THE IMPACT OF PHYSICAL AND MENTAL TASKS ON PILOT MENTAL WORKLOAD

Scott L. Berg and Thomas B. Sheridan
Man-Machine Systems Laboratory
Room 3-346
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Abstract

Seven instrument-rated pilots with a wide range of backgrounds and experience levels flew four different scenarios on a fixed-base simulator. The Baseline scenario was the simplest of the four and had few mental and physical tasks. An Activity scenario had many physical but few mental tasks. The Planning scenario had few physical and many mental tasks. A Combined scenario had high mental and physical task loads. The magnitude of each pilot's altitude and airspeed deviations was measured, subjective workload ratings were recorded, and the degree of pilot compliance with assigned memory/planning tasks was noted.

Mental and physical performance was a strong function of the manual activity level, but not influenced by the mental task load. High manual task loads resulted in a large percentage of mental errors even under low mental task loads. Although all the pilots gave similar subjective ratings when the manual task load was high, subjective ratings showed greater individual differences with high mental task loads. Altitude or airspeed deviations and subjective ratings were most correlated when the total task load was very high. Although airspeed deviations, altitude deviations, and subjective workload ratings were similar for both low experience and high experience pilots, at very high total task loads, mental performance was much lower for the low experience pilots.

Research Supported by NASA Ames Research Center

C-2

I. INTRODUCTION

Cockpit design practices of the last 15 years share a common thread: the degree and complexity of automation is increasing and accelerating. Current state-of-the-art designs such as the Boeing 757, 767, and Airbus Industries A310 have radically changed flight deck activities. Future designs, such as the U.S. Air Force's proposed Advanced Technology Fighter and the Navy's Advanced Combat Aircraft will demand far greater levels of automation because of the requirement to operate in an extremely hostile, changing environment.

Expert systems and artificial intelligence will reduce or eliminate certain types of pilot workload. However, in some instances they may simply change the type of workload. Pilots are operating less as manual controllers and more as supervisory controllers.

The increased time and effort expended in monitoring aircraft equipment has raised concerns that in automating aircraft we may be raising the pilot's mental workload to unacceptable levels (or conversely, lowering it to undesirable levels). Thus, there is great interest in measuring this mental workload and its effects. However, measuring mental workload has been a difficult problem to solve.

Different researchers and different segments of the engineering and design communities have defined mental workload differently. Systems engineers, psychologists, and physiologists all have their own models of mental workload and their own methods of measuring it.

However, over the last decade, there has been a growing consensus that: a) mental workload is multidimensional in nature; and b) because of this multidimensionality, the "best" approach to measuring mental workload is to combine objective performance measures and subjective rating measures.

II. OBJECTIVES

This research examines several issues relating to mental workload. First, how does automation affect pilot mental workload? Since mental workload is multidimensional, automation may affect each dimension differently. Second, how does the level of mental workload affect physical and mental performance? Third, is the magnitude of a pilot's mental workload a function of the time between receiving instructions and executing them?

III. SIMULATOR CONFIGURATION

Figure 1 pictures the laboratory flight simulator environment for this project. The volunteer pilot subjects manipulate controls and switches on a control box while getting aircraft state information from a MEGATEK high resolution cathode ray tube (CRT) display (Figure 2). The MEGATEK displays flight instruments, aircraft and equipment configuration, and a forward

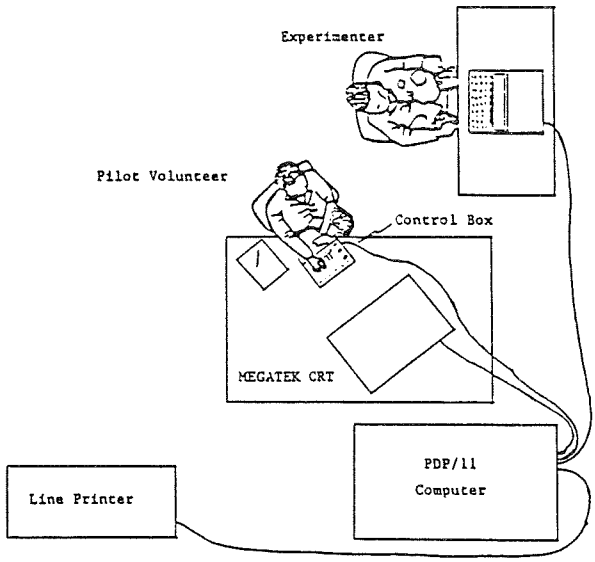

Figure 1: The Laboratory Environment

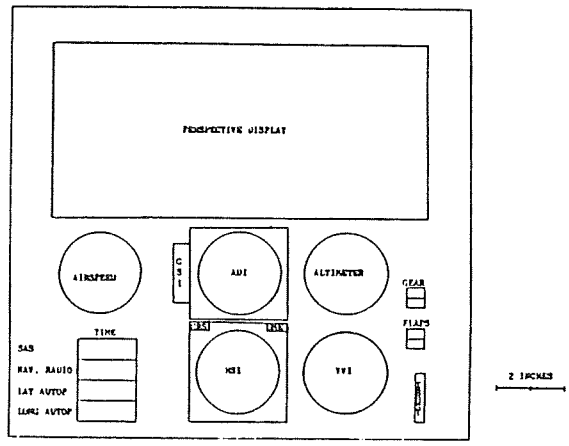

Figure 2: Diagram of the Simulator's high resolution MEGATEK CRT display

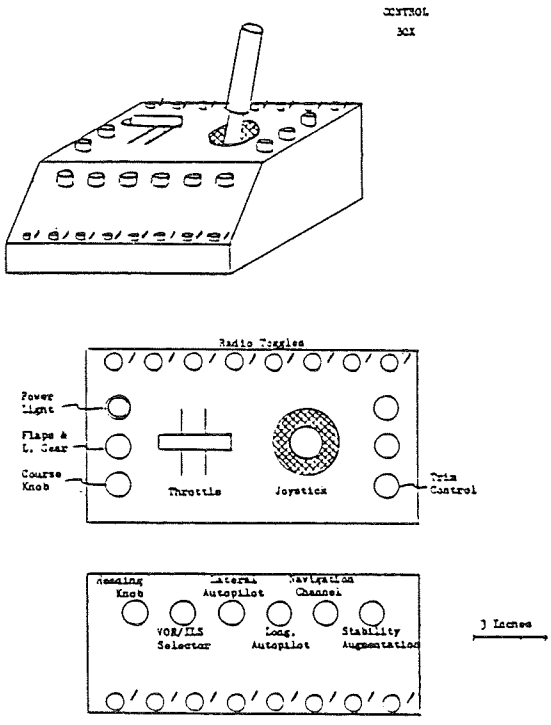

Figure 3: Layout of the Control Box for the flight simulator

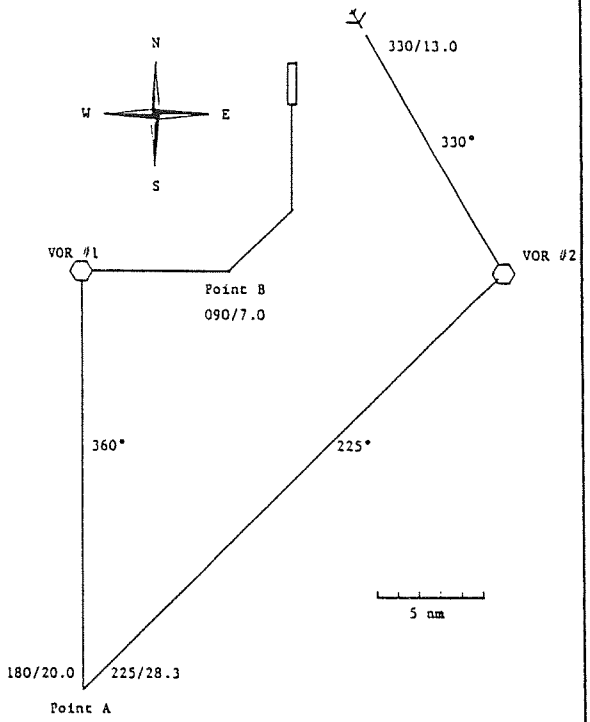

Figure 4: Navigation Chart for the Experiment

perspective view. The investigator has his own video display terminal (VDT) and keyboard for controlling the system.

A drawing of the Control Box is shown in Figure 3. The subject interprets the flight information displayed on the MEGATEK and manipulates the controls and switches on the Control Box to make the "aircraft" respond in a desired fashion. Control Box signals are fed to a PDP/11 Computer. The Computer's simulation program takes the present aircraft state information, Control Box inputs, and the investigator's Keyboard commands to determine aircraft dynamics and a new aircraft state. The information is used to update the MEGATEK and VDT displays.

A great deal of experimental trial and error went into making the simulator's response as close as possible to the response of an actual aircraft. A number of pilots came to the lab, flew the simulator, and evaluated its handling qualities. Eventually, the simulation fidelity was brought to a high level, including realistic stall and landing characteristics.

The Computer stores all Control Box switch or control manipulations and stores aircraft state data every 10.0 seconds. This data can be displayed on the investigator's VDT or printed out on a Line Printer.

