1999

An assessment of the modified subjective workload assessment technique (ModSWAT) metric using pilot-in-the-loop simulation

George Scott Boucek
University of Dayton

An Assessment of the Modified Subjective Workload Assessment Technique (ModSWAT) Metric Using Pilot-In-The-Loop Simulation

Thesis

Submitted to

The School of Arts and Sciences of the

UNIVERSITY OF DAYTON

In Partial Fulfillment of the Requirements for

The Degree

Masters of Arts in Psychology

By

George Scott Boucek

UNIVERSITY OF DAYTON

Dayton, Ohio

May 1999

APPROVED BY:

[redacted]
David W. Biers Ph.D. (Faculty Advisor)

[redacted]
F. Thomas Eggemeier Ph.D. (Faculty Reader)

[redacted]
William F. Moroney Ph.D. (Faculty Reader)

[redacted]
John A. ~~Hassoun (Reader)~~

CONCURRENCE:

[redacted]
F. Thomas Eggemeier, Ph.D.
Chairperson

ABSTRACT

AN ASSESSMENT OF THE MODIFIED SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE (MODSWAT) METRIC USING PILOT-IN-THE-LOOP SIMULATION

Name: Boucek, George, Scott
University of Dayton, 1999

Advisor: Dr. D. W. Biers

The present study sought to determine the validity of using a percentage measure of workload based on the unweighted sum of the three SWAT rating scales (ModSWAT) in a pilot-in-the-loop aircraft simulation. Two separate simulation experiments to determine pilot workload associated with alternative cockpit configurations were re-analyzed using ModSWAT and then compared to the original workload results using traditional SWAT conjoint scaling. Results indicated that the ModSWAT and SWAT conjoint measures were highly correlated and equally sensitive, leading to the same conclusion about workload. These results further strengthen the case for the validity of using ModSWAT in place of the traditional SWAT metric, thereby maintaining the benefit of real-time collection of workload data while eliminating the cost (i.e., preparation time and materials, data collection time, support assets, and money) associated with performing a card sort.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The purpose of the present study was to validate an alternative scaling method used for developing Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988) workload composites. SWAT data collected during pilot-in-the-loop simulation experiments conducted at Wright-Patterson Air Force Base (WPAFB) were used for this evaluation. The alternative scaling method, developed and studied by Biers and his colleges at the University of Dayton (Biers, 1995; Biers & Masline, 1987; Biers & McInerney, 1988; Moroney, Biers, & Eggemeier, 1995), utilizes the unweighted sum of the workload ratings to develop the composite rather than relying on conjoint procedures traditionally employed for SWAT. If the alternative method is successful in replicating the original results obtained using the SWAT conjoint procedure, then time saved by eliminating the need for conjoint scale development will increase the efficiency of collecting workload data, translating into lower research costs. Details concerning SWAT conjoint and the alternative scaling technique (ModSWAT) are provided in the following paragraphs.

Background

In both commercial and military aviation, pilot workload is continually evaluated in an effort to index the relationship between the demands of the environment and the

capacity of the operator (Kantowitz and Casper, 1988). Research conducted within the Advanced Cockpits Branch of Wright Laboratory, located at WPAFB, frequently includes workload as a variable to measure the effects of design changes on the pilot-vehicle-interface (PVI). Methods by which workload metrics are collected and analyzed vary depending on the type of research being conducted. However, one method of choice for pilot-in-the-loop simulations, is the Subjective Workload Assessment Technique (SWAT). Based on criteria outlined by Williges and Wierwille (1979), SWAT is very appropriate for in-flight environments. SWAT is highly portable, conducive to sound experimental control, causes minimal intrusion, maximizes safety, streamlines data transmission and recording, and is generally accepted by the pilots.

Reid and Nygren (1988) define workload for SWAT as being composed of Time Load, Mental Effort Load, and Psychological Stress Load. Time Load is the total amount of time available to perform a task as well as the extent to which tasks overlap; Mental Effort Load refers to the amount of attention or concentration required to perform a task; and Psychological Stress Load is the presence of confusion, frustration, and/or anxiety associated with task performance (Reid, Potter, and Bressler, 1989). For each of the three primary dimensions, SWAT employs a three point rating scale (see Table 1). The three primary dimensions in conjunction with the three point scale make up the 27 possible rating combinations of SWAT (3 time load values X 3 mental effort values X 3 psychological stress values).

Table 1.

Rating Scale Definitions for Each SWAT Dimension

Dimension/Scale	Definition
Time Load	
1	Often have spare time. Interruptions or overlap among activities occur infrequently or not at all
2	Occasionally have spare time. Interruptions or overlap among activities occur frequently
3	Almost never have spare time. Interruptions or overlap among activities are frequent or occur all the time.
Mental Effort Load	
1	Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.
2	Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.
3	Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.
Psychological Stress Load	
1	Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.
2	Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.
3	High to very intense stress due to confusion, frustration, or anxiety. High to extreme determination and self-control required.

Note. Definitions provided by Reid, Potter, and Bressler (1989).

SWAT has two distinct phases: Scale Development and Event Scoring (Reid et al., 1989). The Scale Development phase introduces the descriptors of the SWAT dimensions and associated rating scales. In addition, data are obtained to determine how the dimensions combine to create an individual's personal impression of workload (Reid et al., 1989). This is accomplished by requiring a subject to rank-order 27 cards, each representing a single SWAT rating combination, in terms of perceived workload. Figure 1 illustrates 1 of 27 cards that subjects are required to rank-order from lowest to highest perceived workload based on the situation described by each card.

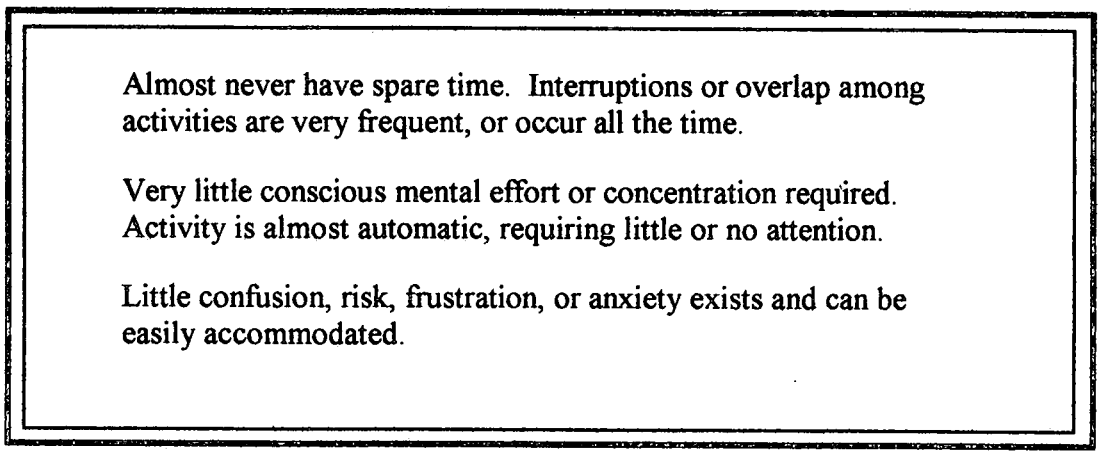

Figure 1. 1 of 27 Cards that Must be Rank-Ordered from Lowest to Highest Perceived Workload

Employing SWAT conjoint techniques, the sorted cards are used to construct a single interval scale (ranging from 0 to 100) of workload composite ratings (Reid and Nygren, 1988). This scale is then used to transform the three part SWAT ratings, given during the Event Scoring phase, into a single workload rating for each task performed. For example, a pilot reports a SWAT rating of 2-3-1 (Time Load, Mental Effort, and

Psychological Stress respectively) following the completion of an acrobatic maneuver. This rating of 2-3-1 corresponds to a single composite rating (e.g. 33) based on the interval scale established from conjoint analysis of the card sorts.

The Event Scoring phase occurs during test situations when the investigator requests a subjective report of workload experienced by the pilot or operator while performing a specific task. This subjective judgement is formulated, and ratings are given, based on the three primary categories of SWAT described previously.

In theory, SWAT conjoint provides face validity because it relies on empirically developed measurement models to produce a workload scale based on actual workload orderings (Reid and Nygren, 1988). However, utilizing the conjoint measurement approach, internal consistency must be achieved within the workload orderings provided by the subjects during scale development. According to Reid and Nygren (1988), this criterion is very impractical because people do not give error free data very often. Because of this, either the card sort is repeated until consistency is achieved or a relaxation of the criteria must be made. Since scale development using SWAT conjoint is time consuming, taking approximately 45 minutes to 1 hour to complete (Biers and McInerney, 1988), and the process of sorting the 27 rating scale combinations can be a cognitively demanding task; performing multiple card sorts is not a practical alternative within an applied setting. Therefore the latter approach of relaxing the requirement for strict internal consistency may be more desirable.

Because an error theory for conjoint measurement that formally addresses the issue of internal consistency has not yet been formulated, Reid, Potter, and Bressler

(1989) established "rules of thumb" based on experience with card sort data. These rules allow up to a 5 to 10 percent violation rate of the independence axioms established for SWAT and used by the scale measurement models. These violations are acceptable as long as the inconsistencies involve adjacent or near-adjacent axiom pairs. For example, one such rule of thumb states that if an overall Kendall's Coefficient of Concordance, used to determine the level of agreement among a particular group of subjects, is .75 or higher, there is sufficient agreement to make a single scale that will represent all of the subjects without incurring a large chance of misrepresenting any single subject (Reid and Nygren, 1988). Since there is usually high agreement among raters, a group solution to forming a single workload scale is typically used (Biers and Masline, 1987).

