THE EFFECT OF PREVIOUS LEVELS OF WORKLOAD IN A SIMULATED FLIGHT TASK

Thesis

Submitted to

The Graduate School of Arts and Sciences

UNIVERSITY OF DAYTON

In Partial Fulfillment of the Requirements for

The Degree

Master of Arts in Psychology

by

Douglas Scott Fischer

University of Dayton

Dayton, Ohio

April, 1995

APPROVED BY:

William F. Moroney, Ph.D.
Chairperson, Thesis Committee

David W. Biers, Ph.D.
Thesis Committee Member

F. Thomas Eggemeier, Ph.D.
Thesis Committee Member

CONCURRENCE:

F. Thomas Eggemeier, Ph.D.
Chairperson, Department of Psychology

ABSTRACT

THE EFFECT OF PREVIOUS LEVELS OF WORKLOAD IN A SIMULATED FLIGHT TASK

Name: Fischer, Douglas, Scott
University of Dayton, 1995

Chairperson, Thesis Committee: William F. Moroney, Ph.D.

The effect of workload contexts on subsequent performance and reported levels of workload ratings has crucial implications regarding workload transition. However few studies have examined workload context effects; and those that have, report contradictory results. This study attempts to determine if the failure to find evidence of workload context effects might be attributable to methodological factors such as task duration, task difficulty, and experimental design.

Twelve subjects flew three sessions of three trials each on a computer-based flight simulator, and rated the workload after each trial. A pre-post experimental design presented the first and third trials at a medium level of difficulty while the second (experimental) trial was of low, medium, or high difficulty. Crosswinds of 2, 12, and 22 knots created the levels of low, medium, and high task difficulty.

Analyses of the performance and workload data did not reveal significant differences in Trial 3 as a function of prior task context presented in Trial 2. The inability to find workload context effects in the present study suggests that previous inconsistent findings can not be attributed to differences in task duration and experimental design. Rather, it appears that contradictory results may be attributable to differences in the range of task difficulty employed, the workload measurement tool, or both.

ACKNOWLEDGMENTS

I would like to thank Dr. William F. Moroney, my advisor, for providing the direction to complete this thesis. I appreciate his expertise and patience; he is a mentor and good friend. I also appreciate my thesis committee for their influence in my education and professional career.

I would also like to thank everyone who helped me with this work. Special appriciation goes to my wife, Maureen E. Reilly, who offered her support throughout.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Understanding the carry-over effects of previously completed tasks in complex environments is important for the accurate assessment of operator performance and system evaluation, and has crucial implications regarding workload transition. Carry-over effects occur when a previous treatment alters performance in a subsequent treatment. A workload context effect, similar to carry-over, occurs when previous levels of workload difficulty influence an operator's perception of workload on subsequent tasks. Performance and workload context effects may occur under conditions in which a pilot or operator progresses from one level of workload to another. Previous investigations of workload context effects, however, have provided contradictory evidence regarding context effects on simulated tasks. Using the NASA-Task Load Index (TLX), Moroney, Reising, Biers, and Eggemeier (1993) reported that prior trial difficulty did not carry over to workload ratings of subsequent trials in a simulated flight task. However, Hancock, Williams, Miyake, and Manning (1992), using a compensatory tracking task, reported that both Subjective Workload Assessment Technique (SWAT) and TLX ratings increased following performance of the easier task and decreased following performance of the more difficult task.

History of Workload Context Effect Research

The purpose of this experiment was to determine if the contradictory results of Moroney et al. and Hancock et al. might be due to paradigm differences. This section

provides a summary of the two previous studies, followed by a comparison of the paradigms employed by both studies.

Moroney, Reising, Biers, and Eggemeier (1993). Moroney et al. required pilots to "fly" a simulated flight task under low, medium, or high workload levels. It was hypothesized that the different levels of task difficulty would carry-over to subsequent flight trials, effecting a subject's performance and perception of workload. Figure 1 shows the instrument panel and external view from a simulated Cessna 182 used to complete all flights.

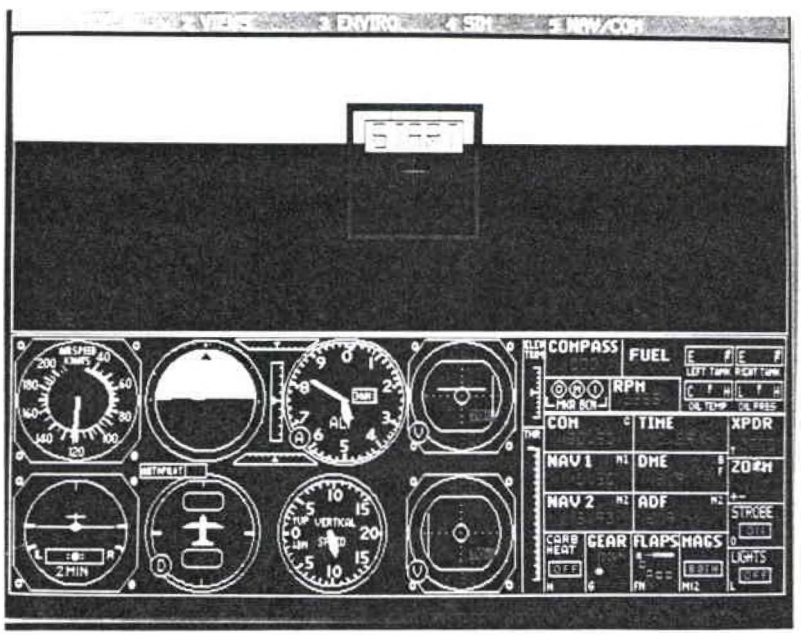


FIGURE 1: View of Simulator used in Moroney et al. (1993)

Subjects flew through a series of 10 mid-air gates (squares 200 ft X 200 ft) centered at a constant altitude of 6000 ft. Each gate was separated by approximately 0.7 miles (See Figure 2).

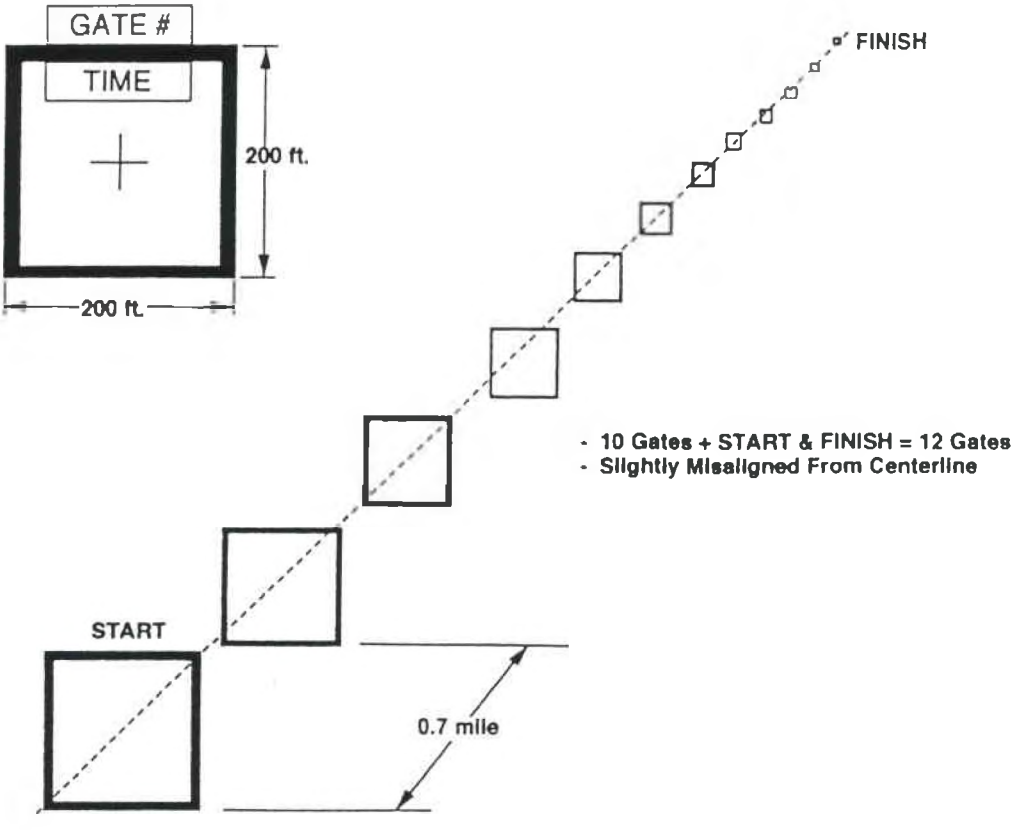


FIGURE 2: Flight Path used in Moroney et al. (1993)

Moroney et al. used twelve subjects between the ages of 18 and 22. Two of the subjects had previous flight experience. Before the experiment began, subjects learned "basic flight maneuvers" in the simulated Cessna 182. To qualify for the experiment, subjects were required to fly through ten mid-air gates with no crosswind, and attain a

minimum criterion score. Subjects were not allowed to vary either the power setting (initially 120 knots) or the aircraft's trim and gear position.

The experiment required subjects to "fly" the identical course with the addition of crosswinds. Crosswinds at 2, 12, and 22 knots (from a heading of 270 degrees) created the levels of low, medium, and high workload, respectively. Subjects were instructed to fly through a crosshair located in the center of each gate as rapidly as possible and to avoid missing any gates. The TLX was administered according to the procedures specified by NASA (1986). Weights, however, were not assigned to the dimension scales based upon previous research that found no statistical difference between overall workload ratings based on weighted and unweighted TLX scores (Moroney, Biers, Eggemeier, and Mitchell, 1991).

Moroney et al. tested subjects in three experimental sessions with a 24-hour interval between sessions. Each session consisted of an Experimental and Context phase consisting of three trials each. All subjects were independently exposed to one of the three levels of crosswind conditions on the three experimental trials for that day and then to three trials at the medium (12 knot) level of difficulty (context effect trials). The TLX was administered at the conclusion of each trial. Each subject received all combinations of flying conditions. The experimental conditions and experimental design are shown in Table 1.

Analyses revealed that performance scores under the low difficulty conditions were significantly greater than scores under the medium difficulty conditions. Similarly, the medium difficulty performance scores were significantly greater than the high difficulty scores. Thus, mean performance scores decreased as difficulty increased. There was no effect of task difficulty on performance during the context effect trial, indicating that subjects obtained similar performance scores during subsequent trials. Using the unweighted TLX scoring technique, analyses further revealed that the level of

previous task difficulty did not significantly influence workload ratings during the context effect trial (Moroney, et al., 1993).

TABLE 1: Comparison of Paradigms Used by Moroney, Reising, Biers, and Eggemeier (1993) and Hancock, Williams, Miyake, and Manning (1992)

Variables	Moroney et al.		Hancock et al.		
Task	Flight course--straight, no crosswind		Compensatory tracking		
Design	Two Phases		Three Phases		
	Experimental	Context Effect*	Baseline	Experimental	Context Effect*
	Low	Med	Med	Low	Med
	Med	Med	Med	Med	Med
	High	Med	Med	High	Med
Number of Trials	3 per phase		1 per phase		
Trial Duration	3 minutes		5 minutes		
Intertrial Interval	30-40 seconds		220 seconds		

* Context Effect phases consisted of those trials in which the context effect, induced during the experimental trials was expected to occur.

Hancock, Williams, Miyake, and Manning (1992). An analogous study conducted by Hancock, Williams, Miyake, and Manning (1992) also investigated the effects of prior task loadings on subsequent workload and performance. Twelve subjects controlled a small animated plane on the display screen, and attempted to keep the plane aligned at the center of a sight circle.

Three levels of tracking task difficulty (i.e., low, medium, or high) were used to determine if previous levels of task difficulty influenced the perceived workload on subsequent tasks. The low, medium, and high levels of difficulty were determined by varying the cutoff frequency and amplitude of the forcing function.

Table 1 also shows the experimental design of Hancock et al. Each subject participated in the Baseline, Experimental, and Context phases. Each phase consisted of three trials of five minute durations where the first and third trials were under medium level of difficulty and Trial 2 was either low, medium, or high level of difficulty (i.e., M-L-M, M-M-M, M-H-M). There was a minimum of two days between each session.

Before the experiment began, subjects practiced the first session of three medium difficulty levels (M-M-M). Each phase was followed by a 200-second interval during which physiological and subjective data were gathered. Both the SWAT and TLX ratings were used to measure subjective workload. Tracking task performance was assessed through both root mean square error (RMSE) and combined lead time (CTL) measures. Hancock et al. indicate that the CTL was a two dimensional time lead that translates joystick input movements to operator output.

To measure context effects, analyses were conducted on percent changes of performance scores and workload ratings. The results indicate that, under the high difficulty manipulation, CTL percent change scores decreased significantly from the Baseline to the Context phase. This means that either the subjects became slower or their predictions on the task were worse, as a result of Trial 2 high difficulty manipulation. However, after the low difficulty level, the CTL increased slightly but not significantly. Root mean square error (RMSE) decreased after lower level task and increased after high level task, even though no significant changes in RMSE performance were found. These results indicate that performance as measured by CTL became better after the easier tasks and performances became worse after more difficult tasks.

