

2006

Context effects on subjective workload assessments using the multi-attribute task battery tracking task: an extension of Voorheis

Michael A. Petkosek
University of Dayton

Follow this and additional works at: https://ecommons.udayton.edu/graduate_theses

Recommended Citation

Petkosek, Michael A., "Context effects on subjective workload assessments using the multi-attribute task battery tracking task: an extension of Voorheis" (2006). *Graduate Theses and Dissertations*. 4920.
https://ecommons.udayton.edu/graduate_theses/4920

This Thesis is brought to you for free and open access by the Theses and Dissertations at eCommons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of eCommons. For more information, please contact mschlangen1@udayton.edu, ecommons@udayton.edu.

**CONTEXT EFFECTS ON SUBJECTIVE
WORKLOAD ASSESSMENTS USING THE MULTI-ATTRIBUTE TASK
BATTERY TRACKING TASK:
AN EXTENSION OF VOORHEIS**

THESIS

Submitted to

**The Department of Psychology of the
UNIVERSITY OF DAYTON**

In Partial Fulfillment of the Requirements for

The Degree

Master of Arts in Psychology

by

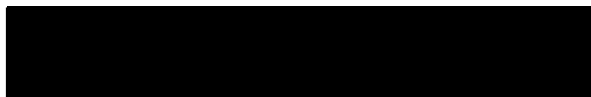
Michael A. Petkosek

UNIVERSITY OF DAYTON

Dayton, Ohio

August, 2006

APPROVED BY:



William F. Moroney

Faculty Advisor



David W. Biers

Committee Member and Department Chairperson



F. Thomas Eggemeier

Committee Member



David. W. Biers

Chair, Department of Psychology

ABSTRACT

CONTEXT EFFECTS ON SUBJECTIVE WORKLOAD ASSESSMENTS USING THE MULTI-ATTRIBUTE TASK BATTERY TRACKING TASK: PART II

Name: Michael A. Petkosek
University of Dayton 2006

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

This study was designed to determine if prior levels of task difficulty impact performance scores and workload ratings on the subsequent task. Earlier research by Hancock, et al. (1995) demonstrated the presence of context effects on subjective workload ratings. Research conducted by: Moroney, et al. (1993), Fisher (1995), Rensch (2002), and Voorheis (2004) was unable to observe similar context effects on workload. This study, a partial replication of Voorheis, was designed to mitigate possible practice effects that may have hindered the emergence of context effects in prior studies.

Thirty-six participants completed one session of the Multi-Attribute Task Battery tracking task. The first and fourth trials (Baseline and Critical respectively) were medium difficulty tasks; the second and third trials (Context Inducing) were low, medium (control condition), or high difficulty tasks. Difficulty was presented between subjects. Workload ratings were obtained after each trial with the National Aeronautics and Space Administration-Task Load Index.

Performance scores and workload ratings on the Baseline and Critical Trials were compared. There was no interaction effect of Difficulty or Trial; Critical Trial workload ratings did not differ for difficulty groups. Thus, no context effects were observed.

ACKNOWLEDGMENTS

In growing and learning, the effort of my hard work is only to return to God a fraction of what I have been given. To my parents, who equipped me with lessons needed to grow and learn, set me in directions I did not always understand, and met every day of my life with patience, this is just one paper but it signifies an entire era. To Dr. Moroney at the University of Dayton, who believed in me before I even thought to myself, I have been set on a course for which you are responsible. To Richard Moss (Ret.) and Gregory Barbato of the Air Force Research Laboratory (AFRL) for providing me career opportunities and responsibilities, and allowing me time to complete my thesis while beginning a career. To Chris Voorheis, for helping me to continue this work and for his patient support in keeping me calm throughout this endeavor. To Dr. Biers and Dr. Eggemeier for the time they devoted to this effort. And to all of my friends and co-workers at AFRL, remember who we are working for, and that everything important will probably come at a price to our time, patience, and self desires. In the words of President George W. Bush, "...it's going to take hard work."

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iv
LIST OF FIGURES	vii
LIST OF TABLES.....	viii
CHAPTER	
I. INTRODUCTION.....	1
Research Review	2
Hancock, Williams, Manning, and Miyake (1995)	2
Moroney, Reising, Biers, and Eggemeier (1993)	4
Fischer (1995)	5
Rench (2002)	7
Voorheis (2004)	9
Current Study.....	12
Hypotheses.....	14
II. METHOD.....	16
Participants	16
Software and Apparatus.....	16
Experimental Design	17
Independent Variable.....	17
Dependent Variables.....	17
Procedure	18

III. RESULTS	20
Baseline Trial (M1).....	20
Performance Scores	20
Workload Ratings	21
Context Inducing Trial (C2)	22
Performance Scores	22
Workload Ratings	23
Context Inducing Trial (C3)	25
Performance Scores	25
Workload Ratings	26
Critical Trial (M4)	28
First Analysis	28
Second Analysis.....	30
Subscale Analyses	31
IV.DISCUSSION.....	33
Performance.....	33
Workload	35
TLX Subscales.....	38
Conclusion	38
Additional Consideration.....	39
REFERENCES	41
APPENDICES	
A. MATB TRACKING INSTRUCTIONS	44
B. TLX MATERIALS	46
C. INFORMED CONSENT	51
D. EXPERIMENTER'S CHECKLIST	53
E. DESCRIPTIVE STATISTIC TABLES REFERRED TO IN CH. III	60

LIST OF FIGURES

1. Mean performance scores as a function of difficulty for the Baseline Trial (M1)..21
2. Mean workload ratings as a function of difficulty for the Baseline Trial (M1)	22
3. Mean performance scores as a function of difficulty for the First Context Inducing Trial (C2)	23
4. Mean workload ratings as a function of difficulty for the First Context Inducing Trial (C2)	24
5. Mean performance scores as a function of difficulty for the Second Context Inducing Trial (C3)	26
7. Mean performance scores as a function of difficulty for the Critical Trial (M4).....	29
8. Mean workload ratings as a function of difficulty for the Critical Trial (M4).....	30
9. Comparison of the subjective workload ratings obtained for each TLX subscale in the Baseline (M1) and Critical (M4) Trialsl	32
10. Mean performance trends for each difficulty group across all trials.	34
11. Mean workload trends for each difficulty group across all trials.	35

LIST OF TABLES

1. Experimental design used by Hancock, et al. (1995)	3
2. Experimental design used by Moroney, et al. (1993).....	5
3. Comparison of the experimental designs of Moroney, et al. (1993) and Fischer (1995).....	6
4. Comparison of the experimental designs of Fischer (1995) and Rench (2002)	8
5. Comparison of the experimental designs of Hancock, et al. (1995) and Voorheis (2004)	9
6. Comparison of participant exposure to the medium level of difficulty in Voorheis (2004) and Petkosek (2005).....	13
7. Comparison of the experimental designs of Hancock, et al. (1995), Voorheis (2004), and Petkosek (2005)	14
E1. Means and Standard Deviations for Baseline Trial (M1) Performance scores	57
E2. Means and Standard Deviations for Baseline Trial (M1) Workload ratings.....	57
E3. Means and Standard Deviations for the first Context Inducing Trial (C2) Performance scores.....	57
E4. Means and Standard Deviations for the first Context Inducing Trial (C2) Workload ratings	58
E5. Means and Standard Deviations for the second Context Inducing Trial (C3) Performance scores	58
E6. Means and Standard Deviations for the second Context Inducing Trial (C3) Workload ratings	58
E7. Means and Standard Deviations for Critical Trial (M4) Performance scores.....	59
E8. Means and Standard Deviations for Critical Trial (M4) Workload ratings	59
E9. Means and Standard Deviations for the NASA-TLX Subscales on the Baseline Trial (M1) and Critical Trial (M4)	59

CHAPTER I

The measurement of operator workload is an important component of system design and evaluation. If the workload level is too high, a system can overload an operator's ability to perform necessary tasks; if the level of workload is too low, operator attention may suffer due to the effects of boredom and fatigue (Wickens & Hollands, 2000). Improper levels of workload can result in decreased performance.

Workload is a complicated construct. A single definition has been elusive; workload has several sources, and can be measured by a variety of methods. Common sources of workload include: the demands imposed by a task, the level of performance an operator can achieve, mental effort exerted by an operator, physical effort exerted by an operator, and the operator's perception of a task (Huey & Wickens, 1993). The most common measurements of workload include: performance-based, physiological, and subjective measures (O'Donnell & Eggemeier, 1986). Performance measures quantify an operator's performance, physiological measures quantify an operators physical state, and subjective measures quantify an operator's perception of workload. A common method of gathering subjective data is the use of rating scales, such as the National Aeronautical and Space Administration Task Load Index (NASA-TLX: Hart & Staveland, 1988). The NASA-TLX is a multidimensional rating technique that requires an operator to assign values on six different dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland).

Workload context effects occur when operator performance reflects task difficulty but the perception of workload is affected by the difficulty of a preceding task. A 1995 study by Hancock, Williams, Manning, and Miyake detected this effect. Performance reflected task difficulty but an increase in workload ratings followed a low difficulty trial and a decrease followed a high difficulty trial. Han-

cock et al. evaluated subjective workload with the NASA-TLX and the Subjective Workload Assessment Technique (SWAT: Reid & Nygren, 1988). However, other studies using only the NASA-TLX were unable to detect these effects (Moroney, Reising, Biers, & Eggemeier, 1993; Fischer, 1995; Rench, 2002; and Voorheis, 2004). In the following section the experimental designs of these studies are presented. Hancock et al. (who used both the NASA-TLX and the SWAT to evaluate workload and found workload context effects) is reviewed first. The studies that used only the NASA-TLX are then reviewed in chronological order.

Research Review

Hancock, Williams, Manning, and Miyake (1995)

Hancock, et al. (1995) investigated the effects of task difficulty on subsequent tasks using a compensatory tracking task. Twelve participants controlled an animated aircraft icon on a display screen to keep it centered in a sight circle, which was a two-dimensional task.

Participants took part in three experimental sessions where the difficulty of control for the icon was varied. Three levels of difficulty (low, medium, and high) were evaluated within-subjects; each participant received one level of difficulty within a session. Sessions consisted of three trials (understood as Baseline, Context Inducing, and Critical). Baseline and Critical Trials were given at medium difficulty (M1 and M3 respectively) and the Context Inducing Trial was given at either low, medium, or high difficulty (L2, M2, or H2 respectively). The presentation of a difficulty manipulation within medium trials created a pre-post design; workload context effects were evaluated in the Critical Trial.

Subjective workload was measured after each trial by paper-and-pencil versions of the NASA-TLX and SWAT. Performance was scored for each trial by root mean square error (RMSE). RMSE was computed by the following formula: $RMSE = \sqrt{1/N \sum (Y_i - \hat{Y}_i)^2}$ where Y_i is the optimal value (in this case the optimal value was zero pixel units away from the center of the target) and \hat{Y}_i is the recorded value (the actual number of pixel units that the aircraft icon was away from the target). The experimental design is reviewed in Table 1.