Based on the previous discussion, it appears that SWAT conjoint is often relegated to the level of "best-fit" by relying on "rules of thumb" when relating scaled variables to the observed data. There are also potential sensitivity issues that arise when employing a group solution to forming a workload composite for each subject. Specifically, using a group solution with less than perfect levels of agreement alters each subjects interval scale of workload composite ratings and thus may not accurately reflect each subjects impression of workload for a given task. Because of this, alternative measurement models that would provide the same workload results without the complexity involved with conjoint scale development and analysis are desired. One such alternative is the Modified Subjective Workload Assessment Technique (ModSWAT). According to Moroney et al. (1995), ModSWAT employs a simple model of forming a workload composite by taking the unweighted sum of the three rating scales (range = 3 to

9) and then converting each sum to a scale from 0 to 100 by using the following formula: ((Sum-3)/6)*100. This formula was developed so that an unweighted sum of 3 derived from a rating of 1-1-1 would equal 0, representing very low workload. Conversely, an unweighted sum of 9 derived from a rating of 3-3-3 would equal 100, representing very high workload. Because rating scale combinations of 3-2-1 would receive the same sum as 2-1-3, criticism of this technique can be made based on its psychometric soundness. Also, the ModSWAT measure yields only seven workload values compared to 27 with conjoint measurement (Moroney et al., 1995).

Despite these criticisms, if the investigator is only interested in an overall sense of workload related to a specific function or task, then the ModSWAT measure may be a more efficient approach than the conjoint procedure, particularly if it can be shown that the SWAT conjoint measure and the ModSWAT measure yield equivalent results.

Several studies have been conducted in an effort to validate the ModSWAT measure. Biers and Masline (1987) applied ModSWAT to workload data previously collected using SWAT conjoint procedures. These data were obtained from a study conducted by Masline (1986) in which subjects performed a card sort prior to performing a continuous memory task. Specifically, subjects were presented with a 1, 2, or 3 digit number, at a slow or fast rate, with a number back of 1 or 2. This factorial combination created twelve different levels of task difficulty. Immediately following performance under each condition, subjects rated workload using the 3-point scale within each dimension of SWAT.

The original workload composite measure was created from the above data using

SWAT conjoint methods. Biers and Masline (1987) created a second composite measure from the same workload data using ModSWAT. Both composites were compared and were found to be sensitive to the same task manipulations. Measures of the strength of the effect (Eta-Squared) and power indicated that SWAT conjoint and ModSWAT were equally sensitive with a median difference in Eta-Squared being 1.8%. In all cases, both composites were highly correlated, ranging from 0.9913 to 0.9991.

One criticism of the Biers and Masline study is that all subjects performed the card sort prior to performing the memory task. This raises the issue that perhaps the card sort influenced the subject's perception of workload and thus affected their rating behavior. To address this, Biers and McInerney (1988) conducted a two group study in which subjects performed the same continuous memory task under the same 12 levels of task difficulty as in the Masline (1986) study. The only difference being that one group (Pre-Task) performed the card sort prior to the memory task while the second group (Post-Task) performed the card sort after performing the memory task. Results indicated no significant interactions of group with any of the task manipulations suggesting that the placement of the card sort did not affect individual scale ratings.

The studies described previously begin to establish the validity of the ModSWAT measure. However, results from these studies were derived from a single laboratory task in which the size of the effects were extremely large, and the subjects were well practiced in both the task itself and in rating workload (Biers, 1995). Further research into the utility of an unweighted sum composite measure within a more real-world context is necessary to generalize the validity of ModSWAT.

Work in this area began with Biers (1995) re-analyzing SWAT data collected from four active duty US Air Force fighter pilots participating in a role playing exercise conducted at the Wright Laboratory's Cockpit Integration Division of WPAFB for the Integrated Mission/Precision Attack Cockpit Technology (IMPACT) program. The purpose of this IMPACT experiment was to identify workload associated with flying an air interdiction mission at night in adverse weather with either a baseline dual-seat F-15E, a conceptual single-seat F-15E, or a conceptual single-seat Advanced Technology Cockpit (ATC). Traditional SWAT conjoint procedures were used for this study and a group solution was employed for data analysis.

While flying ingress, attack, and egress portions of an air interdiction mission, pilots role-played flying each cockpit configuration mentioned previously. The Mission Tasks included: Flying Only Task, requiring pilots to fly the mission route using only the head-up display (HUD); Head-Down Task, requiring pilots to "step through" the mission performing tasks using only the head-down displays (HDD); and a Dual Task, requiring pilots to fly the simulator and manipulate HDD frames to complete the mission. The selected combination of the cockpit configuration and mission task variables formed the six conditions created for this IMPACT study. SWAT ratings were collected after each of six critical mission events, requiring pilots to project what the workload would have been in the conceptual cockpits. This is often referred to as projective SWAT or PROSWAT. The critical mission events included: input mission change, engage ground threat, obtain patch map, weapons delivery, engage air threat, and for overall mission.

Biers (1995) reanalyzed workload data using ModSWAT. Both the original

SWAT conjoint metric and the ModSWAT metric were subjected to a correlation analysis and one way (condition) analyses of variances for each event. Results indicated that the two metrics were highly correlated (0.977) and both were equally sensitive in measuring workload. In fact, a closer inspection of the effect size measures revealed that the effect sizes of condition across the six events were similar with the maximum difference of 0.039 (Biers, 1995).

The research conducted by Biers (1995) is yet another case for the mounting evidence in establishing the validity of ModSWAT within the applied research environment. However, the sample size of only 4 subjects, the use of PROSWAT, and the fact that all subjects performed a card sort prior to participating in data collection, are grounds upon which the Biers (1995) study can be criticized.

The Present Study

In an effort to address the criticisms of the Biers (1995) study, ModSWAT was used in the present study to re-analyze workload data collected in a much larger pilot-in-the-loop simulation experiment employing SWAT conjoint procedures. Specifically, two experiments were re-analyzed, each employing methods of collecting traditional SWAT ratings (versus PROSWAT). The first experiment (n=12) was similar to the Biers and Masline (1987) study in that all subjects performed a card sort prior to data collection. The second experiment (n=18) was similar to the Biers and McInerney (1988) study in that subjects were randomly assigned to two groups, one performing the card sort prior to data collection (Pre/Post-Test) and the second performing the card sort after data collection (Post-Test Only). For both studies being re-analyzed, it was expected that the

workload composite developed using ModSWAT would be highly correlated with the SWAT conjoint composite. This was expected both at the level of condition means and at the level of individual event ratings. It was also expected that the ModSWAT and SWAT conjoint composites would be sensitive to the same task manipulations. Finally, it was expected that no interactions would be found between Pre/Post-Test and Post-Test Only card sort groups.

CHAPTER II

METHOD

Workload data collected in two pilot-in-the-loop simulation experiments, employing SWAT conjoint procedures, were re-analyzed using ModSWAT. This was done in an effort to directly compare the two SWAT metrics. The following is a description of the two experiments.

Experiment 1: The C-141 Full Mission Simulation

The purpose of this experiment was to evaluate proposed cockpit upgrades to the C-141 by comparing current flight instruments commiserate with an electro-mechanical cockpit against those of a more modern "glass" cockpit environment. The evaluation was conducted using a pilot-in-the-loop simulation, flying mission scenarios typical of operational C-141s.

Subjects

Twelve C-141 pilots participated in the Full Mission experiment. The C-141 pilots represented a range of experience levels: 5 first pilots, 3 aircraft commanders, and 4 pilots rated at either instructor or flight examiner. On average, each pilot had a total of 2433 hours of flight experience, 1950 of which were in the C-141. All pilots were

required to be at least wing qualified in Station Keeping Equipment (SKE) for formation flying.

Apparatus

The experiment was conducted in the Transport Aircraft Cockpit (TRAC) simulator located in the Crew Station Integration Laboratory (CSIL) of the Aeronautical Systems Center at Wright Laboratory. The simulator was configured to provide cockpit geometry similar to the C-141 aircraft. The cockpit shell contained three crew stations: pilot, copilot, and flight engineer. The head-down display configurations were presented across three 21 inch Cathode Ray Tube (CRT) monitors. An additional 16 inch direct-view CRT was used to display an out-the-window visual scene to the pilot position only.

Procedure

Pilots participated in this experiment for one full week. The first full day and one-half focused on training. Regarding workload, pilots were given a SWAT briefing which defined the three dimensions and explained the scoring procedures. Appendix A provides the instructions given to each pilot. The briefing concluded with the pilots performing a card sort. The 27 cards, each representing a single combination of the SWAT dimensions, were sorted from lowest to highest workload and the subsequent orderings were used to develop a workload composite using conjoint measurement.