During the third phase of the experiment, SWAT and TLX ratings increased relative to baseline following easy task performance and decreased following performance on the difficult task. Thus, subjective scores increased after the easier task and decreased after the more difficult task. Percent changes in SWAT after the difficult task were reliably different from those following both the low and moderate tracking

sequences. Context phase TLX ratings showed significant differences between those conditions which had involved previous low and high difficulty sequences. Consequently, the context provided by phase two tracking significantly affected both tracking task performance and workload ratings during the final medium difficulty phase of the experiment.

Methodological Comparison of Previous Studies

Therefore, previous investigations of workload context effects (i.e., Moroney et al., 1993; Hancock et al. 1992) have provided contradictory evidence in identifying context effects in simulated tasks. The purpose of this experiment was to determine if the contradictory results of these studies might be due to paradigm differences. The present study focuses on differences between the two experiments: task difficulty, trial duration, and experimental design.

Task Difficulty. Hancock et al. and Moroney et al. used different tasks to test for context effects. Hancock et al. utilized a compensatory tracking task, while Moroney et al. utilized a three dimensional flight task on the Microsoft flight simulator. It is arguable which task is more sensitive to context effects.

Task Duration. Both Moroney et al. and Hancock et al. utilized different task durations. The total trial time in the Moroney et al. experiment (18 minutes) was greater than the Hancock et al. experiment (15 minutes). However, the duration of the individual trials was longer in the Hancock et al. experiment (5 minutes) than in the Moroney et al. experiment (3 minutes). Thus, the task duration used by Moroney et al. may not have been sufficient to establish a context effect of sufficient strength to affect subsequent ratings.

Experimental Design. Finally, the two studies differed in experimental design. Moroney et al. required subjects to perform six trials per session. The first three experimental trials were used to induce a context effect which was expected to effect

performance scores and workload ratings obtained during the last three context trials. Hancock et al., alternatively, tested context effects in a pre-post experimental design. Subjects experienced three trials per session (e.g., baseline, experimental, and context trials). The first and last trials (baseline and context) were the same level of difficulty, while the second trial (experimental) varied at either low, medium, or high conditions. Performances in the first and third trials were compared to determine the effect of the second trial. A change in third trial performance compared to the first indicated that the second trial produced a context effect. The question of experimental design relates back to task duration. Since the trials in Moroney et al. were only three minutes long, subjects may have seen each trial as a unique event. Even though subjects experienced twice as many trials, shorter trials times may have inhibited the task difficulty carry-over effects by causing subjects to "compartmentalize" their perceived levels of workload to specific trials. Furthermore, the Hancock et al. experimental design was similar to a pre-post design. This allowed for the direct comparison of two task difficulties separated by either a higher, lower, or identical task difficulty.

Research Strategy

The purpose of this experiment was to determine if the contradictory results of the Moroney et al. (1993) and Hancock et al. (1992) studies might be due to paradigm differences. The present study focuses on two of these differences: trial duration and experimental design. A longer trial duration increases the likelihood of inducing a context effect. A pre-post experimental design affords greater sensitivity because of the within-subject comparison to the baseline condition.

Hypotheses

The conflicting results of Moroney, et al. and Hancock et al. do not provide sufficient information to make directional predictions about the effects of prior task

loadings on subsequent task performance and workload ratings. Therefore, two-tailed statistical tests were used to test the hypotheses.

Performance Measure

1. Performance scores obtained under high levels of task difficulty will be lower than performance scores obtained under the low difficulty trials. Performance scores obtained under medium levels of difficulty will fall between high and low difficulty scores.
2. Previous levels of task difficulty will affect performance scores of subsequent tasks. Based on the findings of Hancock et al., high levels of previous task difficulty will result in lower performance scores on subsequent tasks, while low levels of previous task difficulty will result in higher performance scores.

Subjective Workload Measure (NASA-TLX)

1. Different levels of task difficulty will cause differences in subjective ratings of workload. Specifically, TLX ratings obtained under high levels of task difficulty will be higher than ratings obtained under the low task difficulty trials. Ratings obtained under medium levels of task difficulty will fall between the high and low difficulty ratings.
2. Previous levels of task difficulty will affect a subject's perception of workload on subsequent tasks. Specifically, high levels of previous task difficulty will result in lower workload ratings on subsequent tasks, while low levels of previous task difficulty will result in higher workload ratings on subsequent tasks.

CHAPTER II

METHOD

Subjects

Twelve subjects (eleven male and one female) volunteered to participate in this study. Eleven of the subjects were either graduate students or previous graduate students from the University of Dayton. The twelfth subject was a pilot in the United States Air Force. Four of the subjects had previous flight experience while the remaining eight subjects had previous simulator experience on the Microsoft Flight Simulator.

Before testing, all subjects were required to master the basic flight skills and demonstrate a minimum pilot proficiency level. To qualify, subjects flew through two training courses. The first course required subjects to fly through twelve mid-air gates (Figure 2; no crosswinds) and obtain a score of at least 910 points. Once subjects had met this criterion, they were required to fly an S-Curve flight course of 20 mid-air gates (Figure 3) with no crosswinds and obtain a score of at least 1700 points. These criteria were based on the results of the pilot study described in Appendix A.

Software and Apparatus

The Microsoft Corporation's "Flight Simulator" and "Flight Simulator: Aircraft Scenery Designer" software was used to create the flight simulation. The flight simulator emulated a Cessna 182 single engine propeller aircraft. The aircraft cockpit displays (Figure 1) were concealed with poster board to assure subjects attend only to the exterior field of view. The subjects' view is illustrated in Figure 4.

The software was installed on a DOS-based 386 personal computer, with 8 MB of RAM, a 40 MB hard drive, and a VGA monitor. The experimental task, as shown in Figure 3, was to fly the simulated aircraft through an S-Curve flight course. Figure 3 shows a series of 21 open gates (squares 200 ft x 200 ft) displaced 0.1 mile East or West about a constant altitude (6000 ft) and separated by approximately 0.7 miles. Course gates for all flight conditions had a crosshair in the middle to which subjects were required to fly as close as possible. The gates were placed in an S-Curve flight pattern to provide a task which required attention and good control of the aircraft. Operators sat at a table and controlled the aircraft with a MAXX-yoke. Power was pre-set to attain a speed of 135 knots, trim was set in the neutral position, and gear was set fully up. Subjects were not allowed to vary either power or aircraft configuration and were instructed to avoid missing any gates. All performance data were recorded by the flight simulator (See Performance Measures in Methods section).

