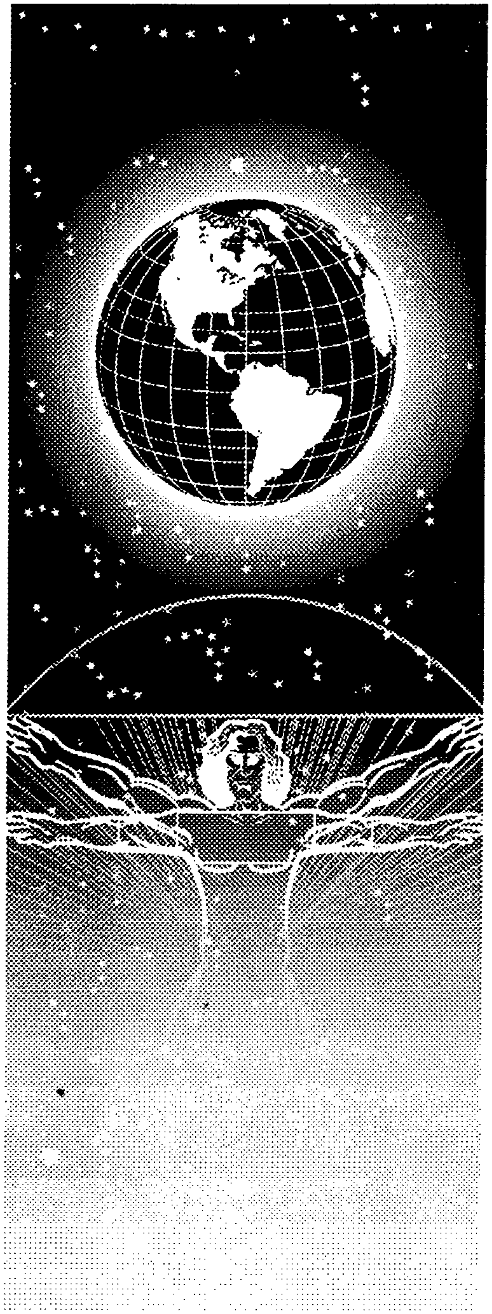
AL/CF-TR-1997-0018



UNITED STATES AIR FORCE
ARMSTRONG LABORATORY

COMPUTER MODELING OF OPERATOR MENTAL WORKLOAD DURING TARGET ACQUISITION: AN ASSESSMENT OF PREDICTIVE VALIDITY (U)

Judi E. See

LOGICON TECHNICAL SERVICES, INC.
P.O. BOX 317258
DAYTON, OH 45437-7258

Michael A. Vidulich

CREW SYSTEMS DIRECTORATE
HUMAN ENGINEERING DIVISION
WRIGHT-PATTERSON AFB, OH 45433-7022

JANUARY 1997

INTERIM REPORT FOR THE PERIOD APRIL 1996 TO DECEMBER 1996

19970909 159

DTIC QUALITY INSPECTED 3

Approved for public release; distribution is unlimited.

Crew Systems Directorate
Human Engineering Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Armstrong Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

TECHNICAL REVIEW AND APPROVAL

AL/CF-TR-1997-0018

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Kenneth R Boff

KENNETH R. BOFF, Chief
Human Engineering Division
Armstrong Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
January 1997

3. REPORT TYPE AND DATES COVERED
Interim, Apr 1996-Dec 1996

4. TITLE AND SUBTITLE
Computer Modeling of Operator Mental Workload During
Target Acquisition: An Assessment of Predictive Validity (U)

5. FUNDING NUMBERS
F41624-94-C-6007
PE 62202F
PR 7184
TA 14
WU 25

6. AUTHOR(S)
*Judi E. See
Michael A. Vidulich

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
*Logicon Technical Services, Inc.
P.O. Box 317258
Dayton OH 45437-7258

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Armstrong Laboratory, Crew Systems Directorate
Human Engineering Division
Human Systems Center
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7022

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AL/CF-TR-1997-0018

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The predictive validity of computer simulation modeling of the operator's mental workload and situational awareness (SA) during a target acquisition mission was assessed in the present study. In Phase I, twelve participants completed a series of target acquisition trials in a laboratory flight simulator and provided subjective ratings of workload (using the Subjective Workload Assessment Technique (SWAT)) and SA (using the Situational Awareness Rating Technique (SART)). In Phase II, computer models of the laboratory task were constructed using the Micro Saint modeling tool. The visual, auditory, kinesthetic, cognitive, and psychomotor components of the workload associated with each task were estimated and used to obtain the measures of average and peak workload. The results from the lab data versus the Micro Saint data were similar but not identical, indicating the computer models were partially, but not completely valid predictors of mental workload and SA. The computer modeling appeared to be a more effective predictor of SA rather than mental workload.

14. SUBJECT TERMS
Operator Workload Measurement, Human Factors,
Human Performance Assessment

15. NUMBER OF PAGES
45

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

PREFACE

This effort was conducted by the Human Interface Technology (AL/CFHP) and the Crew Systems Integration (AL/CFHI) branches of the Armstrong Laboratory at Wright-Patterson Air Force Base, Dayton, Ohio. The project was completed under Work Units 71841425, "Operator Workload Assessment," and 71841044, "Crew-Centered Aiding for Advanced Reconnaissance, Surveillance, and Target Acquisition." Logicon Technical Services, Inc. (LTSI), Dayton, Ohio, provided support under contract F41624-94-D-6000, Delivery Order 0004. Mr. Donald Monk was the Contract Monitor.

The authors wish to acknowledge the support of the Air Force Theater Missile Defense Attack Operations Program Office (ASC/FBXT). In addition, the following individuals should be recognized for their assistance throughout the duration of the project. Gary B. Reid was the task manager, and Gilbert G. Kuperman helped direct the project. Steve Lusk conducted subject training and collected the data during the laboratory flight simulation trials. Jeff Maresh wrote the software for the STORM simulator. Mark Crabtree helped coordinate the activities of all those involved in the project.

The present study represented the initial step toward evaluating the predictive validity of computer simulation modeling of the operator's mental workload and situational awareness (SA) during a target acquisition mission. In Phase I of the study, 12 participants completed a series of target acquisition trials in a laboratory flight simulator and provided subjective ratings of workload and SA, using the Subjective Workload Assessment Technique (SWAT) and the Situational Awareness Rating Technique (SART), respectively. The basic design of the experiment was a 2 (display type) x 2 (threat status) x 2 (target type) repeated measures design. The target was either a transporter-erector-launcher (TEL) or a radar surface-to-air missile (SAM), the latter of which posed more of a threat because it was capable of launching a missile at the operator's aircraft. Threat status referred to the presence or absence of an additional ground threat in the form of an infrared (IR) SAM that, when present, had to be dealt with at the same time as the primary target. Finally, the display was designated as either "high information" or "low information." In the high information condition, a map display showed the locations of the target and the participant's aircraft, while a radar warning display provided information regarding target type. Further, the out-the-window view of the target was red to enhance its salience against the desert background. In the low information condition, the two displays were present but did not provide information regarding target location or type. In addition, the out-the-window view of the target was a less noticeable brown in color.

In Phase II of the study, computer models of the laboratory task were constructed using the Micro Saint modeling tool. The visual, auditory, kinesthetic, cognitive, and psychomotor workload associated with each task comprising a typical trial was estimated via the McCracken-Aldrich 7-point interval level scales. Following model execution, the five component workload estimates were combined to obtain two different measures of the workload during a simulated trial in each experimental condition: the overall or "average" workload (OW) and the maximum or peak workload (PW).

In an attempt to assess the validity of the computer simulation modeling, the results of analyses of variance of the laboratory data were compared with those of the Micro Saint output. Correlational analyses were also conducted. In brief, the results from the lab data versus the Micro Saint data were similar but not identical, indicating that the computer models were partially but not completely valid predictors of mental workload and SA. The nature of the outcomes revealed that the computer modeling may be a more effective predictor of SA rather than mental workload. The results further indicated that the validity of the computer modeling approach might be enhanced by the addition of a "stress" factor that modifies the workload derived from the McCracken-Aldrich scales. Plans for a future study to begin addressing these implications are currently underway.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

One method for assessing system performance that has witnessed recent widespread growth in popularity is computer task network simulation (Hendy, 1994a). In essence, this technique involves decomposing an activity into individual tasks and simulating their completion via computer so that the impact of proposed modifications on system and operator performance can be evaluated. The modeling approach is advantageous in part because the effects of proposed modifications can be evaluated before the alterations are made; hence, if the model indicates that performance or operator workload might be adversely affected, potentially disastrous situations can be averted. Second, the computer model can be executed without the expense of constructing a prototype and running experimental tests with human subjects. Third, the computer model can be much more easily modified than a physical model. Inputs to the computer model can readily be altered to reflect either additional information (e.g., performance data, task durations, etc.) that becomes available or proposed modifications to the system.

Task network simulation has become a particularly widely used technique within the Department of Defense (DoD). In fact, in 1991 the Deputy Secretary of Defense sought to strengthen the application of modeling and simulation in the DoD to promote the effective use of modeling and simulation in training and military operations and in research and development (Kameny, 1995). As part of this initiative, the Defense Modeling and Simulation Office (DMSO) was created in June of 1991 to serve as a center for information concerning DoD modeling and simulation activities. Numerous examples of defense related applications of task simulations testify to the growing recognition of the utility of modeling and simulation to the DoD. Two tools that are frequently used to model crewmember activities and their concomitant performance/workload demands are Task Analysis/Workload (TAWL; Hamilton, Bierbaum, & Fulford, 1991) and the microcomputer version of the Systems Analysis of Integrated Networks of Tasks (Micro Saint, 1996).

