC) yielded fairly low reliability values across all conditions.

Very low reliability estimates were observed for the remaining task measures. Correlation coefficients for percent correct attention acknowledgments during the visual search and vector projection task (VATNPC) were generally quite low in all cases. Low values were also obtained for the percent correct attention acknowledgments during the continuous spatial memory task (MATNPC). The very low correlations for these two variables are due to the ceiling effect discussed previously.

Air Traffic Scenarios Test

In general, none of the performance measures of the Air Traffic Scenarios Test (ATST) provided the unquestionable levels of reliability coefficients that were seen for several of the candidate RTP tasks. However, it should be remembered that the ATST and the Multi-Attribute Task Battery were selected as *job performance* tasks, not RTP tasks. Thus, they ought to have broader demands on performance resources and, likewise, broader variability (see Section 6.0). These job performance tasks were carefully selected because they did have fairly well-defined criterion measures for performance. Then again, it may be unrealistic to expect that more global tasks that integrate broader combinations of cognitive and psychomotor skills would be able to provide highly refined and highly reliable outcome (job performance) measures. For example, the complex nature of the ATST task would naturally lead to considerable variation, even within subjects. Also, even though each scenario is scripted, the "downstream" outcome can be considerably different given its stochastic nature. Finally, the session-to-session correlations are based on similar, but not identical, scenarios. These various factors would help to explain why the test-retest correlations for various measures of performance on this task are not as high as those for tasks assessing more basic processes, such as the candidate RTP tests.

The 24-hour reliability coefficients for the percentage of planes at destination (PCDEST) were relatively high for the first two weeks (0.72, 0.82) and then decreased from 0.65 for week three to 0.18 for week five. This reduction can be explained by a ceiling effect in which subjects achieved 95% to 98% during the last week. This effect probably accounts for the relatively low correlations obtained for a number of the longer test-retest intervals (see Table 7) and average correlations in later weeks (see Table 8). In general, poor differential stability was observed for the PCDEST variable. This is because most of the subjects reached high levels of performance with respect to this measure and then only occasionally committed errors.

On the contrary, correlations for the delay score for planes arriving at the destination (DELAY) improved from a low 0.38 for week one to a high 0.90 for week five (see Table 6). During week one, the simple, short scenarios were such that the delay score was more dependent on the scenario characteristics than on subject skill. This changed as scenarios became more difficult and individual subject skill emerged. Test-retest correlations over longer intervals were more encouraging, especially for weeks three, four, and five, as the 24-hour data would suggest (see Table 7). This was also one of the few ATST measures that began to show evidence of differential stability (see Table 8).

Poor correlation coefficients were also demonstrated for the "number of crashes" variables. Once again, these low correlations are due to the floor effect discussed previously. Poor correlations were also found for most of the "error" measures as well.

Relatively high 24-hour test-retest correlations were obtained for separation errors for aircraft (SEPAC) in the first few weeks, and for the measures reflecting the number of control actions taken by the subject (mouse clicks). For the number of direction changes (NDIR), the correlation for week one was 0.61, and for the later weeks ranged from 0.57 to 0.83. For the number of altitude changes (NALT), marginal correlations were obtained only for weeks two and three. For the remaining weeks, correlations were rather low. This indicates that the number of altitude changes became more consistent across all subjects. Many other variables simply failed to have any real pattern of correlations or correlations of sufficient magnitude to suggest reasonable levels of reliability.

Multi-Attribute Task Battery

Monitoring Task: The 24-hour test-retest correlation coefficients for response times were in general low for week one and higher for the other weeks. For mean response time for lights (LTSRT), correlation coefficients attained variable, but encouraging, levels for weeks two through five. For mean response time for dials (DLSRT), relatively good correlations were obtained over the same period. A similar pattern of coefficients was obtained for mean response time for lights and dials combined (MONRT). Very poor correlations were derived for false alarm errors for lights (LTSFA). For this measure, the highest correlation coefficient was only 0.32. This resulted because almost all subjects were able to achieve zero false alarms. However, correlations for false alarm errors for dials (DLSFA) were considerably higher. Except for week one, for which the correlation was 0.51, correlations were between 0.85 and 0.99. These correlations were a result of the subjects who generated numerous false alarms throughout many sessions as discussed in Section 5.4. As a result, correlations for false alarms for lights and dials combined (MONFA) were also high (except for week one). For this same reason, very high correlations were obtained for all errors combined for lights and dials (MONER). For this variable, correlation coefficients increased from 0.59 for week one to 0.97 for week five.

The longer interval test-retest reliabilities were generally similar to the trends for the 24-hour values. LTSRT, DLSRT, MONRT, DLSFA, MONFA, AND MONER all had very good 48-hour reliabilities. A few of these variables demonstrated greater variability over longer test-retest intervals (see Table 7), but most of these measures also showed very good differential stability over weeks two through five (see Table 8).

Communications Task: Low reliability correlations were generally demonstrated for the measure of mean response time for correct responses (COMCRT). For this measure, few correlation coefficients ever exceeded the 0.5 level, except for some of the longer test-retest intervals, which may have been nothing more than sampling error. On the contrary, correlation coefficients for total number of errors (COMER) were consistently high across all cases due to the number of subjects who consistently forgot to press the "ENTER" key.

Tracking Task: The 24-hour intertrial correlation coefficients for root mean square (TRKRMS) were high (see Table 6), as were the longer interval test-retest values (see Table 7). The average correlation values were high but variable, signifying a reasonable level of differential stability across some weeks. For weeks three and five, however, correlations were unacceptable (0.65 and 0.62 respectively).

Resource Management Task: High correlation coefficients were obtained in nearly all cases for the mean absolute deviation of tanks A and B from 2500 (TNKMAD) and the measure of tank activity (TNKACT). These measures showed some of the highest and most consistent levels of reliability and differential stability of all measures, including the candidate RTP measures.

6.0 DISCUSSION

The main objective of this project was to provide the FAA with a large-scale, highly controlled, laboratory investigation exploring the use of RTP testing. The major issues addressed by this volume of the report were the establishment of learning rate information for the candidate RTP tests and job tasks, and the examination of both candidate RTP test and job task reliability. This information was important for two reasons. Little is known about the nature of skill acquisition (or learning rate) for many of the RTP measures that are currently available. Many of the RTP tests that are commercially available provide little empirical data on training requirements or reliability. Even many laboratory tasks that are conceptually related to RTP tests do not have well-established data on training requirements. Thus, this information was viewed as important for establishing a clear understanding of the basic integrity and the dynamics that regulate the learning process for the various tasks used in this project. This information was also important because it validates the integrity of the basic laboratory model approach to this project. Only if it can be established that the tasks used in this study were well practiced and provided reasonable levels of reliability and stability could confidence be placed in the overall results of the RTP laboratory investigation.

The results presented in this volume summarize an exceptionally large data collection effort. Considerable time was needed to simply inspect, review, and reduce the data set to a form that could be analyzed. Additional time was needed to analyze the data and transform these findings into figures and tables. This phase alone required over 500,000 statistical calculations. Subsequently, even more time was then needed to distill these results into a comprehensible form that would not require hundreds of pages of narrative. This process required the visual inspection and summarization of nearly 60,000 statistical values. The

result of this analysis was the selection of a subgroup of task measures. Basic statistics and a more thorough description of the learning process based on visual analyses were provided for each of the task measures in the subgroup. (More extensive descriptive statistics for all (N=150+) task measures are included in the appendices.) This analysis included a number of test-retest reliability estimates that were calculated for each task measure, as well as differential stability analyses.