Table 1

Experimental design used by Hancock, et al. (1995)

	Hancock et al
Number of participants	12
Task type	Tracking
Number of experimental sessions	3
Number of trials per session	3
Presentation of trial types	Baseline, Context Inducing, Critical
Trial difficulty by session	M1-L2-M3
(L=low, M=med, H=high)	M1-M2-M3
	M1-H2-M3
Trial duration/Session duration	5 minutes/15 minutes
Workload measurement tool	NASA-TLX, SWAT
Design	Pre-Post

Hancock, et al. (1995) hypothesized that previous task difficulty would influence subsequent workload ratings. Context effects were analyzed by computing the percent change of both performance scores and workload ratings. There was some ambiguity as to how the percent change was calculated. The terms *difference score* and *percent change* were used interchangeably in the Hancock text. For the purpose of follow-on studies, percent change has been defined as: $\{(\text{Critical Trial (M3)} - \text{Baseline Trial (M1)}) / \text{M1}\} * 100$ (Voorheis, 2004).

Percent change analyses of performance scores demonstrated a non-significant increase in RMSE for the high difficulty condition and a non-significant decrease in RMSE for the low difficulty condition (higher RMSE values indicate greater error). No context effects of difficulty were found for performance scores.

Percent change analyses of SWAT workload ratings increased significantly after the low difficulty Context Inducing Trial and decreased significantly after the high difficulty Context Inducing Trial. The SWAT percent change ratings were significantly different for each of the Context Inducing manipulations (low, medium, and high difficulty). Similarly, there were significant differences between the SWAT ratings of the Baseline and Critical Trials after the low and high difficulty manipulations.

Percent change analyses of NASA-TLX workload ratings indicated that the low difficulty Context Inducing Trial produced a significant difference when

compared to the high difficulty Context Inducing Trial. A significant difference was seen in NASA-TLX ratings between the Baseline and Critical Trials for the low difficulty Context Inducing Trial. Specifically, after the low difficulty Context Inducing Trial workload ratings increased, and after the high difficulty Context Inducing Trial workload ratings decreased. Thus, context effects from the low and high Context Inducing Trials significantly affected workload ratings of the Critical Trials.

Moroney, Reising, Biers, and Eggemeier (1993)

Moroney, et al. (1993) investigated the effects of task difficulty on subsequent tasks using a desktop flight simulator (Microsoft Flight Simulator). Twelve participants flew a course consisting of ten aerial gates (200 ft x 200 ft) that were centered about a constant heading (000) and altitude (6 000 ft). Each gate was 0.7 miles apart and used a crosshair to indicate its center. Participants received initial training with the simulated aircraft (Cessna 182) and were tested against a time and accuracy criterion.

Participants took part in three experimental sessions where the difficulty of control for the simulated aircraft was varied. Three levels of difficulty (low, medium, and high) were given within-subjects; each participant received one level of difficulty within a session. Sessions consisted of six trials. Trials were presented in two alternating phases: Context and Experimental. No baseline of difficulty was achieved because participants began with a Context Trial. Context Trials were always medium difficulty; Experimental Trials were either: low, medium, or high difficulty. Level of difficulty was created by crosswinds; the speeds of 2, 12, and 22 knots represented the levels of low, medium, and high respectively. Participants flew the same course on which they were trained and were not allowed to adjust aircraft power and trim settings.

Subjective workload was measured after each trial with a paper-and-pencil version of the NASA-TLX. Weights were not assigned based on research that showed a .97 correlation between weighted and un-weighted ratings (Byers, Bittner, & Hill, 1998 and Nygren 1991). Performance was evaluated by an algorithm that calculated accuracy of flight through the center of each gate, flight speed, and

added a twenty-five point bonus for navigating through all the gates. The experimental design can be seen in Table 2.

Table 2

Experimental design used by Moroney, et al. (1993)

	Moroney et al
Number of participants	12
Task type	Flight simulation
Number of experimental sessions	3
Number of trials per session	6
Presentation of trial types	Context, Experimental
Trial difficulty by session (L=low, M=med, H=high)	L1-M2-L3-M4-L5-M6 M1-M2-M3-M4-M5-M6 H1-M2-H3-M4-H5-M6
Trial duration/Session duration	3 minutes/18 minutes
Workload measurement tool	NASA-TLX
Design	Post

Analysis of performance scores showed that prior task difficulty did not affect performance on the subsequent task. Performance scores decreased as difficulty increased; performance scores on low difficulty trials were significantly higher than performance scores on medium and high difficulty trials. Medium and high difficulty trials did not significantly differ.

Moroney, et al. (1993) reported no detectable context effects of prior task difficulty. Analyses of the NASA-TLX ratings showed that previous trial difficulty did not significantly affect ratings during the Context Trial. The previous level of task difficulty did not significantly influence subjective workload ratings of the Experimental Trials of this study.

Fischer (1995)

Fischer (1995) investigated the effects of task difficulty on subsequent tasks using a desktop flight simulator (Microsoft Flight Simulator). Twelve participants flew a simulated aircraft (Cessna 182) through twenty aerial gates (200 ft x 200 ft) along a course of S-curves. The gates were displayed 0.1 miles east or west, fixed at a constant altitude (6 000 ft), and were separated by 0.7 miles.

Participants took part in three experimental sessions where the difficulty of control for the simulated aircraft was varied. Three levels of difficulty (low, medium, and high) were given within-subjects; each participant received one level of difficulty within a session. Similar to the pre-post design of Hancock, et al. (1995), Fischer (1995) sessions consisted of three trials (Baseline, Context Inducing, and Critical). Fisher created levels of difficulty with the same crosswinds as Moroney, et al. (1993) (2, 12, and 22 knots). Fischer used five minute trials (an increase from three minute trials used in Moroney et al.) on the assumption that participants of Moroney et al. were unable to experience context effects in a three minute trial.

Subjective workload was measured after each trial with a paper-and-pencil version of the NASA-TLX. Weights were assigned to replicate the procedure of Hancock, et al. (1995). Performance was scored with the same algorithm used by Moroney, et al. (1993). The design Fischer (1995) is compared to Moroney et al. in Table 3.

Table 3

Comparison of the experimental designs of Moroney, et al. (1993) and Fischer (1995)

	Moroney et al	Fischer
Number of participants	12	12
Task type	Flight simulation	Flight simulation
Number of experimental sessions	3	3
Number of trials per session	6	3
Presentation of trial types	Context, Experimental	Baseline, Context Inducing, Critical
Trial difficulty by session (L=low, M=med, H=high)	L1-M2-L3-M4-L5-M6 M1-M2-M3-M4-M5-M6 H1-M2-H3-M4-H5-M6	M1-L2-M3 M1-M2-M3 M1-H2-M3
Trial duration/Session duration	3 minutes/18 minutes	5 minutes/15 minutes
Workload measurement tool	NASA-TLX	NASA-TLX
Design	Post	Pre-Post

Note. Darkened cells indicate differences between Fischer and Moroney et al.

Fischer (1995) hypothesized that a high difficulty trial would result in lower workload ratings on the Critical Trial. Likewise, the low difficulty trial would result in higher workload ratings on the Critical Trial.

Fischer (1995) found a significant main effect of difficulty for performance scores and workload ratings in the Context Inducing Trial. Performance scores for low difficulty were significantly different from medium and high difficulty (two statistically different subgroups: L,M/H). Workload ratings for low difficulty were significantly different from medium and high difficulty (the same two subgroups: L,M/H). Fischer did not observe a significant difference of percent change from Baseline to Critical Trial for either performance scores or workload ratings. Similar to Moroney, et al (1993), Fischer did not detect subjective workload context effects.

Rench (2002)

Rench (2002) investigated the effects of task difficulty on subsequent tasks using a desktop flight simulator (Microsoft Flight Simulator). Twelve participants flew a simulated Cessna 182 through twenty aerial gates along a course of S-curves. The dimensions of the course and gates were the same as Fischer (1995).

Participants took part in three experimental sessions where the difficulty of control for the simulated aircraft was varied. Three levels of difficulty (low, medium, and high) were given within-subjects; each participant received one level of difficulty within a session. As in Moroney, et al. (1993) and Fischer (1995), level of difficulty was created with crosswinds. Rench (2002) used a larger difference between difficulty groups compared to Moroney et al. and Fischer on the assumption that their participants were unable to experience context effects with the given levels of difficulty. It was hypothesized that the levels of difficulty employed by Moroney et al. and Fischer were not different enough to elicit subjective differences in the perception of workload. Rench presented 0, 30, and 60 knot crosswinds for the levels of low, medium, and high respectively. This provided a thirty-knot difference between levels of difficulty, which was twenty knots more than the difference between the difficulty levels of Moroney et al. and Fischer (who used ten-knot differences: 2, 12, and 22 knot crosswinds). Rench also varied the direction of the crosswinds. The medium difficulty level (30 knots) crosswind originated from a heading of 270 degrees, the high difficulty level (60

knots) crosswind originated from a heading of 90 degrees. Moroney et al. and Fischer presented all crosswinds from a heading of 270 degrees.

The design and analysis were identical to Fischer (1995). Subjective workload was measured after each trial with a paper-and-pencil version of the NASA-TLX, weights were assigned. Performance was scored with the same algorithm used by Moroney, et al. (1993) and Fischer. The design of Rensch's study is shown with Fischer's study in Table 4.

Table 4

Comparison of the experimental designs of Fischer (1995) and Rensch (2002)

	Fischer	Rensch
Number of participants	12	12
Task type	Flight simulation	Flight simulation
Number of experimental sessions	3	3
Number of trials per session	3	3
Presentation of trial types	Baseline, Context Inducing, Critical	Baseline, Context Inducing, Critical
Trial difficulty by session (L=low, M=med, H=high)	M1-L2-M3 M1-M2-M3 M1-H2-M3	M1-L2-M3 M1-M2-M3 M1-H2-M3
Trial duration/Session duration	5 minutes/15 minutes	5 minutes/15 minutes
Workload measurement tool	NASA-TLX	NASA-TLX
Design	Pre-Post	Pre-Post

Note. Rensch used the same design and analysis as Fischer. Rensch increased the difference between levels of difficulty to facilitate subjective workload context effects.

Rensch hypothesized that previous level of task difficulty would inversely influence workload ratings. Specifically, the high level of task difficulty would result in lower workload ratings on a medium task (Critical Trial) relative to the Baseline rating. Likewise, the low level of task difficulty would result in higher workload ratings on a medium task (Critical Trial) relative to the Baseline rating; no change was expected for the control condition of medium.

Rensch (2002) found a significant main effect of difficulty for performance scores and workload ratings on the Context Inducing Trial. Specifically, performance scores were significantly higher for the low level of difficulty than the medium and high levels, and scores were significantly higher for the medium level than the high level. Thus, three distinct subgroups were seen ($L2 > M2 > H2$). Workload ratings were significantly lower for both the low and medium levels of difficulty compared to the high level ($L2 = M2 < H2$). Rensch did not observe a sig-

nificant difference or percent change from Baseline to Critical Trial for either performance scores or workload ratings. Like Moroney, et al. (1993) and Fischer (1995), Rensch did not observe context effects.

Voorheis (2004)

Voorheis (2004) investigated the effects of task difficulty on subsequent tasks using a compensatory tracking task. Twelve participants performed the tracking task contained in Version 4.0 of the Multi-Attribute Task Battery (MATB: Comstock & Arnegard, 1992). This tracking task was essentially the same used by Hancock, et al. (1995).