IV. SUBJECTS

Initially, approximately 30 pilots volunteered to participate. Although we had hoped to use at least a dozen pilots of varied background, the list of 30 was eventually reduced to 7. Unfamiliarity with the flight characteristics of high performance aircraft and the simulator's ADI/HSI display, and the inability to devote the time needed for qualifying on the simulator and flying the data runs eliminated most of the pilots.

All seven subjects were good pilots, and there was a good mix of experience. Three subjects were Air Force pilots with 2400 to 3200 hours of flight time. Two pilots were Certified Flight Instructors with instrument ratings. The four civilian pilots ranged in experience from 300 total hours to 3000 total hours and had between 50 and 250 hours of instrument time.

V. EXPERIMENTAL DESIGN

Four different scenarios were flown using one basic route, illustrated in Figure 4. The four scenarios were labeled Baseline, Activity, Planning, and Combined. The Baseline scenario was the easiest. It simulated a "normal" flight and the pilots were encouraged to use the autopilot to keep workload at a minimum. There were no directed deviations from the basic course, and airspeed and altitude changes were rare. Also, there were very few assigned memory or planning tasks.

A data session consisted of a Baseline run followed by one of the other scenarios. The Baseline scenario was used as a warm-up data run and as a calibration run. Each second run's data was compared to that session's Baseline run. Baseline performance and ratings for different sessions could then be compared to adjust the data for variations due to day-to-day differences such as fatigue, stress, emotional state, et cetera.

The Activity scenario was loaded with a large number of manual-control tasks, but like the Baseline scenario, had a light planning task load. The pilots flew this scenario without using the autopilot.

The Planning scenario was very different from the Activity scenario. It was almost identical in manual activity to the Baseline scenario, (and thus, had a low activity level) but instead of being directed to perform actions immediately, the pilots were directed to perform these actions at a certain time in the future. These instructions often involved overlapping time periods, and the requests were not ordered chronologically. For example, prior to 2:00 minutes the pilot might be told to descend 1000 feet at 5:00 minutes, then told to turn to 300 degrees heading at 13:30 minutes, then to slow to 190 knots at 8:00 minutes. Therefore, the pilots had to "plan ahead".

The Combined scenario was designed to be the most difficult of all. It combined the manual activity of the Activity scenario with the planning requirements of the Planning scenario. This was an effort to saturate the pilots. The pilots were allowed to use the autopilot for help, but the pace of this scenario usually limited its use.

Figure 5 lists the order in which each pilot flew each of the non-Baseline scenarios. Different pilots flew the various scenarios in different orders. However, they all began each session's data runs with a Baseline run. The other three scenarios were not truly order randomized, but they were mixed. No pilot flew the Combined scenario in the first session. It was so unusually difficult, it was felt that this scenario might create an impossible workload for any pilot flying it first.

A Navigation Chart (Figure 4) and a note pad were provided for each pilot's use. Also, special placards were displayed beneath the instrument display to give configuration/airspeed data and help the pilots with the various lateral and longitudinal autopilot modes.

Ground tracks, altitude profiles, and airspeed profiles provided in Figures 6 through 9, clearly illustrate some of the differences and similarities of the various scenarios. Those three items were nearly identical for the Baseline and Planning scenarios, and for the Activity and Combined scenarios. Figure 6 shows the ground track for the Baseline and Planning scenarios while Figure 7 shows the ground track for the Activity and Combined scenarios. Note the large number of heading changes for the Activity/Combined scenarios. In the Activity and Combined scenarios the subjects were given new headings, altitudes, and airspeeds each 2 minutes for the first 5 minutes, each minute for the next 10 minutes, and each 30

SCENARIO	PILOT						
	A	B	C	D	E	F	G
Activity	1	2	3	3	1	2	1
Planning	2	1	1	1	3	1	2
Combined	3	3	2	2	2	3	3

Figure 5: Session number in which pilots flew each scenario

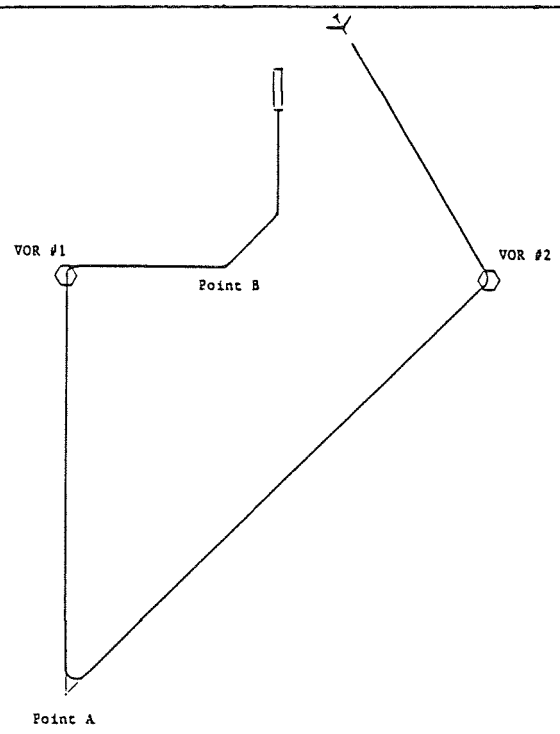

Figure 6: Nominal ground track for the Baseline and Planning scenarios

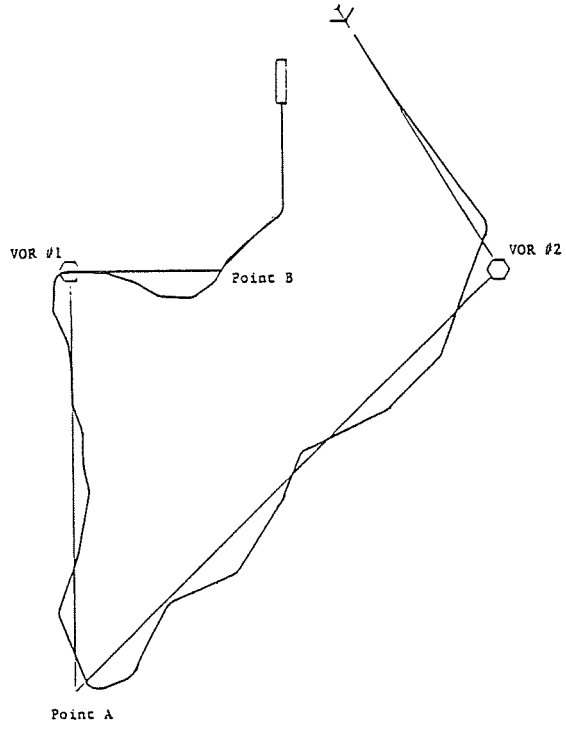

Figure 7: Nominal ground track for the Activity and Combined scenarios

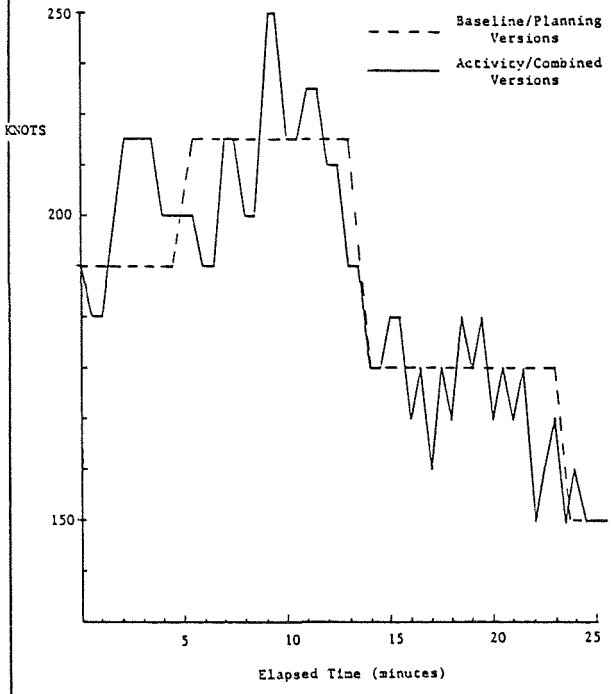

Figure 8: Planned airspeed versus elapsed time

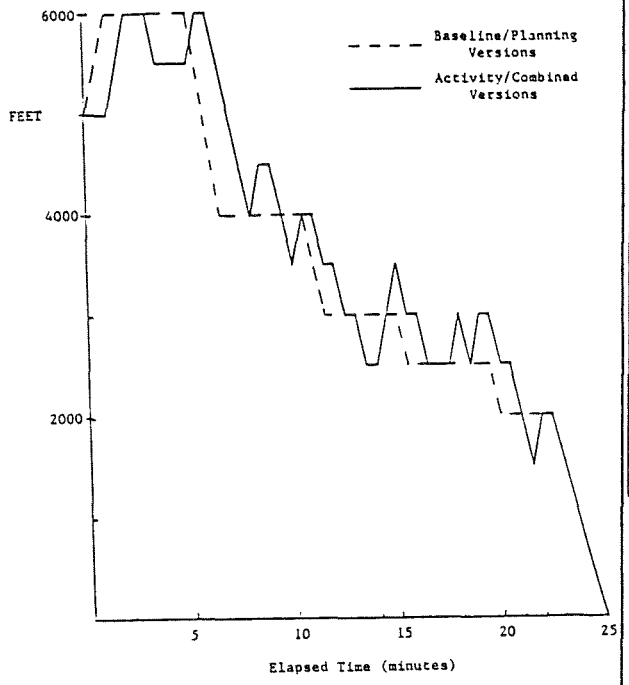

Figure 9: Planned altitude versus elapsed time

	Scenario			
	Baseline	Activity	Planning	Combined
Total Planning WU's	43	47	253	254
Total Number of Planning Tasks	3	3	23	24
Short-term Planning Tasks	0	0	14	16
Medium-term Planning Tasks	3	3	6	5
Long-term Planning Tasks	0	0	3	3
Total Activity WU's	28	150	29	142