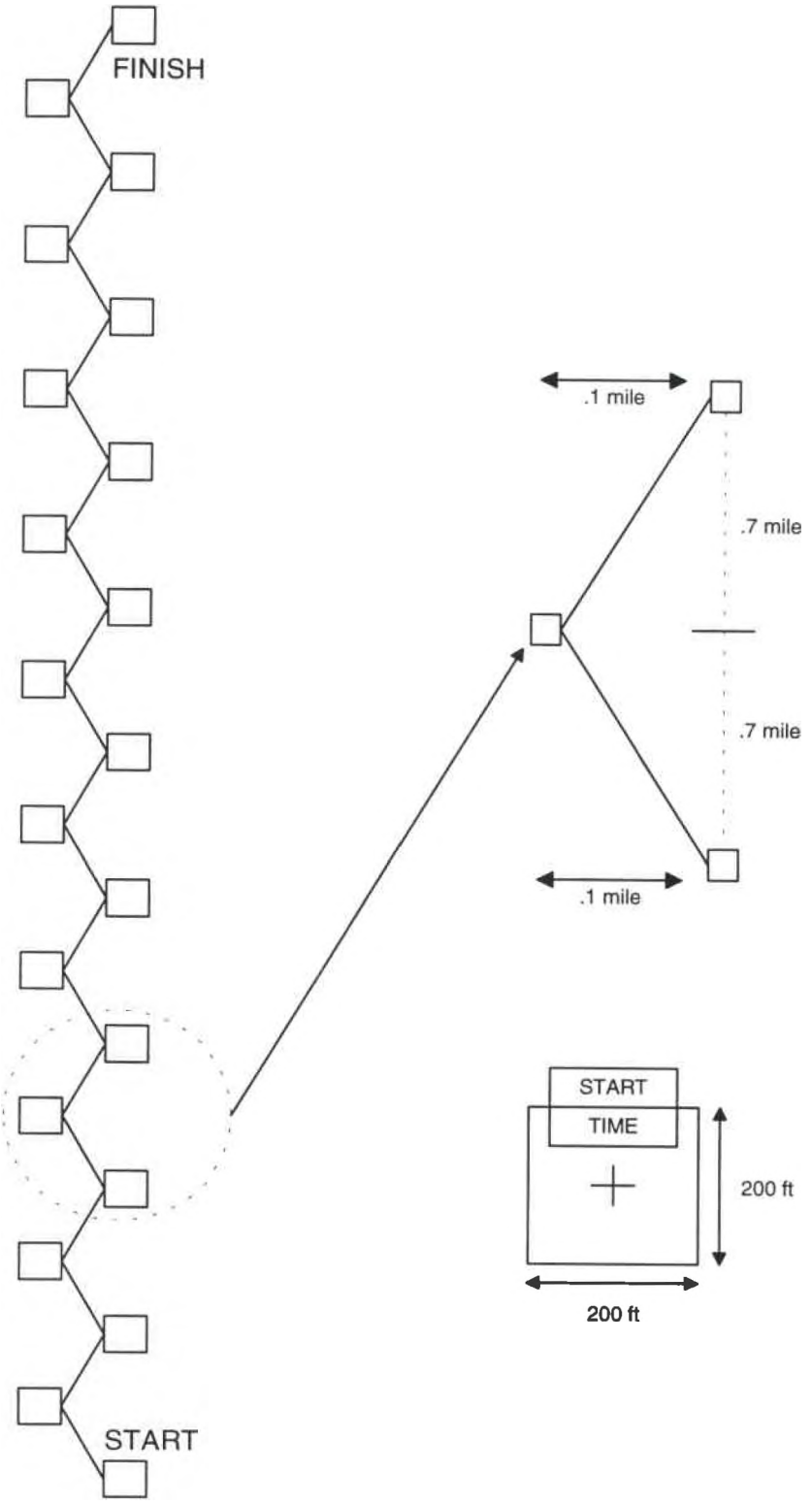


FIGURE 3: S-Curve Flight Path

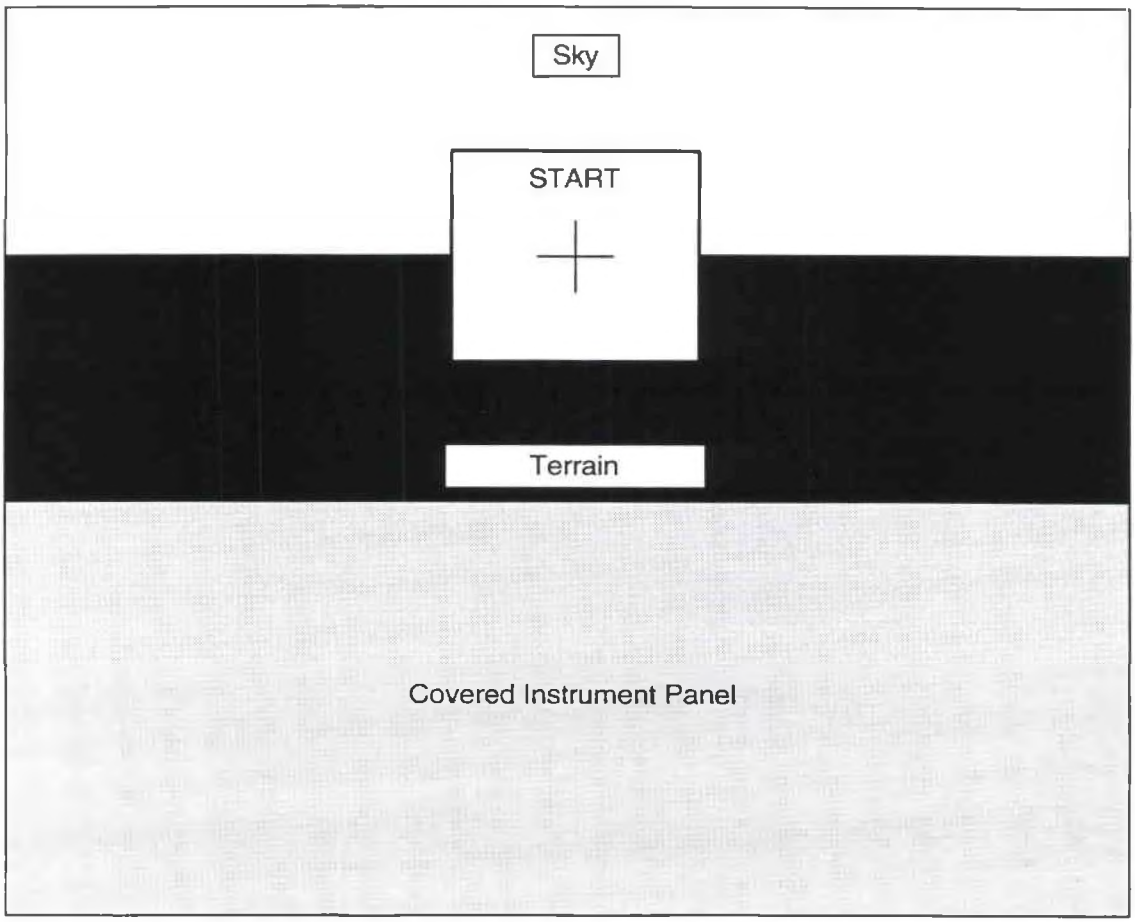


FIGURE 4 Subjects' View From Simulated Cessna 182

Experimental Design

The independent variables include Trial (Baseline, Experimental Manipulation, Context), Difficulty (Low, Medium, High), and Order (1-6). The experimental design consisted of three sessions of three trials each. Within each session there was a baseline trial (Trial 1), an experimental manipulation trial (Trial 2), and a context effect trial (Trial 3). Trial 1 established a performance score and workload rating baseline, while Trial 3 served as a context effect trial, which subsequently was compared to Trial 1. Trial 1 and Trial 3 were always at the medium level of task difficulty, while Trial 2 (the experimental manipulation trial) was at low, medium, or high task difficulty. The low, medium, and high levels of task difficulty were established by varying crosswinds at 2, 12, and 22 knots from a heading of 270 degrees. There were three trial combinations (i.e., M-L-M, M-M-M, M-H-M), and subjects received one of the three trial combinations per session. Trial duration lasted for 5 minutes, requiring approximately 30 - 45 minutes per session. The order of presentation was counterbalanced providing six possible orders. Table 2 shows the experimental conditions and design used in this study, while Table 3 shows the counterbalanced design.

TABLE 2: Experimental Design Used In This Study*

Session	Baseline	Experimental Manipulation	Context
Practice	Medium	Medium	Medium
1	Medium	Low	Medium
2	Medium	Medium	Medium
3	Medium	High	Medium

* There was a minimum of one day off between sessions; order of experimental sessions was counter-balanced; intertrial interval was 60 seconds.

TABLE 3: Between-Subjects Counterbalanced Design

Subject	Session 1	Session 2	Session 3
1	MLM	MMM	MHM
2	MLM	MHM	MMM
3	MMM	MLM	MHM
4	MMM	MHM	MLM
5	MHM	MLM	MMM
6	MHM	MMM	MLM
7	MHM	MMM	MLM
8	MHM	MLM	MMM
9	MMM	MHM	MLM
10	MMM	MLM	MHM
11	MLM	MHM	MMM
12	MLM	MMM	MHM

Dependent Variables

There were two dependent variables: Performance and Perceived Workload. Pilot performance scores were objective measures recorded by the computer. Perceived workload was obtained through the use of the NASA-TLX.

Performance Measure. Subjects were instructed to fly through the crosshair, in the center of each gate, as rapidly as possible and avoid missing any gates. The performance algorithm available on Flight Simulator (4.0) was used to calculate the performance scores. The same algorithm was used by Moroney et al. (1993) and Moroney et al. (1992).

The scoring algorithm is as follows:

(25 points x # of gates passed through) +

(Mean speed x # of gates passed through/ 4) +

(Closeness to Center Bonus) = Final Score in points.

The performance algorithm calculated scores based on flight speed, accuracy to center of gate, and a 25 point bonus for successfully navigating through all gates.

As indicated by Moroney et al. (1992), the first element in the scoring algorithm was awarded only if all gates were successfully flown through. The second element rewarded speed and gross accuracy. Finally, fine accuracy was reflected in a "closeness to the center" bonus. Flying the aircraft through the center of the crosshair earned a bonus of 20 additional points. Bonus points decreased as a function of distance from center reaching a low of 6 points if the aircraft just penetrated the corner of the gate.

Subjective Workload Measure. The TLX was administered immediately upon completion of each flight trial to test for context effects of subjective workload. Subjects were given as much time as necessary to complete the form.

Trial Ratings. Subjects flew the flight course and provided ratings on the six subscales following all trials.

Weights. Weighting was used to replicate the procedures of Hancock et al. There were 15 possible pair-wise comparisons of the six subscales (Appendix C). Each pair was presented on a card. After completing all trials, subjects circled the member of each pair that they felt contributed more to the workload of the tasks. The number of times that each factor was selected was then tallied. The tallies ranged from 0 (not relevant) to 5 (more important that any other factor). One set of cards was prepared for each subject in advance of the experiment. The pairs of factors were separated and presented

individually in a different, randomly selected order to each subject as recommended in NASA-TLX, V 1.0 (1986).

Procedure

Prior to Session 1, subjects received a packet of materials. The materials included: the NASA-TLX Subject Instructions (Appendix B), NASA-TLX Sources-Of-Workload Evaluation (Appendix C), Subject Task Instructions (Appendix D), and an Informed Consent Form (Appendix E).

Task instructions, the NASA-TLX procedures, and dimension definitions were read to all subjects from the experimental materials distributed prior to the beginning of the experiment. All subjects were briefed and given the opportunity to ask questions before the experiment began.

The following steps were taken in administering the TLX:

Subjects received TLX Subject Instructions: Sources-Of-Workload Evaluation prior to the experiment (Appendix C). Subjects were asked to read the scale definitions and instructions prior to the Practice session. All instructions and dimension definitions were read to all subjects in the beginning of the Practice session.

Rating Familiarization. Subjects completed the rating scales after performing the warm-up flight task. This ensured that all subjects developed a standard technique for using the scales. A sample rating sheet is provided in Appendix C.

With the exception of Session 1, each session took approximately 45 minutes to complete. Session 1, took approximately one hour to complete to ensure that the subjects signed the informed consent form, understood the procedures, and met the minimal criterion. Upon completing each of the flight trials, subjects used the TLX to rate workload. All procedures were completed using procedure checklists (See Appendix F).

CHAPTER III

RESULTS

The results of this experiment were divided into four sections. The first three sections describe the performance and workload results specific to each of the three trials (i.e., Trial 1 Baseline, Trial 2 Experimental Manipulation, and Trial 3 Context Effect). For the Experimental Manipulation (Trial 2) and Context Effect (Trial 3) results, workload was further assessed for each of the individual subscales. The TLX subscales were analyzed in an effort to identify the effects of previous levels of workload on the specific subscale dimensions. The fourth section of results describes analyses that replicate the percent change performance and TLX analyses as performed by Hancock et al.

The analytic strategy used in this study first examined the statistical differences existing between performance scores, and between workload ratings under Trial 1 (Baseline) conditions. The results of the Baseline analyses indicate the subjects' performance and workload perception prior to the task difficulty experimental manipulation period (Trial 2). Analyses were next conducted on Trial 2 (Experimental Manipulation) performance scores and workload ratings to determine the effect of the low, medium, and high levels of task difficulty. Finally, analyses compared performance and TLX ratings on Trial 1 and Trial 3 to determine which of the dependent measures differed as a function of the Trial 2 experimental manipulation.

The outcomes of all statistical tests were evaluated at an alpha level of 0.05. To correct for the positive bias of the F-test associated with the within-subjects effects, the

Geisser-Greenhouse correction was used on all analyses. The degrees of freedom (df) reported in the analyses have been adjusted to reflect the Geisser-Greenhouse correction.

Trial 1 Baseline

Performance scores and TLX ratings in Trial 1 were assessed to ascertain that all subjects started from the same performance level and rated workload equally. Since all subjects only received the medium level of difficulty in Trial 1, it was anticipated that the performance scores and workload ratings would be similar.

Performance Scores and Workload Ratings. As expected, the 3 (Difficulty) X 6 (Order) mixed ANOVA revealed that both the interaction of Difficulty X Order of task difficulty presentation, and the main effect of Difficulty was not significant for either the performance scores (F (1.44, 8.67) = .43, p = .603) or TLX ratings (F(1.25, 7.51) = .01, p = .963) respectively. Since Trial 1 performance scores and TLX ratings were equivalent, subjects entered the experimental manipulation trial (Trial 2) at the same level of proficiency and workload perceptions. See Appendix G, Table G-1, for the Trial 1 ANOVA summary table. Since Order was not significant for performance scores (F(5.00, 6.00) = 0.34, p = 0.872) and TLX ratings (F(5.00, 6.00) = 0.80, p = 0.590), the counterbalancing procedure was successful.

Trial 2 Experimental Manipulation

A 3 (Difficulty) X 6 (Order) mixed-design ANOVA was performed to assess the effect of Difficulty and Order on Trial 2 performance scores and workload ratings. The purpose of this analysis was (1) to check the subjects' performance scores under the manipulated levels of task difficulty; and (2) to determine whether subjects perceived the manipulated levels of task difficulty differently.

Performance Scores. It was hypothesized that different levels of task difficulty would affect pilot performance scores, and that Order (i.e., the sequence in which the level of Trial 2 task difficulty was manipulated) would have no effect. Specifically, it was expected that a low level of task difficulty would produce the highest performance scores, medium slightly lower, and the highest level of difficulty the lowest performance scores.

As illustrated in Figure 5, the mean performance scores based upon the Trial 2 level of task difficulty. The means, standard deviations, and standard error of the mean are presented in Table G-2.

As illustrated in Table G-3, the main effect of Difficulty was significant ($F(1.60, 9.57) = 25.79$, $p < .001$) indicating that the subjects scored differently on the low, medium, and high levels of task difficulty. As shown in Figure 5, comparisons of the levels of difficulty indicate that subjects Trial 2 performance scores were similar for the low and medium levels of difficulty ($F(1,6) = 3.74$, $p = .101$) while both low and medium scores were significantly different from those obtained at the high level of difficulty ($F(1,6) = 37.46$, $p = .001$; and $F(1,6) = 47.70$, $p < .001$ respectively). Specifically, performance scores were significantly higher under the low and medium difficulty condition than under the high difficulty conditions. No other effects were significant.

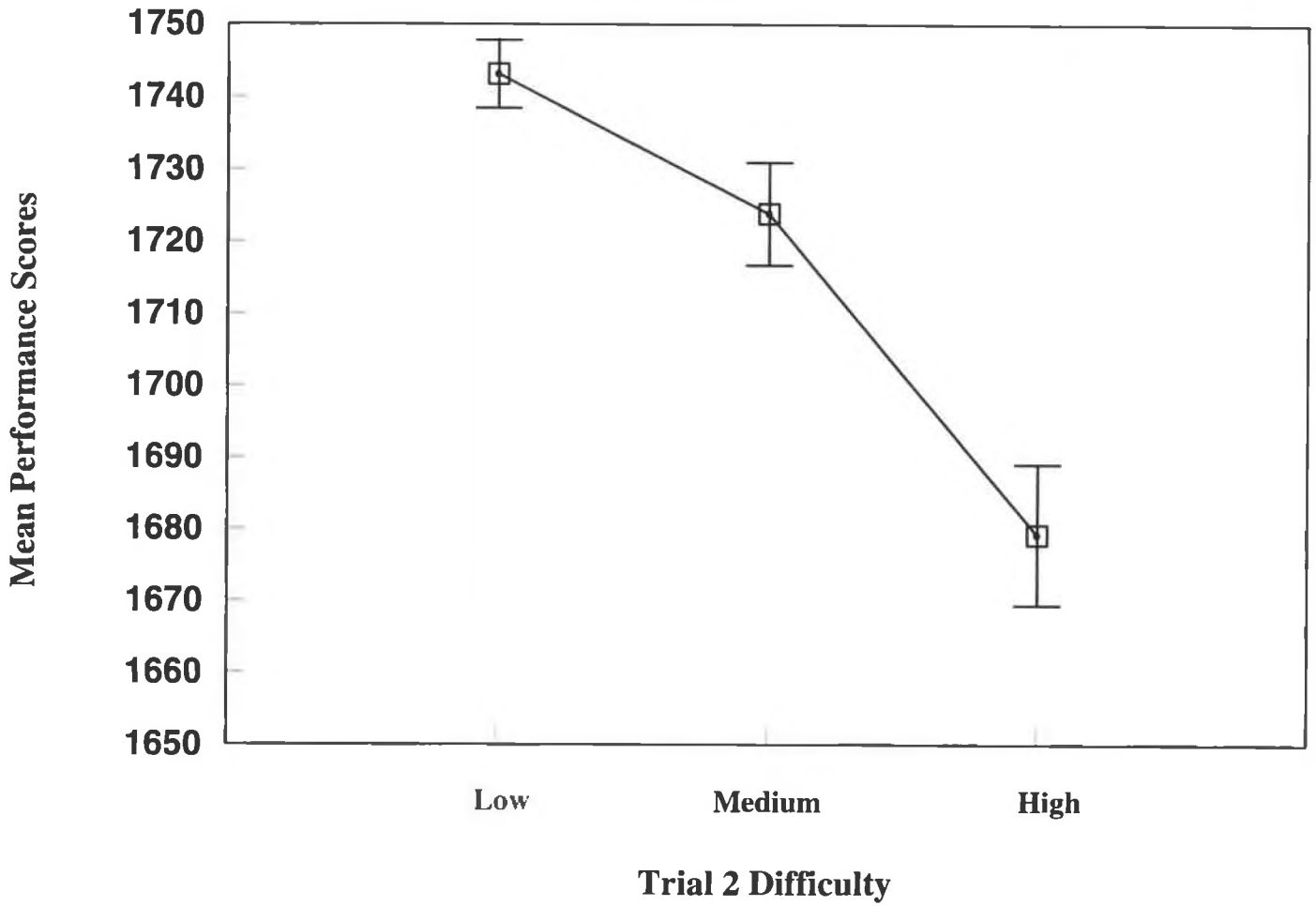


FIGURE 5: Mean Performance Scores as a Function of Trial 2 Level of Task Difficulty

Workload Ratings. It was hypothesized that different levels of task difficulty would effect subjective ratings of workload, and that Order would have no effect. Further, it was expected that a low level of task difficulty would produce the lowest workload ratings, medium slightly higher, and the highest scores under high difficulty.