TAWL Modeling Tool

The TAWL methodology was originally developed during the concept exploration and definition phase of the system development process for the Army's Light Helicopter Family (LHX) aircraft to compare the workload of one- and two-crewmember configurations of the

LHX. It was specifically equipped to predict operator workload using the techniques developed by McCracken and Aldrich (1984). Their approach to workload is similar to Wickens' multiple resource theory because it proposes humans have not just one but several different information processing resources that can be tapped simultaneously in the completion of a task (Wickens, 1984). Under the McCracken-Aldrich approach, workload is viewed as a multidimensional construct that can be divided into sensory, cognitive, and psychomotor components. The sensory component refers to the complexity of the visual, auditory, or kinesthetic stimuli to which the operator must attend. The cognitive component refers to the level of information processing required from the operator. Finally, the psychomotor component refers to the complexity of the operator's behavioral responses. At any given time, the workload experienced by an operator may stem from one or more of these five distinct sources (i.e., the visual, auditory, kinesthetic, cognitive, and psychomotor components). The workload associated with a given task can be estimated by rating each of the five components separately on interval scales developed by Bierbaum, Szabo, and Aldrich (1987) that range from 0 (low workload) to 7 (very high workload). For a task, any combination of ratings can result, such that the workload associated with some components might be very high while the workload for others might be low or nonexistent.

Prior to executing a model in TAWL, the user must identify a mission of interest and decompose it into progressively smaller units referred to as phases, segments, functions, and tasks. The task represents an event or activity that can be specified in terms of a verb-noun combination (e.g., check gauge, select sensor, set range). It is the fundamental unit of analysis in TAWL. Performance times for each task are estimated as is the workload experienced by the crewmember who completes the task. The model is developed by delineating function decision rules that control the sequencing of tasks within each function as well as segment decision rules that govern the sequencing of functions within segments. Finally, the model is executed using the TAWL Operator Simulation System (TOSS) computer software. The simulation produces estimates of each crewmember's visual, auditory, kinesthetic, cognitive, and psychomotor workload during each half-second period of the mission. When multiple tasks are performed simultaneously, the workload for a particular component is the sum of the ratings across the tasks being completed at that moment in time. Hence, so-called overload conditions with ratings that exceed 7.0 may occur throughout the mission. In this way, the TAWL/TOSS system can be used to identify periods of high workload, crewmembers who experience excessive workload, and components with unusually high workload. This information can subsequently be used to determine the feasibility of adjusting the distribution of tasks throughout the mission, among crewmembers, or among components in an attempt to moderate workload levels.

Micro Saint Modeling Tool

Micro Saint is another modeling tool that has frequently been applied in defense-related assessments. Of the many computer software packages that support task network modeling, it has proven to be one of the most popular (Hendy, 1994a). The development of Micro Saint began in 1984 when the U.S. Army Medical Research and Development Command sponsored Micro Analysis and Design to develop a user-oriented simulation system that could be run on a microcomputer (Laughery, 1989). What evolved was a general purpose modeling tool targeted primarily for a human engineering audience. While it was not designed for the specific purpose of analyzing operator workload, Micro Saint's versatility makes it perfectly amenable to such analyses. Micro Saint's basic operator interface is a graphical interface which allows information to be input via typing, pointing and clicking with the mouse, or selecting options from available menus. Briefly, a model is constructed in Micro Saint by (1) drawing the tasks on the screen with the tools provided by Micro Saint, (2) entering task attributes such as workload and the mean, the standard deviation, and the shape of the distribution (e.g., normal, gamma, exponential) of the task completion times, and (3) establishing pathways to connect the tasks and control their sequencing. The task attributes are used to depict operator or system performance, whereas the pathways represent the relationships between the tasks in the network. Many different routes through the network become possible as a result of both the user-defined branching (probabilistic or tactical) between tasks and the variability in task completion times. Hence, each execution of the model will yield different results. Because variability is built into the network, the results of repeated simulations are likely to be indicative of the performance of real-world systems which are themselves characterized by human operator variability.

As stated earlier, use of either the TAWL or Micro Saint computer modeling tools is advantageous because the models are relatively easy to construct, modify, and execute. In general, they are easier to implement than experimental studies which require the participation of human subjects. Hence, computer modeling can save time and effort. One major problem blocking more widespread usage of computer modeling approaches to experimental and system design is the paucity of evidence regarding their predictive validity.

Predictive Validity of Computer Modeling

To date, relatively few investigations of the predictive validity of computer modeling have been conducted. Some data were obtained in a study wherein the TAWL/TOSS methodology was used to complete a task analysis of a UH-60 combat mission (Bierbaum, Szabo, & Aldrich, 1989; Iavecchia, Linton, Bittner, Jr., & Byers, 1989). Nine phases, 34 segments, 48 functions, and 138 tasks were included in the analysis. The resulting baseline model was used to evaluate the total workload experienced by each crewmember for the current UH-60 aircraft so that the impact of proposed modifications to the aircraft on crewmember workload could later be evaluated. Elements of the model were later incorporated into an investigation designed to assess the predictive validity of computer modeling (Iavecchia, Linton, Bittner, Jr., & Byers, 1989). In the ensuing validation study, operator workload in a UH-60A Black Hawk simulator was compared to the workload estimates derived from the TAWL/TOSS computer simulation during each segment of the mission. The analysis was conducted by computing and comparing two measures of workload derived from either operator ratings or TAWL output: Overall Workload (OW) and Peak Workload (PW). Following the flight simulation, operators were asked to provide both a rating of the overall amount of workload (OW) and the peak workload (PW) they had experienced during each segment on scales ranging from 0 (very low workload) to 100 (very high workload). In terms of the TAWL/TOSS computer simulation, OW was derived for each half-second interval in the mission by *averaging* across all five component workload estimates; a segment OW measure was then obtained by averaging all of the means within a segment. PW was derived by *summing* the five component workload estimates at each half-second interval and then selecting the maximum or peak workload within the segment. The results revealed that correlations between TAWL-based predictions and crew results were substantial for OW ($r = .82, p < .01$), but somewhat lower for PW ($r = .62, p < .05$). Further, despite the high degree of association, TAWL-based predictions of OW consistently underestimated the ratings provided by human crewmembers.

In a more recent attempt to assess the validity of computer simulation modeling, Lawless, Laughery, and Persensky (1995) studied the human performance effects of nuclear power plant modifications. Specifically, they used Micro Saint models to examine the difference between the "paper procedures" currently followed in the control room and the new "computerized procedures" that were under consideration but had not yet been implemented. At the same time, traditional experimental tests with human subjects were being conducted in a nuclear power plant control room environment at North Carolina State University to evaluate whether "paper procedures" differed from "computerized procedures." The primary goal of the study was to establish the predictive validity of task network modeling by determining whether the results of the Micro Saint simulations matched those from the experimental tests.

Both paper and computerized procedures for a normal regulatory maneuver and two different accident scenarios were evaluated in both the experimental study and the Micro Saint simulation, providing a total of six conditions in each study. The normal operating conditions involved a routine change of power operation. The two accident scenarios represented a small break loss of cooling accident (LOCA) and a steam generator tube rupture (SGTR). In all three cases, the dependent variable of interest was the time required by the team to complete the preliminary and final phases of the task. Task performance times for the "paper procedures" Micro Saint model were generated from available empirical data. Comparable times for the proposed "computerized procedures" were developed via expert judgment based on the estimated impact of the new procedures on each of the tasks. Each Micro Saint model was executed 5000 times.

A direct comparison of the "computerized" task performance times from the experimental study and those predicted by the Micro Saint simulation for both the preliminary and final procedures of the three scenarios revealed that the two sets of results were significantly different only in the case of the LOCA accident scenario (both preliminary and final procedures). In both cases, the model's predicted performance times underestimated the response times observed in the experimental study. In the two remaining scenarios, the average performance times predicted by the Micro Saint model did not differ from those actually obtained in the empirical study. Thus, the model values matched the empirical values in four of the six possible conditions. The authors concluded that while task network models are easily constructed and readily modified, their predictive validity is not yet sufficiently high to permit a definitive declaration of the success of the modeling approach.

The Present Investigation

The purpose of the current study was to assess the validity of computer simulation modeling for predicting mental workload and situational awareness (SA) during target acquisition in a simulated air-to-ground combat scenario. The project was completed in two stages. In Phase I of the study, 12 subjects participated in a series of target acquisition trials in a laboratory flight simulator, each of which lasted approximately 100 seconds. On each trial, the subject was instructed to fly to and hit a waypoint before taking a pre-specified heading to acquire and destroy the primary target. Factors that might be expected to influence mental workload or SA were manipulated from trial to trial. These included display type, the presence of ground threats, and target type. At the conclusion of selected trials, workload and SA were assessed by means of subjective rating scales. In Phase II of the study, computer models of each experimental condition from the laboratory task were constructed using the Micro Saint modeling tool. The workload associated with each task was estimated via the McCracken-Aldrich approach and used to derive measures of OW and PW in each experimental condition. In an attempt to assess the validity of the computer simulation modeling, the results of statistical analyses of the laboratory data were compared with the results from the Micro Saint output. If computer simulation modeling is valid, the two sets of outcomes should be comparable.

METHOD

Laboratory Study

Participants.

The participants included seven males and five females recruited from local universities in Dayton, OH. They were between the ages of 20 and 32 years, and all of them reported having normal or corrected-to-normal 20/20 vision and normal hearing.

Design.

The basic design was a 2 (display type) x 2 (threat status) x 2 (target type) repeated measures design. The target was either a radar surface-to-air missile (SAM) or a transporter-erector-launcher (TEL), the former of which posed more of a threat because it was capable of launching a missile at the participant's aircraft. Individuals were required to execute a "jink" maneuver to evade the missile on Radar SAM trials, an action that was not necessary on TEL trials. A second independent variable, threat status, referred to the potential presence of an additional ground threat in the form of an infrared (IR) SAM on some trials that had to be dealt with at the same time as the primary target. Finally, the display was designated as either "high information" or "low information." In the "high information" condition, subjects were provided not only with a map display of the area showing the locations of the waypoint and target but also with information regarding target type. Further, the out-the-window view of the target against the desert background was red. In the "low information" condition, the map was present and contained the symbology for the waypoint but no additional information regarding target location or type. In addition, the out-the-window view of the target was brown, making the vehicle more difficult to detect in the sandy terrain.

Apparatus.