The results of this extensive data reduction and analysis effort yielded important findings regarding both the candidate RTP measures and the more complex job tasks. Based on the learning curve analysis, it was found that considerable amounts of learning took place for most of the candidate RTP tasks by the tenth training session and, in many cases, even sooner. A few tasks required a few additional sessions, but certainly major learning effects were overcome for nearly all of these additional task measures by the middle of the second week. It was also the case that nearly all task measures showed some continued improvement, even after five full weeks of experience. For the most part, this continued improvement was considerably less than the improvement seen during the early learning period. Also, task measures varied considerably in the amount of continued learning. Some of the simpler laboratory-based tasks, such as Spatial Processing, Tracking, and the Dual Tasks, showed only modest additional improvement. As might be expected, measures of tasks requiring more complex or integrated cognitive skills saw more learning over the latter sessions in the study (i.e., Mathematical Processing, NovaScan™ subtests, and the ATST), as compared to the simpler tasks.

Of serious concern was whether this additional learning occurring in later test sessions compromised the reliability of these task measures or their ability to be used for the comparative purposes needed in the laboratory model portion of the project. The reliability and differential stability analyses provided clarity in this regard.

An examination of the test-retest correlation coefficients for the various task measures revealed that many of the RTP tests provided surprisingly reliable performance measures. All four of the laboratory-based candidate RTP measures provided multiple, highly repeatable measures that appeared to be both reliable and stable over the latter four weeks of the study. In general, the same was true for the two commercially-based candidate RTP tests. The Switching task provided a large number of reliable and stable measures, and the NovaScan™ test also had a number of measures that were at least acceptable or marginally acceptable with regard to reliability and stability.

The job performance tasks also provided a remarkable number of reliable and stable task measures. This was especially true of the MATB. Eight of the twelve major task measures on the MATB demonstrated acceptable to very good levels of reliability and stability, and two others had encouraging trends. The ATST had no measures that yielded the unequivocal levels of reliability and stability seen in other task measures, but the nature of the ATST may have played an important role in that result. Further comments on the nature of the ATST reliability and stability results are made below.

Aside from highly controlled laboratory studies (e.g., Bittner et al., 1986), human performance measures have a somewhat poor history of reliability. This is particularly true if one considers test-retest correlations across a longer time frame than the typical laboratory study, as would be the case in evaluating job performance measures. Given this background, the results of this study were impressive. Many of the measures used were highly reliable and stable across the critical four-week testing period in this study. This finding provided considerable support for the integrity of the laboratory model approach used in this study. These results suggested that this laboratory approach can provide the basic reliability and stability in measurement to investigate RTP testing. Whether the candidate tasks were effective as RTP tests is, of course, left to further analyses (see Volume II of this report).

One of the more valuable unforeseen benefits of this study was that, as the data were analyzed, additional important insights and findings emerged. These unanticipated "spin-off" results include new insights into criterion measurement issues, possible subject-perceived differences between laboratory tasks and job performance tasks, new approaches to conceptualizing job performance assessment, greater understanding of the relationship between reliability/stability and sensitivity, and the possibility that both phasic and tonic sources of variability may be important in assessing performance. Each of these topics is discussed briefly below.

The identification and accurate assessment of appropriate criterion measures is essential for establishing valid laboratory or job performance evaluations. In analyzing the results of this project, some interesting insights into the nature and process of identifying criterion measures of task performance emerged. The general nature of the ATST led to some interesting problems in criterion measurement. Of all the tasks, the ATST was the most complex over time and required a demanding set of integrated cognitive and psychomotor skills. The learning dynamics of the ATST were also quite complex. To resemble the task

of air traffic control in real life, the ATST was designed to be fairly consistent with the safety-sensitive nature of the job *as it is actually performed*. That is, the ATST did not create a situation in which the average person would commit a large number of critical errors (i.e., crashes). Certainly, the possibility is high that novices or untrained people will have high error rates as they acquire the skills for this task, but well-trained individuals were able to complete even higher difficulty level scenarios without producing high error rates. Thus, the task was structured to be challenging, but after practice, many of the dependent measures quickly developed ceiling or near-ceiling effects that dramatically reduced variability. For example, when first encountering the ATST, subjects typically had to exert considerable effort to manage the complex nature of the task, and many subjects failed to direct all their assigned aircraft to their destinations. This led to considerable variability among subjects and test-retest correlations in the range of $r = 0.70$ to 0.80 during early trials. Following additional practice, most subjects could get all or a very high percentage of the planes to their appointed destinations satisfactorily, which led to correlations of $r = 0.18$ during later trials. This reduction in variability across well-trained subjects significantly compromises the meaningfulness of correlational analyses by deflating correlation coefficients.

Another interesting process also became clear. Other task measures that might reflect more complex skills often took more time to refine. As a result, their learning curves extended longer and did not provide particularly stable measures across the entire testing period. Performance on these measures of quite complex performance typically demonstrated poor and variable reliability in early stages of the study but very high or acceptable levels during the last few weeks. Good examples of this type of measure were the number of directional changes (NDIR) and the measure of overall delay at destination (DELAY). This is analogous to real work skills, which are developed over periods of months and years.

In the above manner, the analysis of the ATST data in this study provided a unique opportunity to evaluate complex task performance from new perspectives. Another additional advantage of this data set is that, in the future, it will provide the ability to explore skill acquisition on complex tasks, an area where there is very little information. The fact that one of these complex tasks (the MATB) was quickly learned and exhibited reliable outcome measures, and the other task (the ATST) had extended learning curves and complex skill acquisition dynamics, will provide a unique opportunity to explore the intricacies of skilled performance more like those seen in the workplace.

The experience gained through conducting this project also inspired some insights into the job prediction dilemma. For example, the low reliability of job performance measures often reported in the workplace was certainly replicated, in part, in these results. Given the nature of job performance, as compared to highly controlled laboratory tasks, it is probably unrealistic to expect the same level of reliability and stability. This is not to say that people do not conduct their jobs in a reliable and stable fashion. Rather, the results of this study suggest that laboratory tasks and real jobs are probably performed differently, but are typically assessed at the same level of analysis. The general lack of success in predicting job performance has not just been a failure to operationalize good criterion measures. Any successful attempt to obtain predictive measures of job performance will probably require more insight into the dynamic differences in the way people perform jobs, as compared to laboratory tasks.

The differences in the two job performance tasks in this study demonstrate this point. Both provided very complex performance requirements, but the performance measures yielded very different outcomes. To conclude that the MATB was the better task because its performance measures were more reliable might be both short-sighted and unfair. The MATB is a well-constructed synthesis of laboratory tasks that can be decomposed easily. It is a fine example of what might be considered a "bottom-up" task. The ATST might be considered a "top-down" task in that it was developed as a direct analogue (near simulation) of a job. It appears that the closer one gets to simulating a job through task performance, the closer one may also get to the critical differences of assessment between job performance and laboratory task performance.

Jobs are typically more global and stochastic in nature—one event leads to multiple layers of decisions and, subsequently, many avenues to a more global end product. The time course may be a few minutes to complete an assembly-line operation or longer time frames to complete large-scale projects. Within either time frame, there are multiple choice points that can provide even those tasks that appear rapid and routine with considerable variability. In fact, even those tasks that are routine may be forced to greater variation by the worker. For example, fast-paced, repetitive tasks are often varied by workers to avoid boredom and to retain higher levels of performance.

By contrast, consider the basic laboratory task (even a complex one) that is learned fairly quickly, performed for a fairly short duration, and often with implicit and explicit demands for a high level of repeatability—to say nothing of the environmental

specter of continual monitoring and evaluation. Compare that to job demands that are no less imposing, but usually involve slower development of skilled performance, are performed over long periods of time, are more oriented toward global performance criteria, and more often are assessed with larger units of outcome. It is not surprising that job performance analysis has often failed to provide the level of reliability and validity seen in laboratory tasks.