Figure 6 compares the mean workload ratings based upon the Trial 2 level of task difficulty. The means, standard deviations, and standard error of the mean are presented in Table G-4.

The main effect of Difficulty was significant ($F(1.60, 8.29) = 36.26$, $p < .001$) which indicates that the subjects perceived workload differently across the levels of three levels of task difficulty (See Table G-5). Pairwise comparisons of the levels of difficulty indicate that Trial 2 workload ratings were similar for the low and medium levels of difficulty ($F(1,6) = 5.52$, $p = .057$) while both low and medium scores and ratings were significantly different from those obtained at the high level of difficulty ($F(1,6) = 154.61$, $p = .001$ and $F(1,6) = 22.15$, $p = .003$). Specifically, workload ratings were significantly lower under low and medium difficulty than under high difficulty. The main effect and interaction with Order were not significant.

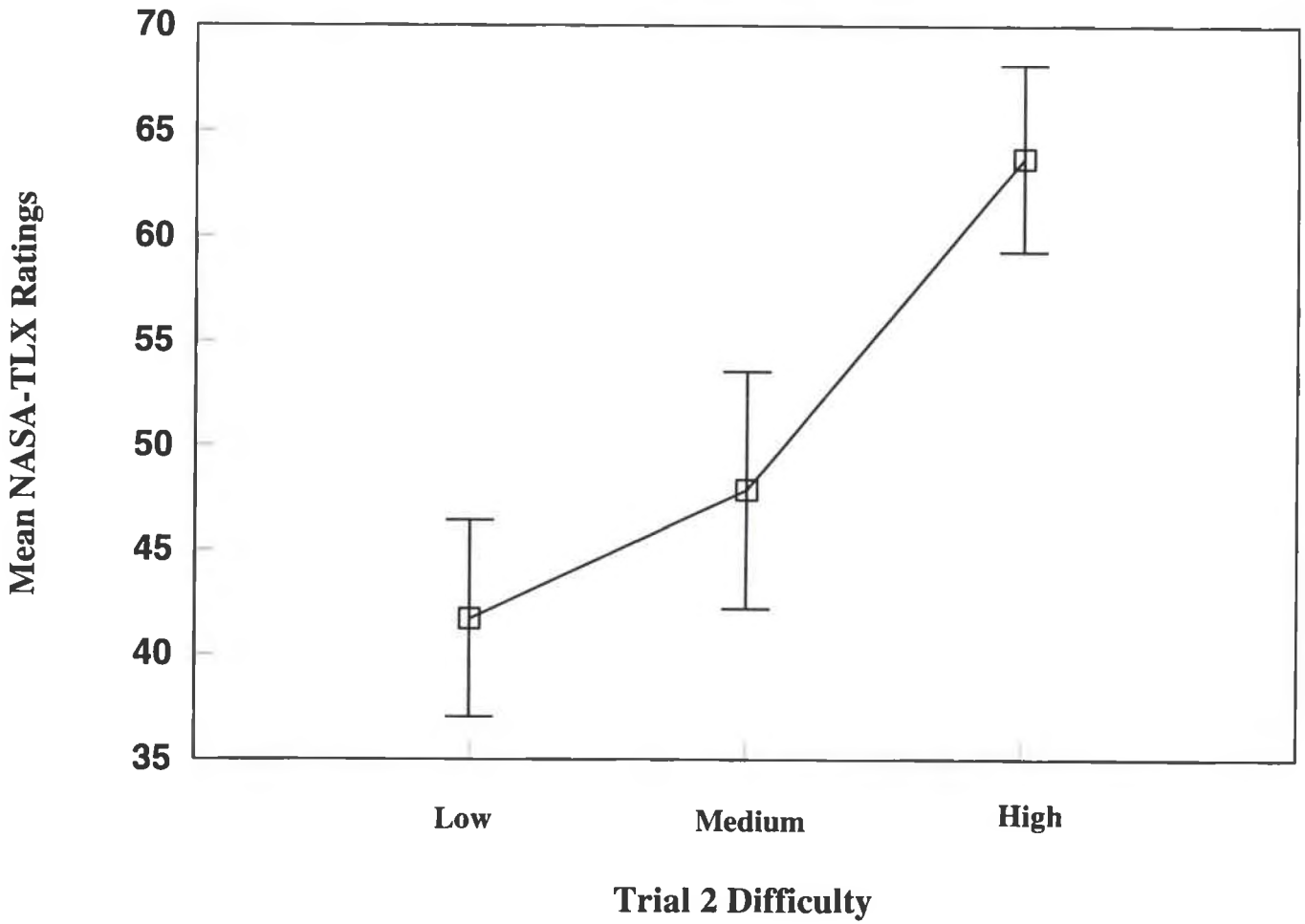


FIGURE 6: Mean TLX Ratings as a Function of Trial 2 Level of Task Difficulty

TLX Subscales. A 3 (Difficulty) x 6 (Order) mixed ANOVA revealed that the following Trial 2 TLX subscales effects were significant: Mental Demand, Physical Demand, Performance, Effort, and Frustration. Thus, the levels of Trial 2 task difficulty manipulation produced significant effects for five of the six subscales rated in Trial 2. No reliable differences were noted for Temporal Demand. The ANOVA summary for the subscales is presented in Table G-6. Table 4, below, shows the mean TLX subscale ratings as a function of Trial 2 level of difficulty. The subscale ratings increased significantly (as the level of task difficulty increased) for all subscales except Temporal Demand.

TABLE 4: Mean TLX Subscale Ratings as a Function of Trial 2 Level of Difficulty

TLX Subscale	Low	Medium	High
Mental	46.667	54.583	67.083
Physical	35.417	40.417	50.833
Temporal	34.583	38.750	43.333
Performance	31.667	37.083	60.833
Effort	46.583	55.000	65.833
Frustration	36.667	44.583	59.583

Trial 3 Context Effect

A 3 (Difficulty) X 2 (Trial) X 6 (Order) mixed-design ANOVA was used to determine if Trial 2 task difficulty influenced performance and perception of workload during Trial 3. Thus, Trial 1 and Trial 3 were compared to determine if a context effect existed for performance scores and workload ratings due to the levels of Trial 2 task difficulty.

Performance Scores. It was hypothesized that the different levels of Trial 2 task difficulty would carry-over to Trial 3, producing a performance context effect. Thus, it

was anticipated that previous high task difficulty would result in lower performance scores on subsequent tasks, and low levels of previous task difficulty would result in higher performance scores on subsequent tasks. The performance scores for the M-M-M conditions were expected to be similar and unaffected by order of Trial 2 task difficulty manipulation.

As illustrated in Figure 7, there was no significant interaction of Difficulty x Trial (Trial 1 vs. Trial 3) indicating no context effect ($F(1.74, 10.45) = .15$, $p = .839$). The Trial 3 scores, thus, were not different from Trial 1 scores across all levels of task difficulty. Additionally, the main effects of Trial and Difficulty were not significant.

In summary, there was no significant change in performance from Trial 1 to Trial 3 for any levels of difficulty, nor were there any differences among the difficulty conditions on Trial 1 and Trial 3. Table G-2 presents the means, standard deviations, and standard error of the mean, while Table G-7 provides the performance ANOVA summary. All other effects were also not significant.

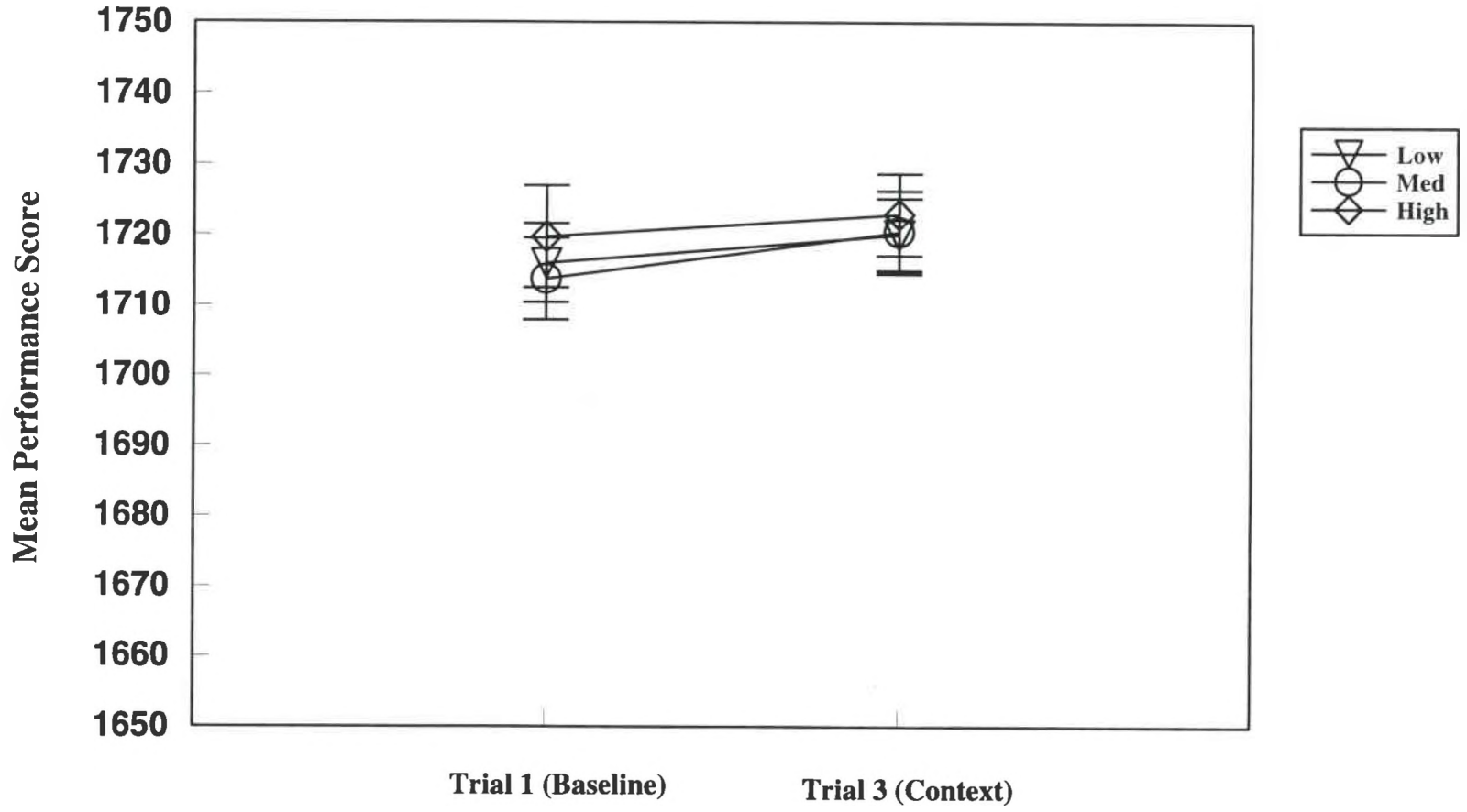


FIGURE 7: Performance Scores Obtained Under Baseline and Context Trials For Low, Medium, and High Levels of Task Difficulty

Workload Ratings. It was hypothesized that the different levels of Trial 2 task difficulty would carry-over to Trial 3, producing a workload context effect. Thus, it was anticipated that workload ratings would increase after exposure to the easier level of task difficulty but decrease after exposure to the more difficult level of task difficulty. The NASA-TLX ratings for the M-M-M conditions were expected to be similar, and Order was not expected to have a significant effect.

