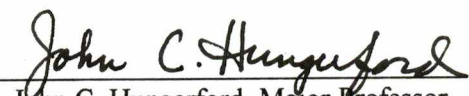
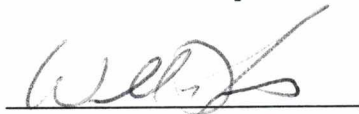


To the Graduate Council:

I am submitting herewith a thesis written by Thomas Joseph Plumb entitled "The Use of Subjective Measures for Sensor Assignment in a Two-Seat Fighter Aircraft." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Aviation Systems.

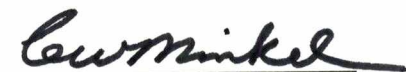

John C. Hungerford, Major Professor

We have read this thesis and
recommend its acceptance:



Charles J. N. Paludan

Accepted for the Council:


Associate Vice Chancellor and
Dean of the Graduate School

**THE USE OF SUBJECTIVE MEASURES FOR SENSOR
ASSIGNMENT IN A TWO-SEAT FIGHTER AIRCRAFT**

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Thomas Joseph Plumb
December 1996

ABSTRACT

This thesis explores the use of subjective measures for determining sensor assignment in a two-seat fighter aircraft, the F-15E. The F-15E is the USAF's newest multi-role fighter aircraft, and it has a formidable air-to-air capability which is centered around its radar. The air-to-air mode of an infrared targeting pod (TGP) provides a complement to radar operations. The aircrew of this aircraft consists of a pilot and a weapons system officer (WSO). Sensor assignment is of critical importance to the success of the mission and the survival of the aircrew. Historically, the pilot primarily performed basic aircraft control functions until weapons employment, while the WSO performed targeting functions with the radar. But with the availability of two sensors, the assignment of radar and TGP responsibility was uncertain.

Direct mission success performance measures would not be adequate for sensor assignment conclusions because of the excessive number of combat related variables. Since both the ability for the pilot to effectively operate a sensor during critical piloting periods, and the WSO to operate two sensors simultaneously was in question due to potentially high workloads, workload assessment was considered crucial. Physiological workload measurement systems were not considered viable, so a subjective measurement system, a modified Bedford scale, was chosen. For relating aircrew preference comparisons between three tasking assignments, the analytical hierarchy process was used. Limited postflight availability and aircrew experience using subjective measures were important considerations leading to the selection of these methods. A questionnaire comments section, one-on-one interviews, and test team discussions were also accomplished.

The overall test methodology was successful in developing working considerations for sensor assignment for various critical tactical conditions. Neither workload or comparison ratings alone would have been successful in reaching the conclusions drawn from combining both methodologies. Interviews and test team meetings were the major source of data used in reaching conclusions. Because of the large number of variables in this test environment, workload and comparison ratings had reduced effectiveness and increased data reduction requirements. While this limited their usefulness in determining conclusions to a secondary role, their inclusion in the test raised aircrew awareness, and forced aircrews to make sensor assignment preference decisions. Overall, their inclusion in the test was helpful but not critical to reaching assignment conclusions.

TABLE OF CONTENTS

CHAPTER	PAGE
1.0 INTRODUCTION.....	1
1.1 Background.....	1
1.2 Scope.....	1
1.3 Overview	2
1.4 Author Participation in Testing.....	2
1.5 Security Classification.....	2
2.0 DESCRIPTION OF F-15E WEAPON SYSTEM.....	3
2.1 General.....	3
2.2 F-15E Systems.....	3
2.3 Overview of Operations and Missions.....	5
2.4 Aircrew Duties and Responsibilities.....	7
3.0 METHODOLOGY.....	8
3.1 General.....	8
3.2 USAF Tactics Development and Evaluation.....	8
3.3 Test Environment.....	8
3.4 Variable Selection.....	10
3.5 Subjects.....	12
3.6 Data Collection.....	13
3.7 Data Analysis.....	24
4.0 TEST PLAN AND EXECUTION.....	25
4.1 General.....	25
4.2 Data Collection Plan.....	25
4.3 Detailed Test Methodology.....	26
4.4 Aircrew Test Preparation.....	27
4.5 TD&E Mission Phase Plan.....	29
4.6 Mission Scheduling.....	30
4.7 Data Compilation.....	30
5.0 RESULTS, CONCLUSIONS, AND RECOMMENDATIONS.....	32
5.1 General.....	32
5.2 Data Analysis and Reduction Results.....	32
5.3 Individual Methodologies Conclusions.....	34
5.4 General Conclusions.....	36
5.5 Recommendations.....	36
LIST OF REFERENCES.....	38
APPENDIX.....	41
VITA.....	45