How does one intervene in the process of a job to evaluate performance more accurately? Aggregate production levels are often too global and consistent to provide the variability needed for accurate assessment. Entering into the stream of the process opens one to measuring the wide range of variability that workers interject into their jobs. To be more accurate in job assessment and prediction, one may have to be more creative in both conceptualizing what people do in their work and in measuring those activities. Again, because this project provided what appeared to be examples of complex work of two types, it may have also provided the opportunity for future analyses to explore this problem. No one would doubt that the subjects performed reasonably well when they engaged the ATST. Yet, the outcome variables were confusing. It may have been that the outcome variables were directed at the laboratory task level of analysis and missed the rich dynamics of the work environment that the task more accurately represented. The data from the ATST provided the opportunity to explore, in more creative ways, methods for analyzing this problem.

This line of reasoning raised another interesting issue. If laboratory tasks and job tasks are fundamentally different in nature, then applying methods of analyzing laboratory tasks to the workplace ought to create not only difficulty in measurement accuracy (not to mention interpretability), but also difficulty in worker acceptance. The well-known stories of worker dissatisfaction with performance assessments may extend beyond the normal disenchantment most people feel in being evaluated. In part, they may also be expressing an intuitive rejection of the reductionistic "bottom-up" methods of job analysts—methods that may not capture the more integrated and sophisticated functions they associate with their work. This could be a unique "reactivity of measures" problem that deserves greater attention in the future.

Another interesting issue that emerged from the analysis of these data was the relationship between performance consistency, reliability, and task sensitivity. The percent correct (PC) measure for many tasks was an interesting example of the problem. Performance in terms of PC was actually quite consistent and high for many tasks, as is the case for real work in safety-sensitive jobs. However, it was the uniformly high levels of performance (near 100%) that caused the lack of variability in the measure and subsequently, the low reliability coefficients. This situation pointed out that, while the PC measures were very consistent, they obviously failed to provide another prerequisite for reliability/stability assessment, that is, enough variance for discrimination between subjects—or *sensitivity* to individual differences.

However, such measures should not necessarily be ignored. These measures were undoubtedly insensitive under conditions of baseline testing. The introduction of risk factors may dramatically change this situation. The response characteristics of subjects may be so dramatically changed after risk factor exposure (such as degrading their performance) that these measures that had exhibited a ceiling effect during baseline conditions may then show differentiation between subjects. This is a matter of test sensitivity and must be a factor in assessing reliability and differential stability. If a test is to be effective, it must have the ability to register change under the circumstances it is intended to assess, and those may or may not be comparable to baseline conditions.

A final issue raised by these data was that, with all due caution, high reliability in the form of test-retest correlations should not be looked upon as a singular goal in performance assessment. There is an interesting problem in the measurement of tonic versus phasic change that seemed relevant to the present methods of RTP testing. Reliability and, more specifically, stability are terms that are most often reserved for the assessment and characterization of tonic variables, those that are fairly stable over time and resistant to the vagaries in the environment that bring about phasic changes. An analogy is seen in the difference between traits (behaviors that are consistent over time) and states (behaviors that are believed to vary more with changes in the situation or environment). Much has been written about the importance of reliability in measures of traits because of the presumed relationship to constancy in trait-related behavior. Less is made of state measurement reliability because it has been assumed that state measures will be volatile. A single-minded drive for high reliability for RTP measures should not blind one to the fact that what is often being assessed with RTP tests are changes at the state-like level. Among the many performance measures recorded in this study, there were some that had low reliabilities for artifactual reasons (e.g., being based on rare events) and probably many that had low reliabilities because they were simply less meaningful in the overall scheme of performance assessment.

However, there may have been some with low reliabilities because they were more sensitive to state-like fluctuations that influence only part of overall performance. Taken to an extreme, a very stable, trait-like performance measure may provide very high levels of reliability and stability, but therein may be a problem for its use as an RTP test. This measure may be so stable that it is resistant to any of the more subtle effects of risk factors.

In summary, the results of the data collected throughout the five weeks of this study that bear directly on the learning and stabilizing of RTP and job task performance suggest clear evidence of reliable and stable performance measures. In addition, these results support the integrity of the laboratory model approach proposed in this study. These data provide the foundation needed to explore the concept of RTP testing as a means of preventative screening for the behavioral variations that often accompany risk factor exposure. Finally, the results of this analysis raise many interesting questions about conceptualizing job measures, operationalizing job measures, and the concepts of reliability, stability, and consistency as they have been applied to performance assessment.

REFERENCES

Aerospace Sciences, Inc. (1991). *Air Traffic Control Specialist Pre-Training Screen preliminary validation: Final report* (Final Report, FAA Contract DTFA-01-90-Y-01034). Washington, DC: Federal Aviation Administration.

Barlow, M.C. (1928). A learning curve equation as fitted to learning records. *Psychological Review, 35*, 142-160.

Benson, A.J., and Gedye, J.L. (1963). *Logical processes in the resolution of orienting conflict* (RAF Rpt. 259). Farnborough, UK: Royal Air Force Institute of Aviation Medicine.

Bittner, A.C., Jr., Carter, R.C., Kennedy, R.S., Harbeson, M.M., and Krause, M. (1986). Performance evaluation tests for environmental research (PETER): Evaluation of 114 measures. *Perceptual and Motor Skills, 63*, 683-708.

Broach, D., and Brecht-Clark, J. *Validation of the Federal Aviation Administration Air Traffic Control Specialist Pre-Training Screen*. Washington, DC: 1994; FAA publication no. DOT/FAA/AM-94/4.

Campbell, D.T., and Stanley, J.C. (1971). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Comstock, J.R., and Arnegard, R.J. (1992). *The Multi-Attribute Task Battery for human operator workload and strategic behavior research* (NASA-TM-104174).

Damos, D.L. (1991). *Multiple task performance*. London, UK: Taylor & Francis.

Damos, D.L., and Wickens, C.D. (1980). The identification and transfer of timesharing skills. *Acta Psychologica, 46*, 15-39.

Freudenheim, M. (1988, December 13). Workers' substance abuse increasing, survey says. *New York Times*, 2.

Gilliland, K., and Schlegel, R.E. (1992). *Evaluation of extended practice effects on the Air Traffic Scenarios Test*. (Final Report, FAA Contract DTFA-02-92-P-13359). Oklahoma City, OK: FAA Civil Aeromedical Institute.

Gilliland, K., and Schlegel, R.E. *Readiness-to-perform testing: A critical analysis of the concept and current practices*. Washington, DC: 1993; FAA publication no. DOT/FAA/AM-93/13.

Gilliland, K., and Schlegel, R.E. (1995, January). Readiness-to-perform testing and the worker. *Ergonomics in Design*, 14-19.

Guilford, J.P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Gulliksen, H.A. (1934). A rational equation of the learning curve based on Thorndike's law of effect. *Journal of General Psychology, 11*, 395-434.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Guttman, L. (1955). Reliability formulas for noncompleted or speeded tests. *Psychometrika, 20*, 113-124.

Hamilton, J. (1991). A video game that tells if employees are fit for work. *Business Week*, June 3.

Hart, S.G., and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), *Human mental workload*. New York: Elsevier Scientific Publishers.

Hegge, F.W., Reeves, D.L., Poole, D.P., and Thorne, D.R. (1985). *The Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB) II: Hardware/software design and specification*. Fort Detrick, MD: U.S. Army Medical Research and Development Command.

Jones, M.B. (1980). *Stabilization and task definition in a performance test battery*. (NBDL Monograph No. M-001). New Orleans, LA: Naval Biodynamics Laboratory.

Jones, M.B., Kennedy, R., and Bittner, AC. (1981). A video game for performance testing. *American Journal of Psychology, 94*, 143-152.

Kennedy, R.S., Carter, R.C., and Bittner, A.C. (1980). A catalogue of performance evaluation tests for environmental research. *Proceedings of the Human Factors Society, 24* (1), 8-12.