The primary apparatus used in the laboratory study was the Simulator for Tactical Operations Research and Measurement (STORM). The STORM apparatus was designed to simulate the cockpit of a one-seater aircraft engaged in air-to-ground attack. The simulator contained a force stick, a throttle, a tactical situation display (TSD), a radar warning display, and an 8 ft x 6 ft projection screen presenting the out-the-window view of the scene in front of the aircraft as well as a head-up display (HUD). The **force stick**, which was located to the right of the pilot's seat, was used during flight for controlling attitude (pitch and roll); in addition, various switches and triggers on the stick were also used to select and fire the weapon, to zoom/unzoom the TSD, and to select right/left views out-the-window when necessary. The **throttle**, positioned to the left of the pilot's seat, was used mainly during flight to control the speed of the aircraft; however, it also contained a button that participants used to dispense flares whenever an IR SAM was encountered. The **TSD** was displayed on an 11 in. x 8 in. Unisys VGA monitor, which was positioned above and to the right of the pilot's seat. The TSD provided a map of the surrounding terrain and a yellow "+" symbol designating the location of the waypoint. A generic aircraft symbol also appeared on the display to provide continuous feedback regarding the location of the aircraft. On "high information" trials, the TSD further contained a red triangle representing the location of the target. The map was a track-up map which rotated and translated so that the aircraft symbol was always located in the center of the display and the current heading was positioned towards the top of the TSD. The **radar warning display** was part of the F-16 Air Intercept Trainer system from which the STORM simulator was adapted. It consisted of a circular green monochrome display with a 5 in. diagonal viewing area located to the pilot's left.

This display was used to identify target type on the "high information" trials. Specifically, if the target was a Radar SAM, an encircled "9" appeared on the radar warning display on "high information" trials but not otherwise. Finally, the **HUD** presented all necessary flight information, including attitude, altitude, speed, heading, and weapon selected as well as a weapon aimsight box and weapon in-range indicators. The simulated HUD display was superimposed on an out-the-window, daytime, desert view, which was rear projected onto the screen by an Electrohome ECP4000 system.

Two rating scales were used to collect data regarding the mental workload and SA associated with performing the tasks comprising selected simulated flight trials on the STORM simulator. The Subjective Workload Assessment Technique (SWAT; Reid & Nygren, 1988), a subjective rating scale developed by the U. S. Air Force Armstrong Aeromedical Research Laboratory, was used to assess mental workload. The scale is comprised of three factors, each of which has three levels ranging from low to high workload. Time Load refers to the total amount of time available to accomplish a task as well as overlap of tasks or parts of tasks. Mental Effort Load is the amount of attention or concentration needed to perform a task. Finally, Psychological Stress Load refers to the presence of confusion, frustration, or anxiety during task performance. Application of the SWAT proceeds in two phases: Scale Development and Event Scoring. The Scale Development phase is used to produce an interval-level workload scale that is normalized for each participant. In this stage, the participant rank orders the 27 possible combinations of the three different levels of Time Load, Mental Effort Load, and Psychological Stress Load from lowest to highest workload by means of a card sorting technique. The rankings are then subjected to a series of tests in order to identify a rule for combining the dimensions for each individual. The resulting interval-level workload scale ranges from 0 (low workload) to 100 (highest workload possible). The Event Scoring phase represents the data collection stage in which individuals evaluate the workload associated with task performance by providing ratings for each of the three dimensions. These ratings are later converted to the normalized workload scale using the rule identified during Scale Development. Workload can be assessed by analyzing the overall score from the normalized scale or by evaluating the ratings on each of the three dimensions.

Another subjective rating scale, the Situational Awareness Rating Technique (SART; Selcon & Taylor, 1990; Taylor, 1990), was used in the laboratory task to assess SA. This scale was developed by asking experienced aircrew to identify factors that affect SA. Thus, SA was defined empirically, rather than a priori, on the basis of the knowledge and experience of the aircrew. Ten independent bipolar dimensions of SA, which could be grouped into three major categories, emerged from analysis of their responses. These dimensions included:

(A) Demand on attentional resources
  (1) Instability of situation
  (2) Variability of situation
  (3) Complexity of situation
(B) Supply of attentional resources
  (4) Arousal
  (5) Spare mental capacity
  (6) Concentration
  (7) Division of attention
(C) Understanding
  (8) Information quantity
  (9) Information quality
  (10) Familiarity

The utility of the SART in assessing human performance has been demonstrated in a variety of skill, rule, and knowledge-based tasks, including tracking and monitoring aircraft HUD flight parameters, unusual aircraft attitude recovery, comprehension of aircraft warnings, and aircraft flight simulation (Selcon & Taylor, 1990; Selcon, Taylor, & Koritsas, 1991; Taylor & Selcon, 1990). The scale can be administered in either its ten dimensional (10-D SART) or three dimensional (3-D SART) form, depending upon the degree of intrusiveness permitted by the task. The 3-D SART is most appropriate when it is necessary to minimize interference with dynamic tasks, such as flight simulation and flight trials. The 10-D SART is most useful when specificity and diagnosticity are important. The results of three studies reported by Selcon and Taylor (1990) indicated that the 3-D SART provides a meaningful, low-intrusive alternative to the more time-consuming 10-D SART. In the current study, the 3-D version of the SART was applied to avoid prolonging the already lengthy experimental sessions. Individuals were asked to supply ratings for each of the three dimensions on 7-point scales. The ratings were subsequently used to derive a single measure of SA with the following formula:

$$SA(c) = U - (D - S) \qquad [1]$$

where *SA(c)* represents calculated SA; *U* is rated Understanding; *D* is rated Demand; and *S* is rated Supply.

Procedure.

All participants received between 16 and 22 one-hour blocks of training in order to meet a basic set of skill requirements prior to completing the data collection trials. The specific flight skills that each individual was required to master are enumerated in detail by Crabtree, Marcelo, McCoy, and Vidulich (1993). These included basic aircraft control, navigation, target acquisition and destruction, missile evasion, and the use of multiple displays. Participants were first given a brief description of the nature of the experiment before receiving instruction on the HUD, the control stick and throttle, and the switches and buttons on each device. Naive participants were told that the primary objective for the first few hours of training was simply familiarization. Pressure to perform well at this time was minimal or absent. Individuals were instructed as needed to execute basic flight tasks (e.g., roll, pitch, coordinated turns) but were also encouraged to "experiment" with the simulator to get a feel for its characteristics. Following this period of familiarization, training proceeded in three stages representing a progression of sortie difficulty.

Stage 1 consisted of a series of 10 min sorties in which participants flew at a leisurely pace from site to site and attempted a gun strike. A Radar SAM and a nearby IR SAM comprised each site. The assigned altitude between sites was 3500 ft, and the airspeed upon weapon delivery was approximately 450 knots indicated air speed (KIAS). The display configuration was identical to the "high information" condition in the subsequent experimental session (i.e., red targets and full tactical information). The aircraft was nearly invulnerable. Once their skill level had increased, participants were instructed to use the High-Speed Anti-Radiation Missile (HARM) against the Radar SAM.

Stage 2 of training was an exposure to several of the "warm-up" scenarios that were used as practice trials during each experimental session. In essence, participants were introduced to the basic framework of the experimental trials during this phase of training. Their goal was to acquire and destroy a waypoint and then follow the assigned heading to acquire the primary target. In contrast to Stage 1, Stage 2 required that subjects begin active, aggressive control immediately. Upon destroying the waypoint (or doing a close fly-by), the rules from Stage 1 for altitude and airspeed applied. A post-waypoint heading was given before the trial began. There were two levels of display configuration to match the experimental design (i.e., low information and high information). Individuals were also introduced to the concept that a red triangle on the TSD could represent either a Radar SAM or a TEL.

Finally, Stage 3 of training was an exposure to a few of the "maintenance" scenarios that were also included in the experimental trials. Everything that subjects had experienced up to this point during Stages 1 and 2 was included; in addition, multiple threats as well as benign targets were now present. Subjects were instructed to complete the mission as before; however, they were warned that they would encounter heavy resistance in the form of ZSU-23 AAAs.

Following training, each participant completed the 56 trials comprising the main experimental session, which was subdivided into 4 blocks of 14 trials. Three types of trials were included in each block: two warm up trials, four maintenance trials, and eight experimental trials. A block always commenced with the two warm up or practice trials (one with a Radar SAM target and one with a TEL target), followed by the presentation of the maintenance and experimental trials in a unique random order for each participant. Workload and SA data were collected only during all 16 of the experimental trials comprising Blocks 2 and 3--the SWAT on half of the trials and the SART on the remaining half. The eight experimental trials in each block differed in terms of display type (low versus high information), threat status (absence versus presence of IR SAMs), and target type (Radar SAM versus TEL). One trial per experimental condition was presented in each block. Maintenance trials were included solely to introduce additional variety for the participants and prevent boredom with the task. On these trials, the primary target was again either a Radar SAM or a TEL, but many additional threats in the form of AAAs, outbuildings, and tanks were also present in the path from the waypoint to the target. Participants were instructed to strike as many threats as possible on maintenance trials without preventing attack of the primary target.

On each trial the basic task was to fly to and hit a waypoint before taking a pre-specified heading to acquire and destroy the primary target, which was either a Radar SAM or a TEL. At the start of each trial, the individual was verbally informed of the heading to be taken after the waypoint. On the low information trials in which there was no tactical display, this was the only information regarding target location that was available to the participant. On the high information trials, a red triangle on the TSD provided additional information to help guide the individual to the target. While airspeed and altitude prior to the waypoint were arbitrary, participants were instructed to maintain 450 knots and 3500 ft thereafter. Enroute to the target, participants were engaged in a number of activities, one of which involved scanning the out-the-window view for their first glimpse of the target. In the high information condition, the target was red in color, making it readily noticeable against the sandy terrain. In the low information condition, on the other hand, the target was dark brown, causing it to blend in more with the background and making it difficult to detect. Hence, the task of locating the target was exacerbated both by the absence of the TSD and by the color of the out-the-window view of the target in the low information condition. Also enroute to the target, participants attempted to determine as early in the trial as possible what type of weapon they should select--the HARM missile was to be used for the Radar SAM, whereas the gun was to be used for the TEL. This involved monitoring the radar warning display for the Radar SAM symbology and remaining alert for the Radar SAM's auditory detection and launch warnings. Finally, individuals also needed to remain alert for the occurrence of IR SAMs between the waypoint and the target. The IR SAM did not appear on either the radar warning display or out-the window. The only indicator of the presence of an IR SAM was an auditory warning in the form of a rapid beeping. This launch warning required the participant to respond by dispensing flares.