Figure 10: Scenario characteristics

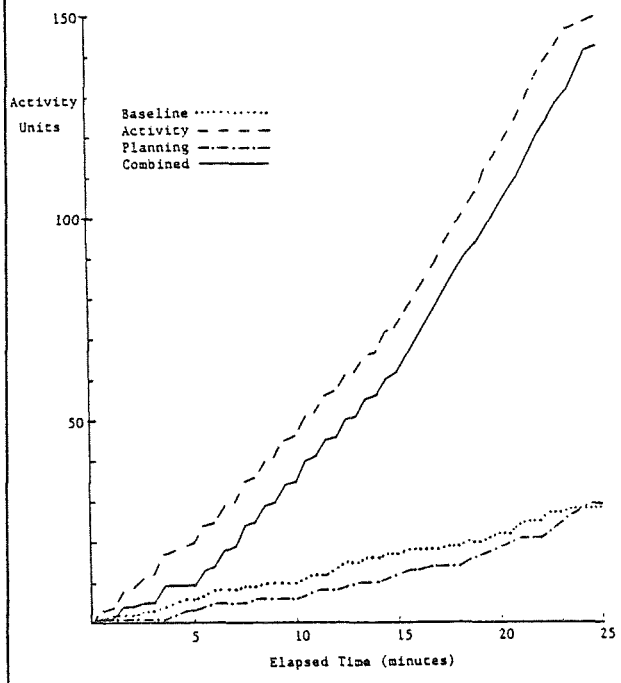

Figure 11: Accumulated Activity Workload Units versus elapsed time

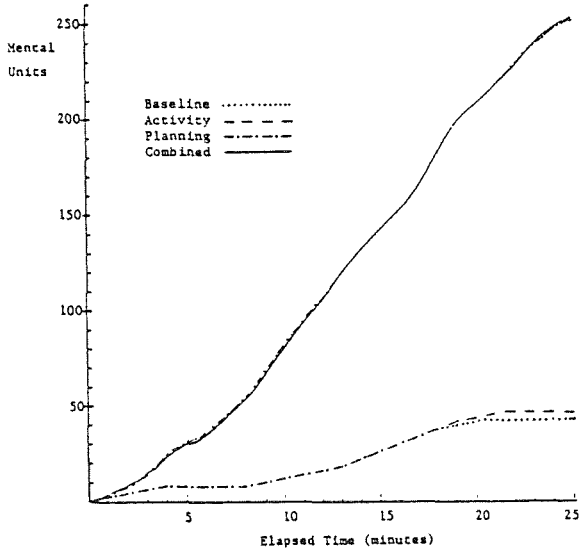

Figure 12: Accumulated Mental Workload Units versus elapsed time

seconds for the final 10 minutes. At several points, pilots were given instructions to contact ARTCC rather than perform some task. Figure 8 is an airspeed versus time plot for the various scenarios. There are 31 airspeed changes for the Activity and Combined scenarios and 3 for the Baseline and Planning scenarios. Finally, Figure 9 shows altitude versus time. The Activity and Combined scenarios have 21 directed altitude changes to 5 for the Baseline and Planning scenarios.

Each mental or physical task was evaluated and assigned a number of "workload units". The total number of workload units (WU's) and the workload unit rate were used to compare the four scenarios. An extensive explanation of the method used to calculate these workload units can be found in Berg and Sheridan, 1984.

Each scenario had a number of planning tasks. These planning tasks were categorized as either Short-term, Medium-term, or Long-term. We arbitrarily defined a short-term planning task as lasting from 0 to 4 minutes, a medium-term task lasting from 4 to 12 minutes, and a long-term task lasting over 12 minutes. The average short-term task was 2.6 minutes long, the average medium task was 7.2 minutes, and the average long-term task was 16.6 minutes.

Figure 10 summarizes the information for all four scenarios. Note that the Planning and Combined scenarios have about 5 times as many planning WU's as the Baseline and Activity scenarios. Also, the Activity and Combined scenarios have roughly 5 times as many activity WU's as the Baseline and Planning scenarios. Finally, the Planning and Combined scenarios have almost 8 times as many planning tasks as the Baseline and Activity scenarios.

In recognition of Miller's (1956) findings about human limits on immediate memory, the number of simultaneous planning tasks never exceeded 9. Although the Planning and Combined scenarios had what seemed to the subjects to be an intense level of simultaneous planning tasks, the mean number of simultaneous planning tasks was only 5.0, with a standard deviation of 1.8.

Figures 11 and 12 portray some of this workload data graphically. Figure 11 is a plot of the accumulated number of activity WU's as a function of time. Figure 12 is a plot of the accumulated number of planning WU's as a function of time. Note not only the difference between dissimilar scenarios, but also the similar workload rates for scenarios with similar types of workload.

VI. TRAINING AND INSTRUCTIONS

In addition to the initial screening sessions, each pilot participated in 4 to 10 hours of additional training. Three of the four pilots had flown the simulator before, but had never used the autopilot. They required about 4 hours of additional practice.

This autopilot is different from most commercial equipment. Longitudinal and Lateral modes must be engaged separately, adding one additional step in

selecting some autopilot functions.

Before a session's data runs, pilots "warmed up" by flying instrument approaches, turns to headings, etc., for 20 to 30 minutes. After this warm up period, the pilots were handed an Instruction Sheet, the Subjective Ratings/Comments Sheet shown in Figure 13, and a sheet which explained the scale to be used in making the subjective ratings.

In the instructions, pilots were told to fly "as well as you can" and follow all directions "to the best of your ability". They were also told that they would be scored on their ability to "follow instructions and comply with requests". Thus, they had no idea which parameter(s) would be measured. Any or all might be scored.

As explained in the instructions, the simulation was "frozen" for subjective ratings at 5:00, 16:00, and 27:00 minutes elapsed time. The desired method for scoring subjective ratings was explained, and the subjects warned that only one minute would be allowed for making the ratings during each break. Preliminary experiments had shown that the pilots only required about 20 to 30 seconds to make these ratings.

After each run, the pilots were debriefed and asked to put any comments or explanations on the rear of the Rating Sheet.

VII. DATA

Every 10 seconds, the computer recorded the aircraft's airspeed and x, y, and z position. This data yielded a ground track, and by comparing position and elapsed time, desired altitudes and airspeeds were determined. This information was then compared with the actual airspeeds and altitudes to derive altitude and airspeed error. Altitude errors were not computed during directed climbs and descents and airspeed errors were not computed during directed airspeed changes. Pilots were expected to climb or descend at a minimum of 1000 feet per minute and accelerate or decelerate to the desired airspeed within 30 seconds or at a rate of at least 50 knots per minute for airspeed changes greater than 25 knots. These rates of change are consisted with recommended piloting techniques.

Ground tracks were plotted for reference, but deviations from the nominal ground track were not scored.

Altitude deviations seemed to be the "best" objective measure to use. However, with only one objective measure, it was possible that pilots might give higher priority to one aspect of aircraft control than another. Thus, airspeed deviations were scored to serve as a check. Both variables were scored with mean absolute and RMS deviations.