Lord, F.M., and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Maltby, L.L. (1990). Put performance to the test. *Personnel, 67*, 30-31.

Mazur, J.E., and Hastie, R. (1978). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin, 85*, 1256-1274.

McNair, D.M., Lorr, M., and Droppleman, L.F. (1971). *Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Service.

McRuer, D.T., and Jex, H.R. (1967). A review of quasi-linear pilot models. *IEEE Transactions on Human Factors in Electronics, 8*, 231-249.

Miller, J.C., Takamoto, G.M., Bartel, G.M., and Brown, M.D. (1985). Psychophysiological correlates of long-term attention to complex tasks. *Behavior Research Methods, Instruments, and Computers, 17*(2), 186-190.

Nesthus, T.E., Schiflett, S.G., Eddy, D.R., and Whitmore, J.N. (1991). *Comparative effects of antihistamines on aircrew performance of simple and complex tasks under sustained operations* (AL-TR-91-104). Brooks AFB, TX: USAF Armstrong Laboratory, Crew Technology Division.

O'Donnell, R.D. (1991). *Scientific validation of the Novascan (tm) tests: Theoretical basis and initial validation studies*. NTI Report to Nova Technology, Inc., 19460 Shenango Drive, Tarzana, CA: NTI, Incorporated.

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

Pennebaker, J.W. (1982). *The psychology of physical symptoms*. New York: Springer-Verlag.

Perez, W.A., Masline, P.J., Ramsey, F.R., and Urban, K.E., (1987). *Unified Tri-Services Cognitive Performance Assessment Battery: Review and methodology* (AAMRL-TR-87-007). Wright-Patterson AFB, OH: Armstrong Aerospace Medical Research Laboratory.

Reader, D.C., Benel, R.A., and Rahe, A.J. (1981). *Evaluation of a manikin psychomotor task* (USAFSAM-TR-81-10). Brooks AFB, TX: USAF School of Aerospace Medicine.

Restle, F., and Greeno, J.G. (1970). *Introduction to mathematical psychology*. Reading, MA: Addison-Wesley.

Santucci, G., Farmer, E., Grissett, J., Wetherell, A., Boer, L., Gotters, K., Schwartz, E., and Wilson, G. (1989). *AGARDograph #308, Human performance assessment methods* (ISBN 92-835-0510-7). Seine, France: North Atlantic Treaty Organization Advisory Group for Aerospace Research and Development, Working Group 12.

Schlegel, R.E., and Gilliland, K. (1990). *Evaluation of the Criterion Task Set - Part I: CTS performance and SWAT data - baseline conditions* (AAMRL-TR-90-007). Wright-Patterson AFB, OH: USAF Armstrong Aerospace Medical Research Laboratory.

Schlegel, R.E., and Gilliland, K. (1992). *Development of the UTC-PAB normative database* (AL-TR-92-0145). Wright-Patterson AFB, OH: USAF Armstrong Laboratory.

Schlegel, R.E., and Storm, W.F. (1983). Speed-accuracy tradeoffs in spatial orientation information processing. *Proceedings of the 27th Annual Meeting of the Human Factors Society*, Santa Monica, CA: Human Factors Society.

Schlegel, R.E., Shehab, R.L., and Gilliland, K (1994). *Microgravity effects on cognitive performance measures: Practice schedules to acquire and maintain performance stability* (AL-CF-TR-1994-0040). Brooks AFB, TX: USAF Armstrong Laboratory, Crew Technology Division.

Shingledecker, C.A. (1984). *A task battery for applied human performance assessment research* (AFAMRL-TR-84-071). Wright-Patterson AFB, OH: Air Force Aerospace Medical Research Laboratory.

Spears, W.D. (1985). Measurement of learning and transfer through curve fitting. *Human Factors, 27*, 251-266.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30*, 276-315.

Thorne, D., Genser, S., Sing, H., and Hegge, F. (1985). The Walter Reed Performance Assessment Battery. *Neurobehavioral Toxicology and Teratology, 7*, 415-418.

Thurstone, L.L. (1919). The learning curve equation. *Psychological Monographs, 26*, No. 114.

Weltin, M., Broach, D., Goldbach, K., and O'Donnell, R. (1992). *Concurrent criterion-related validation of Air Traffic Control Specialist Pre-Training Screen*. (Final Report, FAA Contract DTFA-01-89-Y-01019). Washington, DC: Federal Aviation Administration.

APPENDIX A

PERFORMANCE MEASURES

Reference Guide for Task Variables and Codes

General Information

The identification scheme for the antihistamine trials is the following four-character code:

$c_1c_2c_3c_4$

c_1	h-antihistamine; p-placebo; r-refresher
c_2	(1,2,3,4,5,6)-refresher session number; d-daytime testing; n-nighttime testing
c_3	a-first group tested; b-second group tested; x-not used
c_4	(1,2,3)-test trial; x-not used

Examples: r5xx - fifth refresher session
hdb3 - antihistamine, daytime, second test group, third test trial (dose)

General Variables Used in Many or All Tasks

ID	Subject identification number; subjects for antihistamine study were: 201, 204, 206, 211, 216, 217, 218, 223, 224, 225, 226, 227, 229, 230, 232, 233
SESSION	Session number; antihistamine study consisted of sessions 51 through 68
DATE	Date of session
TIME	Time of session
TASK	Task name
INST	Whether or not instructions were included (indicated by -N1 for instructions)
LENGTH	Program option to specify task length

Antihistamine State Scale (ASH)

TOTAL	Total score for antihistamine symptom impact

Mood Scale (MOO)

xxxN	Total number of adjective responses in category xxx
xxxSUM	Sum of scores for adjectives in category xxx
xxxMN	Mean of scores for adjectives in category xxx
xxxPCT	Percent score for category xxx; (**xxxPCT = [xxxMN - 1]/2**)
xxxRT	Average response time for responses to adjectives in category xxx
RTALL	Overall response time for all responses

xxx	**Category**
ACT	Activity
HAP	Happiness
DEP	Depression
ANG	Anger
FAT	Fatigue
FER	Fear

Activity State Questionnaire (ASK)

PHYSICAL	Total (weighted) score for physical state
PREP	Total (weighted) score for preparedness

Spatial Processing (SPA)

MNCORRT	Mean correct response time
SDCORRT	Standard deviation of correct response times
N	Number of stimuli
PC	Percent correct stimuli
PINC	Percent incorrect stimuli
PLAPSE	Percent lapsed (i.e., timed-out) stimuli
NC	Number of correct stimuli
NINC	Number of incorrect stimuli
NLAPSE	Number of lapsed (i.e., timed-out) stimuli
MNCRTPOS	Mean correct response time for positive stimuli
SDCRTPOS	Standard deviation of correct response times for positive stimuli
NPOS	Number of positive stimuli
PCPOS	Percentage correct for positive stimuli
PINCPOS	Percentage incorrect for positive stimuli

Critical Tracking (TRK)

MAXL	Maximum lambda during trial
CTLOSS	Number of control losses
RMS	Average root mean square error
MEANL	Mean of lambda's at control losses

Dual Task - Individual Lambda and Group Lambda (DULI/DULG)

SET	Positive memory set
NULLSET	Negative memory set
VIEWRT	Memory set viewing time
PCRESP	Percent of responses that were correct (excluding time-outs)
MNALLRT	Mean overall response time
MNCORRT	Mean correct response time
MNINCRT	Mean incorrect response time
MAXL	Maximum lambda during trial
CTLOSS	Number of control losses
RMS	Average root mean square error
MEANL	Mean of lambda's at control losses
PC	Percent correct of all stimuli
SPEED	Responses per minute (60,000/**MNALLRT**)
THRPUT	Throughput (**SPEED** * **PC**)