Participants were required to execute a different series of activities, depending on which type of target was present on a given trial. On **Radar SAM** trials, the symbology appeared on the radar warning display at about 45,000 ft from the target, but only on the high information trials. Thus, on the high information trials, participants received early warning that the target was a Radar SAM and could select the appropriate weapon early in the trial. On low information trials, when the radar warning display was inactive, they did not have sufficient information to determine target type until they were at close range to the target. Specifically, at a distance of approximately 15,000 ft, a detection warning in the form of a slow pulsating tone occurred, regardless of display type, to indicate that the Radar SAM was tracking the aircraft. Around 10,000 ft, a faster tone sounded, indicating that the Radar SAM had launched its missile. At this point, the participant was required to engage in the "jink" maneuver (i.e., a sharp descent followed abruptly by a sharp ascent) in order to evade the missile. Following successful evasion, the individual attempted to re-acquire the target and fire the HARM missile. The trial ended shortly after the attempted strike, regardless of whether it was successful or not.

In contrast to the Radar SAM trials, there were no visual or auditory warnings on **TEL** trials (other than those associated with the IR SAM), and the TEL did not fire on the participant's aircraft. Participants ascertained target type either by the absence of the Radar SAM symbology on the radar warning display (on high information trials) or by the absence of the Radar SAM auditory detection and launch warnings (on both low and high information trials). Once they had selected the gun, they monitored the HUD for the weapon-in-range indicators and attempted to center the aiming reticle over the target. Regardless of target type, participants could determine that they were in range by detecting any one of three weapon-in-range indicators: a change in the color of the aiming reticle from white to blue; the appearance of the text "In range" on the HUD; or a sufficiently small straight line distance indicator on the HUD. As on the Radar SAM trials, a TEL trial ended shortly after the attempted strike, regardless of its outcome.

In summary, participants were cognizant of the factors that might vary from trial to trial, but they were not certain which type of trial they would encounter when it began. Thus, their task was to determine target type as soon as possible so that they might select the appropriate weapon as well as to remain alert for the presence of IR SAMs. In the high information condition, they were assisted in this task by the presence of the tactical display, which designated the location of the target, and by the radar warning display, which designated target type. In addition, in the high information condition, participants were further assisted by a more readily visible red target in the out-the-window view. In the low information condition, the task of locating the target and determining its type was more difficult--there was no symbology on the TSD to indicate target location and no symbology on the radar warning display to indicate target type. Further, the brown out-the-window view of the target made it harder to detect, requiring prolonged visual search.

Computer Simulations

Model construction.

Micro Saint models representing each experimental condition were constructed by first identifying the tasks comprising each type of trial, the actions necessary for their completion, and their sequence throughout a trial. This step was accomplished by observing several participants as they completed the various data collection trials and by interviewing the experimenter to obtain additional detail when needed. Descriptions of each task were then written and used in conjunction with the descriptions of the McCracken-Aldrich interval level scales to obtain estimates for the auditory, visual, kinesthetic, cognitive, and psychomotor workload components. In addition to the component workload estimates, execution times were determined for each task. Means and standard deviations were obtained from the experimental data where possible as well as from other reports documenting similar task analyses (e.g., Hendy, 1994b). Some estimates of task timing and duration (e.g., the frequency of monitoring the TSD for the target location) were also derived on the basis of information provided during post-experimental interviews with five of the participants. Finally, the type of distribution from which the task completion times were sampled during each simulation run was determined for each task. The gamma distribution was used for all tasks involving discrete activations (e.g., pressing the button to fire the gun). This type of distribution is ideal for tasks such as discrete activations that generally cannot be performed much more quickly than the mean but could potentially take much longer. The normal distribution was used for all other types of tasks, which could conceivably be completed either more slowly or more quickly than average (e.g., scanning the TSD; searching out-the-window for the target).

For the sake of convenience, the model was subdivided into five smaller subnetworks and entered in Micro Saint. The five segments included (1) from start to waypoint, (2) from waypoint to the out-the-window view of the target, (3) from the out-the-window view to the target, (4) during target destruction, and (5) after the target had been fired upon. This was done so that the duration of each segment would approximate that of the laboratory trial as closely as possible.

Model execution.

Each model was executed 25 times, a value that was selected to achieve power of at least .80 in subsequent statistical analyses. Component workload estimates were obtained for each half-second interval of each model. The resulting data file was edited and transported to a PC-based version of the Statistical Analysis System (SAS, 1992), where OW and PW were computed from the component workload estimates.

RESULTS

Laboratory Study

Performance results.

While the major dependent variables of interest in this study are workload and SA, several performance results will be considered first, as they may contribute toward understanding the workload and SA findings. Specifically, we examined both the probability of target kill and the probability of crashing before trial completion in each experimental condition. Means and standard deviations for target kills and crashes are depicted in Table 1 for target type, threat status, and display type. The figures in the table reveal that neither target kills nor crashes appeared to be dependent upon display type or threat status. In both cases, however, the probabilities did differ depending on target type. The probability of target kill was greater when the target was a TEL, and the probability of the aircraft crashing was greater for the Radar SAM. These outcomes are not surprising since the Radar SAM was able to fire on the participant's aircraft, increasing the chances that the individual might crash or be killed while attempting to evade the Radar SAM's missile and decreasing the likelihood of successful target destruction.

Table 1

Means and Standard Deviations (in Parentheses) for Probability of Target Kill and Probability of Aircraft Crash by Target Type, Display Type, and Threat Status

	P(TARGET KILL)	P(AIRCRAFT CRASH)
TARGET TYPE		
RADAR SAM	.36 (.28)	.22 (.24)
TEL	.85 (.18)	.07 (.13)
DISPLAY TYPE		
HIGH INFORMATION	.62 (.34)	.13 (.21)
LOW INFORMATION	.59 (.34)	.15 (.22)
THREAT STATUS		
IR SAMs ABSENT	.64 (.34)	.14 (.23)
IR SAMs PRESENT	.57 (.34)	.14 (.19)
MEAN	.61 (.34)	.14 (.21)

The statistical significance of the means in Table 1 was tested via separate 2 (target type) x 2 (display type) x 2 (threat status) repeated measures analyses of variance (ANOVAs). For both target kill and terrain crashes, the main effect for target type was significant, $F(1,11) = 64.45, p < .001$, and $F(1,11) = 5.89, p < .034$, respectively. In addition, for target kill, the interaction between target type and threat status was also significant, $F(1,11) = 6.06, p < .032$. No other sources of variance in either analysis were statistically significant ($p > .05$). The interaction, which is portrayed graphically in Figure 1, indicated that the probability of target kill did not differ depending on threat status when the target was a Radar SAM, possibly due to floor effects (i.e., the low overall probability of target kill for Radar SAM trials could not be significantly reduced by the presence of the IR SAM threats). However, when the target was a TEL and the overall probability of target kill was much greater, participants were even more likely to destroy it successfully when the IR SAMs were absent.

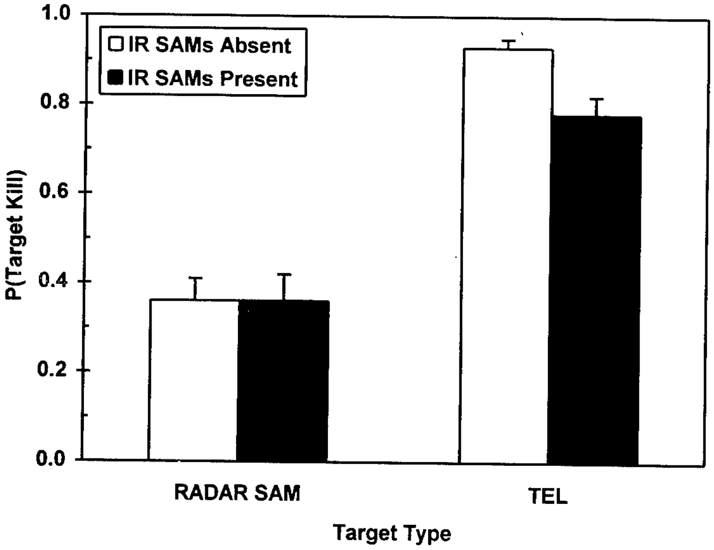


Figure 1. Probability of target kill for Radar SAMs and TELs in the absence and presence of IR SAM threats. (Note: error bars represent the standard error of the mean.)

SWAT workload ratings.

Each participant's SWAT and SART ratings were used to determine the effects of target type, display type, and threat status on workload and SA, respectively. Means and standard deviations for the overall SWAT score as well as the three subscales appear in Table 2 for target type, threat status, and display type. As can be seen in the table, the participants reported experiencing considerably higher workload when the target was a Radar SAM as opposed to a TEL. Both the overall SWAT rating and the ratings for the individual subscales were higher for the Radar SAM. Further, workload was slightly higher in the low information condition. The mean scores indicated that this condition was associated with higher ratings for EFFORT and STRESS but lower ratings for TIME than was true for the high information condition. Finally, the mean SWAT score as well as the ratings for all three subscales were greater when IR SAMs were present.

Table 2

Means and Standard Deviations (in Parentheses) for Time, Effort, Stress, and Overall SWAT Rating by Target Type, Display Type, and Threat Status

	TARGET TYPE		**DISPLAY TYPE**		**THREAT STATUS**	
	RADAR SAM	TEL	HIGH INFORMATION	LOW INFORMATION	IR SAMs ABSENT	IR SAMs PRESENT
TIME	1.54	1.25	1.44	1.35	1.25	1.54
	(0.58)	(0.44)	(0.58)	(0.48)	(0.44)	(0.58)
EFFORT	1.77	1.52	1.56	1.73	1.56	1.73
	(0.59)	(0.54)	(0.65)	(0.49)	(0.54)	(0.61)
STRESS	1.79	1.29	1.44	1.64	1.38	1.71
	(0.65)	(0.46)	(0.58)	(0.63)	(0.60)	(0.58)
SWAT	33.66	16.09	23.40	26.34	17.90	31.84
RATING	(22.64)	(18.31)	(23.76)	(20.89)	(20.36)	(22.17)

The mean overall SWAT ratings were subjected to a 2 (target type) x 2 (display type) x 2 (threat status) repeated measures ANOVA. The main effects for target type and threat status were statistically significant: $F(1,11) = 15.71, p < .0022$ and $F(1,11) = 11.26, p < .0064$, respectively. However, the effect for display type did not attain statistical significance, $F(1,11) = .50, p > .05$. Of the two-way and three-way interactions, only the interaction between display type and threat status was significant, $F(1,11) = 8.06, p < .0161$. The nature of the Display Type x Threat Status interaction is portrayed graphically in Figure 2. As can be seen in the figure, the effect of IR SAM presence on workload was minimal in the low information condition. On the other hand, when the tactical display was present, there was a relatively large difference in workload, depending upon whether IR SAMs were absent or present.