After completing training, the next two and one-half days were spent on testing. Pilots flew an Airdrop mission and an Airland mission using either the current C-141 cockpit configuration or the upgraded configuration, resulting in a total of four

operational missions, with each mission being a different profile. The two Airdrop mission profiles consisted of a Sicily scenario and a Luzon scenario. The two Airland mission profiles consisted of a Pope scenario and a Fayetteville scenario. Presentation of mission and cockpit configuration were counterbalanced and each mission and configuration combination was replicated three times over the course of the entire experiment. SWAT ratings were collected throughout each mission following significant mission events (see Table 2). In addition, an overall SWAT rating was collected at the conclusion of each mission. The final day of participation was spent filling out questionnaires and conducting interviews. For more details regarding the apparatus, mission scenarios, and procedures used for the C-141 experiment, refer to Toms, Cone, Gier, Boucek, and Brown (1995).

Table 2.

Significant Events for the Airdrop and Airland Missions

Airdrop Mission Event	Airland Mission Event
Cruise	Departure
Drop Descent 1st Pass	Cruise
Run-In and Drop 1st Pass	DAMU Fail*
Escape 1st Pass	1st Approach (non-precision / NDB)
Drop Descent 2nd Pass	INS Fail
Run-In and Drop 2nd Pass	2nd Approach (non-precision / TACAN)
Recovery 2nd Pass	
ILS Approach	

Note: Table provided by Toms et al. (1995).
* Not included in SWAT Analysis of Variance.

Design

A 3 factor (Cockpit Configuration, Mission, and Mission Event) repeated measure design was employed for this experiment. However, due to differences in the number of mission events across missions, the study was analyzed separately for each mission. For the Airdrop Mission, the design represents a 2 (Cockpit Configuration) by 8 (Mission Event) repeated measures factorial. Cockpit Configuration consisted of the current C-141 display configuration and the upgraded display configuration. See Table 2 in the Procedure section of this experiment for a list of the eight Mission Events. For the Airland Mission, the study was a 2 (Cockpit Configuration) by 6 (Mission Event) repeated measures factorial design. Cockpit Configuration was the same as for the Airdrop Mission analysis and Table 2 provides a list of the six Mission Events.

Experiment 2: The IMPACT Simulation

This IMPACT experiment was the follow-on to the initial experiment described previously. The purpose of this experiment was to evaluate the integration of advanced technologies into a single seat, multi-role fighter aircraft performing precision strikes at night and in adverse weather. Advanced technologies selected for this experiment included a Helmet-Mounted Display (HMD), Directional Audio, and Large-Screen Displays.

Subjects

Eighteen pilots participated as subjects in this experiment. They were pilots assigned to WPAFB and local Air Force Reserve/Air National Guard units. Total

operational aircraft flight time for the pilots ranged from 770 to 3470 hours (mean = 2618). Sixteen of the pilots had fighter aircraft experience and two had B-52 experience.

Apparatus

The experiment was conducted in the Manned Combat Station (MCS) simulator located in the Crew Station Integration Laboratory (CSIL) of the Aeronautical Systems Center at Wright Laboratory. The cockpit simulator was reconfigurable as either a baseline F-15E front cockpit or an IMPACT cockpit with the capability of incorporating an HMD, Directional Audio, and a 10" x 10" Tactical Situation Display (TSD).

Procedure

A minimum of one full day was required for each pilot's participation. Each pilot reported to the laboratory at 0800 for introductory briefings and training on experimental procedures and equipment. Regarding workload, pilots were given a SWAT briefing which defined the three dimensions and explained the scoring procedures. Again, Appendix A provides the instructions given to each pilot. Pilots were randomly assigned to one of two card sort groups (n=9) which differed only in terms of when they performed the card sort. Half of the pilots performed the card sort both prior to and after data collection and the other half performed the card sort after data collection. The briefing concluded after instructions were given or, if necessary, a card sort had been completed.

After the training session was completed, each pilot was given a short practice session in the simulator to provide familiarity with the mission profile and procedures.

Eight data collection sessions then followed requiring the pilots to fly an air interdiction mission using either the baseline configuration patterned after current weapon systems or the IMPACT configuration incorporating advanced technology. The mission was segmented into four phases which included: Medium Altitude Cruise, requiring pilots to maintain an altitude of 10,000 feet while acquiring a single threat; Terrain Following Descent, requiring pilots to descend from 10,000 feet to 300 feet while acquiring a single threat; High-Speed Terrain Following Ingress, requiring the pilot to maintain 300 feet above the ground while acquiring a single threat presented on two separate occasions; and Weapon Delivery, requiring the pilot to perform a low angle/low drag dive bomb weapon delivery on a stationary SCUD missile launcher. Threat difficulty was manipulated based on the angular position of the threat with respect to the aircraft. Target difficulty was manipulated based on the angular position of the target with respect to the run-in course. SWAT ratings were collected throughout each mission following each of the four threats and immediately after the weapon delivery. For more details regarding the apparatus, mission scenario, and procedures used for this IMPACT experiment, refer to Boucek et al. (1995).

Design

A 4 factor (Cockpit Configuration, Mission Phase, Threat Difficulty, and Target Difficulty) repeated measures design was employed for this experiment. Because the first three mission phases focused on threat acquisition and the fourth on weapon delivery, two separate analyses were performed. Threat acquisition workload data represents a 2 (Cockpit Configuration) by 2 (Threat Difficulty) by 3 (Mission Phase)

within subjects factorial design. Cockpit Configuration consisted of the baseline and IMPACT configurations, Threat Difficulty was divided into high and low, and Mission Phase consisted of the medium altitude cruise, terrain following descent, and high-speed terrain following ingress. Weapon delivery workload data, however, can be conceptualized as a 2 (Cockpit Configuration) by 2 (Target Difficulty) repeated measure design. Cockpit Configuration was the same as for the threat acquisition analysis and Target Difficulty was also divided into high and low.

CHAPTER III

RESULTS

Experiment 1: The C-141 Full Mission Simulation

To test the equivalence of the ModSWAT and SWAT metrics, a general analytic approach was used on workload data collected during each mission. First, correlational analyses were performed between ModSWAT and SWAT using both individual ratings and condition means. Second, a top level three-factor ANOVA (Cockpit Configuration (C) x Mission Event (E) x Metric (M)) was conducted to assess overall differences in sensitivity, focusing on interactions that would indicate differences as a result of the Metric utilized. Including Metric in the top level analysis is a sound approach because the added degrees of freedom increase the sensitivity of the analysis and the approach directly tests differences between metrics.

Third, a separate two-factor ANOVA (C x E) was performed for each metric. This third analysis represents the approach which would have been used had only a single metric been utilized. One would expect to reach the same statistical conclusion about the independent variables based on separate analyses of the two metrics if they are equally sensitive. The rational for conducting separate ANOVAs in addition to the top level analysis is as follows: The top level analysis is expected to contain highly correlated measures (SWAT and ModSWAT) which will contribute to reduced error variance.

Under this condition, very small differences for the within subject factor (Metric) may end up being significant. By conducting separate analyses of these two metrics, one avoids the potential problem of very small differences being significant. Results from these analyses are presented in the following paragraphs.

Airdrop Mission

The correlation between the ModSWAT and SWAT composites was 0.9963 (R^2 = 0.9926) when computed at the level of individual ratings and 0.9984 (R^2 = 0.9969) using condition means. These results indicate that the two metrics are measuring the same phenomenon and this is consistent with the findings of Biers and Masline (1987).

Table 3 summarizes the results from the top level 2 (C) x 11 (E) x 2 (M) analysis of variance as well as subsequent 2 (C) x 11 (E) analyses performed for each metric.

Table 3.

Airdrop mission results from overall and separate analyses as a function of Metric

	Top Level Analysis	Separate Analyses	
Source	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.832	.826	.838
Event (E)	.001	.000	.001
Metric (M)	.000	-	-
CxE	.625	.624	.623
CxM	.873	-	-
ExM	.157	-	-
CxExM	.410	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis. Shaded cells indicate significant results of $p<.05$.

As Table 3 indicates, there were no significant interactions with Metric using a .05 significance level. Cockpit Configuration was not significant (see Figure 2) whereas Mission Event was significant (see Figure 3). Based on the separate analyses performed for each metric, the same conclusion would be made in both cases regarding the effect of Cockpit Configuration and Mission Event (see Figures 2 & 3).