Switching Task (NTI)

xxxCORRT	Mean correct response time for xxx task
xxxPC	Percent correct of all stimuli for xxx task
xxxTP	Throughput for xxx task; (60,000/xxxcorrt) * xxxpc
xxxCORTX	Mean correct response time for xxx transition trials (xxx trials preceded by trial from other task)
xxxPCX	Percent correct of all stimuli for xxx transition trials

xxx	**Task**
MAN	Manikin Task
MTH	Mathematical Processing Task

NovaScan™ FAA Task (NSF)

xxxCOR	Number of correct responses for xxx task
xxxPC	Percent correct for xxx task
xxxCRT	Mean correct response time for xxx task
xxxCSD	Standard deviation of correct response times for xxx task
xxxINC	Number of incorrect responses for xxx task
xxxPI	Percent incorrect for xxx task
xxxTO	Number of time-outs for xxx task
xxxPTO	Percent time-outs for xxx task
xATNREQ	Number of attention requests during xxx task
xATNACK	Number of attention acknowledgments during xxx task
xATNFA	Number of false alarms during xxx task

(x)xx	**Task**
(V)EC	Visual Search and Vector Projection
(M)EM	Continuous Spatial Memory

Air Traffic Scenarios Test (ATC)

SCEN	Scenario
CRSHAC	Number of crashes with other aircraft
CRSHBD	Number of crashes into air space boundary
CRSHAP	Number of crashes into the airport
SEPAC	Number of separation errors with other aircraft
SEPBD	Number of separation errors with air space boundary
ERRAPSPD	Number of speed errors at airport
ERRAPALT	Number of altitude errors at airport
ERRGTSPD	Number of speed errors at boundary gates
ERRGTALT	Number of altitude errors at boundary gates
ERRDEST	Number of destination errors
NDEST	Number of planes at destination
PCDEST	Percentage of planes at destination
DELAY	Delay score in routing planes for planes arriving at destination
NDIR	Number of direction changes
NALT	Number of altitude changes
NSPD	Number of speed changes

TLX for ATST

MENTAL	Rating of mental workload
PHYSICAL	Rating of physical workload
TEMPORAL	Rating of time-related workload
PERFORM	Rating of performance
EFFORT	Rating of required effort
FRUST	Rating of frustration level

Multi-Attribute Task Battery (MTB)

SCRIPT Specific MATB run script

Systems Monitoring

LTSRT	Mean response time for lights
DLSRT	Mean response time for dials
MONRT	Mean response time for lights and dials
LTSSD	Standard deviation for lights
DLSSD	Standard deviation for dials
MONSD	Standard deviation for lights and dials
LTSTO	Time Out errors for lights
DLSTO	Time Out errors for dials
MONTO	Time Out errors for lights and dials
LTSFA	False Alarm errors for lights
DLSFA	False Alarm errors for dials
MONFA	False Alarm errors for lights and dials
LTSER	Time Out and False Alarm errors for lights
DLSER	Time Out and False Alarm errors for dials
MONER	Time Out and False Alarm errors for lights and dials
MONKR	Key Repeats (See explanation for COMRPT under Communications dependent variables below.)

Communications

COMCRT	Mean response time for correct responses
COMCSD	Standard deviation for correct responses
COMORT	Mean overall response time
COMOSD	Standard deviation for overall responses
COMER	Total number of errors (This includes othership false alarms, othership accuracy errors, unexplained errors, ownship accuracy errors, and ownship time-outs. It does not include repeated ENTERs, described below.)
COMYFA	Othership false alarms (correct radio and frequency, but message was for other ship)
COMYAC	Othership accuracy errors (Message was for other ship; either radio or frequency were incorrect.)
COMYIG	Othership messages correctly ignored.
COMAC	Accuracy errors (response to ownship message, but either radio or frequency incorrect)

COMTO Time out errors

COMUNER Unexplained errors
(some response without identifiable cause, possibly false alarm)

COMRPT Repeated ENTERs (Number of times ENTER was pressed within 5 seconds of a previous ENTER press. Some subjects hold the ENTER key down for several seconds during this task. Matproc does not count these repeats as errors, but reports them with this dependent variable.)

Tracking

TRKRMS Root Mean Square (calculated for the each entire epoch)

Resource Management

TNKMAD Mean absolute deviation of tanks A and B from 2500

TNKAMN Mean of Tank A

TNKBMN Mean of Tank B

TNKACT Tank activity (number of pump changes ON or OFF)

Workload Rating Scale

TLXOMN Overall mean of subscales

TLXMEN Mean for Mental Demand subscale

TLXPHS Mean for Physical Demand subscale

TLXTMP Mean for Temporal Demand subscale

TLXPER Mean for Performance subscale

TLXEFT Mean for Effort subscale

TLXFRU Mean for Frustration subscale

TLXDUR Mean for duration of rating screen presentation

APPENDIX B

RELIABILITY AND DIFFERENTIAL STABILITY COEFFICIENTS

		24-Hour Correlations				
		Week 1	Week 2	Week 3	Week 4	Week 5
Task	Measure	8-10	14-15	19-20	24-25	29-30
Spatial	mncorrt	0.85	0.92	0.89	0.88	0.93
Processing	sdcorrt	0.58	0.77	0.70	0.64	0.66
	n	0.87	0.90	0.86	0.88	0.93
	pc	0.34	-0.13	0.48	0.45	0.75
	pinc	0.34	-0.13	0.48	0.45	0.75
	piapse					
	nc	0.73	0.54	0.67	0.63	0.76
	ninc	0.29	-0.16	0.48	0.48	0.76
	niapse					
	mncrtpos	0.86	0.83	0.88	0.87	0.88
	sdcrtpos	0.65	0.73	0.33	0.51	0.67
	npos					
	pcpos	0.32	0.28	-0.15	0.14	0.06
	pincpos	0.32	0.28	-0.15	0.14	0.06
Critical	maxl	0.76	0.74	0.71	0.72	0.77
Tracking	ctloss	0.76	0.68	0.89	0.81	0.74
	rms	0.56	0.46	0.89	0.80	0.78
	meanl	0.84	0.68	0.88	0.85	0.81
Dual Task	pcresp	0.16	0.36	0.39	0.35	-0.06
(Group)	mnallrt	0.95	0.96	0.81	0.91	0.87
	mncorrt	0.95	0.96	0.81	0.91	0.87
	mnincrt	0.32	0.36	0.19	-0.20	-0.17
	maxl					
	ctloss	0.48	0.66	0.75	0.75	0.69
	rms	0.64	0.86	0.81	0.89	0.81
	meanl	0.19	0.29	0.52	0.65	0.63
	pc	0.16	0.36	0.38	0.34	-0.08
	speed	0.94	0.96	0.86	0.92	0.89
	thruput	0.92	0.95	0.82	0.90	0.86
Dual Task	pcresp		0.15	0.44	0.21	-0.01
(Individual)	mnallrt		0.93	0.75	0.90	0.84
	mncorrt		0.93	0.76	0.90	0.84
	mnincrt		-0.04	-0.12	0.06	-0.34
	maxl					
	ctloss		0.46	0.71	0.74	0.53
	rms		0.65	0.75	0.73	0.90
	meanl		-0.12	-0.06	-0.11	0.48
	pc		0.13	0.43	0.32	-0.03
	speed		0.90	0.87	0.90	0.89
	thruput		0.87	0.84	0.90	0.84
Switching	mancorrt	0.96	0.97	0.96	0.91	0.96
Task	manpc	0.46	0.28	0.30	0.33	0.33
(Manikin)	mantp	0.97	0.97	0.97	0.96	0.96
	mancortx	0.90	0.95	0.89	0.93	0.92
	manpcx	0.31	-0.25	-0.14	0.46	0.76
Switching	mthcorrt	0.90	0.96	0.93	0.90	0.95
Task	mthpc	0.46	0.13	0.46	0.48	0.11
(Math)	mthtp	0.95	0.97	0.94	0.96	0.97
	mthcortx	0.87	0.93	0.92	0.94	0.91
	mthpcx	0.15	0.26	-0.03	0.57	0.20
NovaScan™	veccor	0.46	0.66	0.64	0.92	0.91
FAA Task	vecpc	0.46	0.66	0.64	0.92	0.91
	veccrt	0.73	0.86	0.78	0.91	0.90
	veccsd	0.65	0.69	0.73	0.80	0.69
	vecinc	0.46	0.66	0.64	0.92	0.91
	vecpi	0.46	0.66	0.64	0.92	0.91
	vecto					
	vecpto					
	vatnreq					
	vatnack	0.22	-0.12	0.23	0.17	-0.15
	vatnpc	.	-0.10	-0.03	0.61	0.12
	vatnfa	-0.03	0.36	0.02	0.58	0.57
	memcor	0.20	0.14	0.40	0.49	0.69
	mempc	0.20	0.14	0.41	0.49	0.69
	memcrt	0.75	0.65	0.88	0.92	0.91
	memcsd	0.76	0.45	0.69	0.77	0.65
	meminc	0.20	0.14	0.40	0.49	0.69
	mempi	0.20	0.14	0.41	0.49	0.69
	memto					
	mempto					
	matnreq					
	matnack	0.26	-0.16	0.18	0.01	-0.18
	matnpc	0.40	-0.11	0.29	0.62	-0.07
	matnfa	0.02	0.29	0.32	0.57	0.24