Five experimentally proven subjective ratings were used in order to examine the multi-dimensionality of the mental workload. These ratings were ACTIVITY LEVEL, COMPLEXITY, DIFFICULTY, STRESS, and WORKLOAD. Ratings were

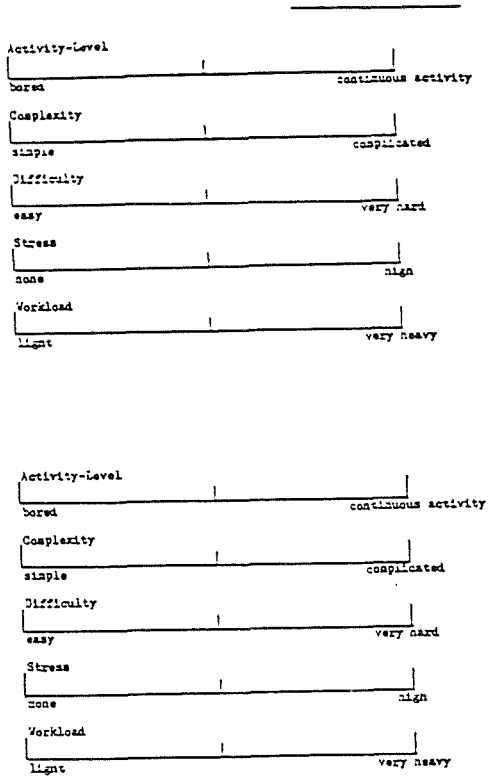

Figure 13: Subjective Ratings/Comments Sheet for Primary Experiment

SCENARIO	SEGMENT	MEAN	STD DEV	RMS
Baseline	I	39.1	18.7	50.6
	II	41.4	24.0	51.0
	III	30.6	13.8	41.4
	Overall	37.0	19.0	47.7
Activity	I	114.4	110.5	147.8
	II	97.7	24.8	153.3
	III	138.0	36.6	199.0
	Overall	116.7	67.3	166.7
Planning	I	19.5	23.0	23.5
	II	47.8	19.1	56.6
	III	55.6	34.0	60.4
	Overall	41.0	29.5	46.8
Combined	I	93.5	85.6	131.7
	II	122.4	77.5	198.2
	III	154.8	81.8	204.9
	Overall	123.6	81.7	178.3

Figure 14: Overall mean absolute and rms altitude deviations (feet)

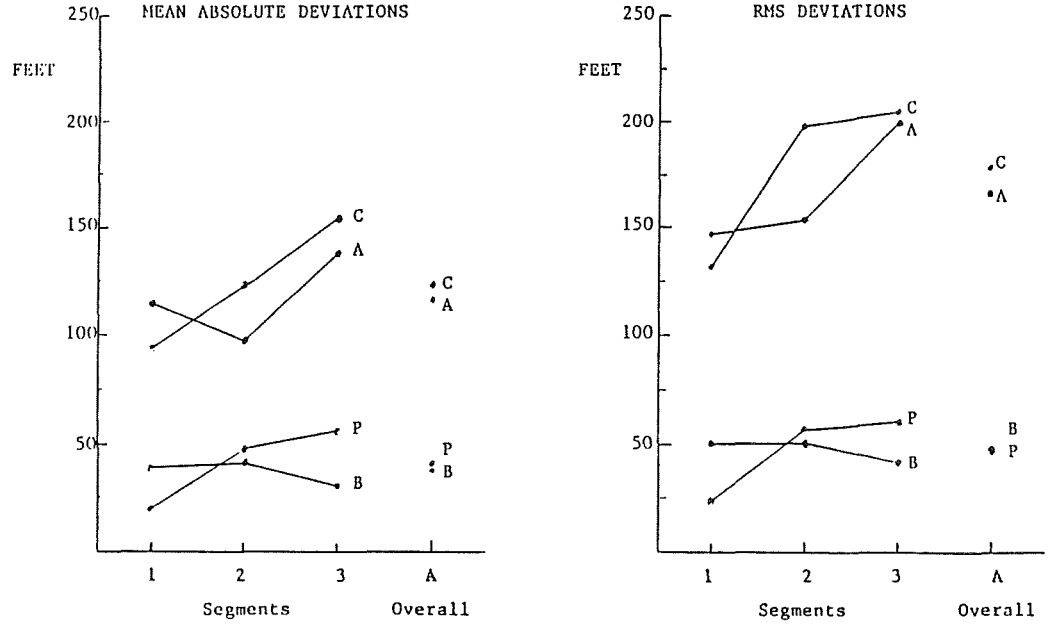

Figure 15: Average altitude deviations for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

made at three points during each run. Subjects were not asked to make an overall rating because overall ratings made during previous experiments were nearly identical to the arithmetic mean of the segment ratings and we believed the same would be true here.

The distance from the left edge of each scale to each pilot rating was measured, divided by the total scale length, and multiplied by ten. This gave subjective ratings with a possible range of 0 to 10.

An integral aspect of this set of experiments was an investigation into not only the degree of mental workload, but also the effect this effort had on observable pilot behavior. Thus, in addition to the aircraft control measures and subjective ratings just discussed, other aspects of pilot behavior were also measured.

During each run, notes were made on the pilot's compliance in carrying out assigned planning or memory tasks. All pilots were assigned specific elapsed times (clearly displayed on the instrument panel) at which to perform these tasks. Each pilot was given $\pm$ 15 seconds from the designated time in which to begin the task. If a task was begun outside these limits, it was noted. When a task was performed improperly, for example climbing to a wrong altitude or accelerating 10 knots instead of climbing 1000 feet, this was also noted. A third type of mental error was forgetting or missing an item entirely.

A final source of information was post-run debriefings. The pilots had many interesting and useful insights into mental workload, stress, and their affect on performance.

VIII. RESULTS

Learning effects

The Objective and Subjective data was examined for "learning effects". Using Student t-test and F-test techniques, we found no significant learning effect for altitude or airspeed deviations for any of the four scenarios.

Each session's Baseline run acted as a "warm up" run and served as a day-to-day metric for the Subjective ratings. For each Subjective rating, the Baseline run ratings were averaged across all seven pilots and all three runs for each pilot. This yielded an overall mean baseline rating. This mean rating was added to the difference of a session's Baseline rating and second run (Activity, Planning, or Combined) rating. This gave an "adjusted" second run rating. The intent was to compensate for day-to-day differences in emotional state, stress, fatigue, et cetera.

Using these adjusted subjective ratings, there was no "learning effect" for any of the ratings for the Activity scenario. For the Planning scenario, only the WORKLOAD ratings showed a learning effect (80 percent confidence level).

So, the extensive training, the modified counterbalancing of scenarios and subjects, and "adjusting" the subjective ratings appears to have minimized learning effect for the Activity and Planning scenarios.

However, there was some evidence of learning effect for the Combined scenario. Three subjective ratings were lower for the third sessions than the second sessions. The effect was at an 80 percent confidence level for COMPLEXITY ratings. Since post-run debriefings showed that COMPLEXITY ratings were closely tied to the pilots' ease with the autopilot, this may be due to greater familiarity with the device. Learning effect was at a much stronger 95 percent confidence level for the DIFFICULTY and WORKLOAD ratings. None of the practice rounds were nearly as intense as the Combined scenario. Furthermore, the Combined scenario was a combination of the Activity and Planning scenarios. Thus, subjects who had seen both the Activity and Planning scenarios before flying the Combined scenario had an advantage over those who flew the Combined scenario after flying only one of the others.

Finally, an analysis of variance showed no statistically significant difference for planning task performance for any scenario.

Objective activity performance results

Altitude and Airspeed error data was synthesized from the computer's output. Altitude error data is summarized in Figure 14. Note the standard deviation data in Figure 14. The bulk of pilot deviations tended to lie near the mean. However, there was usually some pilot whose deviations took an extreme, isolated jump, inflating the standard deviation for the group.

In general, just as the WU rate increased from Segment I to Segment III, so did altitude deviations (see Figure 15). Segment-to-segment mean absolute error differences were significant at a 90 percent confidence level for the Combined scenario, 95 percent for the Baseline and Activity scenarios, and 99 percent for the Planning scenario. The larger spread of individual performance in the Combined scenario was responsible for its lower confidence level.

As Figure 15 shows, there was a considerable difference (99 percent confidence level) between the manually controlled Combined and Activity scenarios and the autopilot controlled Planning and Baseline scenarios. The average deviation was 3.1 times greater (120.2 feet versus 39.0 feet) under manual control, and the rms deviation was 3.6 times greater (172.5 feet versus 47.3 feet). However, it should be noted that the manually controlled Combined and Activity scenarios also had much more difficult altitude profiles than the autopilot controlled scenarios. (See Figure 9)

Interestingly, the magnitude of mental tasking had no significant impact on the magnitude of the altitude deviations. The Baseline scenario's altitude deviations were statistically similar to those of the Planning scenario, the latter differing from the former solely in having a large number of mental planning tasks. Similarly, the mentally easy Activity and mentally demanding Combined scenarios were statistically identical.

Airspeed error data was also synthesized from the computer's output and is summarized in Figure 16. Like the altitude deviation data, some of the large standard deviations in Figure 16 are due to some pilot's momentary lapse. Most of the deviation data was fairly consistent in magnitude.

Segment-to-segment differences were significant for all four scenarios (See Figure 17). For mean absolute airspeed errors, the segments differed at a 90 percent confidence level for the Activity scenario and a 99 percent level for the Baseline, Planning, and Combined scenarios. RMS airspeed errors differed at a 95 percent confidence level for the Baseline and Activity scenarios and a 99 percent confidence level for the Planning and Combined scenarios.

Like the altitude deviation data, the magnitude of airspeed errors was a strong function of the mode of aircraft control. As shown in Figure 17, when airspeed was under manual control, deviations were much greater than when airspeed was under autopilot control. The difference was statistically significant at a 99 percent confidence level for mean absolute error and a 98 percent level for rms errors. Again, part of this result may be due to the much more difficult airspeed profile for the manually controlled scenarios (See Figure 8). This airspeed deviation data also showed little mental tasking effect. There was no significant difference between scenarios which had similar manual activity levels but different planning workloads.

Both altitude and airspeed deviations were similar for all the pilots. In general, the low experience pilots had slightly higher deviations than the most experienced pilots. However, there was enough scatter in the data to keep the differences statistically insignificant.