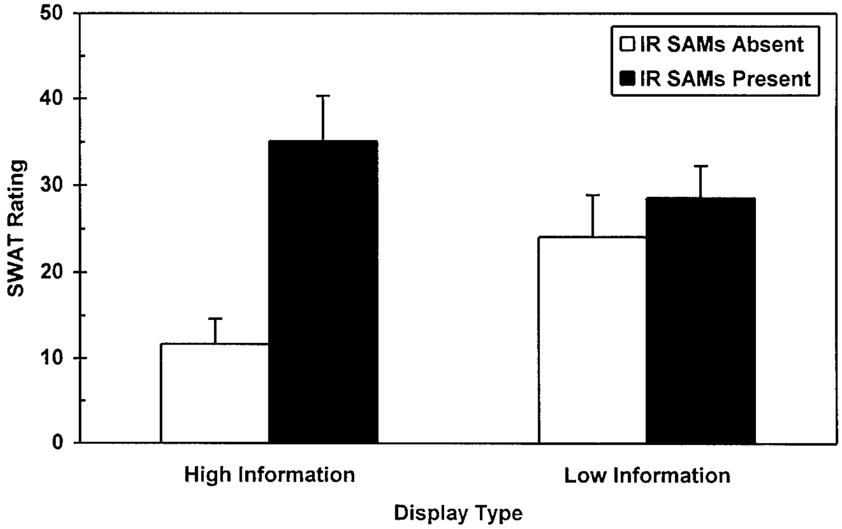


Figure 2. Mean SWAT rating for the high and low information displays when IR SAMs were absent and present. (Note: error bars represent the standard error of the mean.)

Additional 2 (target type) x 2 (display type) x 2 (threat status) repeated measures analyses of variance were conducted on the TIME, EFFORT, and STRESS ratings to determine which of the three subscales contributed to the observed differences in SWAT ratings. First, the analysis of the EFFORT subscale revealed no significant main effects or interactions, $p > .05$, implying that any variations in overall workload among conditions were not due to differences in EFFORT. Second, in the analysis of TIME, the main effects for target type and threat status were significant as was the Display Type x Threat Status interaction: Fs(1,11) = 13.15, $p < .004$; 10.17, $p < .0086$; and 7.33, $p < .0204$. Third, in the analysis of STRESS, only the main effects for target type and threat status attained statistical significance: Fs(1,11) = 11.48, $p < .0061$ and 18.53, $p < .0012$. Hence, the outcomes from the last two analyses indicated that the differences in overall SWAT ratings for target type and threat status were attributable to differences on the TIME and STRESS subscales.

SART situational awareness ratings.

In addition to the SWAT ratings, the SART ratings provided by each participant were used to examine the effects of each independent variable on SA. Means and standard deviations for the overall SART score as well as the three subscales appear in Table 3 for target type, IR SAM presence, and display type. As can be seen in the table, SA was relatively greater when the target was a TEL versus a Radar SAM; when the display provided high versus low information; and when IR SAMs were absent versus present. The means in Table 3 further reveal that the conditions in which SA was enhanced were associated with lower demand (D) scores and higher supply (S) and understanding (U) ratings.

Table 3

Means and Standard Deviations (in Parentheses) for Demand (D), Supply (S), Understanding (U), and Overall SART Rating by Target Type, Display Type, and Threat Status

	TARGET TYPE		**DISPLAY TYPE**		**THREAT STATUS**	
	RADAR SAM	TEL	HIGH INFORMATION	LOW INFORMATION	IR SAMs ABSENT	IR SAMs PRESENT
D	3.94	2.54	3.12	3.35	3.08	3.40
	(1.34)	(1.09)	(1.42)	(1.39)	(1.41)	(1.40)
S	3.69	4.60	4.56	3.73	4.46	3.83
	(1.39)	(1.45)	(1.43)	(1.44)	(1.50)	(1.42)
U	5.04	5.19	5.69	4.54	5.35	4.88
	(1.50)	(1.52)	(1.17)	(1.60)	(1.39)	(1.59)
SART RATING	4.73	5.23	5.38	4.58	5.14	4.81
	(1.08)	(1.29)	(0.98)	(1.30)	(1.22)	(1.20)

A 2 (target type) x 2 (display type) x 2 (threat status) repeated measures ANOVA of the overall SART ratings revealed significant main effects for target type and display type: Fs(1,11) = 5.74, p < .0355, and 23.78, $p < .0005$, respectively. Of the interactions, only the three-way interaction between target type, display type, and threat status was significant, $F(1,11) = 6.60$, $p < .0261$. The nature of the interaction is portrayed graphically in Figure 3. As can be seen in the figure, SART ratings were consistently higher when the target was a TEL rather than a Radar SAM, in all cases except when the low information display was combined with the absence of IR SAMs. In that condition, the SART ratings were similar, regardless of target type.

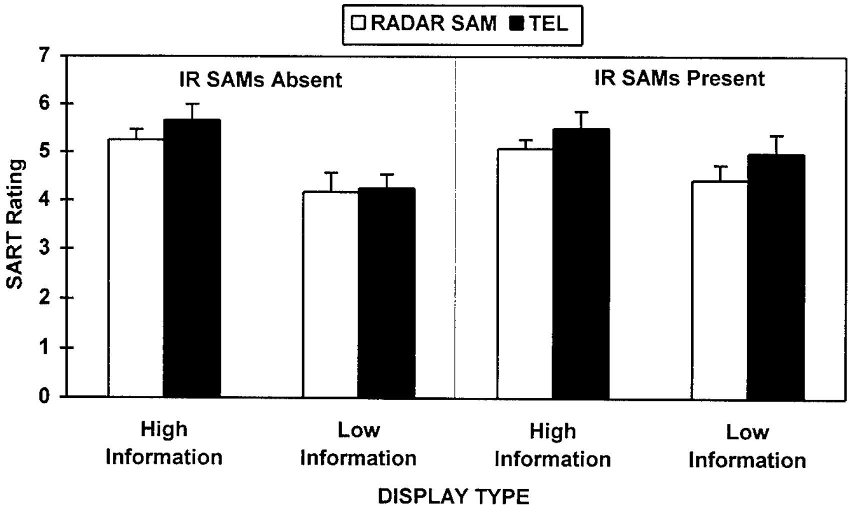


Figure 3. Mean SART ratings for each category of target type, display type, and threat status. (Note: error bars represent the standard error of the mean.)

Additional 2 (target type) x 2 (display type) x 2 (threat status) repeated measures analyses of variance were conducted on the DEMAND, SUPPLY, and UNDERSTANDING ratings to determine which of the three subscales contributed to the observed differences in SART ratings. The analysis of the DEMAND ratings revealed only a significant main effect for target type: $F(1,11) = 19.89, p < .001$. The mean DEMAND rating was higher for the Radar SAM ($M = 3.94$, $SD = 1.34$) than the TEL ($M = 2.54$, $SD = 1.09$). The analysis of SUPPLY revealed significant effects for display type, threat status, and target type: $Fs(1,11) = 11.58, p < .0059$; $22.30, p < .0006$; and $17.99, p < .0014$, respectively. Mean SUPPLY ratings were higher for the TEL ($M = 4.60$, $SD = 1.45$) rather than the Radar SAM ($M = 3.69$, $SD = 1.39$); higher when IR SAMs were absent ($M = 4.46$, $SD = 1.50$) rather than present ($M = 3.83$, $SD = 1.42$); and higher for the high information display ($M = 4.56$, $SD = 1.43$) as opposed to the low information display ($M = 3.73$, $SD = 1.44$). Finally, analysis of the UNDERSTANDING subscale revealed significant effects for display type and threat status as well as a significant interaction between threat status and target type: *F*s(1,11) = 18.87, $p < .0012$; 9.52, $p < .0104$; 9.52, $p < .0104$. The mean UNDERSTANDING rating was higher for the high information display (M = 5.69, SD = 1.17) versus the low information condition ($M = 4.54$, $SD = 1.60$). UNDERSTANDING was also higher when IR SAMs were absent ($M = 5.35$, $SD = 1.39$) rather than present ($M = 4.88$, $SD = 1.59$). The interaction between threat status and target type, which is portrayed in Figure 4, indicated that UNDERSTANDING was similar, regardless of threat status, when the target was a Radar SAM. On the other hand, when the target was a TEL, UNDERSTANDING was higher in the absence of IR SAMs.

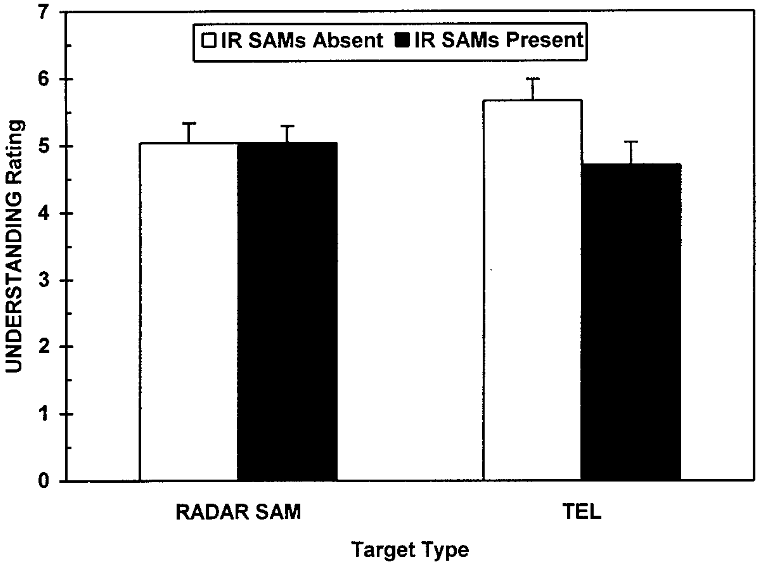


Figure 4. Mean rating on the UNDERSTANDING subscale of the SART for Radar SAMs and TELs when IR SAMs were absent and present. (Note: error bars represent the standard error of the mean.)