Voorheis (2004) used the same experimental design as Hancock, et al. (1995). Three levels of difficulty (low, medium, and high) were created by manipulating the tracking difficulty and gain settings of the icon in the tracking task. Difficulty levels were determined by a pilot study. Performance was scored using RMSE values. The same formula (as Hancock et al.) was applied: $RMSE = \sqrt{1/N \sum (Y_i - \hat{Y}_i)^2}$. Subjective workload was recorded with a software version of the NASA-TLX (Hancock et al. evaluated subjective workload with paper-and-pencil versions of the NASA-TLX and the SWAT). Voorheis replicated the procedure and analyses used by Hancock et al. in calculating percent change and difference scores for performance scores and workload ratings. In addition, Voorheis also analyzed group differences from zero for each level of difficulty. The experimental design of Voorheis is compared to that used by Hancock et al. in Table 5.

Table 5

Comparison of the experimental designs of Hancock, et al. (1995) and Voorheis (2004)

	Hancock et al	Voorheis
Number of participants	12	12
Task type	Tracking	Tracking
Number of experimental sessions	3	3
Number of trials per session	3	3
Presentation of trial types	Baseline, Context Inducing, Critical	Baseline, Context Inducing, Critical
Trial difficulty by session (L=low, M=med, H=high)	M1-L2-M3 M1-M2-M3 M1-H2-M3	M1-L2-M3 M1-M2-M3 M1-H3-M3
Trial duration/Session duration	3 minutes/15 minutes	3 minutes/15 minutes
Workload measurement tool	NASA-TLX, SWAT	NASA-TLX
Design	Pre-Post	Pre-Post

Note. Darkened cells indicate differences between Hancock et al and Voorheis.

Voorheis (2004) analyzed performance scores and workload ratings for all three trials. The following paragraphs describe the findings for each trial.

Baseline Trial

Findings indicated that participants entered the Context Inducing trial at about the same level of proficiency and perceived workload. No effect of difficulty was found for performance or subjective workload. The interaction of difficulty and order was not significant for either performance scores or workload ratings.

Context Inducing Trial

Performance scores varied with difficulty; a significant main effect of difficulty was seen for performance. A two-way mixed ANOVA (Difficulty x Order) revealed significant differences between all three levels of difficulty ($L < M < H$). The same was seen in workload ratings with a significant main effect of difficulty. A two-way mixed ANOVA (Difficulty x Order) revealed the same three subgroups ($L < M < H$).

Critical Trial

Four analyses were performed on the critical trial performance and workload data:

- Evaluation of performance and workload ratings with a three-way mixed ANOVA (Difficulty x Trial x Order).
- Evaluation of percent change with a two-way mixed ANOVA (Difficulty x Order).
- Evaluation of difference scores with a two-way mixed ANOVA (Difficulty x Order).
- Evaluation of differences from zero with one-sample t-tests for each level of difficulty.

Performance and workload ratings. No significant interaction between Difficulty and Trial for performance scores was found using a three-way mixed ANOVA (Difficulty x Trial x Order). This showed no context effect of difficulty on performance due to the difficulty manipulation, which is consistent with Han-

cock, et al. (1995). Likewise, no significant interaction between Difficulty and Trial was found for workload ratings using a three-way mixed ANOVA (Difficulty x Trial x Order). This shows no context effect of difficulty on workload ratings, which is inconsistent with the results of Hancock et al.

Percent change scores. Hancock, et al. (1995) found workload context effects using percent change scores. Voorheis (2004) analyzed percent change scores with a two-way mixed ANOVA (Difficulty x Order) and found no significant differences as a function of difficulty for either performance scores or workload ratings. *This result is contrary to Hancock et al.*, this shows no context effect of difficulty on subjective workload ratings.

Difference scores. Difference scores were analyzed with a two-way mixed ANOVA (Difficulty x Order) for performance scores and workload ratings. No significant difference was found as a function of difficulty between mean difference scores for either performance or workload. Thus, *a significant context effect was not found.*

Differences from zero. One-sample t-tests were used to detect differences from zero in performance scores and workload ratings. Similar to Hancock, et al. (1995) the only significant difference from zero was seen for workload ratings of the high difficulty condition (Voorheis, 2004).

Recommendations for Continued Research

Voorheis (2004) did not detect context effects in a laboratory setting. Aside from Hancock, et al. (1995), no other manipulation of tasks in this series of studies detected such effects in the measure of subjective workload.

Voorheis (2004) noted that Hancock, et al. (1995) used two tools to measure workload: the NASA-TLX and SWAT. There is no indication how the order of presentation for each tool was controlled, thus participants' reflection of a task and subjective rating in one tool may have affected their ratings with the other.

Voorheis (2004) proposed that a between-subjects study would reduce participant exposure and practice with the medium tracking task. This exposure may have lead participants to compare varying levels of difficulty against the medium level, instead of critically evaluating each level individually. Thus, potential

context effects could be undetectable due to the large number of exposures at the medium difficulty relative to the fewer exposures of other difficulty levels. In Voorheis' study participants were exposed to the medium difficulty condition a total of ten times after a familiarization period. Total exposure included: familiarization with the MATB (at the medium difficulty), one practice session (consisting of three medium difficulty trials), and three experimental sessions (that employed seven medium difficulty trials). This created a 5:1 ratio of medium difficulty trials to trials of all other difficulty levels (ten medium difficulty trials, one high difficulty trial, and one low difficulty trial). This may have contributed to not finding context effects. A between-subject presentation of difficulty *and* use of two Context Inducing Trials could reduce this ratio to 1:1, and is the design which will be followed in this thesis.

Current Study

The current study investigated subjective workload context effects. The results of Hancock, et al. (1995) suggest that context effects can be observed in the laboratory; however, an explanation for the contradictory results found by Moroney, et al. (1993), Fischer (1995), Rench (2002), and Voorheis (2004) has not been isolated. The recommendation of Voorheis to implement a between-subjects design was used to decrease exposure to the medium difficulty condition; additionally, two Context Inducing Trails were used to create an equal ratio between medium difficulty trials and trials of the context condition. Table 6 compares the difficulty ratios of Voorheis and the current study.

Table 6

Comparison of participant exposure to the medium level of difficulty in Voorheis (2004) and Petkosek (2005)

	Voorheis	Petkosek
Trail difficulty by session: L=low, M=med, H=high		
Practice trials plus Familiarization	M-M-M Plus familiarization	M-M Plus familiarization
Session1	M1-L2-H3	M1-L2-L3-M4 M1-M2-M3-M4 M1-H2-H3-M4
Session 2	M1-M2-M3	None
Session 3	M1-H2-M3	None
Total exposure to each difficulty level		
Total L	1	2 ^a
Total M	10	4 (L,H groups) ^b 6 (control group) ^c
Total H	1	2 ^a
Ratio of medium to other difficulties		
M:L,H (practice and experimental)	5:1	2:1
M:L,H (experimental only)	7:2	1:1

^a C2 and C3 denote the *context* of the two Context Inducing Trails, which were given at the same level of difficulty within a session. With the between-subjects design each participant received either low, medium, or high for Context Inducing Trials.

^b Only participants in the experimental groups of low or high difficulty received 6 total trials at the medium level of difficulty.

^c The control group (M1-C2-C3-M4, where C=M) received eight trails at the medium level of difficulty.

Thirty-six participants completed one session of four trials using the compensatory tracking task of the MATB. A repeated measures design presented the first trial (Baseline Trial) and fourth trial (Critical Trial) at the medium difficulty (M1 and M4 respectively). The two middle trials (Context Inducing Trials) were presented at either low, medium, or high difficulty.

The Context Inducing Trials were always presented at the same difficulty within a session (C2 and C3 respectively). RMSE scores and NASA-TLX workload ratings were analyzed for all trials. Table 7 compares the experimental designs of Hancock, et al. (1995) and Voorheis (2004) and the design used in this experiment

Table 7

*Comparison of the experimental designs of Hancock, et al. (1995),
Voorheis (2004), and Petkosek (2005)*

	Hancock et al	Voorheis	Petkosek
Participants	12	12	36
Task type	Tracking	Tracking	Tracking
Exp. sessions	3	3	1
Trials	3	3	3
Trial difficulty (L=low, M=med, H=high; C=Context Inducing in Pet- kosek)	Baseline, Context Induc- ing, Critical	Baseline, Context Induc- ing, Critical	Baseline, Context Induc- ing, Context Inducing, Critical
Trial difficulty by ses- sion	M1-L2-M3 M1-M2-M3 M1-H2-M3	M1-L2-M3 M1-M2-M3 M1-H2-M3	M1-C2-C3-M4
Trial/Session duration	5 minutes/15 minutes	5 minutes/15 minutes	5 minutes/20 minutes
Workload tool	NASA-TLX, SWAT	NASA-TLX	NASA-TLX
Design	Pre-Post within subjects	Pre-Post within subjects	Pre-Post between sub- jects

Note. Darkened cells indicate differences between Voorheis and Petkosek .

Hypotheses

Both performance and workload hypotheses were tested in the current study. The following paragraphs break each hypothesis into numbered components for ease of interpretation in the results section.

Performance

Previous levels of task difficulty will not impact subsequent performance scores. The mean performance scores for the first Context Inducing Trial (C2) of all three groups (low, medium, and high difficulty) will be statistically different, which will indicate an effect of difficulty. Specifically, performance scores will increase for the low difficulty trials, decrease for the high difficulty trials, and no change in performance is expected for the medium condition. Performance scores will not change from Baseline to Critical Trials. This hypothesis is based on the results of Hancock, et al. (1995) and Voorheis (2004) who did not report significant changes in performance scores from Baseline to Critical Trials using a tracking task.

Workload

Based on the findings of Hancock, et al. (1995), it is hypothesized that previous level of task difficulty will impact subsequent workload ratings inversely. Specifically:

The mean workload ratings for the first Context Inducing Trial (C2) of all three groups (low, medium, and high difficulty) will be statistically different. This will indicate the effect of difficulty.

1. The mean workload ratings for the first Context Inducing Trial (C2) of all three groups (low, medium, and high difficulty) will be statistically different. This will indicate the effect of difficulty.
2. When participants experience the low difficulty condition after the medium difficulty condition, followed by a return to medium, workload ratings will increase from the Baseline (M1) to the Critical Trial (M4).
3. Conversely, when participants experience the high difficulty condition after the medium difficulty condition, followed by a return to medium, workload ratings will decrease from the Baseline (M1) to the Critical Trial (M4).
4. No change in workload ratings is expected for the control condition of medium difficulty, where participants will not experience a difficulty manipulation.

CHAPTER II

METHOD

Participants

Thirty-six undergraduate students (twenty-six male, ten female) from the University of Dayton participated in this study. Participants took part in the study to obtain research credits in Psychology 101. The number of participants used (twelve) within each of the three levels of difficulty was consistent with the within-subjects designs used by Hancock, et al. (1995) and Voorheis (2004).

Software and Apparatus

Version 4.0 of the MATB was used to create the compensatory tracking task. The software was installed on a Windows-based desktop computer using a Pentium 1.80 GHz processor, 512MB of Ram, and a Samsung Syncmaster 17-inch monitor. Only the tracking task was visible to participants, all other MATB tasks were occluded by a piece of heavy-stock paper.