This objective data showed only a hint of performance degradation due to pilot workload saturation. During the Activity scenario runs, only two pilots out of seven had average mean altitude deviations greater than 150 feet in Segment III, and two other pilots had average mean airspeed deviations greater than 15 knots in Segment III. For the Combined scenario, the number of saturated pilots rose to three for the altitude deviations and remained at 2 for the airspeed deviations.

Within each scenario, there was no significant correlation between airspeed and altitude deviations because different individuals traded-off airspeed and altitude control during all four scenarios. However, overall scenario airspeed and altitude control were correlated. The Baseline and Planning scenarios had low deviations for each score and the Activity and Combined scenarios had high deviations for both scores.

Subjective ratings results

The Subjective Rating data was useful because it illustrated the impression these scenarios were making in the minds of the pilots. Thus, although only an indirect measure, one would expect these ratings to provide a better indication of mental workload than objective performance data.

SCENARIO	SEGMENT	MEAN	STD DEV	RMS
Baseline	I	1.9	0.7	2.9
	II	3.9	0.7	5.0
	III	3.4	1.9	4.4
	Overall	3.1	1.4	4.1
Activity	I	7.9	6.6	9.2
	II	9.5	4.3	12.5
	III	11.9	5.9	15.5
	Overall	9.8	5.7	12.4
Planning	I	0.7	0.4	1.0
	II	3.7	2.4	4.0
	III	3.3	1.9	3.9
	Overall	2.6	2.2	3.0
Combined	I	5.2	2.4	6.2
	II	11.0	4.5	14.3
	III	9.6	4.2	13.2
	Overall	8.6	4.4	11.2

Figure 16: Overall mean absolute and rms airspeed deviations (knots)

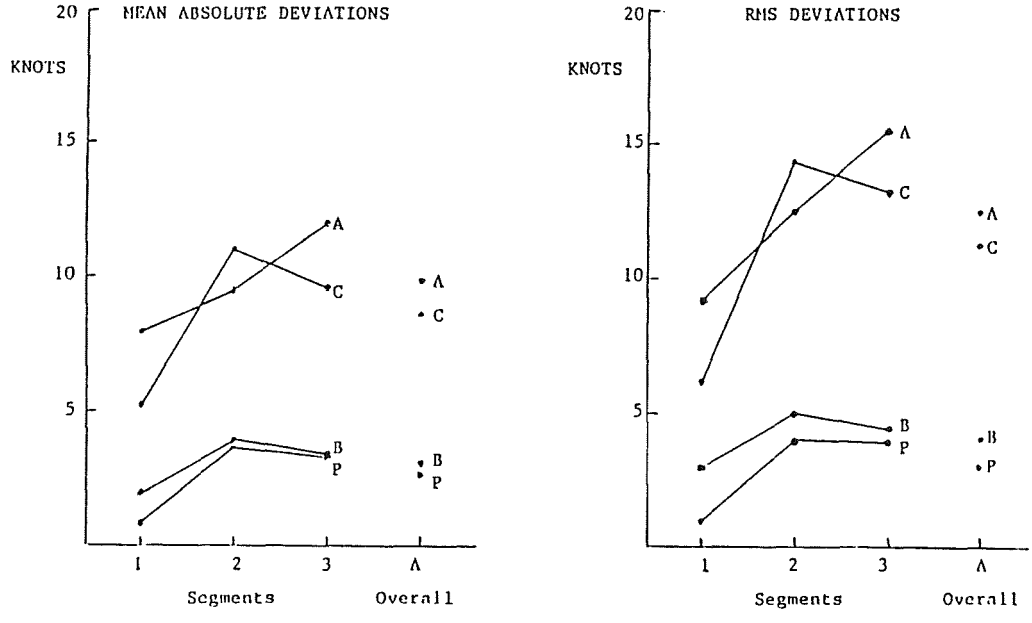

Figure 17: Average airspeed deviations for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

Figure 18 gives the subjective rating data averaged over all the pilots for each segment, scenario, and category. Note that the standard deviation data is very consistent from rating to rating and scenario to scenario. Individual ratings did not exhibit the wide variations present in the altitude and airspeed deviation data.

In general, subjective ratings for the five categories were similar for the Activity and Planning scenarios, but statistically different for those two scenarios and the Combined scenario. The Combined scenario ratings were statistically different from the Activity and Planning scenarios at a 90 percent confidence level for the WORKLOAD and DIFFICULTY ratings, a 98 percent confidence level for the ACTIVITY LEVEL ratings, and a 99 percent confidence level for the COMPLEXITY and STRESS ratings. The averaged ratings for each scenario, segment, and subjective category are plotted in Figures 19, 20, 21, 22, and 23.

The Planning scenario was essentially a Baseline scenario with an added mental task load component. The Activity scenario was a Baseline scenario complicated by a great deal of manual control work. The Combined scenario was a combination of the Activity and Planning scenarios. Therefore, the construction of the scenarios and the results plotted in Figures 19 to 23 led us to investigate whether this construct was reflected in the subjective ratings.

For all five ratings, we found the incremental difference between the Baseline scenario and each of the other three scenarios. We then examined how the sum of these increments for the Activity and Planning scenarios compared with the incremental Combined ratings. For example, suppose that the Baseline rating for DIFFICULTY was 3.0 and the DIFFICULTY ratings for the Activity, Planning, and Combined scenarios were 5.0, 5.3, and 7.5 respectively. The incremental ratings for the Activity, Planning, and Combined ratings would then be 2.0, 2.3, and 4.5. The sum of the Activity and Planning scenario increments would be 4.3. This increment (averaged with the increments for all the other pilot's increments) was compared with the Combined scenario's increment of 4.5 (averaged with the other pilot's Combined scenario increments).

For all five subjective ratings, the sums of the Activity and Planning increments were not statistically different from the incremental Combined ratings.

In view of the well established fact that the magnitude of subjective perception is logarithmically related to stimulus magnitude, this nearly linear response was somewhat surprising. At no point were the pilots ever told that the Combined scenario contained the sum of manual and mental tasks from the Activity and Planning scenarios. However, although this result may be useful when going from low or moderate workloads to high workloads, this linearity must obviously break down when trying to go from high workloads to even greater workloads.

How difficult did the pilots think the three non-Baseline scenarios were?

SCENARIO	SEGMENT I	II	III	Overall	Std Dev
BASELINE					
Activity Level	2.6	2.8	3.5	3.0	0.9
Complexity	2.3	2.5	3.4	2.7	1.0
Difficulty	2.2	2.4	3.1	2.6	0.8
Stress	2.0	2.1	3.0	2.4	0.7
Workload	1.8	2.2	2.8	2.3	0.5
ACTIVITY					
Activity Level	5.4	6.7	7.3	6.5	1.2
Complexity	3.4	5.0	5.7	4.7	1.3
Difficulty	4.5	6.0	6.7	5.7	1.1
Stress	3.7	4.9	6.1	4.9	1.1
Workload	3.9	5.5	7.0	5.5	1.4
PLANNING					
Activity Level	4.1	5.1	7.0	5.4	1.4
Complexity	4.1	4.6	5.9	4.8	1.3
Difficulty	3.3	4.0	6.3	4.6	1.1
Stress	3.3	3.9	5.3	4.2	1.2
Workload	3.9	4.7	6.2	4.9	1.2
COMBINED					
Activity Level	5.9	8.3	9.8	8.0	1.1
Complexity	5.4	6.9	8.5	6.9	1.6
Difficulty	5.9	7.8	9.1	7.6	1.7
Stress	5.5	7.6	8.9	7.3	1.3
Workload	5.7	7.7	9.6	7.7	1.6

Figure 18: Average Subjective Ratings for each Segment (Adjusted)

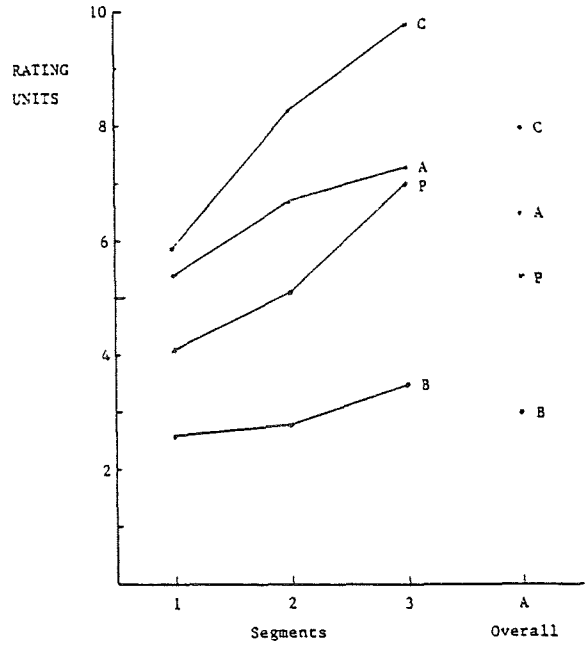

Figure 19: Average subjective ACTIVITY LEVEL ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