In brief, the results of the analyses on the three subscales of the SART indicated that the differences for display type were attributable to differences in SUPPLY and UNDERSTANDING, whereas those for target type were attributable to differences in SUPPLY and DEMAND.

Computer Simulations

The output from the Micro Saint modeling of the laboratory study (i.e., the estimates of visual, auditory, kinesthetic, cognitive, and psychomotor workload during each half second period of the simulated trial) was used to derive averages associated with each experimental condition. These estimates were further used to obtain OW and PW, as described in the Introduction. Means and standard deviations for each of the five workload components, OW, and PW appear in Table 4. The figures for OW and PW in the table indicate that the average and peak workload scores were higher when the target was a Radar SAM, when there was no tactical display, and when IR SAMs were present. Further, in many cases, the component workload scores paralleled these trends. To test the statistical significance of these differences, two different analyses were completed. First, a 2 (target type) x 2 (display type) x 2 (threat status) multivariate analysis of variance (MANOVA) was conducted on the OW and PW scores. Second, a comparable MANOVA was conducted on the five component workload scores themselves.

The MANOVA on OW and PW revealed significant main effects for target type and display type: $Fs(2, 191) = 5.02, p < .0075; 40.18, p < .0001$. Threat status did not attain statistical significance: $F(2, 191) = 1.54, p > .05$. Further, none of the interactions was statistically significant ($p > .05$). Follow-up *t*-tests on OW and PW indicated that the effect for target type was attributable to OW, $t(189.9) = 2.38, p < .0185$; but not PW, $t(176.9) = 1.23$, $p > .05$. As can be seen in Table 4, OW was higher for the Radar SAM than the TEL. The effect for display type was due to differences in both variables: $t(198) = 7.86, p < .0001$ for OW and $t(198) = 2.98, p < .0032$ for PW. Both the average and peak workload were higher for the low information condition as compared to the high information display (see Table 4).

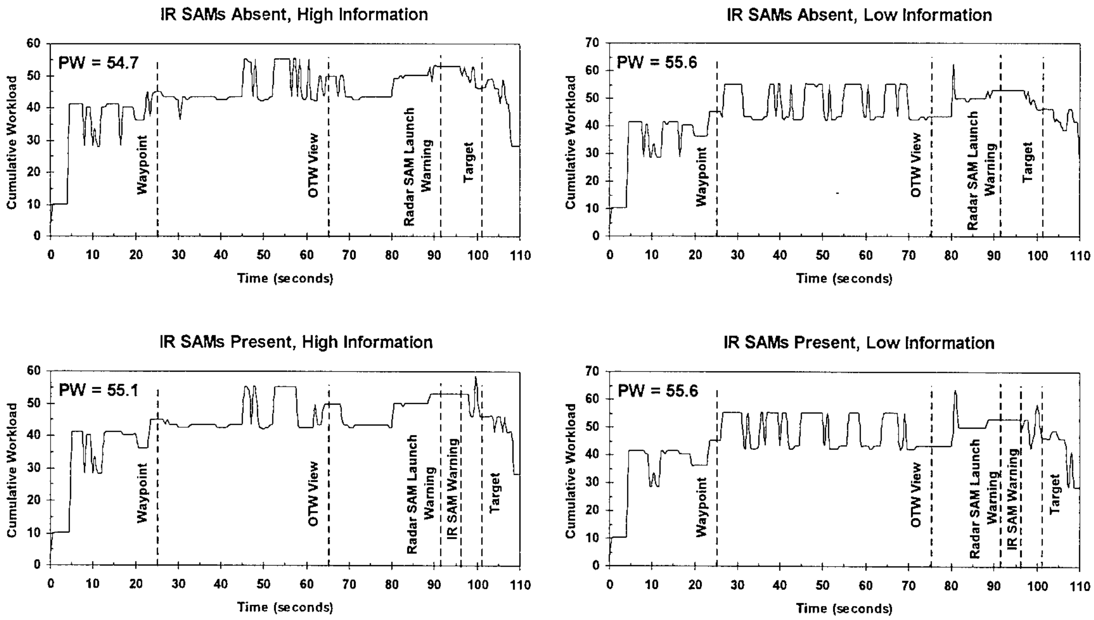
Further inspection of the Micro Saint output revealed that peaks in workload occurred at different time periods, depending primarily upon target type. These peaks can be observed in the eight panels of Figure 5, which depict cumulative workload (i.e., the sum of the five workload components) as a function of time in each experimental condition. As can be seen in the figure, on all TEL trials, the peak workload occurred during the interval prior to the appearance of the target in the out-the-window view. In this time period, the participant would have been engaged in scanning out-the-window to locate the target as well as monitoring the TSD and RWR displays to ascertain target type and location. On Radar SAM trials, the timing of the peak workload was much more variable, depending further upon display type and threat status. When there was no tactical display ("low information") on a Radar SAM trial, the peak workload occurred at the time of the Radar SAM detection warning, regardless of threat status. When the tactical display was present ("high information"), the peak occurred during the out-the-window scanning, provided the IR SAMs were absent. However, on those Radar SAM/high information trials when IR SAMs were present, the peak occurred later in the simulated trial when the participant would have been coping with the IR SAM and simultaneously attempting to lock onto the target and evade its missile.

Table 4

Means and Standard Deviations (in Parentheses) for the Five Workload Components, OW, and PW by Target Type, Display Type, and Threat Status

	TARGET TYPE		**DISPLAY TYPE**		**THREAT STATUS**	
	RADAR SAM	TEL	HIGH INFORMATION	LOW INFORMATION	IR SAMs ABSENT	IR SAMs PRESENT
VISUAL	2.52	2.54	2.21	2.85	2.52	2.54
	(0.33)	(0.35)	(0.10)	(0.12)	(0.34)	(0.33)
AUDITORY	0.56	0.55	0.55	0.56	0.55	0.56
	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)
KINESTHETIC	13.02	13.00	13.02	13.01	13.00	13.03
	(0.21)	(0.26)	(0.22)	(0.25)	(0.24)	(0.23)
COGNITIVE	14.43	14.37	14.30	14.50	14.38	14.43
	(0.33)	(0.39)	(0.38)	(0.31)	(0.38)	(0.34)
PSYCHOMOTOR	5.13	4.90	5.02	5.01	5.00	5.02
	(0.08)	(0.10)	(0.15)	(0.15)	(0.16)	(0.14)
OW	7.13	7.08	7.02	7.19	7.09	7.12
	(0.15)	(0.18)	(0.15)	(0.14)	(0.18)	(0.16)
PW	55.25	55.02	54.87	55.41	55.03	55.24
	(1.52)	(1.06)	(1.27)	(1.31)	(1.38)	(1.24)

RADAR SAM TARGET



TEL TARGET

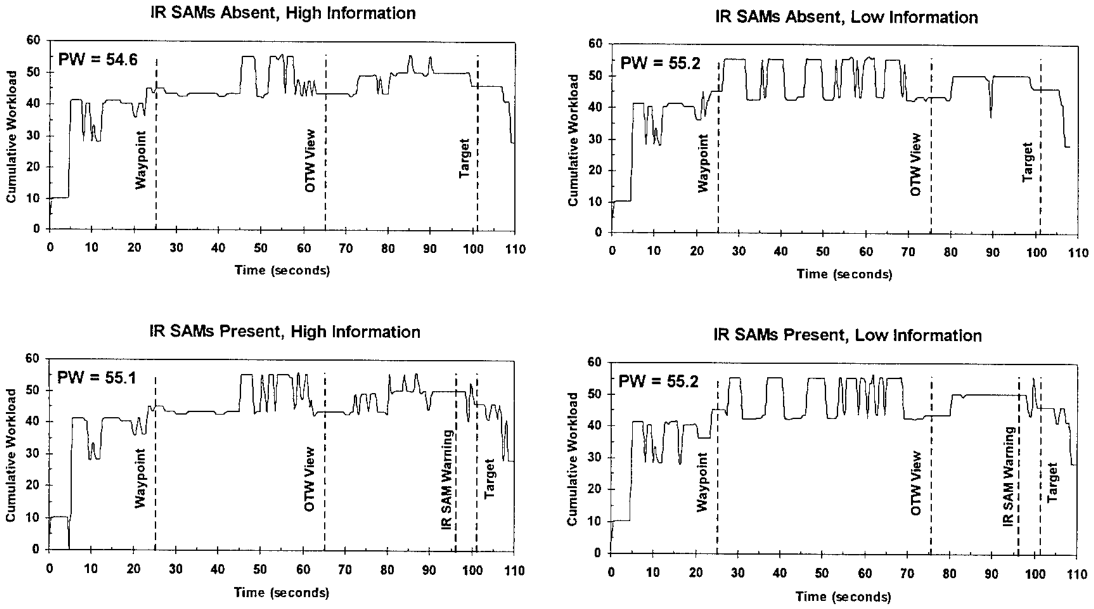


Figure 5. Cumulative workload as a function of time for each combination of target type, threat status, and display type.

Finally, the Micro Saint output was evaluated via a 2 (target type) x 2 (display type) x 2 (threat status) MANOVA on the five workload components. The main effects for target type, display type, and threat status were all statistically significant: *F*s(5,188) = 1212.68, $p < .0001$; 473.25, $p < .0001$; and 3.07, $p < .011$, respectively. Of the interactions, only Target Type x Display Type reached significance: $F(5, 188) = 3.87$, $p < .0023$. Thus, when all five components are considered simultaneously, differences due to target type, display type, and threat status have a significant impact on workload. To determine which components contributed to these differences, univariate ANOVAs were conducted on each component. Only those effects that were significant in the MANOVA (i.e., target type, display type, threat status, and Target Type x Display Type) were included in the univariate tests. The results of these tests are summarized in Table 5.