Participants used a mouse to center the moving circle, 1.0 cm in diameter, on a cross, 0.5 cm x 0.5 cm, in the center of the tracking window. The three levels of difficulty were created by combining two manual tracking difficulty settings and three gain settings. The same difficulty levels were used by Voorheis (2004). The following manipulations were used to create the levels of difficulty:

- Low difficulty trial: tracking = low, gain = 25
- Medium difficulty trial: tracking = high, gain = 15
- High difficulty trial: tracking = high, gain = 1

Participants were instructed to keep the circle as close to the center cross-hair as possible (Appendix A). The experiment took place in an isolated room in St. Joseph Hall at the University of Dayton. Participants were seated at a five-foot

long table directly in front of the monitor. A mouse pad with a wrist support was provided for each participant.

Experimental Design

Participants took part in a practice session and an experimental session. During the practice session, each participant briefly manipulated the cursor at the medium level of difficulty to become familiar with the task. Each participant then completed a practice session of two trials at the medium difficulty level. During the experimental session, each participant completed four trials: Baseline, Context, Context, and Critical. The Baseline and Critical Trials were always presented at the medium level of difficulty; experimental manipulation occurred during the two Context Inducing Trials, which were both presented at one of the three levels of difficulty (low, medium, or high). This design was used to determine if the Context Inducing Trials impacted the performance and subjective workload of the Critical trial. The assignment of participants to a difficulty condition was determined by the order in which they signed-up for the experiment.

Independent Variable

The independent variable was the level of task difficulty (manipulated during the Context Inducing Trials (C2 and C3)). This variable had three levels: low, medium, and high- and was presented between-subjects.

Dependent Variables

The two dependent variables in this study were performance and subjective workload.

RMSE was used to measure performance. It was collected at five-second intervals and averaged. RMSE was computed by the following formula: $RMSE = \sqrt{1/N \sum (Y_i - \hat{Y}_i)^2}$. Where Y_i is the optimal value (in this case the optimal value was zero pixel units away from the center of the target) and \hat{Y}_i is the recorded value (the actual number of pixel units that the aircraft icon was away from the target).

A software version of the NASA-TLX was administered immediately after each trial to measure subjective workload. This software version is essentially the

same as the paper-and-pencil version provided in Appendix B. Participants had an unlimited amount of time to complete each TLX. The TLX score was weighted in calculating the overall index of workload to replicate Hancock, et al. (1995) and Voorheis (2004).

Procedure

The experiment required two sessions, one practice and one experimental. The practice session consisted of four components: welcome and consent, MATB familiarization, NASA-TLX familiarization, and a practice trial of the experimental task. When a participant arrived for the experiment he or she was verbally briefed on the nature of the task to be performed and the NASA-TLX. This was done so that the participant gained an understanding of the tasks. The informed consent (Appendix C) was then read and signed by the participant if he or she wished to take part. The participant was then familiarized with the MATB. Each participant read the MATB instructions, was provided the opportunity to ask questions about the MATB, and then practiced the tracking task for approximately one minute. The participant was then familiarized with the NASA-TLX. Each participant read the TLX instructions, was provided the opportunity to ask questions about it, and then jointly reviewed the subscales with the experimenter. Questions were again fielded after the subscale explanations. The participant then completed a practice session of two medium difficulty trials. This was done to ensure that participants met the performance criterion of $RMSE < 16.0$ at the medium level of difficulty (Voorheis, 2004). If a participant did not reach the criterion level he or she was excused from the experiment but received research credit for Psychology 101 (no data was collected for these participants). The NASA-TLX was administered after each trial. A five-minute break was taken before beginning the experimental session.

The experimental session consisted of four trials. Participants took part in one of three difficulty sessions (M1-L2-L3-M4, M1-M2-M3-M4, or M1-H2-H3-M4). There were twelve participants in each session. The NASA-TLX was ad-

ministered after each trial. These procedures are detailed in the experimenter's checklist (Appendix D).

CHAPTER III

RESULTS

Performance scores and workload ratings were analyzed for the Baseline Trial (M1), each Context Trial (C2 and C3), and the Critical Trial (M4). Performance scores and workload ratings from the Baseline Trial were then compared to the Critical Trial. Finally, the NASA-TLX subscales were analyzed.

Baseline Trial (M1)

The first analysis tested for statistical differences among performance scores and workload ratings in the Baseline Trial (M1). This was conducted to confirm that participants entered the study with comparable performance scores and workload ratings before the introduction of a context.

Performance Scores

A one-way ANOVA was conducted on performance scores. As expected, there was no effect of difficulty: $F(2,33) = 1.21$, $MSE = 4.04$, $p = .310$. Figure 1 shows group means. The means and standard deviations are reported in Table E1 (Appendix E).

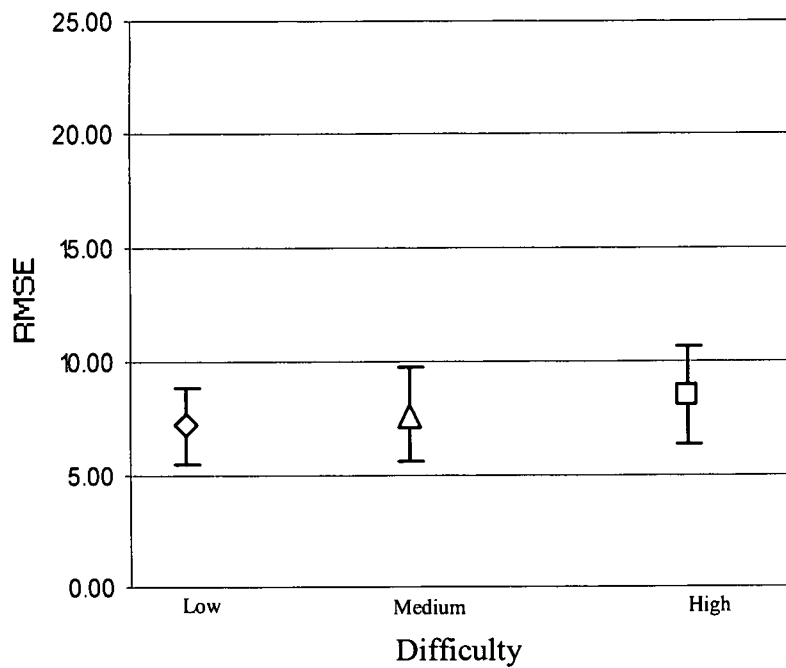


Figure 1. Mean performance scores as a function of difficulty for the Baseline Trial (M1). Bars show standard deviations.

Workload Ratings

A one-way ANOVA was conducted on workload ratings. As expected, there was no effect of difficulty: $F(2,33) = 1.46$, $MSE = 292.44$, $p = .248$. Figure 2 shows group means. The means and standard deviations are reported in Table E2 (Appendix E).

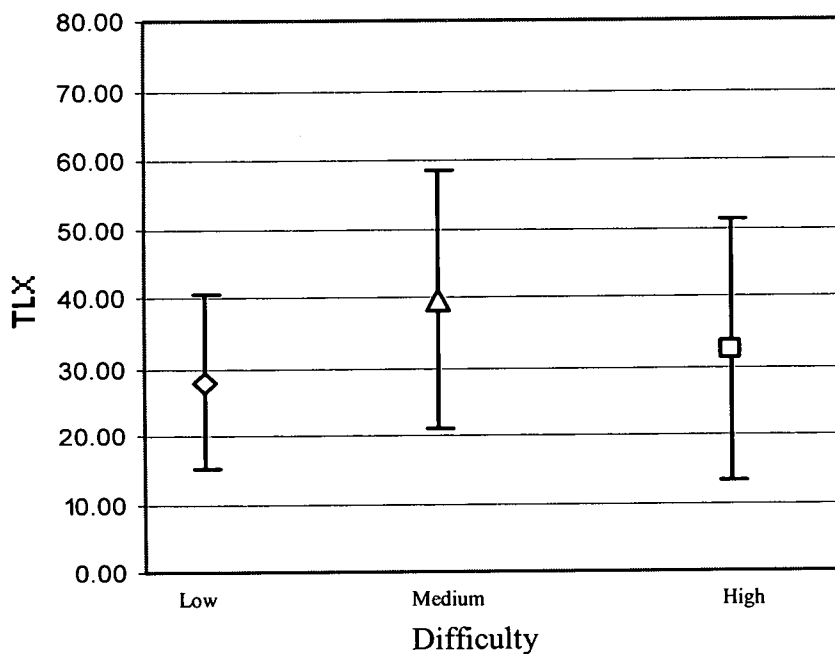


Figure 2. Mean workload ratings as a function of difficulty for the Baseline Trial (M1). Bars show standard deviation.

There was no effect of difficulty on performance scores or workload ratings in the Baseline Trial (M1). Therefore, a common baseline existed from which to compare difficulty (context conditions) manipulations.

Context Inducing Trial (C2)

The second analysis tested for statistical differences among performance scores and workload ratings in the First Context Inducing Trial (C2). This was conducted to confirm that the difficulty manipulation resulted in performance score and workload rating differences.

Performance Scores

A one-way ANOVA was conducted on performance scores. As expected, there was an effect of difficulty: $F(2,33) = 42.57$, $MSE = 10.18$, $p < .001$. Figure 3 shows the group means. The means and standard deviations are reported in Table E3 (Appendix E).

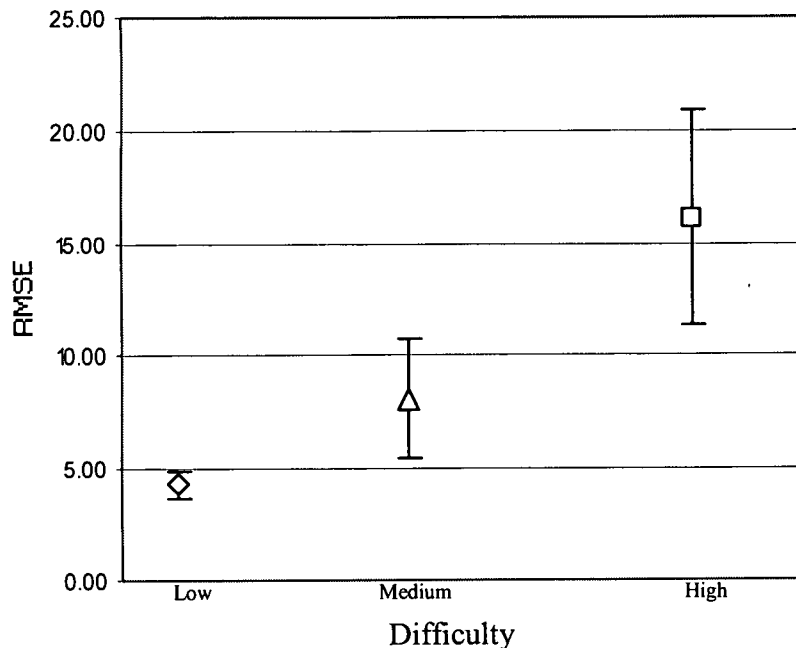


Figure 3. Mean performance scores as a function of difficulty for the First Context Inducing Trial (C2). Bars show standard deviations.