LIST OF FIGURES

FIGURE	PAGE
1. F-15E Aircraft.....	4
2. Cooper-Harper Rating Scale	16
3. Bedford Rating Scale.....	18
4. SWAT Rating Scale	19
5. TLX Rating Scale.....	21
6. Recommended AHP Rating Scale.....	23
7. Tactical Condition Sensor Assignment Methodology	28

APPENDIX

A-1. Aircrew WL Rating Scale	42
A-2. Posflight Aircrew Debrief Form	43
A-3. Preflight Mission Tasking	44

LIST OF ABBREVIATIONS

A/A	air-to-air
A/S	air-to-surface
AGL	above ground level
AHP	analytical hierarchy process
AIM	air intercept missile
AIS	aircraft information system
AVTR	airborne video tape recorder
BVR	beyond visual range
DCA	defensive counter air
DT&E	development test and evaluation
FOT&E	follow-on operational test and evaluation
FOV	field of view
GCI	ground controlled intercept
HOTAS	hands-on-stick-and-throttle
HUD	head-up display
IV	independent variable
IR	infrared
LANTIRN	low-altitude navigation and targeting infrared for night
MOE	measure of effectiveness
MSL	mean sea level
MFD	multi-function display
NM	nautical mile
OA	operations analyst
OCA	offensive counter air
OFP	operational flight program
OT&E	operational test and evaluation
PM	project manager
PO	project officer
PS	project specialist
RFMDS	Red Flag Measurement and Debriefing System
ROE	rules of engagement
RWR	radar warning receiver
SPINS	special instructions
T&E	test and evaluation
TGP	targeting pod
TD&E	tactics development and evaluation
TES	Test and Evaluation Squadron
TWS	track-while-scan
USAF	United States Air Force
WEZ	weapon employment zone
WL	Bedford workload rating
WSO	weapons systems officer

1.0 INTRODUCTION

1.1 Background

The addition of the Lockheed-Martin Aircraft Company's targeting pod (TGP) to the F-15E fighter aircraft has given it an infrared (IR) air-to-air (A/A) mode. Prior to this system being utilized in the A/A combat role, the primary operator of the main A/A sensor, the radar, has been the weapons system officer (WSO). Typically, the pilot performed basic flying duties, and did not become heavily involved in targeting until just before weapons employment. With A/A functions available from the TGP, the sensor assignment roles during A/A operations in various combat scenarios was unclear. It was not known if the pilot would be able to effectively operate one of these A/A sensors, or if the WSO should operate both sensors simultaneously. Additionally, the impact of variables such as air threat, weather, aircraft altitudes, and aircrew experience on sensor assignment was unknown. Optimum assignment of these sensors is critical for the successful completion of the mission, as well as for returning safely after mission accomplishment. A viable method was needed to establish the delineation of sensor assignment.

1.2 Scope

This thesis is a overview of how the author, a project manager and flight test aircrew for an operational test and evaluation (OT&E) unit, researched, planned, executed, and analyzed the application of subjective measures for sensor assignment during a United States Air Force (USAF) tactics development and evaluation (TD&E). Since the thesis objective was a subset of a larger test being accomplished at the same time, the determination and application of the chosen methods were driven by the overriding test. The methodology for this project objective had to fit into the overall constraints of the overriding test. Special compensation was not given to any aspect of this thesis project objective.