Table 5

Results of Univariate ANOVAs on Each of the Five Workload Components (df = 1, 195)

IV	DV	MS	*F*	*p*
Target Type	Visual	0.0184	1.50	NS
	Auditory	0.0008	0.56	NS
	Kinesthetic	0.0024	0.04	NS
	Cognitive	0.1686	1.40	NS
	Psychomotor	**2.6929**	**320.72**	**.0001**
Display Type	**Visual**	**20.4960**	**1672.89**	**.0001**
	Auditory	0.0010	0.69	NS
	Kinesthetic	0.0060	0.11	NS
	Cognitive	**1.9455**	**16.12**	**.0001**
	Psychomotor	0.0015	0.18	NS
Threat Status	Visual	0.0187	1.52	NS
	Auditory	0.0029	1.96	NS
	Kinesthetic	0.0336	0.59	NS
	Cognitive	0.1127	0.93	NS
	Psychomotor	0.0224	2.67	NS
Target Type x Display Type	Visual	0.0167	1.36	NS
	Auditory	**0.0058**	**3.95**	**.05**
	Kinesthetic	0.1008	1.78	NS
	Cognitive	0.1184	0.98	NS
	Psychomotor	0.0100	1.19	NS

As can be seen in the table, the effect for target type was attributable to differences in psychomotor workload. As might be expected, the Radar SAM, which required execution of the evasive jink maneuver, was associated with higher psychomotor workload than the TEL, which did not require evasive maneuvers. The effect for display type was due to visual and cognitive workload, both of which were higher when there was no tactical display. In the absence of this display, participants were forced to engage in more out-the-window scanning to spot the target. On the other hand, when the display was present, the location of the target in the surrounding terrain was readily apparent, as symbolized by a red target triangle on the TSD. As can be seen by the results of the univariate tests, the effect for threat status was not the result of significant differences on any one particular component, but rather was due to the combined effects of small differences on all components. Finally, the Target Type x Display Type interaction was a consequence of differences in auditory workload. Specifically, the mean auditory workload was comparable for the two target types when there was no tactical display (*M* = .56, *SD* = .03 in both instances), but slightly higher for the Radar SAM (*M* = .56, *SD* = .05) than the TEL (*M* = .55, *SD* = .04) when the tactical display was present.

Comparison of Laboratory Data and Micro Saint Data

Univariate and multivariate analyses of variance.

A comparison of the effects that were significant in the analyses of the SWAT and SART data from the laboratory task on the one hand and OW, PW, and the five workload components from the Micro Saint data on the other reveals first that no single interaction was simultaneously significant in both sets of data. With respect to significant main effects only, the ANOVA on the SART data and the MANOVA on OW and PW produced identical results (i.e., target type and display type were significant in both cases). All three main effects were significant in the MANOVA on the five workload components, making it partially comparable to both the ANOVA on the SWAT (where target type and threat status were significant) and the ANOVA on the SART (where target type and display type were significant).

Correlational analyses.

A more direct method of comparing the laboratory data with the simulation data involved computing correlations among the various dependent measures. These analyses were conducted by computing the mean for each measure in each of the eight experimental conditions for both the laboratory data and the Micro Saint data and then obtaining correlations between the two sets of eight means. As can be seen in Table 6, these analyses revealed that neither OW nor PW was significantly correlated with the overall SWAT rating. However, it should be noted that PW was more highly correlated with the SWAT than was OW, although it did not attain statistical significance, in part due to the small sample size. Further, PW was significantly correlated with two of the subscales of the SWAT--EFFORT and STRESS.

Table 6

Pearson Correlation Coefficients among the SWAT Workload Ratings and the McCracken-Aldrich Workload Components

	TIME	**EFFORT**	**STRESS**	**SWAT RATING**
VISUAL	-.15	.40	.31	.12
AUDITORY	.42	.25	.51	.45
KINESTHETIC	.30	-.07	.20	.21
COGNITIVE	.17	.41	.54	.38
PSYCHOMOTOR	.64	.60	.79*	.74*
OW	.11	.51	.57	.38
PW	.42	.78*	.76*	.66

*$p < .05$

Table 7

Pearson Correlation Coefficients among the SART Situational Awareness Ratings and the McCracken-Aldrich Workload Components

	DEMAND	**SUPPLY**	**UNDER-STANDING**	**SART RATING**
VISUAL	.14	-.56	-.84**	-.68
AUDITORY	.36	-.57	-.36	-.28
KINESTHETIC	.12	-.22	-.03	.06
COGNITIVE	.36	-.67	-.73*	-.61
PSYCHOMOTOR	.95**	-.63	-.10	-.41
OW	.42	-.75*	-.82*	-.73*
PW	.54	-.68	-.77*	-.74*

*$p < .05$; **$p < .01$

With respect to the SART, the figures in Table 7 reveal that it was significantly correlated with both OW and PW--higher SA tended to be associated with both lower average and peak workload. The UNDERSTANDING subscale of the SART appears to be the most significant contributor to these relationships. It was correlated not only with both OW and PW but also with the visual and cognitive workload components. Finally, a comparison of the figures in Tables 6 and 7 indicates that there were a greater number of significant correlations with the SART than with the SWAT.

DISCUSSION

The primary purpose of the present study was to assess the validity of computer modeling of mental workload and SA. In Phase I of the study, 12 individuals completed a series of target acquisition trials in a laboratory flight simulator. At the conclusion of certain trials, they were asked to assess either the mental workload (SWAT) or SA (SART) associated with completing the tasks comprising the trial. In Phase II of the study, Micro Saint models for each experimental condition were constructed and executed. The workload associated with each activity necessary for task completion was estimated via the McCracken-Aldrich approach. The output from model execution was used to derive two measures of workload: OW (average or overall workload) and PW (peak workload). If computer simulation modeling is valid, the results of analyses of the laboratory workload/SA data should be comparable to the results of analyses of the Micro Saint workload data.

Before considering the decisions that we reached regarding this issue, it is important to consider the quality of the data that entered into them. A number of internal consistencies in both the laboratory data and the Micro Saint data strongly suggest that the quality of the information used to determine the validity of computer modeling was in fact quite high. Turning first to the laboratory data, the results there indicated that mental workload was higher for the Radar SAM as opposed to the TEL. This outcome is consistent with what might be expected, given that the Radar SAM was capable of launching a missile at the participant's aircraft whereas the TEL was not. The presence of the Radar SAM further meant that individuals were required to engage in additional activity (i.e., the jink maneuver), which might be expected to increase workload relative to that for the TEL target. The elevated workload in the presence of the Radar SAM is also consistent with the finding that participants were less likely to accomplish target destruction and more likely to crash when the target was a Radar SAM. Mental workload was higher also when IR SAMs were present during a trial, an outcome that is not unexpected since additional activity was required when these threats were encountered. Further, this activity was required at the same time that individuals were trying to center the aiming reticle over the target and watch for the weapon in-range indicators on the HUD. Finally, one other noteworthy finding from analyses of the laboratory data was the enhanced SA when the tactical display was present. As with the other outcomes just described, this result conforms to what one might expect since participants were readily able to ascertain the location of the target amid the surrounding terrain when the TSD was present, but not when it was absent. Thus, participants gained greater awareness of where the target was in relation not only to their own aircraft but also to other objects in the environment; consequently, their SART ratings were higher in the tactical or high information condition.

Like the laboratory data, the Micro Saint data also exhibited many internal consistencies. First, as with the laboratory data, workload (OW) was significantly higher for the Radar SAM than the TEL. Further analyses indicated that this effect was due primarily to higher workload on the psychomotor component, an outcome that would be expected since the jink maneuver, a motor activity, was required only when the target was a Radar SAM. Second, the Micro Saint data showed higher workload (OW and PW) for the low information condition as opposed to the high information condition. Because more visual search for the target was required in the absence of the TSD, both visual and cognitive workload were higher in the low information condition, contributing to higher overall and peak workload.

Given that the quality of the data appears to be acceptable, the question of the validity of the computer modeling can now be considered. In brief, we can state at this point that the modeling effort was partially but not completely valid, chiefly because the results of the analyses of the lab data versus the Micro Saint data were similar but not identical. The similarities that did emerge between the two sets of data indicate that the computer modeling approach does have a promising future as a tool for evaluating operator workload and SA. The data also suggest methods by which the validity might be further enhanced.

First, as revealed by inspection of the SWAT, SART, OW, and PW ratings in Tables 2 through 4, the means for the two sets of data were consistently in the same direction. Workload was higher and SA was lower for the Radar SAM versus the TEL; for the low versus high information condition; and for the presence of IR SAMs versus their absence.

Second, and somewhat surprising, the pattern of the results suggests that the McCracken-Aldrich approach to computer modeling might be a more valid predictor of SA rather than mental workload. This conclusion stems primarily from the finding that the same main effects were significant in the analysis of OW and PW from the Micro Saint data and in the analysis of the SART data. Namely, in both cases, significant main effects for display type and target type were evident. Furthermore, in the correlational analyses, the SART but not the SWAT was significantly correlated with OW and PW. The absence of a correlation between model-based predictions of OW and PW and the SWAT ratings from the laboratory data contradicts Iavecchia et al.'s (1989) results, which indicated that both OW and PW were significantly correlated with the workload ratings provided by human operators. However, unlike Iavecchia et al., we also assessed the correlation between model-based predictions of workload and SART situational awareness ratings. Paradoxically, this relationship proved to be much stronger than that between the seemingly more comparable workload measures.

Closer inspection of the analyses of variance and the correlations between (1) the five workload components used to derive OW and PW and (2) the three subscales of the SART reveals some subtle consistencies among the laboratory SA data and the Micro Saint output, which further suggest that the computer modeling was a stronger predictor of SA. First, the psychomotor workload component from the Micro Saint data was significantly correlated with the DEMAND subscale from the laboratory data, and both of these dependent measures varied significantly with target type in their respective analysis of variance tests (i.e., both psychomotor workload and DEMAND were higher for the Radar SAM target). Second, the visual and cognitive workload components were significantly related to the UNDERSTANDING subscale from the laboratory data. For all three dependent variables, analyses of variance revealed significant differences with respect to display type. Visual and cognitive workload were lower and UNDERSTANDING was higher when the high information display was present. Thus, the combined results from the ANOVAS and the correlational analyses indicate a stronger correspondence between the McCracken-Aldrich approach to computer modeling and SA as opposed to mental workload. Nevertheless, it should be noted that a multivariate analysis of the five workload components themselves picked up effects that were significant in *both* the analysis of SWAT and the analysis of SART. This outcome implies that while the McCracken-Aldrich approach to computer modeling may be a better predictor of SA than mental workload, it is still not entirely comparable to either construct.