A Tukey post hoc analysis indicated three statistically different subgroups (L,M,H; low was different from medium: $p = .016$, medium was different from high: $p < .001$, and low was different from high: $p < .001$). This confirms the performance hypothesis: the mean performance scores for the three difficulty groups (low, medium, and high) will be statistically different for the first Context Inducing Trial (C2). Difficulty group performance scores are depicted in Figure 3.

Workload Ratings

A one-way ANOVA was conducted on workload ratings. As expected, there was an effect of difficulty: $F(2,33) = 7.94$, $MSE = 403.36$, $p = .002$. Figure 4 shows the group means. The means and standard deviations are reported in Table E4 (Appendix E).

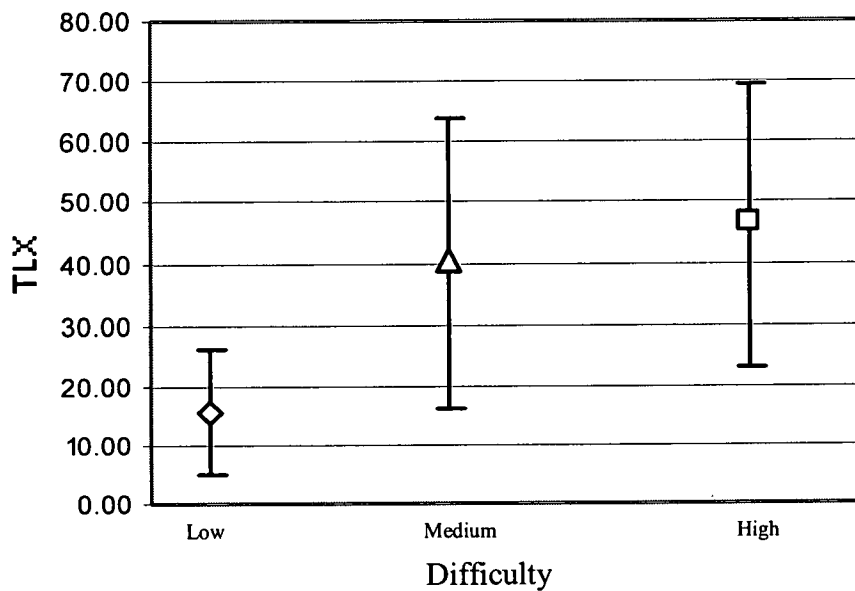


Figure 4. Mean workload ratings as a function of difficulty for the First Context Inducing Trial (C2). Bars show standard deviations.

A Tukey post hoc analysis indicated two statistically different subgroups (L,M/H) where workload ratings obtained under the low difficulty level were different from those obtained under both the medium ($p = .014$) and high ($p = .002$) difficulty levels. However, ratings obtained under the medium difficulty level were not different from high ($p = .722$). Thus the first aspect of the workload hypothesis was not confirmed. In that the mean workload ratings for all three difficulty groups (low, medium, and high) were not statistically different from each other for the first Context Inducing Trial (C2).

The difficulty manipulation produced a Main Effect in performance scores and workload ratings. Post hoc tests indicated statistically different subgroups. The performance hypothesis can be examined with three statistically different groups of performance scores (low, medium, and high). Only two statistically different subgroups were observed for the workload hypothesis (low and medium/high).

Context Inducing Trial (C3)

The third analysis tested for statistical differences among performance scores and workload ratings in the Second Context Inducing Trial (C3). This was conducted to confirm that effects of the difficulty manipulation persisted. This persistence would indicate the absence of learning effects and ensure that two Context Trial Inducing Trials were being presented.

Performance Scores

A one-way ANOVA was conducted on performance scores. As expected, there was an effect of difficulty: $F(2,33) = 23.54$, $MSE = 16.17$, $p < .001$. Figure 5 shows the group means. The means and standard deviations are reported in Table E5 (Appendix E).

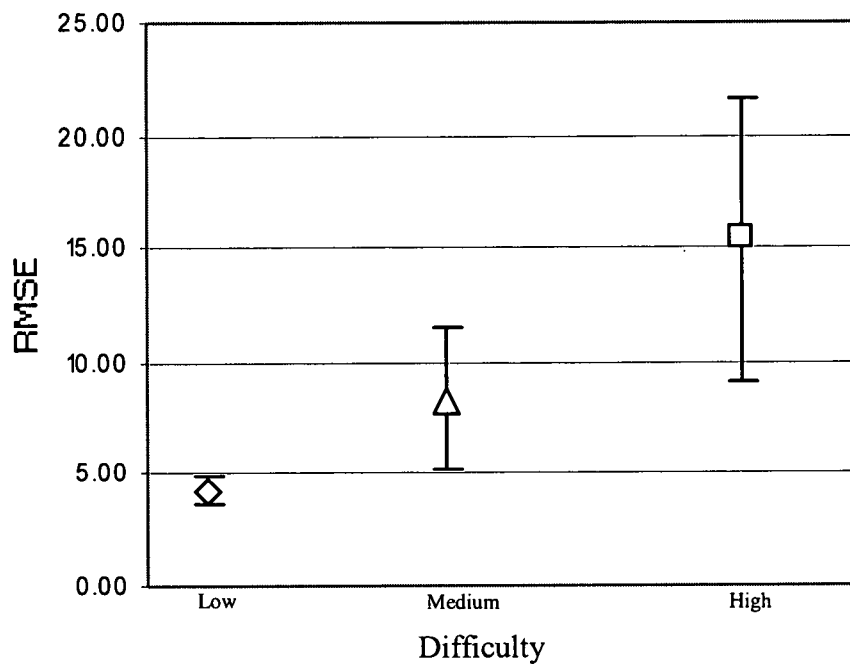


Figure 5. Mean performance scores as a function of difficulty for the Second Context Inducing Trial (C3). Bars show standard deviations.

A Tukey post hoc analysis indicated three statistically different subgroups (L,M,H; low was different from medium: $p = .047$, medium was different from high: $p < .001$), and low was different from high: $p < .001$). Difficulty group performance scores are depicted in Figure 3.

Workload Ratings

A one-way ANOVA was conducted on workload ratings. As expected, there was an effect of difficulty: $F(2,33) = 6.35$, $MSE = 410.05$, $p = .005$. Figure 6 shows the group means. The means and standard deviations are reported in Table E6 (Appendix E).

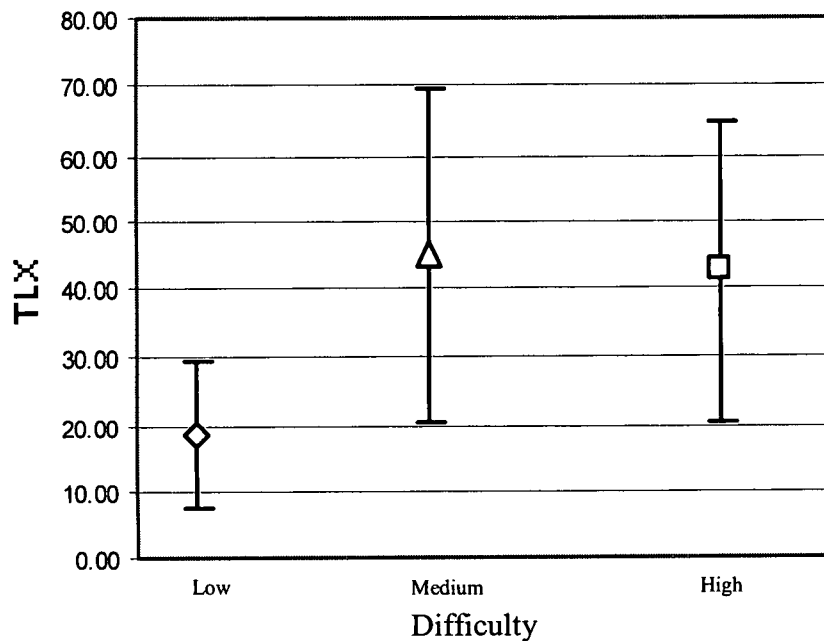


Figure 6. Mean workload ratings as a function of difficulty for the Second Context Inducing Trial (C3). Bars show standard deviations.

A Tukey post hoc analysis of the workload ratings indicated two statistically different subgroups (L,M/H) where ratings obtained under the low difficulty conditions were significantly different from those obtained under the from medium ($p = .008$) and high ($p = .017$) difficulty conditions. Ratings obtained under the medium difficulty condition were not significantly different from those obtained under the high difficulty condition ($p = .944$). Difficulty group workload ratings are depicted in Figure 6.

The difficulty manipulation produced a Main Effect in performance scores and workload ratings in the second Context Inducing Trial. Post hoc tests indicated three statistically different subgroups for performance scores and two statistically different subgroups for workload ratings. These results are consistent with the findings of the first Context Inducing Trial (C2).

Critical Trial (M4)

Two analyses were conducted on the Critical Trial (M4). The first analysis examined difficulty group differences (low, medium, and high); context effects would be indicated by statistically different groups. The groups that received the low or medium difficulty on the Context Inducing Trial are expected to differ from each other on subjective workload ratings; the group that received the medium difficulty on the Context Inducing Trial is expected to differ from both the low and high difficulty groups on subjective workload rating. No difference is expected for performance. The second analysis evaluated the change in performance scores and workload ratings from Baseline to Critical Trials. A significant change from the medium difficulty Baseline Trial to the medium difficulty Critical Trial would indicate the presence of a context effect. Significant changes are expected in both the low and high difficulty groups; no change is expected in the medium difficulty group.

First Analysis

Performance Scores

A one-way ANOVA was conducted on performance scores. As hypothesized, there was no effect of difficulty: $F(2,33) = .07$, $MSE = 7.84$, $p = .930$. Previous levels of task difficulty did not impact subsequent performance scores. Figure 7 shows the group means. The means and standard deviations are reported in Table E7 (Appendix E).

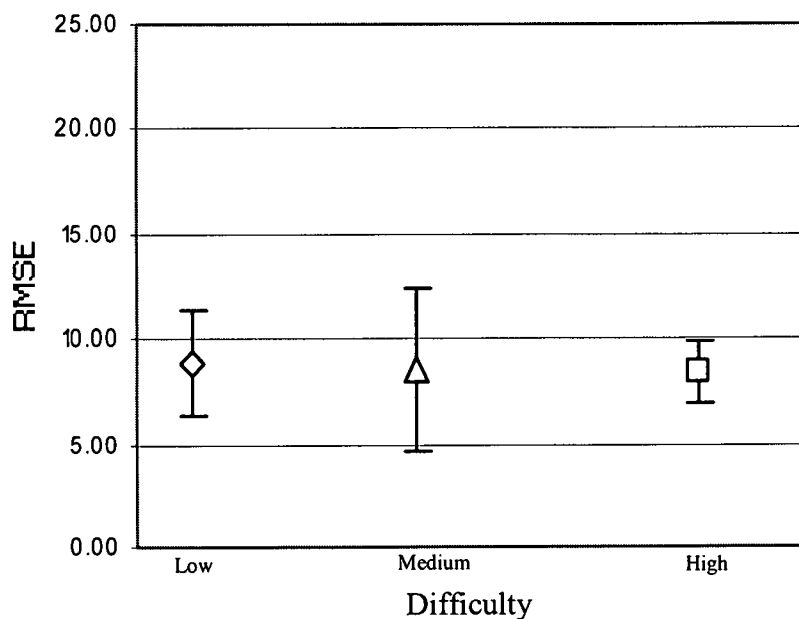


Figure 7. Mean performance scores as a function of difficulty for the Critical Trial (M4). Bars show standard deviations.