The TLX ratings obtained under baseline (Trial 1) and Context (Trial 3) are plotted in Figure 8. The figure illustrates that there also was no significant interaction of Difficulty x Trial (Trial 1 vs. Trial 3) indicating no workload context effect ($F(1.74, 10.42) = .55$, $p = .486$). Additionally, the main effects of Trial and Difficulty were not significant. These results indicate that there was no significant change in workload ratings from Trial 1 to Trial 3 for any levels of difficulty, nor were there any differences among the difficulty conditions on Trial 3. Table G-4 presents the means, standard deviations, and standard error of the mean for the TLX Ratings data, while Table G-8 provides the TLX Ratings ANOVA summary.

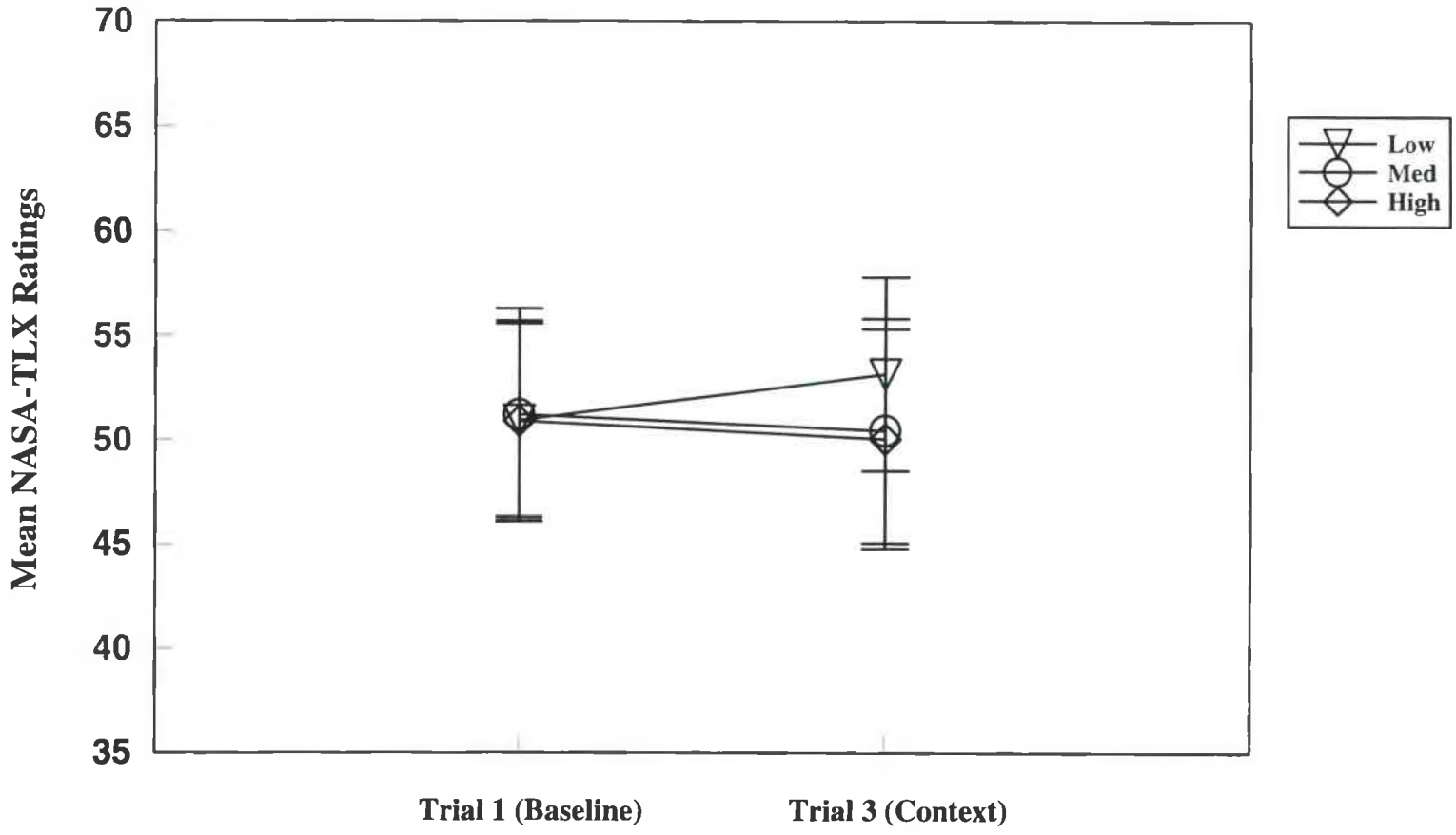


FIGURE 8: NASA-TLX Ratings Obtained Under Baseline and Context Trials For Low, Medium, and High Levels of Task Difficulty

TLX Subscales. Similar to the overall context effect analyses, none of the TLX subscale Difficulty X Trial interactions were significant (Table G-9). This finding indicates that Trial 3 TLX subscale ratings were not affected by task difficulty carry-over from Trial 2. Thus, task difficulty manipulation did not affect a subject's perception of workload for the individual TLX subscales.

Percent Change Performance and Workload

Hancock et al. found workload transition effects using percent change scores. For comparability, an analysis was also conducted on percent change for both performance and workload measures. Percent change was calculated by dividing the Trial 1 score or rating into the score or rating obtained by subtracting Trial 1 from Trial 3. Hence, percent change equals:

$$\frac{\text{Trial 3 - Trial 1}}{\text{Trial 1}}$$

Based on the above formula, percent change was calculated. Separate 3 (Difficulty) x 6 (Order) ANOVAs were performed for the performance scores and workload ratings. No significant differences were found in the percent change scores for either performance ($F(1.76, 10.54) = .15$, $p = .840$) or workload ($F(1.41, 8.48) = .74$, $p = .462$). See Table G-10 for performance and workload ANOVA summary. The minimal changes in performance and TLX ratings are illustrated in Figures 9 and 10. Thus, contrary to the findings of Hancock et al., there was no evidence of performance or workload context effects.

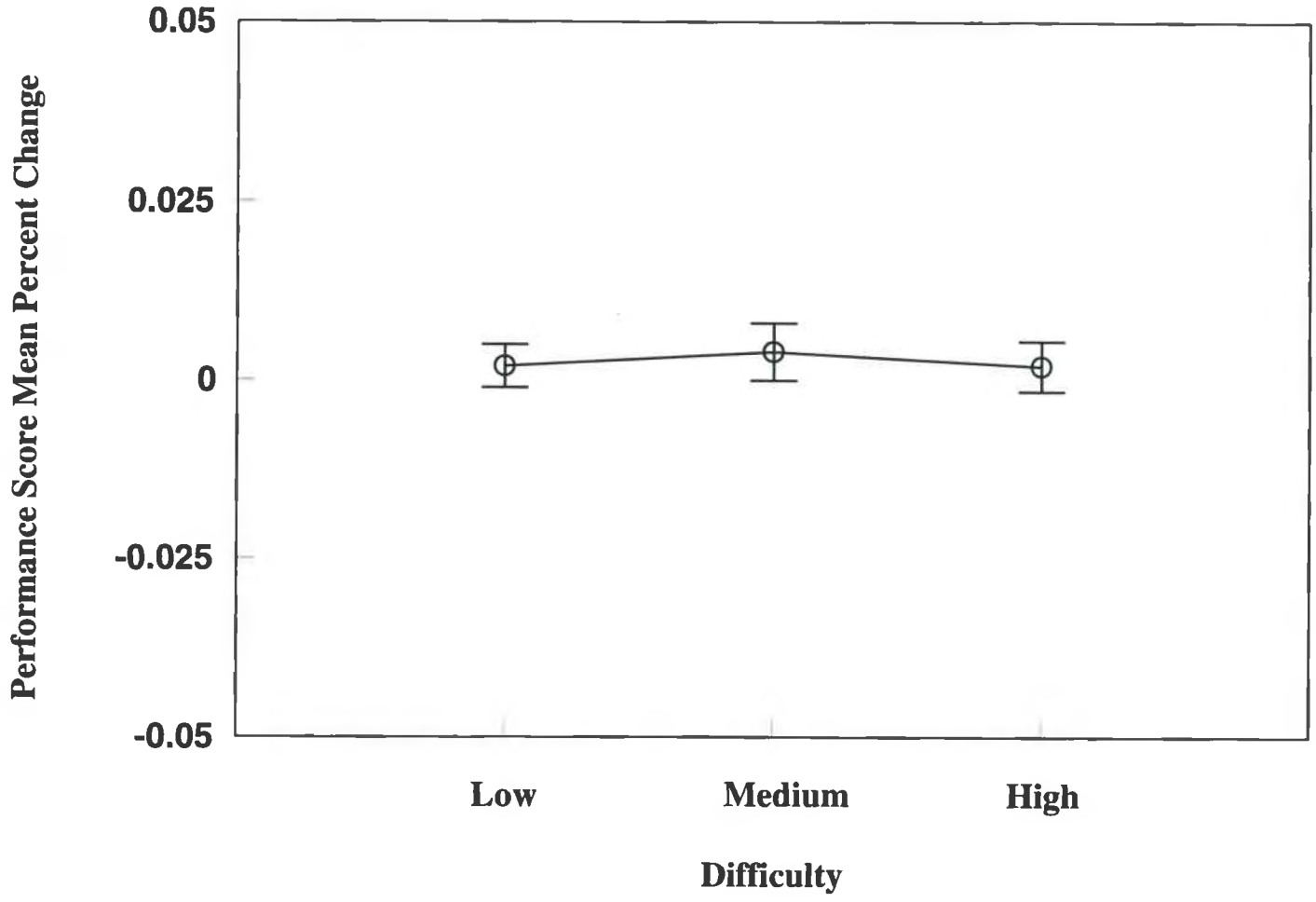


FIGURE 9. Percent Change Performance Scores From Trial 1 to Trial 3 as a Function of Trial 2 Level of Task Difficulty

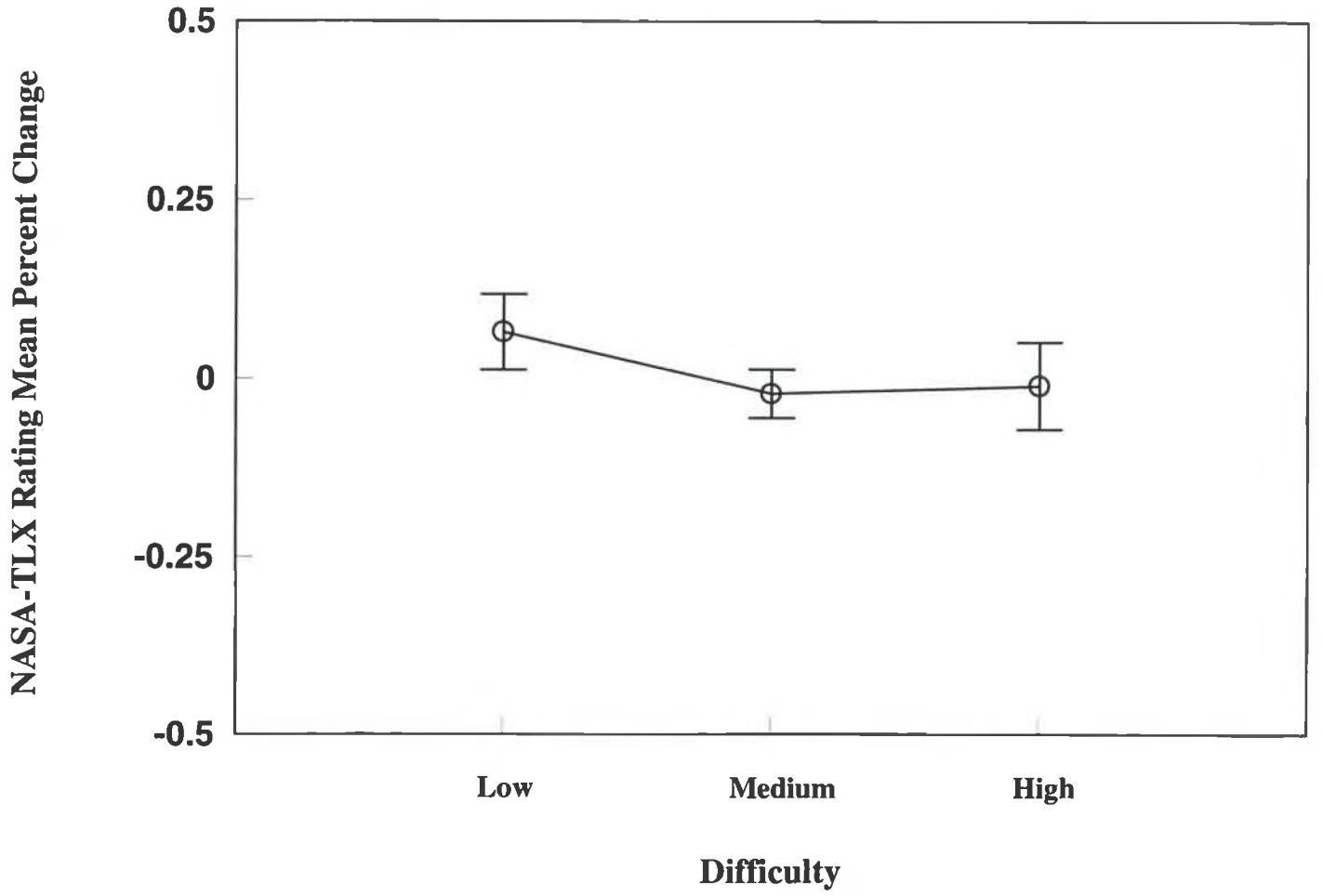


FIGURE 10. Percent Change Reported TLX Ratings From Trial 1 to Trial 3 as a Function of Trial 2 Level of Task Difficulty

CHAPTER IV

DISCUSSION

The purpose of this study was to examine the effects of previous levels of flight task difficulty on subsequent task performance and reported workload ratings. Referred to as context effects, it was expected that differences in Trial 1 and Trial 3 task performance and NASA-TLX reported workload ratings would result from the introduction of different levels of flight task difficulty in Trial 2.