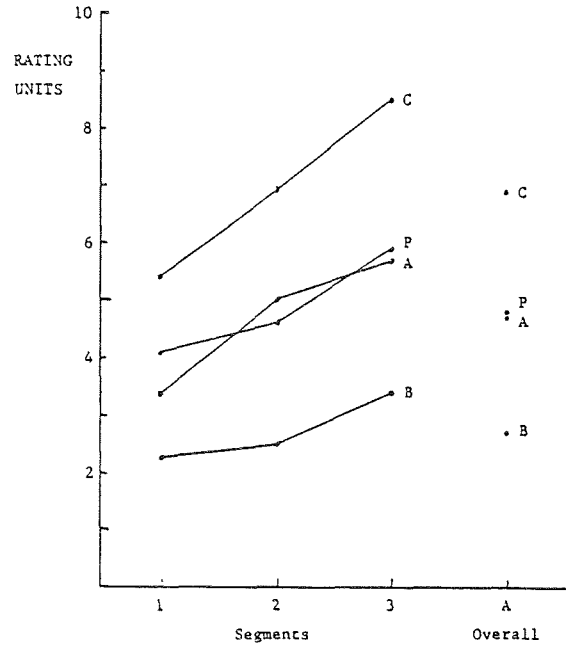

Figure 20: Average subjective COMPLEXITY ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

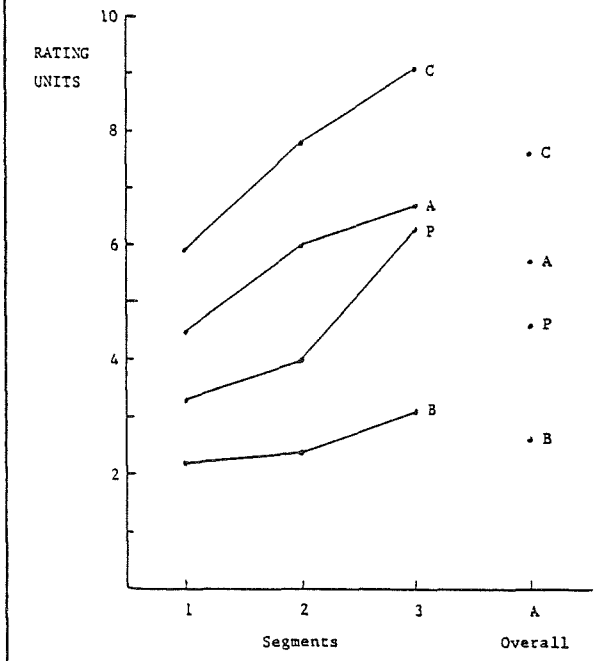

Figure 21: Average subjective DIFFICULTY ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

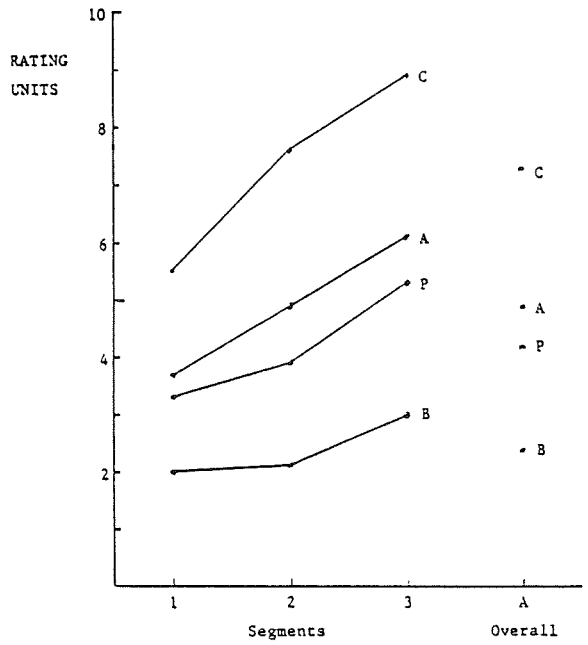

Figure 22: Average subjective STRESS ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

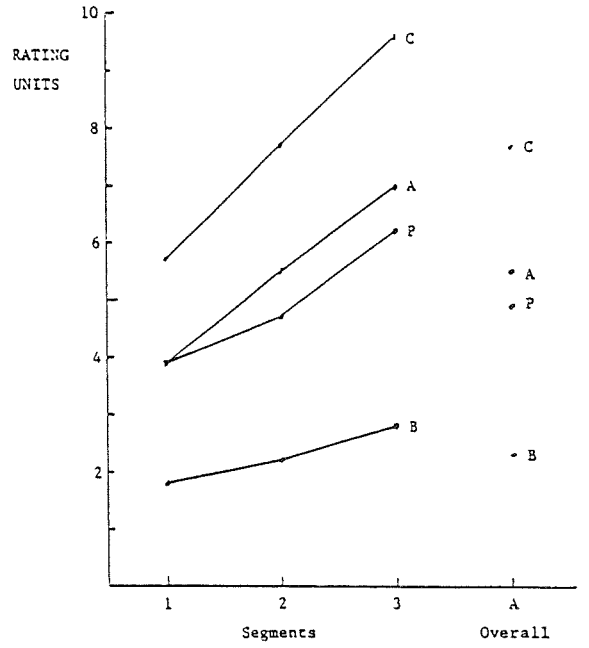

Figure 23: Average subjective WORKLOAD ratings for the Baseline (B), Activity (A), Planning (P), and Combined (C) scenarios

SCENARIO	Activity		Planning		Combined	
DEVIATION TYPE	mean	rms	mean	rms	mean	rms
Activity Level	.401	.805	.880	.782	.986	.953
Complexity	.389	.797	.843	.777	.999	.896
Difficulty	.403	.807	.817	.746	.990	.945
Stress	.583	.911	.428	.792	.986	.954
Workload	.568	.903	.882	.823	.999	.911

Figure 24: Pearson Product-Moment Correlation Coefficient for aggregate Altitude Deviations and Subjective Ratings

	SEGMENT		
	I	II	III
Activity Level	5.8	7.4	9.6
Complexity	6.5	6.8	8.3
Difficulty	4.5	4.1	11.0
Stress	1.1	3.0	3.1
Workload	5.5	5.6	10.0
Altitude Error: Mean	11.9	59.8	110.6
RMS	12.7	60.3	110.6
Airspeed Error: Mean	1.2	4.1	7.3
RMS	2.1	4.2	7.3

Figure 25: Example of related performance deterioration and subjective saturation: Pilot C; Planning Scenario

The only scenario which consistently "saturated" pilots was the Combined scenario. If one defines a "saturated" pilot as one who scores a subjective rating category at 9.0 or higher, the Activity scenario was least likely to saturate pilots. This is interesting because when there were significant differences between the Activity and Planning scenario ratings, the Activity scenario rating was always slightly higher. Thus, certain individuals found the Planning scenario very difficult, while the pilots as a group, found the Planning scenario slightly less demanding than the Activity scenario.

For the Activity scenario, there was one saturated rating for WORKLOAD. For the Planning scenario, there were two saturated ratings for ACTIVITY LEVEL, and one each for DIFFICULTY and WORKLOAD. For the Combined scenario, there were five saturated ratings for ACTIVITY LEVEL and WORKLOAD, four for DIFFICULTY and STRESS, and two for COMPLEXITY.

These experiments verified that on a subjective level, a difficult, purely mental task load can equal a difficult, purely manual task load. In general, all the subjective category ratings were similar for the Planning and Activity scenarios.

There was no consistent correlation between subjective ratings and a pilot's experience level. This is not surprising since there is no universal subjective mental metric. Two persons working equally hard may rate their workloads very differently. They have different utilities, and one person may use a linear scale while another uses a logarithmic, and still another, an exponential scale.

Objective activity performance versus subjective ratings

We looked for a correlation in altitude or airspeed deviations with each pilot's subjective ratings. On an individual basis, objective activity performance data and subjective ratings were uncorrelated. This result was not unexpected, and had been reported previously. See, for example, the short discussion in Kantowitz, Hart, and Bortolussi, 1983.

Nevertheless, in the aggregate, objective performance data was correlated with subjective ratings. Using Pearson's Product-Moment Correlation Coefficient, "r", rms altitude errors weakly correlated with the corresponding subjective ratings for the Activity scenario (See Figure 24). ACTIVITY LEVEL, COMPLEXITY, and DIFFICULTY correlated with an "r" of 0.8 (.805; .797; .807). For the STRESS and WORKLOAD ratings, "r" was about 0.9 (.911; .903).

Correlations were slightly better for the Planning scenario. Mean absolute altitude deviations and ACTIVITY LEVEL had an "r" of .880. COMPLEXITY, DIFFICULTY and WORKLOAD had "r's" of .843, .817, and .882. Mean altitude errors did not correlate with STRESS, but rms errors did: .792. The ability of the rms error data to correlate with STRESS ratings better than the mean deviation data did might be due to the fact that the rms data weights large errors more heavily than small errors. Intuitively, beyond a certain point, stress should be an exponential function of the magnitude of deviations. Thus, large deviations would be better reflected in the rms values and

STRESS ratings.