		24-Hour Correlations				
		Week 1	Week 2	Week 3	Week 4	Week 5
Task	Measure	8-10	14-15	19-20	24-25	29-30
Air Traffic	crshac	0.57	0.76	0.67	0.15	-0.05
Scenarios	crshbd	0.44	0.73	0.56	.	.
Test	crshap	0.33	-0.05	0.14	-0.06	-0.18
	sepac	0.77	0.70	0.92	0.63	0.53
	sepbd	0.33	0.75	0.75	0.28	-0.13
	errapspd	0.42	0.42	0.72	0.56	0.91
	errapalt	0.39	0.40	0.69	0.49	0.18
	errgtspd	-0.10	-0.09	.	.	.
	errgtalt	0.10	-0.08	0.01	-0.04	1.00
	errdest	-0.11	0.04	0.06	-0.11	-0.08
	ndest	0.72	0.82	0.65	0.63	0.18
	pcdest	0.72	0.82	0.65	0.63	0.18
	delay	0.38	0.71	0.59	0.83	0.90
	ndir	0.61	0.57	0.77	0.83	0.71
	nalt	0.49	0.66	0.86	0.61	0.40
	nspd	0.60	0.74	0.79	0.65	0.36
Multi-	ltsrt	0.35	0.71	0.57	0.90	0.61
Attribute	dlsrt	0.41	0.76	0.73	0.91	0.69
Task	monrt	0.47	0.76	0.77	0.95	0.67
Battery	ltssd	0.04	0.56	0.34	0.33	0.27
(MATB)	dlssd	0.19	0.46	0.56	0.75	0.49
	monsd	0.28	0.55	0.68	0.81	0.52
	ltsto	1.00	-0.04	-0.05	0.43	-0.05
	dlsto	0.82	0.82	0.88	0.75	0.60
	monto	0.85	0.76	0.88	0.72	0.60
	ltsfa	-0.03	0.32	-0.13	0.16	0.02
	dlsfa	0.51	0.86	0.89	0.93	0.99
	monfa	0.41	0.85	0.89	0.92	0.99
	ltser	0.79	0.07	-0.20	0.28	0.02
	dlser	0.58	0.87	0.78	0.85	0.97
	moner	0.59	0.83	0.78	0.84	0.97
	monkr	-0.09	0.04	0.35	0.13	0.36
	comcrt	0.20	0.14	0.41	0.49	0.69
	comcsd	0.36	0.32	0.70	0.33	0.64
	comort	0.69	0.85	0.76	0.89	0.92
	comosd	0.30	0.36	0.70	0.38	0.69
	comer	0.93	0.80	0.95	0.90	0.90
	comyfa	.	.	.	.	.
	comyac	.	.	.	.	.
	comylg	.	-0.03	.	.	.
	comac	0.13	0.06	0.09	0.05	0.32
	comto	1.00	0.94	0.99	0.98	0.97
	comuner	-0.06	0.17	0.49	-0.09	0.30
	comrpt	-0.05	0.08	0.90	0.80	1.00
	trkrms	0.95	0.91	0.70	0.84	0.54
	tnkmad	0.82	0.70	0.85	0.99	0.83
	tnkamn	0.38	0.79	0.75	0.94	0.79
	tnkbmn	0.55	0.78	0.82	0.94	0.93
	tnkact	0.84	0.80	0.93	0.95	0.94
	tlxomn	0.87	0.89	0.48	0.82	0.91
	tlxmen	0.93	0.80	0.55	0.82	0.89
	tlxphs	0.92	0.94	0.85	0.81	0.90
	tlxtmp	0.68	0.92	0.40	0.82	0.71
	tlxper	0.88	0.80	0.48	0.48	0.88
	tlxeft	0.88	0.73	0.63	0.63	0.61
	tlxfru	0.84	0.78	0.42	0.76	0.87
	tlxdur	0.69	0.32	0.74	0.75	0.52