Hypotheses for this study were based on the findings of Hancock et al. (1992). It was hypothesized that subjects exposed to low task difficulty on context building trials would score higher and rate workload higher on subsequent tasks, and subjects exposed to high task difficulty would score lower and rate workload lower on subsequent tasks. The Trial 1 versus Trial 3 analyses failed to identify a context effect in either the performance or workload measures, indicating that previous task difficulty experience had minimal effect on subsequent task experience.

The results of this study replicate the findings of Moroney et al. (1993) who did not detect a reliable performance and workload context effects in a simulated flight task. Thus, the results of this study conflict with those of Hancock et al. (1992) who found significant differences with a similar experimental design. Using a compensatory tracking task, Hancock et al. identified differences between the first and third trials, which they attributed to a Trial 2 treatment manipulation in a compensatory tracking task. The Moroney et al. and Hancock et al. experiments produced the same general trend related to context effects on workload ratings. In both experiments, a more difficult task

condition which preceded a moderate condition led to the tendency to decrease workload ratings associated with the moderate difficulty condition, while a less difficult task context tended to inflate ratings of the moderate difficulty trials relative to baseline. However, while the Hancock et al. experiment resulted in some reliable differences in workload ratings between context conditions, the Moroney et al. experiment did not. This study identifies methodological differences specifically task difficulty, task duration, and experimental design as potential factors for differences in results of the previous research.

Despite an attempt to replicate the findings of Hancock et al. (1992) by changing task duration and using their pre-post design, no context effects were obtained for either performance or workload ratings in the simulated flight task. This result was further confirmed when the same percent change analyses as performed by Hancock et al. did not indicate performance or workload carry-over from Trial 2.

The inability to obtain a context effect can, in part, be attributed to the lack of a significant difference in performance and rated workload associated with crosswinds of 2 and 12 knots on the experimental trial. One would not expect a context effect to occur for the M-L-M condition in comparison to the M-M-M condition under these circumstances. However, there was no evidence for a context effect for the condition in which transition was from high to moderate difficulty.

Although both Hancock et al. (1992) and Moroney et al. (1993) varied task difficulty, the two studies utilized different methods of difficulty manipulation: tracking difficulty in Hancock et al., and crosswinds in Moroney et al. and the present study. Thus, the low, moderate, and high difficulty conditions are not comparable across independent variables and across experiments. A major problem inherent in workload measurement is the inability to readily quantify objective levels of task difficulty.

While unable to compare objective levels of difficulty, subjective comparisons are possible in terms of perceived workload. For example, in the Hancock et al. (1992)

experiment, mean values for the low, moderate, and high difficulty conditions were approximately 20, 40, and 70, respectively on the TLX, and 12, 30, 76 on the SWAT (Hancock, et al.). In the present experiment, the same three difficulty conditions received mean TLX workload ratings of 42, 48, and 64 respectively, whereas in the previous Moroney et al. study, the ratings were 33, 36, and 44. Table 5, below, shows the mean TLX rating from Hancock et al., Moroney et al., and the current study.

TABLE 5: Comparison of TLX Ratings From Hancock et al. (1992), Moroney et al. (1993), and Current Study

	Difficulty Level		
Study	**Low**	**Med**	**High**
Hancock et al.	20	40	70
Moroney et al.	33	36	44
Current Study	42	48	64

Three observations are noteworthy from the above data. First, the changes made in the present experiment from a straight course to an S-Curve course increased the task difficulty to a much higher level of workload. Secondly, in this study and the Moroney et al. experiment, the differences in perceived workload between the low and medium difficulty conditions was minimal and thus a workload context effect would not be predicted as the subjects transitioned from a low to medium difficulty task.

Most important, however, was the fact that the range of differences in perceived workload among the low, moderate, and high difficulty conditions was much greater (50 points) for Hancock et al., than for either Moroney et al. (11 points) or the present experiment (22 points). This suggests that the failure to induce a workload context effect in Moroney et al. and this experiment may be due to the limited range between the levels of task difficulty employed.

Another possible explanation for not finding significant workload context effects is that perhaps the TLX is not as sensitive to workload context effects as is the SWAT. In the original Hancock et al. experiment, there were reliable context effects for the TLX only in the transition from low to medium, whereas for the SWAT, reliable effects were found for transition from low to medium and high to medium. Given the limited effect with the TLX in Moroney et al. and this experiment, it is possible that the TLX is not as sensitive to prior task loadings as SWAT.

Conclusion

Assuming that workload context effects are "real," clearly they are elusive. This study suggests that future laboratory studies must utilize a wider range in perceived task difficulty or more sensitive measuring instruments if context effects are to be demonstrated.

REFERENCES

Hancock, P.A., Williams, G., Miyake, S., and Manning, C.M. (1992, unpublished). *The influence of task demand characteristics on workload and performance.* University of Minnesota

Microsoft (1990). *Flight simulator: Aircraft and scenery designer.* Washington: Microsoft.

Microsoft (1990). *Flight simulator.* Washington: Microsoft.

Moroney, W.F., Biers, D.W., Eggemeier, F.T., and Mitchell, J.A. (1992). A comparison of two scoring procedures with the NASA task load index in a simulated flight task. *Proceedings of the 1992 IEEE National Aerospace Electronics Conference, 734-740.*

Moroney, W.F., Reising, J., Biers, D.W., and Eggemeier, F.T. (1993). *The effect of previous levels of workload on the NASA task load index (NASA-TLX) in a simulated flight task.* Proceedings of the Seventh International Symposium on Aviation Psychology, 882-885. Columbus, OH: Ohio State University.

NASA Task Load Index (TLX) Paper and Pencil Package Version 1.0 (1986). Moffett Field, CA: Human Performance Research Group, NASA Ames Research Center.

APPENDIX A

PILOT STUDY

PILOT STUDY

A pilot study was initially conducted to select the experimental flight course, obtain a minimum criterion indicating pilot proficiency, and verify experimental procedures. Three subjects with previous Microsoft Flight Simulator (4.0) experience participated in the pilot study. Prior to the pilot study, the three subjects were informed of the objectives and were asked to evaluate the courses in terms of task difficulty, course configuration, aircraft control, and display size. Subjects were provided time to practice on a course of 15 gates in an "S" pattern with no crosswinds.

Table A-1 lists the four course conditions tested in the pilot study. There were two courses (See Figures A-1 and A-2) with two visual display configurations (i.e., full screen and half screen). After flying through each course condition, subjects provided subjective evaluations on task difficulty, course configuration, aircraft control, and display size. Objective performances scores were gathered to compare subject performances on all courses. Performance scores included elapsed flight time, average course speed, and a total score.

TABLE A-1: Description of Pilot Test Courses

	Course Conditions			
	1	**2**	**3**	**4**
Course Shape	Figure A-1	Figure A-1	Figure A-2	Figure A-2
Visual Display	Small	Full	Small	Full
			See Figure 4	See Figure 4

In general, subjects felt that the easiest course was Course 1 (Figure A-1). Further, subjects received their highest (total) performance scores on Course 1. This course was the easiest because there were not many turns and subjects could maintain straight and level flight for two gates at a time. The video display configuration included eliminating the flight instrument panel and showing a half screen window of the flight course. Course 1 was eliminated because the gate sequence did not increase flight difficulty as desired. Two of the subjects indicated, however, that flight task difficulty was increased by eliminating the flight instrument panel.

Course 2 (Figure A-1) displayed a full-screen window of the flight course used in Course 1. The full screen display caused flight task difficulty to increased dramatically. Control movements of the aircraft became 'jumpy', or 'gross' and were perceived to move to a greater degree. Thus, a small yoke movement would cause an exaggerated gross aircraft movement. This course was eliminated because of subjects' inability to control the aircraft with the enlarged screen.

Similar control movement problems were experienced with Course 4 (Figure A-2). Since Course 4 also displayed an enlarged screen, aircraft movements were jumpy and difficult to control. Course 4 was eliminated due to the troubles subjects experienced controlling the aircraft.

Course 3 (Figure A-2) was selected as the experimental course. Subjects were presented with a display configuration that eliminated the flight instrument panel and showed a half screen window of the flight course (See Figure 3). Compared to Course 1, subjects felt that Course 3 was more difficult because it required more turns and therefore, more predictions of aircraft altitude, attitude, and heading. The control movements were also 'smoother' with the half screen window display, because the aircraft's response was not exaggerated.

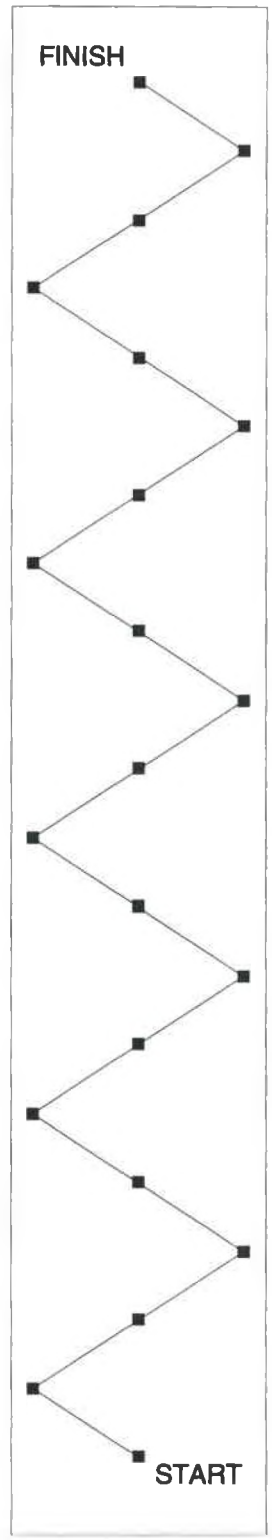


Figure A-1. Pre-Experimental Courses 1 and 3

Figure A-2. Pre-Experimental Courses 2 and 4

APPENDIX B

NASA-TLX SUBJECT INSTRUCTIONS

NASA-TLX SUBJECT INSTRUCTIONS

We are not only interested in assessing your performance, but also the experiences you had during the different task conditions. Right now we are going to describe the technique that will be used to examine your experiences. In the most general sense, we are examining the "workload" you experience. Workload is a difficult concept to define precisely, but a simple one to understand generally. The factors that influence your experience of workload may come from the task itself, your feelings about your performance, how much effort you put in it, or the stress and frustration you felt. The workload contributed by different task elements may change as you get more familiar with a task, perform easier or harder versions of it, or move from one task to another. Physical components of workload are relatively easy to conceptualize and evaluate. However, the mental components of workload may be more difficult to measure.

Since workload is something that is experienced individually by each person, there are no effective "rulers" that can be used to estimate the workload of different activities. One way to find out about workload is to ask people to describe the feelings they experienced. Because workload may be caused by many different factors, we would like you to evaluate several of them individually rather than lumping them into a single global evaluation of overall workload. This set of six rating scales was developed for you to use in evaluating you experiences during different tasks. Please read the descriptions of the scales carefully. If you have questions about any of the scales in the table, please ask me (Doug Fischer) about it. It is extremely important that they be clear to you. You may keep the descriptions with you for reference during the experiment.

After performing each task, you will be given a sheet of rating scales. You will evaluate the task by putting an "X" on each of the six scales at the point which matches your experience. Each line has two endpoint descriptors that describe the scale. Note that "own performance" goes from "good" on the left to "bad" on the right. This order has

been confusing for some people. Please consider your responses carefully in distinguishing among the different task conditions. Consider each scale individually. *Your ratings will play an important role in the evaluation being conducted, thus, your active participation is essential to the success of this experiment and is greatly appreciated!*

APPENDIX C

SUBJECT INSTRUCTIONS: SOURCES-OF-WORKLOAD EVALUATION

6. SUBJECT INSTRUCTIONS: RATING SCALES

We are not only interested in assessing your performance but also the experiences you had during the different task conditions. Right now we are going to describe the technique that will be used to examine your experiences. In the most general sense we are examining the "workload" you experienced. Workload is a difficulty concept to define precisely, but a simple one to understand generally. The factors that influence your experience of workload may come from the task itself, your feelings about your own performance, how much effort you put in, or the stress and frustration you felt. The workload contributed by different task elements may change as you get more familiar with a task, perform easier or harder versions of it, or move from one task to another. Physical components of workload are relatively easy to conceptualize and evaluate. However, the mental components of workload may be more difficulty to measure.