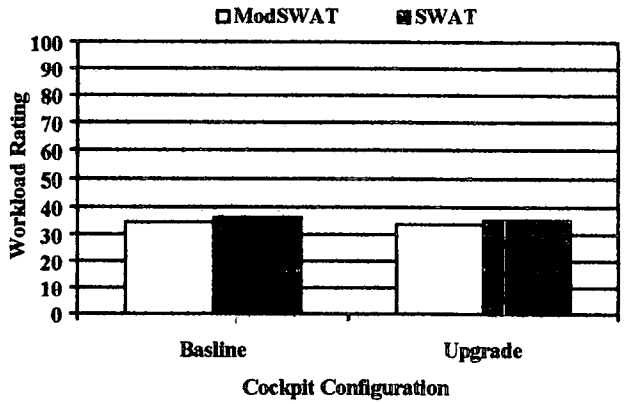

Figure 2. Workload Ratings for Cockpit Configuration as a Function of Metric

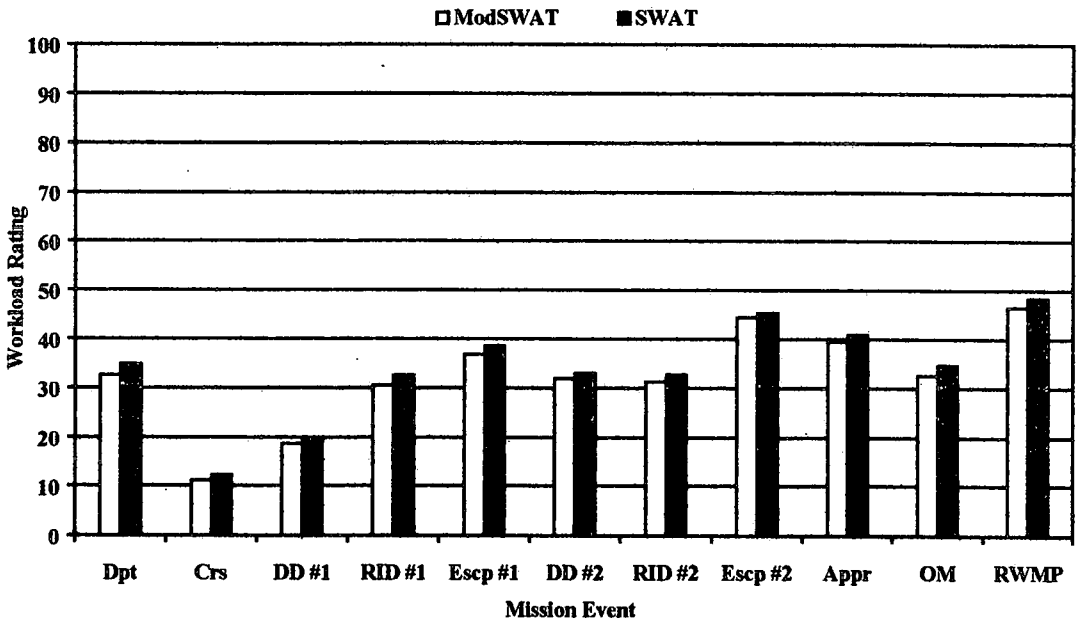

Figure 3. Workload Ratings for Mission Event as a Function of Metric

Table 3 reveals a significant effect for Metric. Mean workload ratings were higher for SWAT ($\overline{X}$ =35.72) than for ModSWAT ($\overline{X}$ =34.16). However, as shown in Figures 2 & 3, this difference was very small (< 2%): The mean square error (MSE =

9.40) for this effect indicates that there was relatively small error variance. Thus the analysis was extremely sensitive to detecting small differences.

Figure 4 presents the frequency distribution of the difference between SWAT and ModSWAT over the 264 measurements taken in this study. In Figure 4, a positive difference indicates higher SWAT rating. As seen in the figure, there was a bias toward the SWAT metric leading to a higher workload rating than ModSWAT with most of the differences being within three rating points of one another.

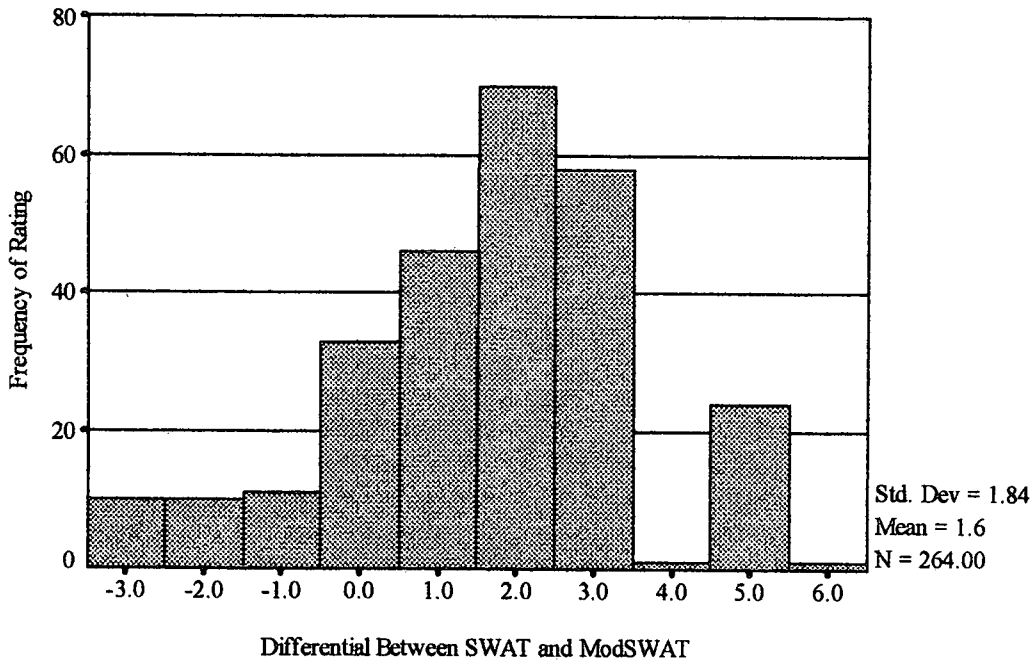

Figure 4. Frequency Distribution of the Difference Between SWAT and ModSWAT for all Cockpit Configuration and Airdrop Mission Event Combinations.

The results presented in Table 3 were evaluated using the .05 significance level. One could argue however that the present researcher is trying to show that there are no interactions with Metric, which is tantamount to proving the null hypothesis. In research attempting to

demonstrate (prove) the null hypothesis, a less stringent significance level is sometime adopted, for example, the .20 significance level. If the effect is not found significant using the .20 level, the researcher is on firmer ground in stating that the conditions are equivalent. Based on using the .20 criteria, the Mission Event by Metric interaction becomes significant (Table 3). Again however, the difference between the Metric means as a function of Mission Event is very small the largest difference being 2.13 and the smallest being .65 (see Table 4). In every case the mean workload value was higher under

Table 4.

Actual Difference Between SWAT and ModSWAT Means for Each Airdrop Mission Event

Airdrop Mission Event	Mean Workload Composite - SWAT Metric	Mean Workload Composite - ModSWAT Metric	Difference Between Workload Composites
Departure (Dpt)	34.77	32.64	2.13
Cruise (Crs)	12.30	11.12	1.18
Drop Descent (DD) 1st Pass	19.35	18.70	0.65
Run-In and Drop (RID) 1st Pass	32.55	30.55	2.00
Escape 1st Pass (Escp)	38.55	36.80	1.75
Drop Descent (DD) 2nd Pass	33.05	31.94	1.11
Run-In and Drop (RID) 2nd Pass	32.72	31.25	1.47
Escape (Escp) 2nd Pass	45.46	44.44	1.02
ILS Approach (Appr)	41.02	39.58	1.44
Overall Mission (OM)	34.75	32.65	2.10
Real World Mission Projection (RWMP)	48.41	46.53	1.88

SWAT than ModSWAT. The differences between the largest and smallest difference (i.e. the interaction -- 2.13 vs. 0.65, of the last column in Table 4) is only 1.48 rating points on a 100 point scale -- an extremely small differential effect by any standard. Thus, if one considers the C x E interaction to be significant (using the .20 level), the effect is of little practical importance.

Airland Mission

The results for Airland exactly paralleled the Airdrop Mission. First, the correlations between the two metrics were extremely high -- 0.997 (R^2 = 0.9932) when computed at the level of individual ratings and 0.9984 (R^2 = 0.9968) using condition means. Secondly, the top level analysis revealed no significant interactions (see Table 5). Finally, in performing separate analyses, the same pattern of results were found with only Mission Event approaching significance.

Table 5.

Airland mission results from overall and separate analyses as a function of Metric

	Top Level Analysis	Separate Analyses	
Source	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.114	.115	.112
Event (E)	.077	.084	.072
Metric (M)	.001	-	-
CxE	.246	.186	.242
CxM	.642	-	-
ExM	.175	-	-
CxExM	.937	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis. Shaded cells indicate significant results at $p<.05$.

As with the Airdrop results, there was a significant effect of Metric in the top level analysis, with the mean SWAT ratings being higher ($\overline{X}$ =28.48) than for ModSWAT ($\overline{X}$ =27.18). Inspection of the frequency distribution of the differences reveals the same small bias in favor of the ratings being higher for SWAT (see Figure 5).