We should at this point recognize some of the weaknesses in the laboratory portion of the present study, which may themselves have served to attenuate the validity of the computer modeling (i.e., some of the fault may lie with the laboratory simulation rather than the computer modeling per se). For example, the display manipulation was not nearly as potent as we had anticipated. Although the high information display was associated with greater situational awareness, display type did not have a significant impact on participants' SWAT workload ratings. This outcome is somewhat surprising since locating the target should have been comparatively easier in the high information condition where the TSD continuously displayed the locations of the target and the aircraft in the surrounding terrain. Further, the red color of the target in the high information condition was selected so that it would be much more noticeable than the brown color in the low information condition. In an earlier study in which the STORM simulator was used in the context of a European environment, red tanks were associated not only with significantly more hits than brown tanks but also with lower DEMAND ratings on the SART scale (Vidulich, Stratton, Crabtree, & Wilson, 1994). Accordingly, it was expected that the red targets in the present study would be associated with lower workload, particularly since they occurred in conjunction with the additional information provided by the TSD; but this was not the case. In fact, during post-experimental interviews, several participants commented that the red target was often as difficult to spot as the brown target, if not more so in some cases. One factor may have been the change from the European background to the desert terrain. While a red target may appear much more salient than a brown target when the background consists of green grass, this apparent difference in target salience may be minimized, or even disappear altogether, when the background consists of desert sand.

In addition to the display manipulation, another factor that may have weakened the validity of the computer modeling approach was the large number of crashes that occurred during flight simulation. Fourteen percent of all trials culminated in either a *g*-load or terrain crash (see Table 1). Nevertheless, participants were still asked to supply subjective ratings of workload or SA on these trials, and they were included in all analyses that were conducted. The Micro Saint models, on the other hand, were designed to simulate in its entirety the "average" or "typical" trial, which ended shortly after the participant had fired upon the target. This disparity between the laboratory and computer modeling phases of the study may have reduced validity.

Other weaknesses of the present study included (1) the limited performance data that could be obtained from the laboratory simulation; (2) the absence of a true multiple resource manipulation; and (3) the inclusion of only a single type of computer modeling. First, with respect to the performance data, the only objective indicators of performance effectiveness that could be meaningfully derived were crashes and kills. At the time that this study was conducted, the software for collecting other types of performance data (e.g., reaction time) had not been completed. In future experiments, a variety of performance metrics will be collected so that the relationships among performance effectiveness, mental workload, and situational awareness can be assessed more fully. Second, the absence of a multiple resource manipulation represented another limitation of the current study. That is, we did not introduce task combinations that would purposely tap a common resource simultaneously and induce mental overload. For example, under a multiple resource approach, two tasks that simultaneously require the visual modality should generate higher workload than two tasks having distinct modalities of input (e.g., visual and auditory). Finally, because we have explored only a single type of computer modeling to date, our conclusions regarding the predictive validity of computer modeling procedures are limited to the McCracken-Aldrich approach. The validity of other approaches is not yet known.

Thus, the next task is to attempt simultaneously to overcome the weaknesses just described by improving the design of future experiments and to enhance the validity of the computer modeling approach. The former will be accomplished in part by modifying the display manipulation and by attempting to reduce the quantity of crashes. Further, additional performance data will be collected in future studies. With respect to validity enhancement, the nature of the results herein suggests that the addition of some factor that accounts for operator stress or uncertainty during task completion might be beneficial. This line of reasoning was prompted by the finding that the significant effects for threat status and target type in the analysis of the SWAT data were due to differences on the TIME and STRESS dimensions of the scale. Computer modeling of workload via the McCracken-Aldrich approach is inherently equipped to handle workload due to time pressure; e.g., workload will be higher whenever multiple tasks must be completed concurrently at any given time. However, it does not directly take into account workload due to the effects of mental stress and uncertainty since there is no "stress" workload component. As evidenced by the elevated STRESS ratings, participants in this task in particular experienced considerable stress and uncertainty. At the beginning of a trial, they did not know whether additional threats would be present, what type of target they would encounter, or where it would be. This stress and uncertainty influenced their experience of mental workload and augmented their final rating, particularly when IR SAMs were actually encountered and when the target was a Radar SAM. This outcome strongly implies that the validity of the computer modeling approach might be enhanced by the addition of a "stress" factor that modifies the workload derived from the McCracken-Aldrich scales. Further studies to address these implications are currently underway.

REFERENCES

Bierbaum, C. R., Szabo, S. M., & Aldrich, T. B. (1987). *A comprehensive task analysis of the UH-60 mission with crew workload estimates and preliminary decision rules for developing a UH-60 workload prediction model* (Draft Technical Report No. ASI690-302-87[B], Vol. I, II, III, IV). Fort Rucker, AL: Anacapa Sciences, Inc.

Bierbaum, C. R., Szabo, S. M., & Aldrich, T. B. (1989). *Task analysis of the UH-60 mission and decision rules for developing a UH-60 workload prediction model: Volume I: Summary report* (Research Product 89-08). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Crabtree, M. S., Marcelo, R. A. Q., McCoy, A. L., & Vidulich, M. A. (1993). Subjective measurement of situation awareness during simulated tactical operations training. In *Proceedings of the 7th International Symposium on Aviation Psychology* (pp. 891-895). Columbus, OH: The Ohio State University.

Hamilton, D. B., Bierbaum, C. R., & Fulford, L. A. (1991). *Task analysis/workload (TAWL) user's guide: Version 4.0* (Research Product 91-11). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD-A241 861)

Hendy, K. C. (1994a). *Survey of national practices in task network simulation for human-machine systems design.* Washington, D. C.: The Technical Cooperation Program, Subgroup U, Technical Panel 7.

Hendy, K. C. (1994b). *Implementation of a human information processing model for task network simulation* (DCIEM No. 94-40). North York, Ontario, Canada: Defence and Civil Institute of Environmental Medicine.

Iavecchia, H. P., Linton, P. M., Bittner, Jr., A. C., & Byers, J. C. (1989). Operator workload in the UH-60A Black Hawk: Crew results vs. TAWL model prediction. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1481-1485). Santa Monica, CA: Human Factors Society.

Kameny, I. (Ed.) (1995). *Defense Modeling and Simulation Office Data and Repositories Technology Working Group (DRTWG) Meetings Held February 7-10, 1995 and Additional Task Force and Subgroup Meetings Held Between July 1994 and February 1995.* RAND National Defense Research Institute.

Laughery, K. R. (1989). Micro Saint: A tool for modeling human performance in systems. In G. R. McMillan, D. Beevis, E. Salas, M. H. Strub, R. Sutton, & L. van Breda (Eds.), *Applications of human performance models to system design* (pp. 219-230). New York: Plenum Press.

Lawless, M. T., Laughery, K. R., & Persensky, J. J. (1995). *Using Micro Saint to predict performance in a nuclear power plant control room: A test of validity and feasibility* (Technical Report No. NUREG/CR-6159). Washington, D.C.: Division of Systems Technology, Office of Nuclear Regulatory Research.

McCracken, J. H., & Aldrich, T. B. (1984). *Analyses of selected LHX mission functions: Implications for operator workload and system automation goals* (Technical Report No. ASI479-024-84). Fort Rucker, AL: U.S. Army Research Institute for the Behavioral and Social Sciences.

Micro Saint [Computer software]. (1996). Boulder, CO: Micro Analysis & Design Simulation Software, Inc.

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* (pp. 185-218). Amsterdam: North-Holland.

SAS (The SAS System for Windows 3.10, Release 6.08) [Computer software]. (1992). Cary, NC: SAS Institute Inc.

Selcon, S. J., & Taylor, R. M. (1990, April). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In AGARD-CP-478, *Situational*

Awareness in Aerospace Operations (pp. 5-1 to 5-8). Neuilly Sur Seine, France: Advisory Group Aerospace Research & Development. (AD-A223939)

Selcon, S. J., Taylor, R. M., & Koritsas, E. (1991). Workload or situational awareness?: NASA TLX versus SART for aerospace systems design. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 62-66). Santa Monica, CA: The Human Factors Society.

Taylor, R. M. (1990, April). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In AGARD-CP-478, *Situational Awareness in Aerospace Operations* (pp. 3-1 to 3-17). Neuilly Sur Seine, France: Advisory Group for Aerospace Research & Development. (AD-A223939)

Taylor, R. M., & Selcon, S. J. (1990). Cognitive quality and situational awareness with advanced aircraft attitude displays. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 26-30). Santa Monica, CA: The Human Factors Society.

Vidulich, M. A., Stratton, M., Crabtree, M., & Wilson, G. (1994). Performance-based and physiological measures of situational awareness. *Aviation, Space, and Environmental Medicine, 65 (5, Suppl.),* A7-A12.

Wickens, C. D. (1984). *Engineering psychology and human performance.* Columbus, OH: Merrill.

GLOSSARY

3-D SART	Three-dimensional Situational Awareness Rating Technique
10-D SART	Ten-dimensional Situational Awareness Rating Technique
ANOVA	Analysis of Variance
D	Demand
df	Degrees of Freedom
DMSO	Defense Modeling and Simulation Office
DoD	Department of Defense
HARM	High-speed Anti-Radiation Missile
HUD	Head-Up Display
IR	Infrared
KIAS	Knots Indicated Air Speed
LHX	Light Helicopter Family
LOCA	Loss of Cooling Accident
M	Mean
MANOVA	Multivariate Analysis of Variance
OW	Overall Workload
PW	Peak Workload
r	Pearson correlation coefficient
S	Supply
SA	Situational Awareness
SA(c)	Calculated Situational Awareness
Saint	Systems Analysis for Integrated Networks of Tasks
SAM	Surface-to-Air Missile
SART	Situational Awareness Rating Technique
SAS	Statistical Analysis System
SD	Standard Deviation
SGTR	Steam Generator Tube Rupture
STORM	Simulator for Tactical Operations Research and Measurement
SWAT	Subjective Workload Assessment Technique
TAWL	Task Analysis/Workload
TEL	Transporter-Erector-Launcher
TOSS	TAWL Operator Simulation System

TSD	Tactical Situation Display
U	Understanding