Since workload is something that is experienced individually be each person, there are no effective "rulers" that can be used to estimate the workload of different activities. One way to find out about workload is to ask people to describe the feelings they experienced. Because workload may be caused by many different factors, we would like you to evaluate several of them individually rather than lumping them into a single global evaluation of overall workload. This set of six rating scales was developed for you to use in evaluating your experiences during different tasks. Please read the descriptions of the scales carefully. If you have a question about any of the scales in the table, please ask me about it. It is extremely important that they be clear to you. You may keep the descriptions with you for reference during the experiment.

After performing each of the tasks, you will be given a sheet of rating scales. You will evaluate the task by putting an "X" on each of the six scales at the point which matches your experience. Each line has two endpoint descriptors that describe the scale. Note that "own performance" goes from "good" on the left to "bad" on the right. This order has been confusing for some people. Please consider your responses carefully in distinguishing among the different task conditions. Consider each scale individually. Your ratings will play an important role in the evaluation being conducted, thus, your active participation is essential to the success of this experiment and is greatly appreciated by all of us.

7. SUBJECT INSTRUCTIONS: SOURCES-OF-WORKLOAD EVALUATION

Throughout this experiment the rating scales are used to assess your experiences in the different task conditions. Scales of this sort are extremely useful, but their utility suffers from the tendency people have to interpret them in individual ways. For example, some people feel that mental or temporal demands are the essential aspects of workload regardless of the effort they expended on a given task or the level of performance they achieved. Others feel that if they performed well the workload must have been low and if they performed badly it must have been high. Yet others feel that effort or feelings of frustration are the most important factors in workload; and so on. The results of previous studies have already found every conceivable pattern of values. In addition, the factors that create levels of workload differ depending on the task. For example, some tasks might be difficult because they must be completed very quickly. Others may seem easy or hard because of the intensity of mental or physical effort required. Yet others feel difficult because they cannot be performed well, not matter how much effort is expended.

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

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

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

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

RATING SCALE DEFINITIONS		
Title	**Endpoints**	**Descriptions**
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving
PHYSICAL DEMAND	*Low/High*	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	*Low/High*	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
PERFORMANCE	*good/poor*	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
FRUSTRATION LEVEL	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

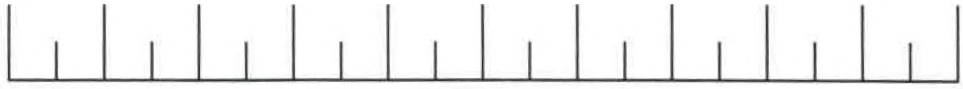
SOURCES OF WORKLOAD COMPARISON CARDS

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

Subject ID: _____ **Task ID:** _____

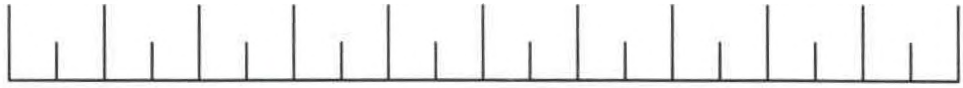
RATING SHEET

MENTAL DEMAND



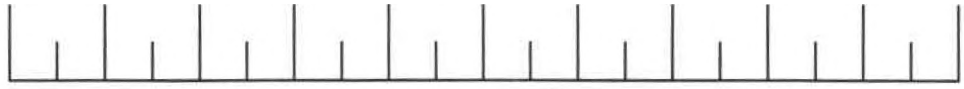
Low High

PHYSICAL DEMAND



Low High

TEMPORAL DEMAND



Low High

PERFORMANCE

Good Poor

EFFORT

Low High

FRUSTRATION

Low High

APPENDIX D

SUBJECT TASK INSTRUCTIONS

SUBJECT TASK INSTRUCTIONS

The purpose of this experiment is to determine how different levels of complexity effect how people perceive task complexity on subsequently performed tasks. The experiment will be performed on the Microsoft Flight Simulator, Version 4.0, and will not impose any dangerous levels of stress. The only stress resulting from this experiment may be some initial frustration as you learn the task. Other than that, you will be free from any stress.

Excluding the training session, the experiment will require approximately four hours within a seven day period.

You will be provided with the opportunity to train in preparation of the experiment. The training sessions will occur prior to the experiment. The training sessions consists of two flight courses. The first course contains 10 gates in a straight line. The second course contains 20 gates in an "S" configuration. You will be required to obtain minimum scores of 710 and 1600 points, respectively, to qualify for the experiment.

Your participation is strictly voluntary and will remain confidential. As part of the data analysis, your data will be combined with that of other individuals and you will no longer be identifiable as a participant.

You will receive 1 credit per hour (maximum 4 credit hours) in your undergraduate psychology course for participating in this study. Thus you will receive 4 extra credit points for completion of this study. Sorry, there will be no monetary reward.

Your participation in this study is extremely important and greatly appreciated. If you have any further questions, please do not hesitate to contact Doug Fischer, 294-2978 (H), 255-4842 (W).

APPENDIX E

INFORMED CONSENT

INFORMED CONSENT

Flight Simulation Experiment

I, _____, have read and retained the attached experimental task description entitled "Flight Simulation Experiment". I have also been provided with the opportunity to ask questions of the investigator. The investigator informed me that, including the training session, the experiment will require approximately four hours within a seven day period.

I understand that this experiment will impose no stress. The only stress I may experience in this experiment may be some initial frustration as I learn the task. As part of the data analysis, my data will be combined with that of other individuals and I will no longer be identifiable as a participant.

I have been informed that I have the right to withdraw from the experiment, and that the experiment monitor may terminate my participation in the interest of safety and the experiment. I also certify that I am at least 18 years of age.

I have also been informed that if additional details are needed, I may contact: Doug Fischer, 294-2978 (H), 255-4842 (W); or Professor William F. Moroney at 229-2766, or at St. Joe's Hall room 305.

Signed: _____

Date: _____ / _____ / 1993

Witness: _____

Date: _____ / _____ / 1993

APPENDIX F

EXPERIMENTAL PROCEDURES

PRACTICE

Experimental Procedures

Subject # _____ **Date:** _____ **Time:** _____

Sequence: _____

Prior to Practice Session: Provide copy of NASA-TLX materials. Ask subject to read the material before starting session.

PRACTICE SESSION

_____ Complete informed consent form.
_____ Bring 3 NASA-TLX rating forms, bring extra pencil

Warm-Up: Familiarization and Criterion:
Gates12, 6000DF (0 crosswinds; 710 points minimum)
DF60S2 , DOUG1 (0 crosswinds; 1600 points minimum)
CRITERION: Must fly through all gates and reach minimum scores
_____ Record data: # of trials, time, speed, score, gates missed

NASA-TLX: Introduce concept, subject should have read material previously.
Provide instructions and answer any questions
_____ One more flight: DF60S2, DOUG1 -- record data, get NASA-TLX rating, must meet criteria.
_____ Examine NASA-TLX form; answer any questions

START

Record Data: # of trials, time, speed, performance score, gates missed, NASA-TLX rating sheets for each trial.

_____ Trial 1: MMM -- DF60S3b, DOUG1 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX
_____ Trial 2: MMM -- DF60S3b, DOUG1 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX
_____ Trial 3: MMM -- DF60S3b, DOUG1 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX
_____ Confirm schedule for next session, at least 24 hrs later (1 day betwn. trials).

SESSION 1

Experimental Procedures

Subject # _____ **Date:** _____ **Time:** _____

Sequence: _____

Session 1

_____ Bring 3 NASA-TLX rating forms, extra pencil

Warm-Up: Familiarization and Criterion:

_____ Gates12, 6000DF (0 crosswinds; 710 points minimum)
_____ DF60S3 , DOUG2 (0 crosswinds; 1600 points minimum)
_____ CRITERION: Must fly through all gates and reach minimum scores
_____ Record data: # of trials, time, speed, score, gates missed

START

Record Data: # of trials, time, speed, performance score, gates missed, NASA-TLX rating sheets for each trial.

_____ **Trial 1:** M -- DF60S3b, DOUG2 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX
_____ **Administer and Collect NASA-TLX**

_____ **Trial 2:** Condition: _____
_____ **Administer and collect NASA-TLX**

_____ **Trial 3:** M -- DF60S3b, DOUG2 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX
_____ **Administer and collect NASA-TLX**

___ **Confirm schedule for next session.** At least 24 hrs later (1 day betwn.)

SESSION 2

Experimental Procedures

Subject # _____ **Date:** _____ **Time:** _____

Sequence: _____

Session 2

_____ Bring 3 NASA-TLX rating forms, extra pencil

Warm-Up: Familiarization and Criterion:

_____ Gates12, 6000DF (0 crosswinds; 710 points minimum)

_____ DF60S3 , DOUG2 (0 crosswinds; 1600 points minimum)

_____ CRITERION: Must fly through all gates and reach minimum scores

_____ Record data: # of trials, time, speed, score, gates missed

START

Record Data: # of trials, time, speed, performance score, gates missed, NASA-TLX rating sheets for each trial.

_____ **Trial 1:** M -- DF60S3b, DOUG2 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX

_____ **Administer and collect NASA-TLX**

_____ **Trial 2:** Condition: _____

_____ **Administer and collect NASA-TLX**

_____ **Trial 3:** M -- DF60S3b, DOUG2 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX

_____ **Administer and collect NASA-TLX**

_____ **Confirm schedule for next session**. At least 24 hrs later (1 day btwn.)

SESSION 3

Experimental Procedures

Subject # _____ **Date:** _____ **Time:** _____

Sequence: _____

Session 3

_____ Bring 3 NASA-TLX rating forms, extra pencil

Warm-Up: Familiarization and Criterion:

_____ Gates12, 6000DF (0 crosswinds; 710 points minimum)

_____ DF60S3 , DOUG2 (0 crosswinds; 1600 points minimum)

_____ CRITERION: Must fly through all gates and reach minimum scores

_____ Record data: # of trials, time, speed, score, gates missed

START

Record Data: # of trials, time, speed, performance score, gates missed, NASA-TLX rating sheets for each trial.

_____ **Trial 1:** M -- DF60S3b, DOUG2 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX

_____ **Trial 2:** Condition: _____

_____ **Administer and collect NASA-TLX**

_____ **Trial 3:** M -- DF60S3b, DOUG2 -- 20 gates -- (12kts @ 270)
Criterion: all gates, administer and collect NASA-TLX

_____ **Administer and collect NASA-TLX**

_____ **Confirm schedule for next session**. At least 24 hrs later (1 day btwn.)

APPENDIX G

ANOVA SUMMARY TABLES*

AND

DESCRIPTIVE ANALYSIS TABLES

* Degrees of freedom (df) reflect the Geisser-Greenhouse correction.

Table G-1: Trial 1 Performance and Workload Analyses

Performance Score Analysis:

Source	SS	df	MS	F	p
O	2096.92	5.00	419.38	0.34	0.872
Ss / O	7415.17	6.00	1235.86		
D	220.50	1.44	110.25	0.43	0.603
O X D	2839.83	7.22	283.98	1.10	0.440
D X Ss / O	3100.33	8.67	258.36		

Workload Rating Analysis:

Source	SS	df	MS	F	p
O	3098.81	5.00	619.76	0.80	0.590
Ss / O	4669.32	6.00	778.22		
D	0.68	1.25	0.34	0.01	0.963
O X D	941.28	6.26	94.13	2.01	0.183
D X Ss / O	562.33	7.51	46.86		

Source Key: O = Order
Ss = Subjects
D = Difficulty

Table G-2: **Mean Performance Scores, Standard Deviations, and Standard Error of Mean Obtained Under Trial 2 as a Function of Session and Task Difficulty.**

	SESSION 1			SESSION 2			SESSION 3		
	MEAN	**SD**	**SEM**	**MEAN**	**SD**	**SEM**	**MEAN**	**SD**	**SEM**
Low:	1715.917	19.332	5.581	1743.167	16.247	4.690	1720.000	17.735	5.120
Med:	1713.667	20.110	5.805	1723.917	24.586	7.097	1720.333	20.340	5.872
High:	1719.667	25.032	7.226	1679.333	34.033	9.824	1722.833	20.099	5.802

Table G-3: ANOVA Summaries for Trial 2 Difficulty Manipulation Check

Performance.