The focus of this work was on the selection and effectiveness of the chosen methodology in answering the objective in the given test environment, not on the specific results of the test. Discussion is centered on how the author and test team attained and utilized the resulting data, not on the actual data and specific results. This work is about the process, and should additionally present an introduction to special considerations of military testing in an OT&E environment.

In terms of overall test scope, a total of eight pilots and nine WSOs participated in the overriding TD&E. The test squadron had five test aircraft, and typically two missions with a total of four to six sorties were flown each duty day. The TD&E duration was approximately three months, with approximately 150 test sorties for the TD&E. Simulators were not available or used during the test. All missions were flown in field representative F-15E aircraft.

1.3 Overview

The primary objective of the thesis was to evaluate the use of subjective measures for determining the best sensor assignment for various tactical conditions. With this in mind, first the specifics of the test are defined, including the aircraft and aircrew. Next, an overview of applicable methodologies is provided, and the methodology which was selected is detailed. Following a description of test execution and data collection, results and conclusions are described, and recommendations are made.

1.4 Author Participation in Testing

The author was the project manager for the overriding TD&E, and the opinions expressed in this thesis are solely his. This work and its contents have not been officially approved by the 57th Test Group or the United States Air Force.

1.5 Security Classification

This report is unclassified. Detailed technical information on modern radar systems can be found in airborne radar textbooks and aviation related periodicals.^{1,2} Some weapon systems capabilities and employment tactics are classified, and will not be addressed in this unclassified forum. The scope of this thesis was specifically defined in order to explore subjective human factors testing techniques without compromising critical technologies or tactics.³

2.0 DESCRIPTION OF F-15E WEAPON SYSTEM

2.1 General.

In order to put the application of subjective measures in context, the conditions under which the aircrew operates the F-15E need to be delineated. For this purpose, a brief overview of the F-15E aircraft weapon system is presented. This information will facilitate comprehension of the operational environment the aircrew operates under and must be understood in order to properly plan, administer, and report on the test methodology. This chapter highlights the aircraft's systems, mission employment concepts, and specifics on aircrew roles in common combat and training scenarios.

2.2 F-15E Systems

The F-15E "Strike Eagle" is the USAF's newest operational fighter aircraft, having become operational in 1990. The aircraft, pictured in Figure 1, is a multi-role fighter which performs capably in both A/A and air-to-ground (A/G) combat roles. While previous operational versions of the F-15 had a single aircrew, this A/G oriented version has a pilot and a weapon systems officer (WSO). Large by jet fighter standards, this twin tailed aircraft is powered by two 29,000 pound thrust Pratt & Whitney F100-229 afterburning turbofan engines. With an empty (basic equipped) weight of approximately 40,000 pounds, fully loaded maximum gross takeoff weight can be as high as approximately 80,000 pounds.³ While the basic F-15 platform maximum speed at altitude (four AIM-7 A/A missiles) is Mach 2.5 (1,653 mph), low level strike airspeed is typically in the 500-600 mph range.⁴

The aircraft is capable of delivering a wide variety of A/A and A/G ordnance. A/A ordnance includes medium range AIM-7 Sparrow semi-active guidance missiles as well as the actively guided AIM-120 Advanced Medium Range Air-to-Air Missile (AMRAAM). Short range heat seeking AIM-9 missiles and a 20mm M61-A-1 gun can be employed for close A/A combat. The Strike Eagle is capable of employing a wide variety of A/G munitions, including free fall general purpose and cluster munitions. Laser guided and other precision weapons such as the TV/IR sensor GBU-15/AGM-130 and AGM-65 Maverick missiles can also be employed.⁵