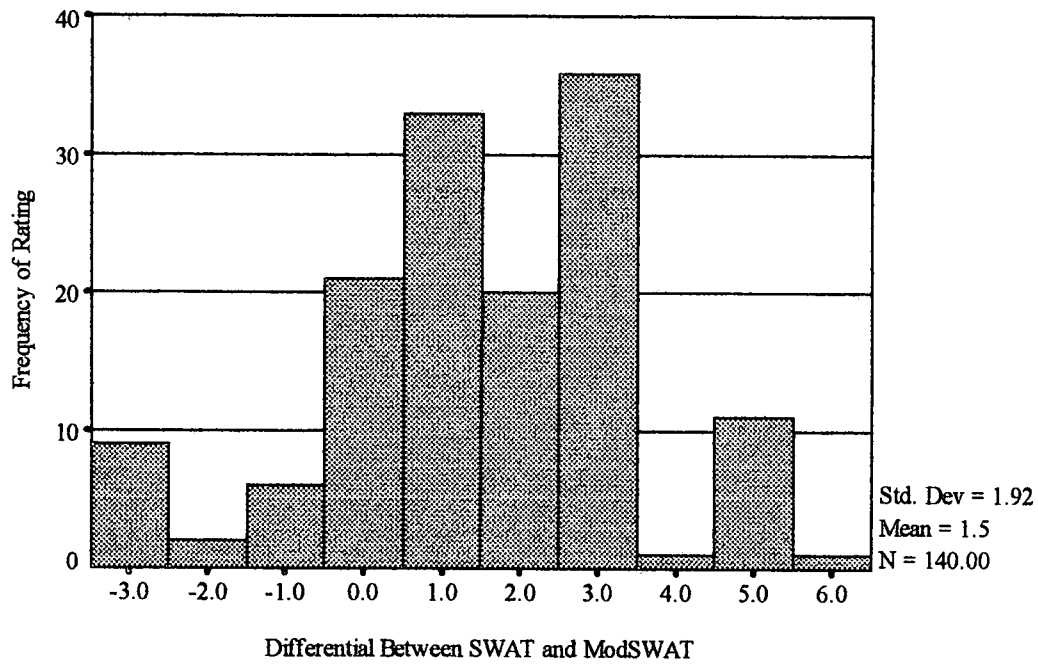

Figure 5. Frequency Distribution of the Difference Between SWAT and ModSWAT for all Cockpit Configuration and Airland Mission Event Combinations.

If one assumes that the researcher is attempting to prove the null hypothesis and thus adopt the .20 significance level, again only the Mission Event by Metric interaction is significant in the top level analysis. Table 6 shows that the differential difference in Metric across Mission Event is extremely small, with the difference between the largest and smallest difference being 1.78 rating points.

Table 6.

Actual Difference Between SWAT and ModSWAT Means for Each Airland Mission Event

Airdrop Mission Event	Mean Workload Composite - SWAT Metric	Mean Workload Composite - ModSWAT Metric	Difference Between Workload Composite
Departure (Dpt)	22.40	21.54	.86
Cruise (Crs)	5.74	5.56	.18
1st Approach	38.56	36.80	1.76
INS Fail	29.45	28.34	1.11
2nd Approach	38.04	36.80	1.24
Overall Mission (OM)	29.05	27.09	1.96
Real World Mission Projection (RWMP)	32.41	30.56	1.85

Experiment 2: The IMPACT Simulation

In addition to evaluating the sensitivity of ModSWAT and SWAT to the same task manipulations, the purpose of re-analyzing workload data collected during the IMPACT experiment was to determine the effect (if any) of performing a card sort on reported SWAT ratings.

To investigate this effect, the original design separated the subjects into two groups (Pre/Post-Test group and Post-Test Only group) as described in the Method section (Chapter 2). However, inadequacies in the experimental design resulted in a confound which rendered the analysis useless. First, true random assignment was not utilized -- subjects were assigned to groups on an alternating basis. Second, and more importantly, threat difficulty was not equally balanced across the Pre/Post-Test and Post-Test Only groups. This resulted in three Hard and one Easy threat being presented during

the first phase (High Level Ingress) for the Pre/Post Test group flying the IMPACT cockpit configuration and three Easy and one Hard threat for the Post-Test Only group. The exact opposite was the case for both groups flying the baseline configuration.

This confound did not impact the results of the original experiment because the Pre/Post Test grouping was ignored. When collapsed across groups, threat difficulty was equally balanced across conditions. Using the conjoint values derived from the combined Post-Test data, it was possible to determine the effect of task manipulations on workload uncontaminated by any difference in threat.

Due to the confound in experimental design, separate analyses were performed on threat acquisition data collected for each group. This was also the case with Weapon Delivery since the order of receiving the two configurations was opposite for the two groups.

The analytical approach taken for the IMPACT experiment was similar to that employed for the C-141 experiment. For threat acquisition, a top level four factor ANOVA (Cockpit Configuration (C) x Threat Difficulty (ThtD) x Mission Phase (P) x Metric (M)) was performed to assess differences in sensitivity. Subsequent three factor ANOVAs (C x ThtD x P) were performed to better interpret slight differences between highly correlated measures. For weapon delivery, a top level three factor ANOVA (Cockpit Configuration (C) x Target Difficulty (TgtD) x Metric (M)) was performed along with subsequent two factor ANOVAs (C x TgtD) for each metric.

Threat Acquisition

Group 1 - Pre/Post-Test. Two different SWAT conjoint values were derived depending upon group. First, conjoint SWAT values were derived using the pretest card sort (Pre-SWAT). This Pre-SWAT metric represents the traditional SWAT measure. Second, since the card sort was also performed subsequent to the experiment, a Post-SWAT conjoint metric was also calculated. The correlation between the pre and post test conjoint value was .992 (correlated the 27 rating scale combinations for Pre-SWAT and Post-SWAT). Next, each of the two SWAT metrics was compared with the single set of ModSWAT values derived from the task (Table 7). The high correlation between the metrics, both for individual ratings and condition means, is consistent with previous findings from the C-141 experiment.

Table 7.

Correlation of threat acquisition results between ModSWAT and SWAT as a function of Pre/Post-Test Group

		Pre-SWAT of the Pre/Post-Test Group	Post-SWAT of the Pre/Post-Test Group
Individual Ratings:	ModSWAT	.998	.998
Condition Means:	ModSWAT	.999	.999

Comparison of ModSWAT with the Pre-SWAT Metric. Table 8 summarizes the results from the top level 2 (C) x 2 (TthD) x 3 (P) x 2 (M) analysis of variance as well as subsequent 2 (C) x 2 (TthD) x 3 (P) analyses performed for each metric. Table 8 shows

Table 8.

Threat acquisition results from top level and separate analyses based on the Pre-SWAT of the Pre/Post-Test Group

	Top Level Analysis	Separate Analyses	
Source	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.004	.004	.004
Mission Phase (P)	.293	.281	.306
Threat Diff. (ThtD)	.007	.008	.007
Metric (M)	.000	-	-
CxP	.232	.225	.240
CxThtD	.102	.110	.093
CxM	.030	-	-
PxThtD	.142	.150	.135
PxM	.684	-	-
ThtDxM	.029	-	-
CxPxThtD	.010	.012	.008
CxPxM	.804	-	-
CxThtDxM	.992	-	-
PxThtDxM	.110	-	-
CxPxThtDxM	.188	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis. Shaded cells indicate significant results of $p<.05$.

the same pattern of significance regarding task manipulation, both the top level analysis and separate analyses of the two metrics -- either both are significant or both non-significant. However, contrary to expectations, three effects involving metric were significant at the .05 level (M, C x M, and ThtD x M). The significant Metric effect is consistent with previous analyses – SWAT resulted in slightly higher workload ratings than did ModSWAT. The significance of the latter two are potentially damaging to the

position that the two metrics are equally sensitive in that the effect for Cockpit Configuration and Threat Difficulty are statistically different for the different metrics. However, inspection of Table 9 reveals that the differential effect is small with the difference between the two cockpit configurations being 14.68 for SWAT and 14.26 for ModSWAT. The differential effect (0.42), although significant, was so small as to again raise the question of practical importance. Moreover, Table 10 reveals the differential effect for Threat Difficulty was very small (0.28) with the difference between the easy and hard threat being 13.52 for SWAT and 13.24 for ModSWAT.

Table 9.

Actual difference between SWAT and ModSWAT Means, based on the Pre-SWAT of the Pre/Post-Test Group, for each Cockpit Configuration

Workload Metric	Baseline Cockpit Configuration	Advanced Cockpit Configuration	Difference in Workload Rating
SWAT	52.17	37.49	14.68
ModSWAT	49.31	35.05	14.26

Table 10.

Actual difference between SWAT and ModSWAT Means, based on the Pre-SWAT of the Pre/Post-Test Group, for Threat Difficulty

Workload Metric	Hard Threat Difficulty	Easy Threat Difficulty	Difference in Workload Rating
SWAT	51.59	38.07	13.52
ModSWAT	48.80	35.56	13.24

Since these small differences are causing significant results at the .05 level, one would expect significant differences to be even smaller when compared at the .20 level. Therefore, significant effects found at the .20 level are no longer presented in this section.

Comparison of ModSWAT with Post-SWAT Metric. The same pattern of results was found between ModSWAT and SWAT composites derived from the post-task card sort of the Pre/Post-Test group (Table 11). First, based upon separate analyses, the researcher would arrive at the same conclusion about the effect of task

Table 11.

Threat acquisition results from top level and separate analyses based on the Post-SWAT of the Pre/Post-Test Group

	Top Level Analysis	Separate Analyses	
Source	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.004	.004	.004
Mission Phase (P)	.297	.281	.315
Threat Diff. (ThtD)	.007	.008	.007
Metric (M)	.000	-	-
CxP	.238	.225	.253
CxThtD	.105	.110	.100
CxM	.189	-	-
PxThtD	.143	.150	.136
PxM	.168	-	-
ThtDxM	.018	-	-
CxPxThtD	.011	.012	.010
CxPxM	.437	-	-
CxThtDxM	.433	-	-
PxThtDxM	.203	-	-
CxPxThtDxM	.524	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis. Shaded cells indicate significant results of $p<.05$.

manipulation -- either they were both significant or both non-significant. Second, however, the top level analysis suggests differential sensitivity in that Metric and Threat Difficulty x Metric are significant at the .05 level. The significant main effect for Metric

represents the workload ratings being approximately 2% greater under SWAT than ModSWAT. The ThtD x M interaction is the result of a small differential effect of 0.11 (13.35 for SWAT vs. 13.24 for ModSWAT). This extremely small effect probably represents the extreme sensitivity of the top level analysis -- small error variance associated with highly correlated measures (see Table 12).