Workload Ratings

A one-way ANOVA was conducted on workload ratings. Contrary to the workload hypothesis, there was no effect of difficulty: $F(2,33) = 59$, $MSE = 437.89$, $p = .559$. Previous levels of task difficulty did not impact subsequent workload ratings. Figure 8 shows the group means. The means and standard deviations are reported in Table E8 (Appendix E).

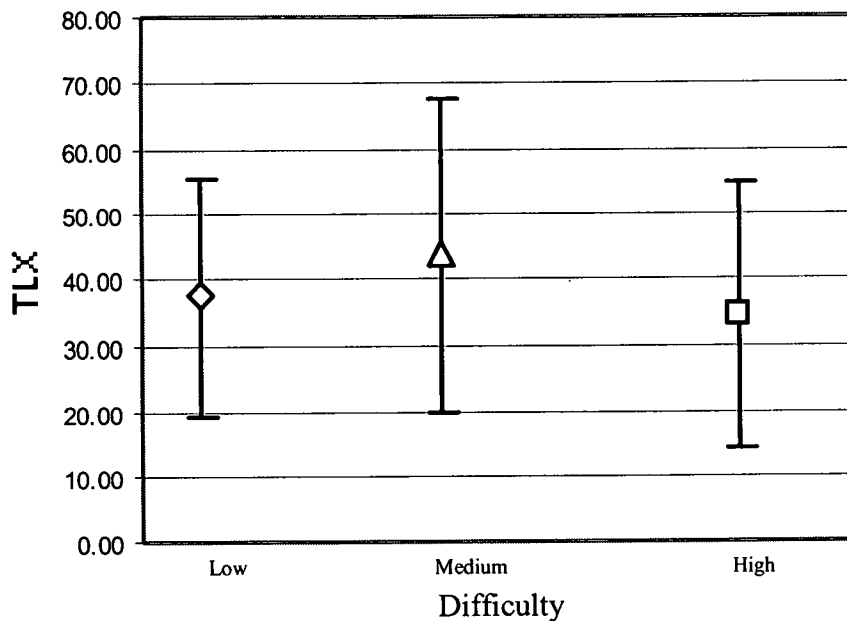


Figure 8. Mean workload ratings as a function of difficulty for the Critical Trial (M4). Bars show standard deviations.

Context effects were not observed for performance scores or workload ratings by comparing group means. This confirms the performance hypothesis that previous levels of task difficulty will not impact subsequent performance scores, but does not confirm the workload hypothesis that previous levels of task difficulty will impact subsequent workload ratings.

Second Analysis

Performance Scores

A 3x2 ANOVA (Difficulty x Trial) for performance scores was conducted to determine if the change from Baseline (M1) to Critical (M4) Trial was at all different for the three context inducing conditions. There was no significant interaction effect: $F(2,33) = 2.14$, $MSE = 4.20$, $p = .134$. Thus, the difficulty of the Context Inducing Trials (C2 and C3), given after the medium difficulty Baseline Trial (M1), did not impact performance on the Critical Trial (M4). This analysis also indicated that there was no effect of difficulty: $F(2,33) = .13$, MSE

= 9.92, $p = .008$; however, there was a significant effect of trial: $F(2,33) = .6.1$, $MSE = 2.0$, $p = .019$.

Workload Ratings

A 3x2 ANOVA (Difficulty x Trial) of workload ratings was conducted to determine if the change from Baseline (M1) to Critical (M4) Trial was at all different for the three context inducing conditions. Thus, there was no significant interaction effect: $F(2,33) = 1.23$, $MSE = 66.21$, $p = .305$. The difficulty of the Context Inducing Trials (C2 and C3), given after the medium difficulty Baseline Trial (M1), did not impact workload ratings on the Critical Trial (M4). This analysis also indicated that there was no effect of difficulty: $F(2,33) = .909$, $MSE = 664.12$, $p = .413$. However, there was a significant effect of Trial: $F(2,33) = 7.73$, $MSE = 66.21$, $p = .009$. That is, subjective workload ratings varied between Baseline (M1) and Critical (M4) trials.

As an alternate to the ANOVA, an analysis of covariance was conducted on the Critical Trial (M4) workload ratings. The Baseline Trial (M1) was treated as the covariate to correct for initial ratings. This indicated that difficulty did not affect participant workload ratings in the Critical Trial (M4): $F(2,32) = 1.21$, $MSE = 136.41$, $p = .312$.

Subscale Analyses

Individual TLX subscales for all three levels of difficulty in the Baseline (M1) and Critical (M4) Trials were examined to determine if a change in workload had been masked by previous analyses of combined workload score. The six subscales are: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). Figure 9 provides a comparison of the TLX subscales means obtained from Trials M1 and M4. In all cases the M4 workload ratings were higher. The means and standard deviations are reported in Table E9 (Appendix E).

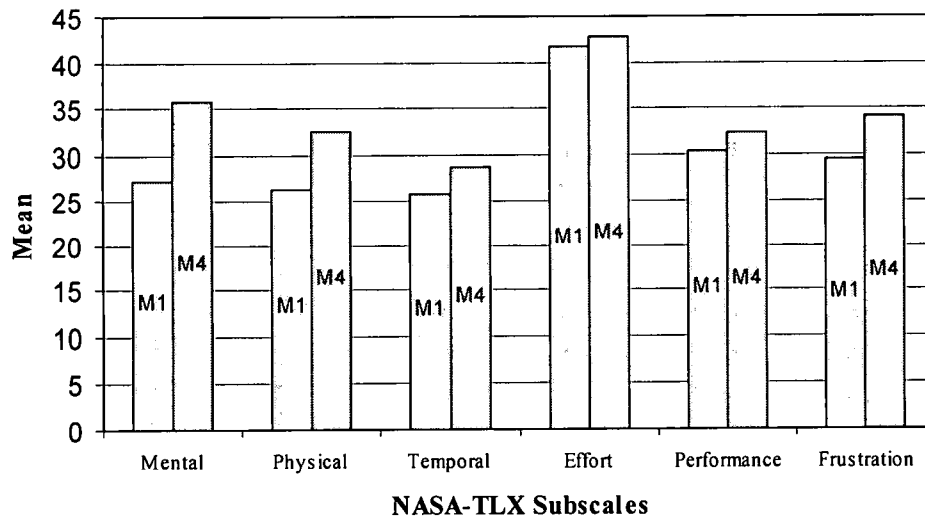


Figure 9. Comparison of the subjective workload ratings obtained for each TLX subscale in the Baseline (M1) and Critical (M4) Trials.

Six 3x2 ANOVAs (Difficulty x Trial) for the NASA-TLX subscales were conducted. This was done to avoid adding a six-level variable of scale (6x3x2 ANOVA) that may have an impact on individual interaction effects. Within each of the six analyses the three effects of Trial, Difficulty, and Trial x Difficulty were evaluated without paying an error rate penalty. Thus, the alpha family wise level for each test was .15 (3 tests x .05). The Modified Bonferroni approach corrected for significance, the alpha per comparison was .025 (.15 alpha family wise / 6 tests). No interaction effect of Difficulty x Trial was significant for any of the six analyses. The interaction effect of difficulty and trial for the effort subscale was the closest to reaching significance: $p = .039$. The interaction effect for the other subscales falls off in this order: frustration $p = .063$, performance $p = .206$, physical demand $p = .698$, temporal demand $p = .789$, and mental demand $p = .901$. Apparently the subscale ratings did not differ significantly between the Baseline (M1) and Critical (M4) Trials. The means and standard deviations are reviewed in Table E9.

CHAPTER IV

DISCUSSION

This study investigated subjective workload context effects. Specifically, it examined the effect of task difficulty on the performance and perceived workload of a subsequent task. Hancock, et al. (1995) demonstrated that when participants were given a low difficulty task on a Context Inducing Trial, their perception of workload on the Critical Trial was higher compared to the Baseline Trial. Inversely, workload ratings after a high difficulty Context Inducing Trial were lower on the Critical Trial compared to the Baseline Trial. These results would suggest that context effects can be observed in the laboratory; however, similar studies by Moroney, et al. (1993), Fischer (1995), Rench (2002), and Voorheis (2004), did not produce similar workload context effects. This study utilized the recommendation of Voorheis (2004) for a between-subjects presentation of difficulty to examine context effects.

Voorheis (2004) noted that his participants had more exposure to the medium difficulty tracking task, compared to the low and high difficulty tasks. The between-subjects design used in the current study reduced the total number of trials, thus exposure to the medium difficulty was reduced from Voorheis. Additionally, two Context Inducing Trials were used in the current study to increase participant exposure to other levels of difficulty. These changes resulted in a 1:1 ratio of medium and the other difficulties (see Table 6). The following sections describe the results of this study and commences by examining the effect of the context manipulation on performance scores and then workload ratings.

Performance

Performance was measured by RMSE score. Group means varied according to difficulty in the Context Inducing Trials (C2 and C3). As expected, per-

formance scores of the Context Inducing Trials (C2 and C3) increased with the high difficulty condition, decreased with the low difficulty condition, and remained relatively constant across the medium difficulty control condition. Performance group means for each difficulty level across all trials are provided in Figure 10.

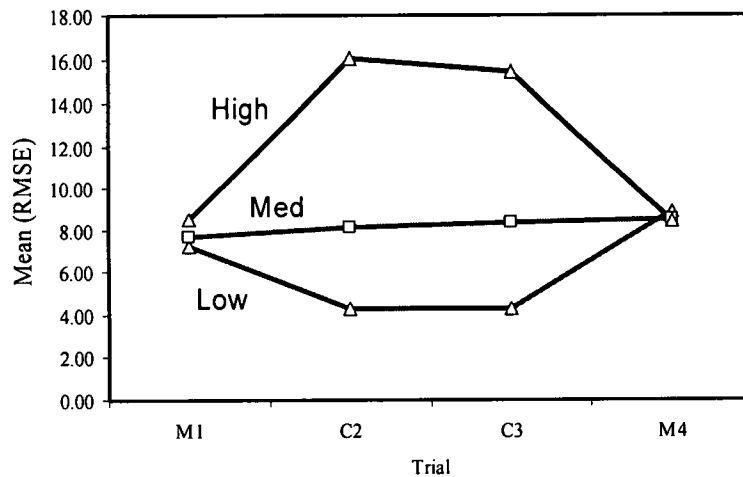


Figure 10. Mean performance for each difficulty group across all trials.

Analyses were conducted on all four trials in order to: confirm that the initial level of performance between the groups was not significantly different, determine if the difficulty manipulation had an effect between groups, and to confirm that participant performance ended at a point of equal comparison. The analysis of the Baseline Trial (M1) indicated that participants began with comparable scores. This was expected because all three difficulty groups began with a medium difficulty trial (M1). The analysis of the Context Inducing Trials (C2 and C3) indicated an effect of difficulty. Post hoc tests showed that low difficulty significantly differed from medium difficulty, which significantly differed from high difficulty. This met the performance hypothesis that mean performance scores would vary between groups and result in three distinct groups. The analysis on the Critical Trial (M4) indicated, as hypothesized, that participants had comparable performance scores for all three difficulty groups.