There was excellent correlation between mean absolute error data and all five ratings for the Combined scenario. The lowest "r" was for STRESS, (.986) with COMPLEXITY having an "r" of .9999. Because the pilots were heavily loaded during the Combined scenario, they may have been operating near their personal limits. This may have lessened differences in proficiency resulting in the good correlation between objective performance data and the subjective ratings.

Tulga and Sheridan, 1980, reported that once a subject passed "saturation", performance deteriorated sharply. While flying the Planning scenario, Pilot C crashed during Segment III. Figure 25 lists relevant data for Segments I, II, and III for this pilot. Although he reported only low STRESS, the other four subjective factors sharply increased from Segment II to Segment III. Likewise, note that his mean absolute and rms altitude errors increased by 85 percent and 83 percent, and the corresponding airspeed errors increased by 78 percent and 74 percent from Segment II to Segment III. Although one can argue about which was cause and which was effect, mental saturation accompanied a severe performance degradation.

Planning/memory task performance

As workload increased, there were a number of ways that each pilot could respond to these requests for some action at a future time. They could fail to perform a task, choosing not to do it or simply forgetting to do it. They could also perform the task incorrectly, do some unrequested task, or perform the required task at some time other than the directed time. Overall planning task error percentages for each scenario are plotted in Figure 26.

Although the planning task load for the Baseline and Activity scenarios was the same, the overall error percentage was much higher for the Activity scenario. Similarly, although the Planning and Combined scenarios had similar planning task loads, the Combined scenario percentage was much higher (and differed at a 99 percent confidence level). The Planning and Activity scenarios had similar Subjective ratings, but their mental task performance data was very different. A high manual workload had a profound effect, increasing errors.

The standard deviations for the overall error percentages varied widely from scenario to scenario. For the Baseline and Planning scenarios where the error percentages were low, standard deviations were only 8.8 and 13.4 percent respectively. The difficult Combined scenario had a standard deviation of 27.2 percent, indicating more variability among the pilots. The Activity scenario showed the greatest variability. The low number of mental tasks and the high error percentages for some pilots resulted in a standard deviation of 51.4.

Figure 27 illustrates the error percentages for each segment and scenario. The performance for the Planning and Combined scenarios was virtually identical for Segment I. However, for Segments II and III, the difference

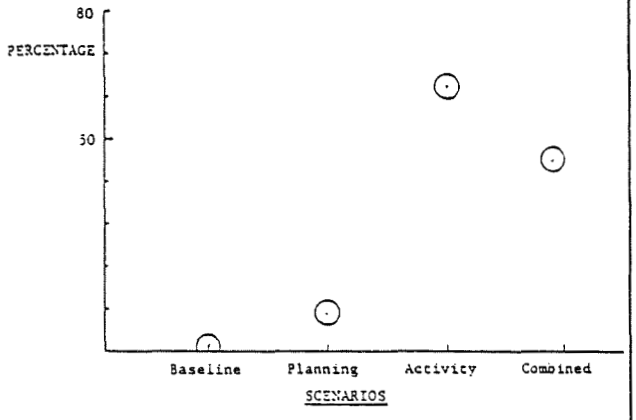

Figure 26: Overall percentage of planning/memory task errors

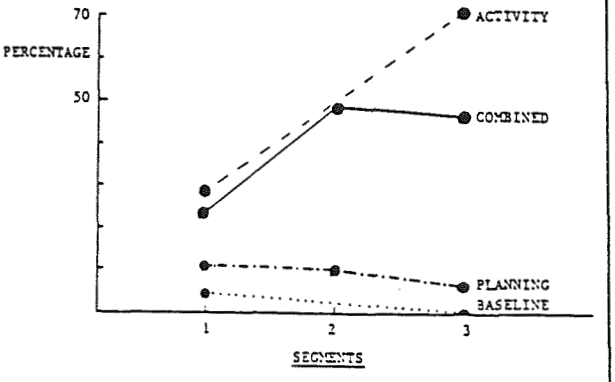

Figure 27: Percentages of planning/memory task errors per segment

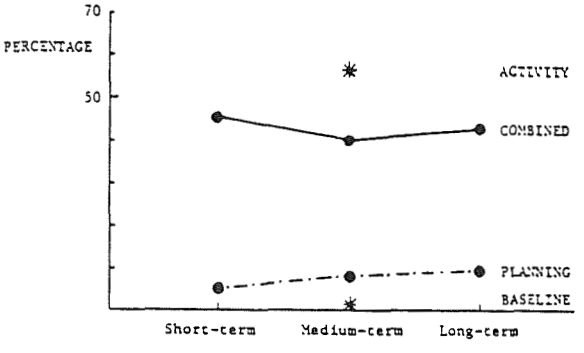

Figure 28: Error percentages for three different planning task time spans

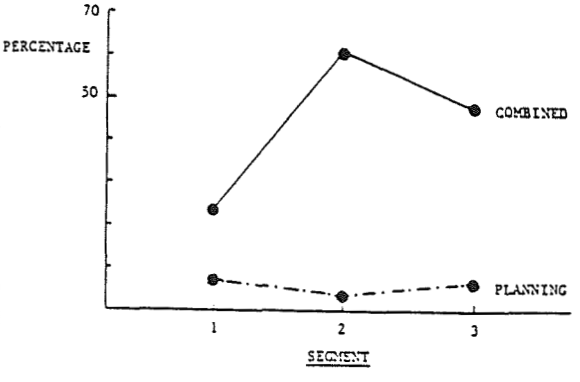

Figure 29: Short-term planning/memory task error percentages by segment

between the two scenarios was significant at the 99.9 percent confidence level. Although individual performance differed a great deal, the data suggests that at low or moderate levels, manual control workload does not affect mental performance. Sufficient cognitive reserve exists to handle all tasks. However, at relatively high manual control levels, cognitive reserves disappear and mental performance deteriorates. Figure 26 suggests that this mental deterioration may even be evident for low levels of mental tasking, such as in the Activity scenario.

The various planning tasks were categorized as Long-term, Medium-term, or Short-term based upon the length of time the pilot had from receiving the task assignment to performing it. When aggregated for each scenario, the data yields the plot shown in Figure 28. Analyzing the error percentages for each scenario, there was no statistically significant difference between the three different task time spans. This was probably because the pilots were allowed to take notes. Additional errors probably arose in the Short-term tasks when the pilots struggled to plan and perform these tasks in a very busy environment. Thus, they would miss some tasks or perform them late. This balanced the errors engendered in the Long-term tasks by the pilots forgetting about tasks.

An analysis of the data supports this hypothesis. There were no Long-term planning errors due to performing an action at the wrong time. However, 33 percent of the Short-term and 53 percent of the Medium-term errors were due to performing an action at the wrong time.

Planning task errors for all three time spans were affected by manual-control activity. Note in Figure 28 that the two low manual workload scenarios (Baseline and Planning) had low error percentages while both high manual workload scenarios (Activity and Combined) had high error percentages. The Activity scenario had a high error percentage even though its planning task load was low.

Looking only at the two scenarios (Planning and Combined) with a high planning task load, the differences between the scenarios was statistically significant for all three time spans. Differences were significant at an 80 percent confidence level for medium-length tasks, at a 95 percent level for long-term tasks, and 98 percent level for short-term tasks. Thus, the level of manual control was again decisive in determining mental performance. The data was too coarse and individual pilot performance was too variable to make standard deviation data useful.

Only the Planning and Combined scenarios had Short-term planning tasks. Examining Figure 29, differences between the Planning and Combined scenarios for Short-term planning tasks were not statistically significant in Segment I. However, the differences were at a 98 percent confidence level for Segments II and III, when workloads were higher.

All four scenarios had Medium-term planning tasks. Looking at Figure 30, there was no statistically significant difference between the scenarios in Segments I or II. However, in Segment III, the highest workload segment,

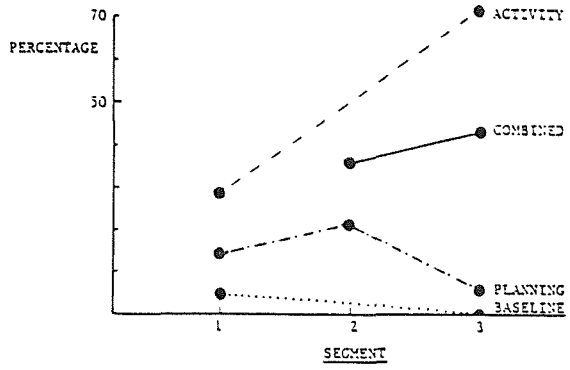

Figure 30: Medium-term planning/memory task error percentages by segment

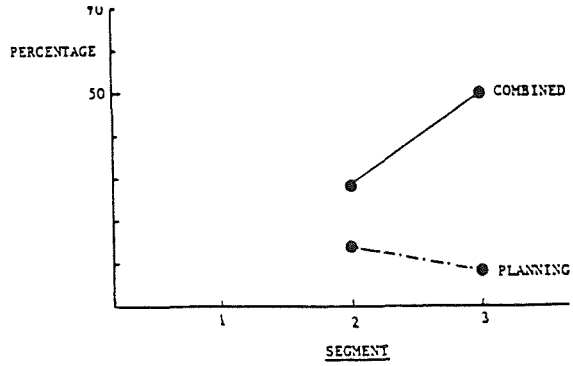

Figure 31: Long-term planning/memory task errror percentages by segment

the Combined scenario errors were higher than the Planning scenario errors (90 percent confidence level). The Planning and Activity difference was even greater (at a 95 percent confidence level). The Activity and the Combined scenarios, and the Planning and Baseline scenarios were statistically similar. Once again, at high overall workload levels, the presence of a high manual task load made a significant difference.