Table 12.

Actual difference between SWAT and ModSWAT Means, based on the Post-SWAT of the Pre/Post-Test Group, for each level of Threat Difficulty

Workload Metric	Hard Threat Difficulty	Easy Threat Difficulty	Difference in Workload Rating
SWAT	50.96	37.61	13.35
ModSWAT	48.80	35.56	13.24

Group 2 - Post-Test Only. The correlation between SWAT conjoint values derived using the post-test card sort (Post-SWAT) of the Post-Test Only group and ModSWAT equaled .990 at the level of individual ratings and .998 for condition means. Again, a high correlation exists between the metrics.

Table 13 summarizes the results from the top level 2 (C) x 2 (ThtD) x 3 (P) x 2 (M) analysis of variance as well as subsequent 2 (C) x 2 (ThtD) x 3 (P) analyses performed for each metric. As Table 13 indicates, there were no significant main effects

Table 13.

Threat acquisition results from top level and separate analyses based on the Post-SWAT of the Post-Test Only Group

	Top Level Analysis	Separate Analyses	
Source	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.001	.001	.000
Mission Phase (P)	.070	.088	.089
Threat Diff. (ThtD)	.000	.000	.000
Metric (M)	.870	-	-
CxP	.001	.002	.001
CxThtD	.188	.172	.260
CxM	.788	-	-
PxThtD	.047	.063	.040
PxM	.133	-	-
ThtDxM	.149	-	-
CxPxThtD	.000	.000	.000
CxPxM	.149	-	-
CxThtDxM	.501	-	-
PxThtDxM	.656	-	-
CxPxThtDxM	.567	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis. Shaded cells indicate significant results of $p<.05$

or interactions with Metric using a .05 significance level. This indicates equal sensitivity of metrics. However, based on the separate analyses performed for each metric, the results are identical with one exception. The P x ThtD interaction was significant for SWAT ($p = .04$) and not significant for ModSWAT ($p = .063$). This reversal in significance indicates that different conclusions may be reached when the effect is marginal (i.e. the p value is close to .05). This result is understandable in that slightly

larger effects were found for SWAT than ModSWAT in a number of previous top level analyses. However, this reversal was the lone exception to reaching identical conclusions when separate analyses were performed.

Weapon Delivery

Group 1 - Pre/Post-Test. Table 14 shows the correlations between the two SWAT metrics and the single set of ModSWAT values, both at the level of individual ratings and condition means. Clearly, the two metrics are highly correlated.

Table 14.

Correlation of weapon delivery results between ModSWAT and SWAT as a function of Pre/Post-Test Group

		Pre-SWAT of the Pre/Post-Test Group	Post-SWAT of the Pre/Post-Test Group
Individual Ratings:	ModSWAT	.998	.998
Condition Means:	ModSWAT	.998	.998

Comparison of ModSWAT with the Pre-SWAT Metric. Table 15 summarizes the results of the top level 2 (C) x 2 (TgtD) x 2 (M) analysis and subsequent 2 (C) x 2 (TgtD) analyses for each metric. As shown in Table 15, one would arrive at the same statistical conclusion about the significance of task manipulation when analyzing the workload metrics separately. However, consistent with other results, the top level analysis reveals

Table 15.

Weapon delivery results from top level and separate analyses based on the Pre-SWAT of the Pre/Post-Test Group

	Top Level Analysis	Separate Analyses	
Source	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.105	.105	.106
Target Diff. (TgtD)	.704	.745	.665
Metric (M)	.000	-	-
CxTgtD	.081	.077	.085
CxM	.339	-	-
TgtDxM	.002	-	-
CxTgtDxM	.838	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis. Shaded cells indicate significant results of $p<.05$.

a significant effect of Metric as well as a interaction of Metric with one of the task manipulations (TgtD). The SWAT metric yielded slightly higher workload ratings than did ModSWAT ($\overline{X}=39.51$ and $\overline{X}=36.57$ respectively). Table 16 shows the small differential difference in Metric across Target Difficulty was 0.97 rating points (3.74 for SWAT vs. 2.77 for ModSWAT). Even though this interaction suggests differential sensitivity, it is probably a function of the analysis being extremely sensitive to small differences.

Table 16.

Actual difference between SWAT and ModSWAT Means, based on the Pre-SWAT of the Pre/Post-Test Group, for each level of Target Difficulty

Workload Metric	Hard Target Difficulty	Easy Target Difficulty	Difference in Workload Rating
SWAT	41.38	37.64	3.74
ModSWAT	37.96	35.19	2.77

Comparison of ModSWAT with the Post-SWAT Metric. For the analysis performed on the Post-SWAT of the Pre/Post-Test Group, Table 17 indicates no significant main effects or interactions with Metric. The same conclusions would be made concerning task manipulation when the metrics were analyzed separately. Here, the top level analysis and separate analyses of the two metrics are consistent in showing no differential sensitivity.

Table 17.

Weapon delivery results from top level and separate analyses based on the Post-SWAT of the Pre/Post-Test Group

	Top Level Analysis	Separate Analyses	
Source	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.097	.105	.092
Target Diff. (TgtD)	.712	.745	.680
Metric (M)	.266	-	-
CxTgtD	.103	.077	.138
CxM	.686	-	-
TgtDxM	.331	-	-
CxTgtDxM	.189	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis.

Group 2 - Post-Test Only. Table 18 summarizes the results of the top level 2 (C) x 2 (TgtD) x 2 (M) analysis and subsequent 2 (C) x 2(TgtD) analyses for each metric. Table 18 provides support for the trend established by previous analyses, the pattern of significance regarding task manipulation is the same for both SWAT and ModSWAT. A main effect for Metric does exist, again, with workload ratings being slightly higher for

SWAT ($\overline{X}$=53.11) than ModSWAT ($\overline{X}$=53.01) with a MSE = 4.71. However, no interactions with Metric were uncovered.

Table 18.

Weapon delivery results from top level and separate analyses based on the Post-SWAT of the Post-Test Only Group

Source	Top Level Analysis	Separate Analyses	
	Overall Fprob	ModSWAT Fprob	SWAT Fprob
Configuration (C)	.027	.027	.027
Target Diff. (TgtD)	.270	.263	.278
Metric (M)	.000	-	-
CxTgtD	.162	.160	.166
CxM	.502	-	-
TgtDxM	.824	-	-
CxTgtDxM	.692	-	-

Note that in the table, a dash (-) indicates that the source of variance was not applicable for that analysis. Shaded cells indicate significant results of $p<.05$.

CHAPTER IV

DISCUSSION

The purpose of this study was twofold: (1) to determine if the ModSWAT and traditional SWAT metrics were sensitive to the same differences in task workload associated with performance within a pilot-in-the-loop simulation environment; and (2) to determine what influence, if any, the act of performing a card sort had on workload ratings provided during data collection. Prior to examining the results in light of these two purposes, it is first necessary to discuss overall differences in workload associated with the two metrics.

Overall Difference between the ModSWAT and SWAT Metrics

Across all analyses in which the differences between ModSWAT and SWAT were directly tested, six out of eight main effects for Metric were found significant at the .05 level. In all cases, the average of the SWAT ratings were slightly higher (≈2%) than those for ModSWAT. The fact that such a small difference was significant probably represents a statistical artifact caused by the high correlation among the two metrics. To illustrate this point, consider that with one degree of freedom, the F for Metric is equal to t^2 where the t for dependent samples equals $\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_{\overline{X}_1}^{\ 2} + S_{\overline{X}_2}^{\ 2} - 2(r)S_{\overline{X}_1}S_{\overline{X}_2}}}$.

The denominator portion of the t-test formula, the error term, is known as the standard error of the difference between two dependent sample means. Notice the value of the correlation coefficient appears in the denominator – the value of the error term is reduced by an amount corresponding to the correlation between the levels of the within-subjects factor. Table 19 illustrates how the magnitude of the correlation affected the significance of the effect of Metric for threat acquisition workload data derived from ModSWAT versus the Pre-SWAT of the Pre/Post-Test group.

Table 19.

The effect of varying the magnitude of correlation between ModSWAT and Post-SWAT of the Post-Test Only Group on results obtained from a simple t-test ($t_{crit} = 2.306$).

Magnitude of Correlation	Standard Error of the Difference	t-Test Result
0.1	5.005	0.515
0.2	4.720	0.546
0.3	4.415	0.584
0.4	4.088	0.630
0.5	3.733	0.690
0.6	3.340	0.772
0.7	2.894	0.891
0.8	2.365	1.090
0.9	1.677	1.537
1.0	0.178	14.502

Notice that it is not until the correlation between the two variables approaches 1.0 that Metric becomes significant ($t_{crit.} = 2.306$). This example supports the claim that the high correlation between the ModSWAT and SWAT metrics resulted in very low error variances and caused small differences to be significant. Thus the statistical results of

any top level analyses in which metric was included as a factor may be suspect – an artifact of a extremely sensitive analysis.