Performance context effects were analyzed by comparing the Baseline Trial (M1) to the Critical Trial (M4). There was no interaction of difficulty (low, medium, or high) and trial (Baseline and Critical), which is shown in Figure 10 and the analyses presented in Chapter 3; thus performance in the Critical Trial (M4) was not affected by the difficulty manipulation. Similar to Hancock, et al. (1995) and Voorheis (2004) there was no context effect for performance. These results support the performance hypothesis: the performance on the difficulty manipulation differed for all three groups but did not impact subsequent performance scores. There was no context effect for performance.

Workload

Subjective workload was measured by the NASA-TLX procedure. Group means varied with difficulty in the Context Inducing Trials (C2 and C3). As expected, workload scores increased for high difficulty trials, decreased for low difficulty trials, and remained relatively constant across medium difficulty trials. The workload means for each difficulty level is shown in Figure 11, and with the exception of scores obtained for the medium level of difficulty, appears as it would be if context effects were absent.

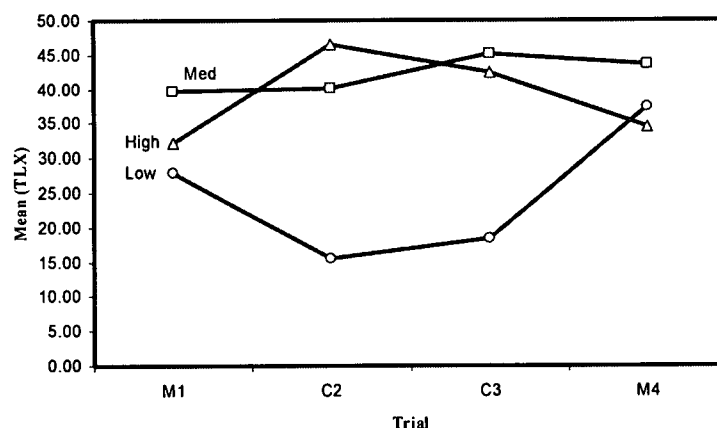


Figure 11. Mean workload trends for each difficulty group across all trials.

Analyses were conducted on all four trials to: confirm that participants began with comparable workload ratings, determine if the difficulty manipulation had an effect between groups, and confirm that participants concluded with significantly different group means. The analysis of the Baseline Trial (M1) indicated that participants began with comparable ratings. This was expected because all three difficulty groups experienced a medium difficulty trial. The analyses of the Context Inducing Trials (C2 and C3) indicated an effect of difficulty. Post hoc tests showed that workload ratings obtained for the low difficulty trials were significantly lower than ratings obtained for the medium difficulty trials, which did not differ significantly from ratings obtained for the high difficulty trials. This unexpected finding does not confirm the first point of the workload hypothesis, which stated that workload ratings would vary between groups and result in three distinct groups (only two were observed: L:M/H). The analysis workload ratings for the Critical Trial (M4) indicated the ratings were not significantly different. This was not expected because the hypothesized workload context effects would have resulted in different group means in the Critical Trial (M4).

Workload context effects were analyzed by comparing the Baseline Trial (M1) to the Critical Trial (M4). There was no interaction of difficulty (low, medium, or high) and trial (Baseline and Critical), thus workload in the Critical Trial (M4) was not affected by the difficulty manipulation. Similar to Voorheis (2004), this analysis showed no context effect for subjective workload. These results fail to confirm the workload hypothesis that workload ratings on the difficulty manipulation will differ for all three groups and will result in a context effect. Thus, there were no context effects for subjective workload.

Despite the absence of a significant context effect, the means of the low difficulty group follow the expected trend: mean workload ratings increase in the Critical Trial (M4) compared to the Baseline Trial (M1). This can be seen in Figure 11. The low difficulty group's mean for trial M1 (medium difficulty level) was 27.9; as expected, the means decrease during the Context Inducing Trials (low difficulty level) to 15.5 and 18.4 (C2 and C3 respectively). The group mean for trial M4 is 37.4, an increase of approximately ten units from the Baseline Trial

(M1), which is a 27% increase in workload ratings in the Critical Trial (M4) over the Baseline Trial (M1). This suggests the potential for a context effect.

The means of the medium difficulty control group remain relatively constant across all trials (39.7, 40.1, 45.2, and 43.7 for trials M1, C2, C3, and M4 respectively). The five unit increase from trial C2 (40.1) to C3 (45.2) is somewhat surprising. Given that all trials in the medium difficulty group were given at the medium level of difficulty (C2 and C3 were not true Context Inducing Trials), it was expected that practice effects would help to maintain, or even reduce, workload ratings. It appears, rather, that fatigue may have resulted in the perception of increased workload- as indicated by the increase in workload rating.

The group means of the high difficulty group do not indicate a trend towards a context effect as the low difficulty group does. Critical Trial (M4) group mean of 34.6 does not drop below the Baseline Trial (M1) mean of 32.1, which would be a prerequisite for a context effect.

The trend of the group means of the low, medium, and high difficulty groups, when taken together, support the possibility of a weak context effect, mitigated by fatigue. The trend in means of the low difficulty group is in the direction required for a context effect. It is possible that, within the low difficulty group, boredom-induced fatigue contributed to workload ratings that were higher than expected in the Critical Trial (M4). The slight increase in workload demonstrated by the medium difficulty control group may also be attributed to increased boredom. Since workload ratings on M4 for the high difficulty group did not fall below their M1 baseline ratings, a trend toward a context effect was not noted. However, the appearance of this trend may have been inhibited by the presence of a learning effect, suggested in Figure 10. The learning effect combined with task-induced fatigue, which one would expect to be greatest for the high difficulty task, may have lead participants to report a slightly higher level of workload than they had on their Baseline (M1) trial.

Readers are reminded that while the argument presented above posits the possible existence of a weak context effect, the results indicate that such an effect

did not attain the required level of significance. Therefore, the above argument is presented as conjecture.

The 3x2 ANOVA used to evaluate context effects by examining the change from Baseline (M1) to Critical Trials (M4) indicated a significant effect of Trial. This finding indicates that workload ratings varied with trial. The variance between NASA-TLX subscale results from Baseline (M1) and Critical Trial (M4) was analyzed to determine if workload components measured by individual scales varied with trial. A significant variation among subscales between Baseline and Critical Trials might reveal that a particular subscale was sensitive to the context effect.

TLX Subscales

In the NASA-TLX procedure a user assigns a value to six different scales that are subsequently weighted to produce a combined score. The six subscales are mental demand, physical demand, temporal demand, performance, effort, and frustration. Descriptions of these subscales, developed by Hart and Staveland (1988), are provided in Appendix B.

The NASA-TLX subscales were also evaluated to determine if changes occurred in a particular subscale after being exposed to the Context Inducing Trials (C2 and C3). Analyses were conducted to evaluate the interaction effect of Difficulty x Trial from Baseline (M1) to Critical Trials (M4). None of the effects reached the required significance level (.025), which had been adjusted to account for a family wise error rate penalty. The interaction effect (Difficulty x Trial) for the effort subscale came the closest to reaching significance ($p = .039$). The interaction effect for the rest of the subscales appeared in this order: frustration ($p = .063$), performance ($p = .206$), physical demand ($p = .698$), temporal demand ($p = .789$), and mental demand ($p = .901$). It was concluded that the Context Inducing Trials (C2 and C3) did not induce a context effect on any of the subscales.

Conclusion

The results of this study did not demonstrate the presence of subjective workload context effects in a laboratory setting. It was hypothesized that three

unique groups of workload ratings would emerge in the difficulty manipulation (low, medium, and high difficulty groups) and would result in context effects in the Critical Trial (M4). However, only two unique difficulty groups were observed (low and medium/high). Workload context effects were not observed in either of two analyses: 1) comparing difficulty levels in the Critical Trial (M4), which showed no effect of difficulty and therefore no workload context effects, or 2) evaluating the change in workload ratings from Baseline (M1) to Critical Trial (M4), which showed no significant changes and therefore no workload context effects.

Performance results were as expected: the difficulty manipulation resulted in three unique groups of performance scores (low, medium, and high difficulty). However, there were no performance induced context effects in the Critical Trial (M4).

These findings support prior research: Moroney, et al. (1993), Fischer (1995), Rensch (2002), and Voorheis (2004) where workload context effects were not observed in the laboratory. However, Hancock, et al. (1995) identified such context effects. The following section offers a possible explanation for the workload findings of Hancock et al. and the elusiveness of such results.

Additional Consideration

Hancock, et al. (1995) used multiple measures of subjective workload (NASA-TLX and SWAT). The other studies reported (Moroney, et al., 1993; Fischer, 1995; Rensch, 2002; and Voorheis, 2004), as well as the current study, only used the NASA-TLX. Because it is not clear how Hancock et al. controlled the administration sequence of these two measures, several possible explanations for the contradictory results exist:

- The use of two subjective workload measures in the Hancock et al. study may have affected participants' understanding of the NASA-TLX criteria, specifically SWAT instructions may influenced the values of the TLX subscales.
- Priming or context effects may have been present because of SWAT administration, specifically participants reflected on their perception of

workload for longer periods of time because of the use of two measures.

- Demand characteristics may have been present, specifically the repetition of workload metrics may have lead participants to expect that change in workload ratings were expected by the experimenter.

It is therefore possible that the use of two measures influenced the results of the Hancock et al. experiment; this is supported by the inability of multiple studies to reproduced the findings of Hancock et al. with one measure of subjective workload. However, additional work employing two measure of subjective workload is required in order to directly evaluate such an hypothesis.

REFERENCES

- Byers, J. C., Bittner, A. C., & Hill, S. G. (1998). Traditional and raw Task Load Index (TLX) correlations: Are paired comparisons necessary? In A. Mital (Ed.), *Advances in industrial ergonomics and safety* (Vol. 1, pp. 481-485). London: Taylor & Francis.
- Comstock, J. R. & Arnegard, R. J. (1992). The Multi-Attribute Task Battery for Human Operator Workload and Strategic Behavior Research. (Tech Memorandum 104174). Hampton, VA: NASA Langley Research Center.
- Fischer, D. S. (1995). *The effect of previous levels of workload in a simulated flight task*. Master's Thesis, University of Dayton, Dayton, Ohio.
- Hancock, P. A., Williams, G., Manning, C. M., & Miyake, S. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, 5(1), 63-86.
- Hart, S. & Staveland, L. (1988). Development of NASA-TLX: Results of empirical and theoretical research. In P Hancock and N. Meshkati (Eds.), *Human mental workload*. Amsterdam: North-Holland.
- Huey, B.M. and Wickens, C.D., (Eds.). Panel on Workload Transition, Committee on Human Factors, Commission on Behavioral and Social Sciences and Education, National Research Council. (1993). *Workload Transition: Implications for Individual and Team Performance*. Washington, DC: National Academy Press

- Moroney, W. F., Reising, J., Biers, D. W., & 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* (p. 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.
- Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload, *Human Factors*, 33, 17-34.
- O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance*: Vol. 2. Cognitive process and performance. New York: Wiley Interscience.
- Reid, G.B., & Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. North-Holland: Elsevier Science Publishers.
- Rench M. E. (2002). *The impact of pervious levels of task difficulty on a flight simulation task*. Master's Thesis, University of Dayton, Dayton, Ohio.