Figure 31 is a plot of the Long-term planning task results. In Segment II, the Planning and Combined scenarios were statistically indestinguishable. However, at the higher workload level of Segment III, the error percentage for the Combined scenario was clearly greater (90 percent confidence level).

The Activity and Planning scenarios had moderate manual or mental workloads, respectively. At these levels, error percentages were similar for all of the pilots. However, some differences arose in the high workload Combined scenario. The low experience pilots averaged 14.0 task errors while the high experience pilots averaged 7.3 task errors. Thus, there were signs of experience related saturation in this mental performance data which was much less obvious in the objective performance data and subjective rating data. This difference was verified at a 95 percent confidence level.

The number of individual planning errors and individual altitude or airspeed deviations were not correlated. Nor were planning errors and subjective ratings. However, in the aggregate, altitude and airspeed deviations, subjective ratings, and the number of planning errors all increased with increasing task loads.

Pilot comments

The planning task instructions given to the pilots were seldom in chronological order. This was done to make the planning function more difficult. This strategy apparently worked, since several subjects mentioned that instructions "mixed in time" were difficult to organize.

Some pilots considered the autopilot a hindrance while others found it a useful aid. Several pilots stated that when things "really got busy", the autopilot was the only thing which kept workload at a manageable level. But, several pilots reported that having to plan how to use the autopilot was worse than the demanding manual control work. An oft-reported result is once again clear: if the initial set-up or programming of a "pilot aid" is difficult or unduly time consuming, pilots will use manual procedures and avoid its use.

A number of the pilots stated that planning and memory items tended to get second priority to immediate task demands. This is consistent with the finding that a high activity workload significantly increased planning task errors. Pilots were obeying the prime directive taught every student pilot: "First, fly the aircraft!" These statements and results are also consistent with Tulga and Sheridan's (1980) finding that subjects don't plan ahead when they're very busy.

Finally, the pilots mentioned four items which increased their mental stress and workload. One was the "annoyance" factor caused by having too many things to do or by being interrupted before completing a task. This type of problem is common on final approach when the need to fly and/or monitor equipment, clear for other aircraft, look for the runway, interact with ATC, and run aircraft checklists, combine to make the flight deck a busy, stressful environment. A second item was the effect of "getting behind". Again, this is most likely to occur when things get very busy. The stress generated by a lengthening "mental queue", combined with the possible need to modify a former plan, increases the perceived workload. Similarly, abnormal events significantly increase workload, disrupt concentration, and increase the frustration level. These effects have been discussed in the open literature. See, for example, Hart and Bortolussi (1983), Jensen and Chappell (1983), and Tanaka, Buharali, and Sheridan (1983). The fourth item concerned the effect of adding an increment of workload when the workload is already high. As the pilot becomes task saturated, additional tasks must be prioritized, added to a mental queue, or ignored. This increases stress, frustrates the pilot, and increases his mental manipulations. These factors result in lower performance, increased mental workload, and lower safety margins.

IX. FINDINGS AND CONCLUSIONS

1. The number of assigned mental tasks had no statistically significant impact on the degree of aircraft control. The level of manual workload was the decisive factor. When mental tasking was high but manual tasking was at a low level, altitude and airspeed deviations were small. When mental tasking was low but manual tasking was high, altitude and airspeed deviations were large. The level of mental activity affected aircraft control only when mental workload reached "critical" levels.

2. Incremental subjective ratings were calculated relative to the ratings for a Baseline scenario. The incremental rating for a high manual workload scenario added to the incremental rating for a high mental workload scenario was equal to the incremental rating for a scenario which combined both types of workloads.

3. Subjective ratings given by individual pilots during the high manual tasking scenario were very similar. However, there were individual differences in the subjective ratings for the high mental tasking scenario. Some pilots were not stressed by the mental tasks while others significantly increased their subjective ratings. Subjective ratings were more sensitive than aircraft deviation measures in indicating individual mental workloads.

4. At low or moderate levels of manual and mental task loads, aircraft deviations and memory task performance did not correlate with the subjective ratings. At high workload levels, the correlation was very good. It's possible that at lower task loads, there is reserve mental capacity which varies from pilot to pilot, affecting performance and ratings. At high workload levels, all pilots may be tapping most or all of their mental

capacity, resulting in much greater consistency between performance and the subjective ratings.

5. The magnitude of manual task loads was decisive in determining the ability of the pilots to handle mental tasks. A mentally difficult, manually easy scenario resulted in a low percentage of mental errors. A mentally easy, manually difficult scenario resulted in a high percentage of mental errors. The manual activity was presumably consuming a great deal of the pilots' mental processing capacity, even when they were not aware of it. This finding was equally valid for long-term, medium-term, and short-term mental tasks. Thus, pilots flying a highly automated flight control system might be able to more easily handle high mental workloads.

6. Under conditions of high manual and mental workload, the low experience pilots did not perform mental tasks as well as the high experience pilots did. However, objective aircraft performance and subjective ratings were similar for the two groups. Thus, these experiments suggest that monitoring and measuring mental performance might be a more sensitive indicator of mental workload and reserve mental capacity than objective aircraft performance data or subjective ratings.

X. RECOMMENDATIONS FOR FOLLOW-UP STUDIES

1. In future studies of this type or in a re-examination of this study, it might be enlightening to "filter" the data by only considering altitude deviations greater than $\pm$ 50 or $\pm$ 100 feet, or airspeed errors greater than $\pm$ 5 or $\pm$ 10 knots. This might compensate for individual pilots' tolerance boundaries.

2. Subjective Ratings should be used in future studies of mental workload. They provide a useful, if imprecise, measure of the pilot's mental state.

3. The only significant difference found between the low experience and high experience pilots was in their performance of mental planning tasks. This should be further investigated in future studies.

XI. REFERENCES

1. Berg, S. L.; and Sheridan, T. B.. Measuring workload differences between short-term memory and long-term memory scenarios in a simulated flight environment. Proceedings of the twentieth annual conference on manual control; 1984 June 12-14: 397-416.

2. Berg, S. L.; and Sheridan, T. B.. Effects of time span and task load on pilot mental workload. Cambridge, Massachusetts: Massachusetts Institute of Technology; 1985. Master's Thesis.

3. Hart, S. G.; Bortolussi, M. R.. Pilot errors as a source of workload. Paper presented at the second symposium on aviation psychology; 1983 April 25-27; Columbus, Ohio.

4. Jensen, R. S.; Chappell, S.. Pilot performance and workload assessment: an analysis of pilot errors. Report submitted to NASA Ames Research Center in 1983 February.

5. Johannsen. G.. Workload and Workload Measurement. N. Moray, ed. Mental workload; theory and measurement. New York: Plenum; 1979.

6. Kantowitz, B. H.; Hart, S. G.; Bortolussi, M. R.. Measuring pilot workload in a motion-based simulator: asynchronous secondary choice-reaction task. Paper submitted to the IEEE Transactions on Systems, Man, and Cybernetics. 1983.

7. Katz, J. G.. Pilot workload in the air transport environment: measurement, theory, and the influence of Air Traffic Control. Massachusetts Institute of Technology Flight Transportation Laboratory. FTL Report R80-3. 1980 May.

8. Leplat, J.. Factors determining workload. Ergonomics. 21: 143-149; 1978.

9. Miller, G. A.. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review. 63: 81-97; 1956.

10. Rehmann, J. T.; Stein, E. S.; Rosenberg, B. L.. Subjective pilot workload assessment. Human Factors. 25(3): 297-307; 1983.

11. Sheridan, T. B.; Simpson, R. W.. Toward the definition and measurement of the mental workload of transport pilots. Massachusetts Institute of Technology Flight Transportation and Man-Machine Laboratories, Technical Report No. DOT-OS-70055. 1979 January.

12. Tanaka, K.; Buharali, A.; Sheridan, T. B.. Mental workload in supervisory control of automated aircraft. Proceedings of the nineteenth annual conference on manual control; 1983 May 23-25: 40-58.

13. Tulga, M. K.; Sheridan, T. B.. Dynamic decisions and workload in multitask supervisory control. IEEE Transactions on Systems, Man, and Cybernetics. SMC-10 (No. 5): 217-232; 1980.

14. Walden, R. S.; Rouse, W. B.. A queueing model of pilot decision making in a multi-task flight management situation. IEEE Transactions on Systems, Man, and Cybernetics. SMC-8 (No. 12): 867-875; 1978.

15. Williges, R. C.; Wierwille, W. W.. Behavioral measures of aircrew mental workload. Human Factors. 21(5): 549-574; 1979.