One minor implication of the consistently higher workload values yielded by SWAT than ModSWAT deals with the use of a workload value of 40 as a red line. Some researchers have used a SWAT value of 40 as a threshold to indicate an inhibitive level of workload that may adversely affect performance (Reid and Colle, 1988). If future research within the applied environment continues to find a ≈2% difference between the metrics, then adjustments may be necessary to this arbitrary red line value for use with ModSWAT.

Evidence for Equal Sensitivity

The preponderance of evidence indicates equal sensitivity of the two SWAT metrics. First, all of the 14 correlations between ModSWAT and SWAT, whether they be at the level of individual ratings or condition means, were significant and extremely high with the lowest correlation being .990. In the worst case, the two metrics share 98 percent of their variability in common.

Second, the same effects were found to be significant when parallel analyses were performed for each metric separately. Of the 72 effects tested for significance over 16 ANOVAs, the statistical conclusion using the ModSWAT and SWAT metric in separate analyses was the same in 71 (99 percent) of the cases. In only one case was a different statistical conclusion reached -- this difference was found for a marginal effect in which there was a slightly larger effect for SWAT (Fprob = .04) than ModSWAT (Fprob = .06).

Finally, in the top-level analyses, the majority of statistical tests indicated no significant interaction of Metric with the task manipulations. Specifically, 31 of the 35 task x Metric interactions (89 percent) were found to be non-significant at the .05 level indicating the same task effects for the two metrics.

Evidence for Differential Sensitivity

There was some evidence, albeit minimal, for differential sensitivity of the two metrics. However, the differences were so small as to be of little practical importance. Four out of 35 task x Metric interactions were found to be significant at the .05 level. It could be argued that any significant interaction, even one, suggests differential sensitivity. Furthermore, the magnitude of the task differences for these significant effects was consistently greater with SWAT than ModSWAT suggesting that SWAT, in some cases, is more sensitive.

Despite this consistent finding, the average differential difference between SWAT and ModSWAT was 1.25. Within an applied environment, such as pilot-in-the-loop simulation, where stringent experimental control is often difficult to achieve, such a small differential effect would be very difficult to translate into real-world terms, especially for a subjective measure such as SWAT.

Given the small differential effects which were found to be significant at .05, it does not make sense to adopt the alternative .20 significance level as would be the case with testing the Null Hypothesis. Under these conditions, if extremely small differences are found significant at the .05 level, then we would be considering even smaller

differences to be significant at the .20 level. If we question the practical importance of these differential effects at the .05 level, we would most certainly question their practical importance at the .20 level.

Furthermore, in 35 tests of significance for task x Metric interactions, one would expect to find 1.75 significant outcomes just due to chance (using the .05 level). Although the actual number of significant outcomes (n = 4) is greater than the theoretically expected value, a binomial test indicates that it is not significantly above chance level (p = .087). Thus, these four significant outcomes could represent a chance occurrence.

Finally, the fact that very small effects were at all significant can probably be attributed to the overly sensitive statistical analyses (see previous discussion of metric effect) due to the extremely high correlation between metrics. Therefore the author chooses to interpret these significant interactions as a statistical artifact which does not represent a finding of practical importance.

The Effect of Card Sort on Task Ratings

Due to the shortcomings in the IMPACT experimental design caused by the lack of random assignment and an imbalance in the presentation of the threat difficulty and cockpit configuration conditions, the researcher was unable to conduct group comparisons to determine the effect of performing a card sort. The patterns formed by these discrepancies would have caused the results obtained to be suspect and thus difficult to interpret. Unfortunately this prevented the evaluation of card sort effects

within the applied environment. However, the fact still remains that Biers and McInerney (1988) found no interactions for Group (Pre-Test vs. Post-Test) when evaluating the card sort effects, although one must keep in mind that this was done in a highly controlled laboratory environment.

Conclusion and Implications

The present study adds to the data base supporting the validity of using ModSWAT within a less controlled applied environment by replicating laboratory results obtained by Biers and Masline (1987) as well as results obtained from applied research involving a small sample size (Biers, 1995). In a very high percentage of cases, the same conclusion regarding workload can be made regardless of which metric (ModSWAT or SWAT) is used to develop the composite scale.

Two issues still remain that will ultimately determine the applicability of ModSWAT as an alternative composite measure. First, further research is needed to evaluate the effects of performing a card sort on workload ratings. This is necessary to determine whether the act of performing a card sort prior to the Event Scoring phase, within the applied environment, alters the subject's perception of workload. Following the Pre/Post and Post Test group design attempted in this evaluation should address this issue as long as proper counterbalancing is employed. Second, the sensitivity between the ModSWAT and SWAT metrics must be evaluated in cases where the overall Kendall's Coefficient of Concordance is less than .75 and thus requiring custom prototyping. This is necessary to determine whether the ModSWAT metric, in which the composite is always developed in the same manner, maintains equal sensitivity when

compared to the SWAT metric in the event that the group solution is not appropriate and a custom composite is developed based on subject differences between the SWAT dimensions.

Assuming the trend of equal sensitivity between metrics is replicated, researchers will be able to use the ModSWAT metric in place of the traditional SWAT metric, thereby maintaining the benefit of real-time collection of workload data using relatively simple dimensions and rating scale while eliminating the cost (time, money, and resources) associated with performing a card sort. For example, assume a labor rate of $100.00 per hour and the 30 pilots who participated in the C-141 and IMPACT experiments were compensated for their time. One experimenter, one pilot, and one simulator operator/software engineer (often sitting idle while the card sort is being administered) would cost the project $300.00 per hour. Since the card sort takes roughly one hour to complete (assuming it is only done once), the cost to the project would be $9000.00. Add to that, approximately $800.00 (eight hours) needed to construct the composite scale and the total project cost of the SWAT conjoint method would be $9800.00. Since finding qualified subjects is difficult and funding is limited the savings in time and effort afforded by ModSWAT is of practical importance for the applied environment.

REFERENCES

Biers, D. W. (1995). SWAT: Do We Need Conjoint Measurement? Proceedings of the 39th Annual Meeting of the Human Factors Society, 1233-1237.

Biers, D. W., & Masline, P.J. (1987). Alternate approaches to analyzing SWAT data. Proceedings of the 31st Annual Meeting of the Human Factors Society, 63-66.

Biers, D. W., & McInerney, P. (1988). An Alternative to Measuring Workload: Use of SWAT Without the Card Sort. Proceedings of the 32nd Annual Meeting of the Human Factors Society, 1136-1139.

Boucek, G. S., Orr, H. A., Williams, R. D., Montecalvo, A. J., Redden, M. C., Rolek, E. P., & Cone, S. M. (1995). Integrated Mission/Precision Attack Cockpit Technology (IMPACT), Advanced Technology Integration Experiment: Cueing Benefits of Large Tactical Situation Displays, Helmet-Mounted Display, and Directional Audio, Report No. 63797-95U/P61209. Dayton, OH: Veda Inc.

Kantowitz, B. H., & Casper, P. A. (1988). Human Workload in Aviation. Human Factors in Aviation, 164-188. San Diego, CA: Academic Press.

Masline, P. J. (1986). A Comparison of the Sensitivity of Interval Scale Psychometric Techniques in the Assessment of Subjective Workload. Unpublished Master's Thesis. Dayton, OH: University of Dayton.

Moroney, W. F., Biers, D. W., & Eggemeier, F. T. (1995). Some Measurement and Methodological Considerations in the Application of Subjective Workload Measurement Techniques. The International Journal of Aviation Psychology, 5(1), 87-106. Lawrence Erlbaum Associates, Inc.

Reid, G. B., & Colle (1988). Critical SWAT Values for Predicting Operator Overload. Proceedings of the 32nd Annual Meeting of the Human Factors Society, Anaheim, CA.

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. Human Mental Workload, 185-218. North-Holland: Elsevier Science Publishers.

Reid, G. B., Potter, S. S., & Bressler, J. R. (1989). Subjective Workload Assessment Technique (SWAT): A User's Guide, Report No. AAMRL-TR-89-023. Dayton, OH: Armstrong Aerospace Medical Research Laboratory.

Toms, M. L., Cone, S. M., Gier, G. R., Boucek, G. S., & Brown, T. R. (1995). C-141 Cockpit Upgrade Program Volume II: Full Mission Evaluation, Report No. 63803-95U/P61208. Dayton, OH: Veda Inc.

Williges, R. C., & Wierwille, W. W. (1979). Behavioral Measures of Aircrew Mental Workload. Human Factors, 21(5), 549-573.

APPENDIX A

VERBAL SWAT INSTRUCTIONS: PILOT POPULATION

Introduction

Workload Concept

You are probably quite familiar with the concept of mental workload. This is a concept that has become increasingly important in modern high technology aircraft. When we speak of mental workload, we are referring to some sense of mental effort. The basic idea is that we have a finite capacity for performing mental work; and if we exceed this capacity, then we will begin to make a large number of errors or experience total performance breakdown. We can all think of situations where very little effort is required thus leaving us considerable spare capacity for work. Likewise, we can all think of situations that require substantial effort leaving us with little or no spare capacity.

In this study, we are going to measure workload through the use of a scaling approach called the Subjective Workload Assessment Technique, or SWAT. This is a technique that has been developed and extensively tested at the Armstrong Laboratory at Wright-Patterson Air Force Base and has been successfully used in a number of simulation tests, flight tests and OT&Es. This technique is different from most scaling procedures in that there are two parts to it. The first part is called Scale Development and is what we'll doing today; the second part is called Event Scoring which is when you report your workload ratings in the simulator. One of the primary objectives of this technique is to create as little interference as possible during task performance while getting the highest quality data.