		48-Hour			1-Week		2-Week
		Week 3	Week 4	Week 5	Week 3-4	Week 4-5	Week 3-5
Task	Measure	18-20	23-25	28-30	18-23	23-28	18-28
Spatial	mncorrt	0.84	0.85	0.93	0.84	0.89	0.86
Processing	sdcorrt	0.62	0.56	0.76	0.68	0.76	0.77
	n	0.77	0.85	0.91	0.79	0.88	0.83
	pc	0.52	0.32	0.51	0.06	0.37	0.32
	pinc	0.52	0.32	0.51	0.06	0.37	0.32
	plapse						
	nc	0.70	0.49	0.56	0.49	0.67	0.61
	ninc	0.46	0.34	0.54	0.07	0.38	0.32
	nlapse						
	mncrtpos	0.84	0.80	0.86	0.86	0.87	0.84
	sdcrtpos	0.50	0.44	0.56	0.69	0.55	0.59
	npos						
	pcpos	0.31	0.12	0.00	0.19	0.30	0.19
	pincpos	0.31	0.12	0.00	0.19	0.30	0.19
Critical	maxl	0.58	0.73	0.73	0.64	0.65	0.62
Tracking	ctloss	0.77	0.68	0.60	0.66	0.42	0.65
	rms	0.61	0.74	0.55	0.56	0.56	0.33
	meanl	0.81	0.78	0.72	0.70	0.61	0.68
Dual Task	pcresp	0.46	0.01	0.62	0.03	0.38	0.46
(Group)	mnallrt	0.85	0.59	0.79	0.56	0.57	0.72
	mncorrt	0.86	0.59	0.78	0.57	0.58	0.73
	mnincrt	-0.02	-0.04	0.15	0.02	-0.04	0.09
	maxl						
	ctloss	0.66	0.82	0.48	0.75	0.64	0.54
	rms	0.87	0.85	0.87	0.82	0.89	0.81
	meanl	0.39	0.37	0.69	0.51	0.44	0.17
	pc	0.51	0.07	0.63	0.20	0.36	0.46
	speed	0.91	0.85	0.87	0.83	0.79	0.80
	thruput	0.89	0.85	0.87	0.84	0.77	0.78
Dual Task	pcresp	0.52	0.30	0.12	0.18	0.48	0.42
(Individual)	mnallrt	0.72	0.96	0.77	0.30	0.41	0.86
	mncorrt	0.75	0.96	0.76	0.30	0.41	0.86
	mnincrt	0.10	0.00	-0.38	0.20	0.43	0.55
	maxl						
	ctloss	0.35	0.39	0.51	0.32	0.36	0.39
	rms	0.71	0.50	0.81	0.69	0.58	0.64
	meanl	0.05	-0.08	0.33	0.10	0.21	0.36
	pc	0.59	0.21	0.08	0.15	0.51	0.49
	speed	0.78	0.88	0.83	0.68	0.72	0.83
	thruput	0.76	0.85	0.83	0.64	0.71	0.82
Switching	mancorrt	0.94	0.97	0.96	0.89	0.85	0.90
Task	manpc	0.34	0.67	-0.15	0.24	0.31	0.59
(Manikin)	mantp	0.96	0.96	0.96	0.96	0.90	0.91
	mancortx	0.90	0.96	0.93	0.92	0.84	0.81
	manpcx	-0.08	0.60	-0.22	0.21	-0.10	-0.03
Switching	mthcorrt	0.88	0.90	0.95	0.88	0.88	0.89
Task	mthpc	0.39	0.09	0.32	0.25	0.25	0.43
(Math)	mthtp	0.91	0.97	0.97	0.92	0.95	0.92
	mthcortx	0.83	0.87	0.91	0.81	0.86	0.77
	mthpcx	0.06	0.19	-0.17	0.39	0.39	0.67
NovaScan™	veccor	0.75	0.93	0.66	0.84	0.90	0.78
FAA Task	vecpc	0.75	0.93	0.66	0.84	0.90	0.78
	veccrt	0.84	0.87	0.81	0.77	0.83	0.81
	vecced	0.67	0.80	0.73	0.89	0.84	0.87
	vecinc	0.75	0.93	0.66	0.84	0.90	0.78
	vecpi	0.75	0.93	0.66	0.84	0.90	0.78
	vecto						
	vecpto						
	vatnreq						
	vatnack	-0.18	0.16	-0.09	-0.19	-0.02	0.01
	vatnpc	0.08	-0.08	-0.11	-0.07	0.23	0.28
	vatnfa	0.49	0.46	0.35	0.19	0.39	0.68
	memcor	0.60	0.38	0.43	0.61	0.64	0.51
	mempc	0.60	0.38	0.43	0.61	0.64	0.51
	memcrt	0.74	0.68	0.94	0.74	0.71	0.87
	memced	0.41	0.42	0.78	0.49	0.32	0.32
	meminc	0.60	0.38	0.43	0.61	0.64	0.51
	mempi	0.60	0.38	0.43	0.61	0.64	0.51
	memto						
	mempto						
	matnreq						
	matnack	0.00	0.00	-0.18	-0.26	0.03	-0.09
	matnpc	-0.12	0.07	-0.10	0.13	0.32	0.16
	matnfa	0.53	0.28	0.29	0.08	0.35	0.56

		48-Hour			1-Week		2-Week
		Week 3	Week 4	Week 5	Week 3-4	Week 4-5	Week 3-5
Task	Measure	18-20	23-25	28-30	18-23	23-28	18-28
Air Traffic	crshac	0.49	0.01		-0.07		
Scenarios	crshbd	0.61			0.36		
Test	crshap	0.28	0.05	-0.02	-0.11	-0.11	0.10
	sepac	0.97	0.57	0.03	0.92	0.24	0.14
	sepbd	0.87	0.34	-0.13	0.65	0.02	0.11
	errapspd	0.74	0.26	0.92	0.45	0.19	0.34
	errapalt	0.71	0.61	0.42	0.45	0.64	0.54
	errgtspd				-0.04		
	errgtalt	-0.16	-0.06	1.00	0.23	0.81	-0.07
	errdest	-0.11	-0.05	-0.08	-0.05	-0.04	-0.05
	ndest	0.73	0.50	0.08	0.86	0.42	0.38
	pcdest	0.73	0.50	0.08	0.86	0.42	0.38
	delay	0.70	0.79	0.91	0.90	0.56	0.56
	ndir	0.81	0.74	0.72	0.77	0.54	0.48
	nalt	0.75	0.81	0.66	0.68	0.76	0.63
	nspd	0.65	0.44	0.67	0.72	0.27	0.45
Multi-	ltsrt	0.69	0.87	0.73	0.45	0.45	0.53
Attribute	dlsrt	0.73	0.82	0.78	0.71	0.59	0.65
Task	monrt	0.77	0.86	0.75	0.69	0.52	0.66
Battery	ltssd	0.41	0.45	0.63	0.40	0.38	0.31
(MATB)	dlssd	0.58	0.62	0.68	0.57	0.39	0.63
	monsd	0.65	0.68	0.69	0.66	0.48	0.63
	ltsto	-0.03	0.95	-0.05	-0.04	-0.05	-0.05
	dlsto	0.82	0.54	0.52	0.62	0.65	0.36
	monto	0.82	0.54	0.51	0.61	0.66	0.36
	ltsfa	0.08	0.32	0.26	-0.14	0.14	-0.06
	dlsfa	0.95	0.99	0.99	0.68	0.95	0.49
	monfa	0.95	0.99	0.99	0.67	0.95	0.48
	ltser	0.03	0.64	0.26	-0.15	0.13	-0.12
	dlser	0.94	0.98	0.98	0.66	0.93	0.46
	moner	0.94	0.96	0.98	0.65	0.93	0.46
	monkr	0.04	0.48	0.40	0.12	0.68	0.33
	comcrt	0.60	0.38	0.43	0.85	0.86	0.85
	comcsd	0.30	0.10	0.07	0.68	0.28	0.39
	comort	0.81	0.81	0.74	0.87	0.87	0.85
	comosd	0.39	0.14	0.13	0.63	0.27	0.33
	comer	0.93	0.92	0.90	0.97	0.92	0.96
	comyfa						
	comyac						
	comyig						
	comac	0.13	0.05	0.04	0.71	0.10	0.31
	comto	0.98	0.98	0.97	0.99	0.98	0.99
	comuner	0.56	0.02	0.19	0.50	0.24	0.55
	comrpt	0.99	1.00	0.61	0.98	0.59	0.56
	trkrms	0.79	0.87	0.75	0.88	0.83	0.77
	tnkmad	0.95	0.93	0.87	0.80	0.87	0.93
	tnkamn	0.89	0.91	0.83	0.89	0.83	0.69
	tnkbmn	0.90	0.95	0.74	0.84	0.88	0.87
	tnkact	0.95	0.96	0.96	0.88	0.94	0.86
	tlxomn	0.58	0.88	0.86	0.65	0.86	0.61
	tlxmen	0.59	0.76	0.88	0.60	0.85	0.68
	tlxphs	0.86	0.89	0.94	0.77	0.89	0.75
	tlxtmp	0.79	0.86	0.69	0.65	0.85	0.56
	tlxper	0.74	0.53	0.83	0.73	0.81	0.65
	tlxeft	0.71	0.71	0.81	0.61	0.77	0.45
	tlxfru	0.28	0.77	0.79	0.65	0.82	0.60
	tlxdur	0.72	0.79	0.46	0.68	0.77	0.65