Voorheis C. M. (2004). *Context effects on subjective workload assessments using the multi attribute task battery tracking task*. Master's Thesis, University of Dayton, Dayton, Ohio.

Wickens, C. & Hollands, J. G (2000). *Engineering Psychology and Human Performance*. Upper Saddle River: Prentice Hall.

APPENDIX A. MATB TRACKING INSTRUCTIONS

MATB Instructions

Performance on the MATB:

The MATB is a computer simulation of some of the kind of tasks that pilots perform (communication, tracking, systems monitoring, and resource management). For the purposes of this study, we will only focus on the tracking task.

Look at the TRACKING window. While in flight an operator has to keep the plane on a certain course. This TRACKING task will require that you keep the tracking circle as close as possible to the crosshair in the center of the box throughout the experiment. Moving the mouse around does this. If you do not move the mouse and do not make constant adjustments, the tracking cursor will “drift,” simulating movement off course. Try to stay as close to the target center as possible.

General comments about this task:

This task may seem very difficult and even overwhelming at times. It is not expected that you will be able to perform it perfectly or do very well the first time. But please do try to learn it and do the best you can.

Adapted From

Comstock, J. R. & Arnegard, R. J. (1992). The Multi-Attribute Task Battery for Human Operator Workload and Strategic Behavior Research. (Tech Memorandum 104174). Hampton VA: NASA Langley Research Center.

Good luck!

General note about this task:

This task assumes that you have used a computer mouse before and are comfortable with operating one. If this is not true, please inform the experimenter now. This task also assumes that you are accustomed to using a mouse with your right hand. If this is not the case, please inform the experimenter now.

Any Questions?

APPENDIX B. TLX MATERIALS

NASA-TLX Instructions and Questionnaire

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 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 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 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 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 task, you will receive a form with six rating scales. You will evaluate the task by marking each scale at the point that matches your experience. Each line has two endpoint descriptors that describe the scale. Note that "performance" goes from "good" on the left to "poor" on the right. This order has been confusing for some people. Mark the desired location. Please consider your responses carefully in distinguishing among the 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!

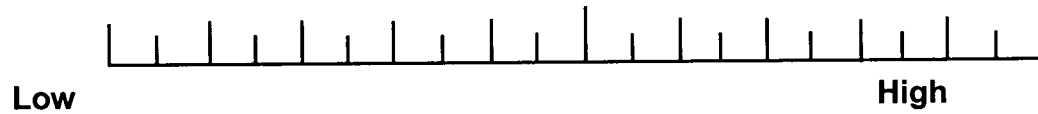
NASA Task Load Index (TLX) Paper and Pencil Package Version 1.0 (1986). Moffett
Field, CA: Human Performance Research Group, NASA Ames Research Center.

NASA TLX RATING SCALE DEFINITIONS

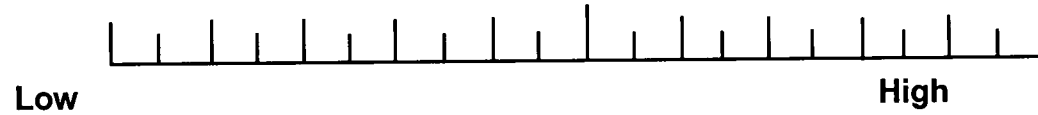
Title	Endpoints	Descriptions
MENTAL DEMAND	Low/High	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	Low/High	How much physical 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?
EFFORT	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
PERFORMANCE	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
FRUSTRATION LEVEL	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

NASA-TLX Rating Scales

MENTAL DEMAND



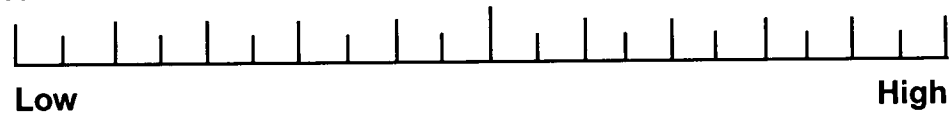
PHYSICAL DEMAND



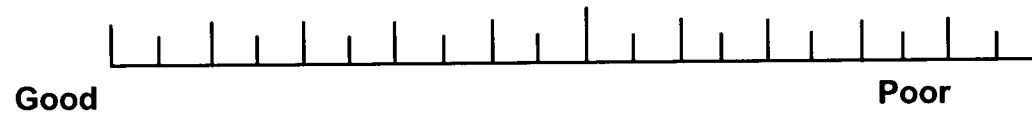
TEMPORAL DEMAND



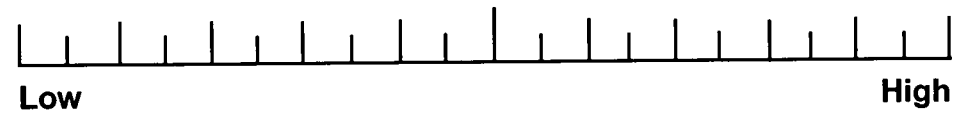
EFFORT



PERFORMANCE



FRUSTRATION



NASA-TLX Rating Scales

APPENDIX C. INFORMED CONSENT

Informed Consent

Project Title:

Flight Tracking Task

Investigator(s):

Michael A. Petkosek and William F. Moroney, Ph.D. (faculty sponsor)

Description of Study:

Participants will perform a tracking task while seated at a personal computer; a mouse will be used to manipulate a tracking circle, participants are to keep the circle as close as possible to the center of the field throughout the whole experiment. Participants will fill out a computer-based survey related to the workload experienced while performing the tracking task.

Adverse Effects and Risks:

No adverse effects have been reported in similar research. Participants may experience some initial frustration with the tracking task.

Duration of Study:

This study will take between 1-2 hours to complete. You will participate in one 2-hr session.

Confidentiality of Data:

Your name will be kept separate from the data. Both your name and the data will be kept in a locked filing cabinet. Only the investigators' names above will have access to the locked filing cabinet. Your name will not be revealed in any document resulting from this study.

Contact Person:

Students may contact William F. Moroney, Ph.D. in SJ 305, 937.229.2767 or Michael A. Petkosek, 440.522.0588. If you have questions about your rights as a participant, please contact the Chair of the Research Review and Ethics Committee, Charles E. Kimble, Ph.D. in SJ 319, 937.229.2167.

Consent to Participate:

I have voluntarily decided to participate in this study. The investigator named above has adequately answered any and all questions I have about this study, the procedures involved, and my participation. I understand that the investigator named above will be available to answer any questions about research procedures throughout this study. I also understand that I may voluntarily terminate my participation in this study at any time and still receive full credit. I also understand that the investigator named above may terminate my participation in this study if s/he feels this to be in my best interest. In addition, I certify that I am 18 (eighteen) years of age or older.

Signature of Student

Student's Name (printed)

Date

Signature of Witness

Date

APPENDIX D. PROCEDURES

Experimental Procedures
Practice Session

Date: ____

Time: ____

Participant #: ____

Difficulty: ____

Checklist:

NASA-TLX Instructions
MATB Instructions
Informed Consent sheet
NASA-TLX Subscales
NASA-TLX Subscale Explanations

1. Welcome/Consent:

- _____ Verbally brief participant on the nature of the tracking task.
- _____ Complete the Informed Consent form in order to proceed.

2. MATB Familiarization:

- _____ Participant reads MATB Instructions.
 Answer any questions that the participant may have.
- _____ MATB Tracking Task familiarization- Briefly practice tracking to become familiar.
 High
 Gain 15

3. NASA-TLX Familiarization:

- _____ Participant reads TLX instructions.
 Answer any questions that the participant may have.
- _____ Detailed explanation of the subscales (joint review by participant & experimenter).
 Answer any questions that the participant may have.

4. Perform Practice Session:

- _____ Simulate experimental conditions- Two trials.
 MM
 High
 Gain 15
 Subscale explanations viewable to participant
 Record Data: (Number of trials, Data file name, TLX rating for each trial)
- _____ Administer and collect TLX, examine the TLX.
 Answer any questions that the participant may have.
- _____ Qualification for experimental session: RMSE<16.0.

5. Finished: Five minute break

Experimental Session

Date: ____

Time: ____

Participant #: ____

Difficulty: ____

Checklist:

NASA-TLX Subscale Explanations

1. Perform Experimental Session:

_____ Administer experimental trials- Four trials.

MCCM

_____ Administer and collect TLX, examine the TLX.

5. Finished: Debrief

**APPENDIX E. DESCRIPTIVE STATISTIC TABLES REFERRED TO IN
CHAPTER III**

Table E1

Means and Standard Deviations for Baseline Trial (M1) Performance scores

Difficulty Level	Mean	Standard Deviation
Low	7.22	1.72
Medium	7.67	2.11
High	8.48	2.17

Table E2

Means and Standard Deviations for Baseline Trial (M1) Workload ratings

Difficulty Level	Mean	Standard Deviation
Low	27.92	12.78
Medium	39.67	18.79
High	32.08	19.00

Table E3

*Means and Standard Deviations for the first Context Inducing Trial (C2)**Performance scores*

Difficulty Level	Mean	Standard Deviation
Low	4.29	0.62
Medium	8.10	2.63
High	16.06	4.82

Table E4

Means and Standard Deviations for the first Context Inducing Trial (C2) Workload ratings

Difficulty Level	Mean	Standard Deviation
Low	15.50	10.28
Medium	40.08	23.69
High	46.42	23.31

Table E5

Means and Standard Deviations for the second Context Inducing Trial (C3) Performance scores

Difficulty Level	Mean	Standard Deviation
Low	4.27	0.65
Medium	8.35	3.10
High	15.40	6.21

Table E6

Means and Standard Deviations for the second Context Inducing Trial (C3) Workload ratings

Difficulty Level	Mean	Standard Deviation
Low	18.42	11.21
Medium	45.17	24.61
High	42.50	22.33

Table E7

Means and Standard Deviations for Critical Trial (M4) Performance scores

Difficulty Level	Mean	Standard Deviation
Low	8.86	2.46
Medium	8.53	3.90
High	8.44	1.50

Table E8

Means and Standard Deviations for Critical Trial (M4) Workload ratings

Difficulty Level	Mean	Standard Deviation
Low	37.42	18.14
Medium	43.67	23.76
High	34.58	20.49

Table E9

*Means and Standard Deviations for the NASA-TLX Subscales on the Baseline
Trial (M1) and Critical Trial (M4)*

Scale	Mean	Standard Deviation
Mental Demand (M1)	27.22	20.05
Mental Demand (M4)	35.69	25.41
Physical Demand (M1)	26.39	21.50
Physical Demand (M4)	32.50	25.06
Temporal Demand (M1)	25.69	19.08
Temporal Demand (M4)	28.61	21.10
Effort (M1)	41.67	25.83
Effort (M4)	42.78	25.79
Performance (M1)	30.28	17.81
Performance (M4)	32.36	17.09
Frustration (M1)	29.31	20.74
Frustration (M4)	34.17	22.54

R002588820