Before I start a more detailed explanation of the procedure, I would like for you to quickly read the written instructions. Don't labor over these as I'm going to repeat much of it anyway. I want you to read the instructions first to be sure that I don't forget something important and to provide you with a preview making this easier to follow.

(LONG PAUSE)

For the purposes of SWAT, workload has been defined as being composed primarily of three things: Time Load, Mental Effort, and Psychological Stress. Each of these three factors or dimensions has had three levels defined resulting in a total of 27 possible combinations. Your task today, through a card sort procedure, is to help us determine how these dimensions combine to create your concept of workload. The deck of cards in front of you has a card for each of the possible combinations. Each card has three descriptors written on it; one for time load, one for mental effort, and one for psychological stress. By arranging this deck in an order that represents which combination you think describes the lowest workload condition to the combination that you think represents the highest workload condition and the 25 steps in between, you are helping us create a scale that will reflect the way you think these dimensions combine to create the impression of workload. This is not going to be the same for everyone. Some people think that time is the only element that has any importance in determining workload, while others will say that the only thing of importance is managing the psychological stress. Still others will believe that task difficulty drives workload. This card sort will tell us what your personal interpretation of workload is.

Definitions of Dimensions

Before we start, let me define these three dimensions. Time load is the amount of time pressure experienced in performing your task. This includes the fraction of total time available that you are busy as well as the degree to which different aspects of the task overlap or interfere with one another. Under extreme time load, you are unable to complete the task due to a shortage of time or interference created by an overlap of activities. For example, in an emergency situation, especially in a situation with multiple emergencies, the required actions may be relatively simple and well practiced. The only real problem may be that things happen so fast that you just cannot get everything accomplished before things go from bad to worse.

Mental Effort is the amount of attention and/or concentration required to perform a task. Things that are considered as mental effort include recalling things from long-term memory, decision making, performing calculations. Storing and retrieving things from short-term memory, and problem solving. High levels of mental effort are required in a situation which demands total concentration. While during low levels of mental effort, your mind may wander or your attention may easily be shared with several relatively simple tasks [For example, mental effort could involve such things as recalling a radio frequency that must be selected after passing some navigation point or having to make a decision regarding which of several potential targets should be attacked and what direction to approach a target from on each pass] Another example of mental effort might be the memory load associated with remembering a complex procedure needed to activate a particular piece of equipment. This situation might be intensified if

employment of the equipment is a rare event and, therefore, not as thoroughly learned as something performed routinely.

Psychological Stress refers to the presence of confusion, frustration, and/or anxiety which hinders completion of your task. Psychological stress refers to the feelings of apprehension and tension one usually thinks of when the term stress is discussed. In addition, other factors, such as fatigue, motivation, and physical stresses may also contribute to the feeling of psychological stress. It is well known that physical stresses such as G forces, vibration, temperature, and noise can, when existing in sufficient magnitude, interfere with task performance. At low levels, these stresses may not actually interfere but may provide enough of an annoyance that some of a person's capacity to cope will be expended just to keep the irritation pushed into the background. This level of capacity expenditure would be attributed to the psychological stress dimension of workload.

Description of Levels Within the Dimensions

Now that we have some idea what is meant by the three dimensions, we can begin to discuss the levels within each dimension. Level one is associated with the lowest degree of each dimension, level two is associated with a moderate degree of load for each dimension. and level three is the highest degree of each dimension. Descriptions have been written to precisely define each of the levels for each of the dimensions. The numbers associated with each of these levels will be used by you later when you are doing the event scoring. You've been introduced to the descriptors of each level since you've read the written instructions. Now, as you arrange the card deck in order form

lowest workload situation to the highest workload situation. you will probably refer to these descriptors several times. This will help you become familiar with the meaning associated with each level of each dimension. You are asked to try to think of the wording of the descriptors when you do your ordering rather than trying to use the numbers associated with the levels. The ordering information is very important in helping to define your personal scale but equally important to us is the training value associated with carefully considering the relationships of the meanings of the levels of each of the three dimensions.

Several points need to be made at this stage. Remember that there is not a correct answer. You are making judgments about conditions in teens of the degree of workload associated with an event. This is a communication process that we use which provides a vehicle for you to express the way you view workload in terms that allow us to put numbers on your judgments. There is no right or wrong in this procedure. However, try to be consistent when giving your judgments about events. Because people differ, it is best that you not "compare notes" with anyone. Do not discuss things in an attempt to form a consensus.

As you do the card sort, try to think of an experience that you have had that each card (or set of descriptors) would describe. Then put the cards in order by deciding which of the experiences had the higher workload. Remember, you provide the event so make it something you are familiar with. This process of recalling events helps to establish a scale that is representative of the pilot population's opinion.

Some of the combinations may not remind you of a particular event. It may be

very difficult to think of how you could have the highest level of one dimension while having the lowest level on the other two dimensions. It is true that in most cases the levels of the dimensions will go in the same direction.

However, as this technique has been developed and used, it has been determined that subjects can think of events in which the odd combinations of levels have existed. We suggest that these combinations do exist, but they are rather rare in occurrence. If you simply cannot think of an event for a particular combination of descriptors, then treat it as a hypothetical situation; that is, if it *did* exist, where would it fit in the order. Also remember, we are asking you to judge "how much" work is associated with each card -- not which combination you would prefer to have. It might be clear that one task has a very low level of workload associated with it. In fact, this task might be so low in demand that in your judgment it would be intolerably boring. Someone with a low tolerance for boredom might be tempted to think, "I know this is a low workload task, but I really hate to be bored. This will stress me out, so I'm going to move this task up the order." Remember, we are asking you to rank the relative amount of workload that exists for each situation.

You may use whatever strategy seems best for you to accomplish the card sort. A strategy that has proved useful for many people is to divide the deck into three categories (low, medium, and high), order each of these smaller decks, recombine the decks, and finally, fine tune the resulting order. This strategy is not mandatory.

This is not an easy task. It will probably take you from 30 minutes to an hour to finish and some of the discriminations are going to be difficult. Please concentrate and

give us the best sort possible. Even though this is a laborious process, it will pay off in our analysis. Also when you get to the simulator, the rating task will be easier and more meaningful because of the effort you'll put into this card sort.

If there are no questions, you may go ahead and start. If you have a question now, or a question develops later during your sort, please feel free to ask.

SWAT CARD SORT INSTRUCTIONS: PILOT POPULATION

During the course of this experiment, you will be asked to quantify the mental workload required to complete the mission you'll be flying. Mental workload refers to how hard you work to accomplish some tasks a group of tasks, or an entire job. The workload imposed on you at any one time consists of a combination of various dimensions which contribute to the subjective feeling of workload. The Subjective Workload Assessment Technique (SWAT) defines these dimensions as (1) Time Load, (2) Mental Effort, and (3) Psychological Stress.

For the purposes of SWAT. the three dimensions have been assigned three levels. Each dimension and its three levels are defined below:

Time Load

Time load refers to the amount of spare time that you have available (that fraction of total time that you are busy). When time load is low, sufficient time is available to complete all of your mental work, with some time to spare. As time load increases. spare time diminishes and some aspects of performance overlap and tasks interrupt one another. This overlap and interruption can come from Performing more than one task or from different aspects of performing the same tasks At high levels of time load. Several aspects of performance often occur simultaneously, you are very busy, and interruptions are very frequent. Time load is rated according to the three point scale below:

1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all

2. Occasionally have spare time. Interruptions or overlap among activities occur frequently.

3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.

Mental Effort

Mental effort load is intended to be an index of the amount of attention or mental effort required by a task. Regardless of the number of task or the time limitations. It is strictly an evaluation of the difficulty of the task. When mental effort is low, the concentration required by the task is minimal and performance is nearly automatic. As the demand for mental effort increases due to task complexity or the amount of information that must be dealt with, the degree of concentration and attention required increases. High mental effort demands total attention or concentration due to task complexity or information processing requirements. Mental effort is rated acceding to the three point scale below:

1. Very little conscious mental effort or concentration requited. Activity is almost automatic, requiring little or no attention.

2. Modem conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention is required.

3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.

Psychological Stress

Psychological stress refers to the contribution to total workload of any conditions that produce anxiety, frustration or confusion while performing a task or tasks. At low levels of psychological stress, one feels relatively relaxed. As stress increases confusion, anxiety, or frustration increases and greater concentration and determination are required to maintain control of the situation. Psychological stress is rated on the three point scale below:

1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodate

2. Moderate stress due to confusion. Frustration or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.

3. High to very intense stress due to confusion, frustration, or anxiety. High to extreme determination and self-control required.

Each of the three dimensions (time load, mental effort. and psychological stress) just described contribute to workload during performance of a task or group of tasks. Note that all three factors may be correlated. But need not be. For example, one can have many tasks to perform in the time available (high time load), but they may require little or no concentration (low mental effort). Likewise, one can be anxious and frustrated (high psychological stress), but have plenty of spare time between relatively simple tasks. Since the three dimensions contributing to workload are not necessarily corrected. please treat each dimension individually and give independent assessments of the time load

R00258399S

mental effort, and psychological stress that you experience in performing the following tasks.