		Weekly Average Correlations				
		Week 1	Week 2	Week 3	Week 4	Week 5
Task	Measure	Ave 8-10	Ave 13-15	Ave 18-20	Ave 23-25	Ave 28-30
Spatial	mncorrt	0.88	0.90	0.87	0.88	0.91
Processing	sdcorrt	0.59	0.69	0.69	0.64	0.71
	n	0.88	0.89	0.81	0.88	0.92
	pc	0.38	-0.03	0.38	0.38	0.57
	pinc	0.38	-0.03	0.38	0.38	0.57
	plapse					
	nc	0.66	0.61	0.66	0.60	0.66
	ninc	0.36	-0.05	0.35	0.42	0.59
	nlapse					
	mncrtpos	0.86	0.84	0.87	0.86	0.85
	sdcrtpos	0.54	0.66	0.44	0.47	0.55
	npos					
	pcpos	0.15	0.13	0.16	0.15	0.01
	pincpos	0.15	0.13	0.16	0.15	0.01
Critical	maxl	0.75	0.63	0.69	0.70	0.70
Tracking	ctloss	0.80	0.68	0.80	0.72	0.60
	rms	0.66	0.38	0.61	0.77	0.73
	meanl	0.87	0.69	0.83	0.80	0.73
Dual Task	pcresp	0.07	0.36	0.44	0.24	0.21
(Group)	mnallrt	0.93	0.86	0.82	0.67	0.82
	mncorrt	0.94	0.87	0.82	0.67	0.82
	mnincrt	0.22	0.08	0.12	-0.09	0.04
	maxl					
	ctloss	0.14	0.74	0.69	0.78	0.68
	rms	0.65	0.85	0.81	0.87	0.84
	meanl	0.26	0.48	0.43	0.41	0.65
	pc	0.10	0.36	0.44	0.22	0.22
	speed	0.94	0.90	0.88	0.86	0.88
	thruput	0.92	0.89	0.86	0.85	0.85
Dual Task	pcresp		0.32	0.38	0.21	0.08
(Individual)	mnallrt		0.90	0.75	0.92	0.81
	mncorrt		0.90	0.76	0.92	0.80
	mnincrt		0.05	0.12	0.01	-0.03
	maxl					
	ctloss		0.38	0.52	0.53	0.54
	rms		0.68	0.71	0.60	0.84
	meanl		-0.07	0.07	0.14	0.41
	pc		0.33	0.40	0.20	0.06
	speed		0.86	0.84	0.88	0.86
	thruput		0.85	0.80	0.87	0.84
Switching	mancorrt	0.94	0.97	0.95	0.92	0.95
Task	manpc	0.50	0.26	0.33	0.46	0.04
(Manikin)	mantp	0.97	0.97	0.96	0.95	0.96
	mancortx	0.92	0.95	0.91	0.93	0.92
	manpcx	0.40	-0.03	-0.05	0.50	0.12
Switching	mthcorrt	0.91	0.95	0.91	0.90	0.95
Task	mthpc	0.57	0.12	0.46	0.19	0.22
(Math)	mthtp	0.96	0.96	0.94	0.96	0.96
	mthcortx	0.83	0.89	0.87	0.89	0.90
	mthpcx	0.39	0.06	0.14	0.29	0.11
NovaScan™	veccor	0.63	0.78	0.73	0.92	0.80
FAA Task	vecpc	0.63	0.78	0.73	0.92	0.80
	veccrt	0.77	0.87	0.85	0.90	0.80
	vecced	0.65	0.71	0.71	0.81	0.69
	vecinc	0.63	0.78	0.73	0.92	0.80
	vecpi	0.63	0.78	0.73	0.92	0.80
	vecto					
	vecpto					
	vatnreq					
	vatnack	0.07	-0.12	0.01	0.20	-0.15
	vatnpc	-0.09	-0.07	0.09	0.15	0.16
	vatnfa	-0.05	0.11	0.30	0.48	0.51
	memcor	0.14	0.28	0.52	0.51	0.51
	mempc	0.14	0.28	0.52	0.51	0.51
	memcrt	0.83	0.72	0.84	0.79	0.93
	memced	0.67	0.57	0.56	0.59	0.69
	meminc	0.14	0.28	0.52	0.51	0.51
	mempi	0.14	0.28	0.52	0.51	0.51
	memto					
	mempto					
	matnreq					
	matnack	0.07	-0.12	0.06	0.03	-0.25
	matnpc	0.19	0.12	0.26	0.17	0.09
	matnfa	0.02	0.35	0.33	0.43	0.38

		Weekly Average Correlations				
		Week 1	Week 2	Week 3	Week 4	Week 5
Task	Measure	Ave 8-10	Ave 13-15	Ave 18-20	Ave 23-25	Ave 28-30
Air Traffic	crshac	0.57	0.54	0.60	0.04	-0.05
Scenarios	crshbd	0.44	0.62	0.64		
Test	crshap	0.33	0.01	0.12	-0.03	-0.03
	sepac	0.77	0.68	0.93	0.70	0.27
	sepbd	0.33	0.53	0.80	0.30	-0.12
	errapspd	0.42	0.50	0.75	0.49	0.92
	errapalt	0.39	0.38	0.66	0.56	0.29
	errgtspd	-0.10	0.03	0.42		
	errgtalt	0.10	0.29	0.05	0.24	1.00
	errdest	-0.11	0.13	0.03	-0.08	0.28
	ndest	0.72	0.82	0.69	0.61	0.14
	pcdest	0.72	0.82	0.69	0.61	0.14
	delay	0.38	0.60	0.64	0.82	0.90
	ndir	0.61	0.68	0.78	0.74	0.74
	nalt	0.49	0.61	0.79	0.72	0.59
	nspd	0.60	0.74	0.71	0.57	0.55
Multi-	ltsrt	0.39	0.74	0.62	0.90	0.61
Attribute	dlsrt	0.28	0.73	0.76	0.86	0.70
Task	monrt	0.39	0.78	0.79	0.90	0.68
Battery	ltssd	0.17	0.60	0.36	0.42	0.42
(MATB)	dlssd	0.09	0.54	0.58	0.69	0.57
	monsd	0.17	0.63	0.67	0.74	0.60
	ltsto	1.00	0.21	-0.04	0.62	0.17
	dlsto	0.78	0.80	0.83	0.69	0.56
	monto	0.82	0.75	0.83	0.67	0.55
	ltsfa	-0.06	0.14	0.05	0.21	0.07
	dlsfa	0.58	0.85	0.92	0.95	0.99
	monfa	0.48	0.85	0.92	0.95	0.99
	ltser	0.20	0.14	-0.02	0.44	0.12
	dlser	0.65	0.84	0.84	0.90	0.98
	moner	0.63	0.79	0.84	0.90	0.98
	monkr	0.13	0.04	0.11	0.21	0.43
	comcrt	0.14	0.28	0.52	0.51	0.51
	comcsd	0.26	0.39	0.51	0.32	0.26
	comort	0.75	0.78	0.82	0.85	0.82
	comosd	0.22	0.40	0.56	0.35	0.29
	comer	0.95	0.84	0.94	0.92	0.92
	comyfa					
	comyac					
	comyig		-0.03			
	comac	0.38	0.10	0.11	0.07	0.20
	comto	1.00	0.95	0.98	0.98	0.98
	comuner	0.13	-0.01	0.48	-0.04	0.18
	comrpt	-0.05	0.57	0.95	0.88	0.74
	trkrms	0.94	0.92	0.65	0.83	0.62
	tnkmad	0.80	0.78	0.88	0.95	0.82
	tnkamn	0.57	0.76	0.76	0.92	0.79
	tnkbmn	0.56	0.79	0.81	0.93	0.82
	tnkact	0.87	0.85	0.94	0.95	0.96
	tlxomn	0.87	0.81	0.59	0.86	0.87
	tlxmen	0.87	0.79	0.59	0.82	0.88
	tlxphs	0.94	0.92	0.87	0.87	0.90
	tlxtmp	0.76	0.83	0.56	0.85	0.74
	tlxper	0.82	0.70	0.63	0.80	0.82
	tlxeft	0.84	0.75	0.66	0.71	0.65
	tlxfru	0.76	0.60	0.44	0.79	0.83
	tlxdur	0.56	0.30	0.72	0.71	0.59

☆U.S. GOVERNMENT PRINTING OFFICE: 1997 - 559-879/40064