Source	SS	df	MS	F	p
Order	2836.47	5.00	567.29	0.48	0.780
Ss / O	7062.50	6.00	1177.08		
D	25731.72	1.60	12865.86	25.79	0.000*
O X D	6407.28	7.98	640.73	1.28	0.352
D X Ss / O	5987.00	9.57	498.92		

Comparison: Low vs. Medium

Source	SS	df	MS	F	p
O	2836.47	5.00	567.29	0.48	0.780
Ss / O	7062.50	6.00	1177.08		
L vs. M	2223.38	1.00	2223.38	3.74	0.101
O X L vs. M	1088.38	5.00	217.68	0.37	0.855
D X Ss / O	3564.75	6.00	594.13		

Comparison: Low vs. High

Source	SS	df	MS	F	p
O	2836.47	5.00	567.29	0.48	0.780
Ss / O	7062.50	6.00	1177.08		
L vs. H	24448.17	1.00	24448.17	37.46	0.001*
O X L vs. H	3342.33	5.00	668.47	1.02	0.479
D X Ss / O	3915.50	6.00	652.58		

Comparison: Medium vs. High

Source	SS	df	MS	F	p
O	2836.47	5.00	567.29	0.48	0.780
Ss / O	7062.50	6.00	1177.08		
M vs. H	11926.04	1.00	11926.04	47.70	0.000*
O X M vs. H	5180.21	5.00	1036.04	4.14	0.056
D X Ss / O	1500.25	6.00	250.04		

* Significant < 0.05

Source Key: O = Order
Ss = Subjects
D = Difficulty

Table G-4. **Mean NASA-TLX Ratings, Standard Deviations, and Standard Error of Mean Obtained Under Trial 2 as a Function of Session and Task Difficulty.**

	SESSION 1			SESSION 2			SESSION 3		
	MEAN	**SD**	**SEM**	**MEAN**	**SD**	**SEM**	**MEAN**	**SD**	**SEM**
Low:	50.917	16.030	4.627	41.722	16.240	4.688	53.139	16.112	4.651
Med:	51.194	17.594	5.079	47.861	19.644	5.671	50.417	18.599	5.369
High:	50.889	16.624	4.799	63.694	15.366	4.436	50.028	18.217	5.259

Table G-5. ANOVA Summaries for Trial 2 Reported Workload

Reported Workload.

Source	SS	df	MS	F	p
O	3334.59	5.00	666.92	0.70	0.641
Ss / O	5683.62	6.00	947.27		
D	3084.62	1.38	1542.31	36.26	0.000*
O X D	214.28	6.91	21.43	0.50	0.808
D X Ss / O	510.41	8.29	42.53		

Comparison: Low vs. Medium Levels of Difficulty

Source	SS	df	MS	F	p
O	3334.59	5.00	666.92	0.70	0.641
Ss / O	5683.62	6.00	947.27		
L vs. M	226.11	1.00	226.11	5.52	0.057
O X L vs. M	188.07	5.00	37.61	0.92	0.527
D X Ss / O	245.77	6.00	40.96		

Comparison: Low vs. High

Source	SS	df	MS	F	p
O	3334.59	5.00	666.92	0.70	0.641
Ss / O	5683.62	6.00	947.27		
L vs. H	2896.66	1.00	2896.66	154.61	0.000*
O X L vs. H	98.96	5.00	19.79	1.06	0.465
D X Ss / O	112.41	6.00	18.74		

Comparison: Medium vs. High

Source	SS	df	MS	F	p
O	3334.59	5.00	666.92	0.70	0.641
Ss / O	5683.62	6.00	947.27		
M vs. H	1504.17	1.00	1504.17	22.15	0.003*
O X M vs. H	34.39	5.00	6.88	0.10	0.988
D X Ss / O	407.44	6.00	67.91		

* Significant at 0.05 level

Source Key: O = Order
Ss = Subjects
D = Difficulty

Table G-6. ANOVA Summaries for Trial 2 NASA-TLX Subscales

Mental.

Source	SS	df	MS	F	p
O	4813.89	5.00	962.78	0.39	0.839
Ss / O	14758.33	6.00	2459.72		
D	2543.06	1.96	1271.53	22.06	0.000*
O X D	748.61	9.81	74.86	1.30	0.331
D X Ss / O	691.67	11.77	57.64		

Physical.

Source	SS	df	MS	F	p
Order	15222.22	5.00	3044.44	2.24	0.177
Within Cells	8150.00	6.00	1358.33		
D	1484.72	1.54	742.36	39.59	0.000*
O X D	1040.28	7.69	104.03	5.55	0.009*
D X Ss / O	225.00	9.23	18.75		

Temporal.

Source	SS	df	MS	F	p
O	8822.22	5.00	1764.44	2.74	0.126
Ss / O	3866.67	6.00	644.44		
D	459.72	1.66	229.86	2.49	0.138
O X D	648.61	8.29	64.86	0.70	0.698
D X Ss / O	1108.33	9.95	92.36		

Performance.

Source	SS	df	MS	F	p
O	3761.81	5.00	752.36	1.92	0.224
Ss / O	2345.83	6.00	390.97		
D	5776.39	1.93	2888.19	20.49	0.000*
O X D	631.94	9.64	63.19	0.45	0.889
D X Ss / O	1691.67	11.56	140.97		

Source Key: O = Order
Ss = Subjects
D = Difficulty

Effort.

Source	SS	df	MS	F	p
O	12191.81	5.00	2438.36	1.55	0.304
Ss / O	9461.83	6.00	1576.97		
D	2235.06	1.23	1117.53	18.79	0.002*
O X D	405.28	6.15	40.53	0.68	0.674
D X Ss / O	713.67	7.38	59.47		

Frustration.

Source	SS	df	MS	F	p
O	14622.22	5.00	2924.44	1.71	0.265
Ss / O	10241.67	6.00	1706.94		
D	3251.39	1.48	1625.69	24.90	0.000*
O X D	1115.28	7.40	111.53	1.71	0.223
D X Ss / O	783.33	8.88	65.28		

* Significance < 0.05

Source Key: O = Order
Ss = Subjects
D = Difficulty

Table G-7: ANOVA Summaries for Medium Condition Performance Scores Under Trial 1 and Trial 3

Performance.

Source	SS	df	MS	F	p
O	1777.24	5.00	355.45	0.17	0.962
Ss / O	12207.25	6.00	2034.54		
Difficulty	283.53	1.67	119.26	0.32	0.695
O X D	2794.97	8.35	279.50	0.75	0.654
D X Ss / O	4461.50	10.02	371.79		
T	387.35	1.00	387.35	1.44	0.267
O X T	1633.90	5.00	326.78	1.21	0.405
T X Ss / O	1617.58	6.00	269.60		
D X T	39.53	1.74	19.76	0.15	0.839
O X D X T	1784.97	8.71	178.50	1.31	0.333
T X D X Ss / O	1629.17	10.45	135.76		

Source Key: O = Order
Ss = Subjects
D = Difficulty
T = Trial

Table G-8: ANOVA Summaries for Medium Condition NASA-TLX Ratings Under Trial 1 and Trial 3

Reported Workload.

Source	SS	df	MS	F	p
O	6177.43	5.00	1235.49	0.71	0.639
Ss / O	10456.960	6.00	1742.83		
D	32.620	2.00	16.31	0.24	0.792
O X D	1028.440	10.00	1.50	1.50	0.250
D X Ss / O	822.52	12.00	68.54		
T	0.68	1.00	0.68	0.03	0.859
O X T	263.61	5.00	52.72	2.67	0.132
T X Ss / O	118.41	6.00	19.73		
D X T	37.03	1.74	18.51	0.55	0.486
O X D X T	311.72	8.68	31.17	0.93	0.538
T X D X Ss / O	403.74	10.42	33.64		

Source Key: O = Order
Ss = Subjects
D = Difficulty
T = Trial

Table G-9: ANOVA Summaries for the Medium Condition TLX Subscale Ratings Under Trial 1 and Trial 3

Mental.

Source	SS	df	MS	F	p
O	7198.96	5.00	1439.79	0.35	0.868
Ss / O	24943.75	6.00	4157.29		
D	27.08	1.60	13.54	0.19	0.780
O X D	827.08	8.00	82.71	1.19	0.395
D X Ss / O	837.50	9.60	69.79		
T	292.01	1.00	292.01	5.08	0.053
O X T	360.07	5.00	72.01	1.43	0.335
T X Ss / O	302.08	6.00	50.35		
D X T	104.86	1.80	52.43	1.19	0.336
O X D X T	124.31	9.02	12.43	0.28	0.966
T X D X Ss / O	529.17	10.82	44.10		

Physical.

Source	SS	df	MS	F	p
O	34356.94	5.00	6871.39	2.75	0.125
Ss / O	15000.00	6.00	2500.00		
D	179.86	1.94	89.93	1.27	0.316
O X D	753.47	9.69	75.35	1.06	0.453
D X Ss / O	850.00	11.63	70.83		
T	68.06	1.00	68.06	2.04	0.203
O X T	506.94	5.00	101.39	3.04	0.104
T X Ss / O	200.00	6.00	33.33		
D X T	4.86	1.26	2.43	0.07	0.848
O X D X T	95.14	6.29	9.51	0.29	0.932
T X D X Ss / O	400.00	7.55	33.33		

Source Key: O = Order
Ss = Subjects
D = Difficulty
T = Trial

Table G-9: ANOVA Summaries for the Medium Condition TLX Subscale Ratings Under Trial 1 and Trial 3

Temporal.

Source	SS	df	MS	F	p
O	19656.94	5.00	3931.39	3.12	0.099
Ss / O	7558.33	6.00	1259.72		
D	134.03	1.11	67.01	0.83	0.407
O X D	899.31	5,56	89.93	1.12	0.438
D X Ss / O	966.67	6.67	80.56		
T	0.00	1.00	0.00	0.00	1.00
O X T	58.33	5.00	11.67	0.37	0.856
T X Ss / O	191.67	6.00	31.94		
D X T	139.58	1.27	69.79	1.50	0.268
O X D X T	402.08	6.37	40.21	0.86	0.564
T X D X Ss / O	558.33	7.65	46.53		

Performance.

Source	SS	df	MS	F	p
O	7330.90	5.00	1466.18	2.02	0.208
Ss / O	4347.92	6.00	724.65		
D	304.86	1.69	152.43	0.82	0.447
O X D	1490.97	8.45	149.10	0.81	0.617
D X Ss / O	2220.83	10.14	185.07		
T	153.13	1.00	153.13	4.28	0.084
O X T	703.13	5.00	140.63	3.93	0.063
T X Ss / O	214.58	6.00	35.76		
D X T	43.75	1.46	21.87	0.23	0.734
O X D X T	968.75	7.32	96.88	1.01	0.487
T X D X Ss / O	1154.17	8.78	96.18		

Source Key: O = Order
Ss = Subjects
D = Difficulty
T = Trial

Table G-9: ANOVA Summaries for the Medium Condition TLX Subscale Ratings Under Trial 1 and Trial 3

Effort.

Source	SS	df	MS	F	p
O	23351.74	5.00	4670.35	1.46	0.325
Ss / O	19143.75	6.00	3190.63		
D	134.03	1.97	67.01	1.17	0.343
O X D	1345.14	9.84	134.51	2.35	0.083
D X Ss / O	687.50	11.81	57.29		
T	0.35	1.00	0.35	0.01	0.931
O X T	235.07	5.00	47.01	1.12	0.440
T X Ss / O	252.08	6.00	42.01		
D X T	150.69	1.75	75.35	1.56	0.253
O X D X T	320.14	8.77	32.01	0.66	0.723
T X D X Ss / O	579.17	10.53	48.26		

Frustration.

Source	SS	df	MS	F	p
O	23489.24	5.00	4697.85	1.33	.363
Ss / O	21114.58	6.00	3519.10		
D	586.11	1.98	293.06	2.30	0.144
O X D	1559.72	9.88	155.97	1.22	0.366
D X Ss / O	1529.17	11.85	127.43		
T	58.68	1.00	58.68	1.94	0.213
O X T	464.24	5.00	92.85	3.07	0.102
T X Ss / O	181.25	6.00	30.21		
D X T	369.44	1.91	184.72	3.94	0.051
O X D X T	1826.39	9.54	182.64	3.90	0.017
T X D X Ss / O	562.50	11.45	46.88		

Source Key: O = Order
Ss = Subjects
D = Difficulty
T = Trial

Table G-10: Percent Change ANOVA Summaries for Medium Condition Performance Scores and Workload Ratings Under Trial 1 and Trial 3

Performance.

Source	SS	df	MS	F	p
O	0.00	5	0.00	1.22	0.401
Ss / O	0.00	6	0.00		
D	0.00	1.76	0.00	0.15	0.840
O X D	0.00	8.78	0.00	1.33	0.326
D X Ss / O	0.00	10.54	0.00		

Workload.

Source	SS	df	MS	F	p
O	0.24	5.00	0.05	2.30	0.170
Ss / O	0.13	6.00	0.02		
D	0.05	1.41	0.03	.74	0.424
O X D	0.22	7.07	0.02	.62	0.733
D X Ss / O	0.42	8.48	0.04	.74	0.462

Source Key: O = Order
Ss = Subjects
D = Difficulty