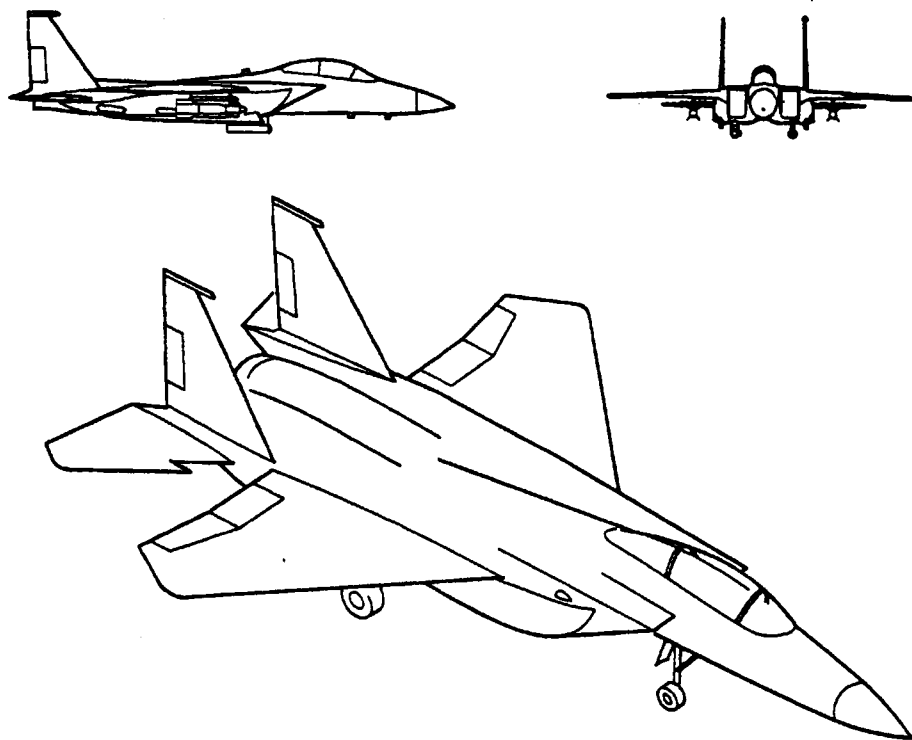


Figure 1. F-15E Aircraft³

The F-15E has an extensive and modern avionics suite. For defensive measures, the aircraft uses the Loral ALR-56 radar warning receiver (RWR), the ALE-45 system for dispensing missile defeating chaff and flares, and the Northrop ALQ-135 internal countermeasures set (ICS) for defensive electronic combat. The F-15E typically carries the two pod externally mounted Low Altitude Navigation and Targeting for Night (LANTIRN) system. This system includes a navigation pod which contains a navigating forward looking infrared (FLIR) system and a terrain following radar (TFR) for night ground collision avoidance.⁵

The second pod of the LANTIRN system contains one of the sensors whose operation is under study in this project. The AN/AAQ-14 targeting pod (TGP) is an infrared sensor and laser target designator. By design an A/G sensor for precision-guided munitions, such as the GBU-10 and the GBU-12, this infrared sensor has multiple fields of view (4X, 10X, and 20X) to allow distant target identification and designation. The TGP has a tactical "combat" laser frequency as well as an "eye safe" training frequency. The TGP also has a basic A/A mode.²

The second sensor system under study in this project is the multi-mode pulse-Doppler Hughes APG-70 radar. While the radar is capable of performing A/G functions, its A/A employment is of concern to this work. The A/A mode of the radar is capable of detecting and tracking aircraft in look-up and look-down scenarios. A track-while-scan (TWS) mode allows monitoring multiple aircraft simultaneously. The radar is used to designate targets for the described A/A munitions.

A hands-on-stick-and-throttle (HOTAS) controls and displays philosophy has been extensively utilized in the aircraft. With HOTAS, a large percentage of radar and TGP functions can be accomplished without requiring the operator to take hands off of the control sticks. In the front cockpit (FCP) the pilot uses the aircraft control stick and throttles to perform HOTAS. In the rear cockpit (RCP) the WSO uses two system control sticks for HOTAS control. The FCP has three multi-function displays (MFDs) plus a heads-up display (HUD) for displaying and controlling systems, while the WSO has four MFDs in the RCP.³

2.3 Overview of Operations and Missions

It was essential to have a good understanding of the planning and briefing requirements of a fighter mission when designing the test. A typical mission might begin two to four hours before the mission briefing. At this point the flight lead aircrew researches the mission specifics, such as aircraft configuration (A/A and A/G), airspace ranges, availability of adversary aircraft, and

requirements of aircrew in the flight. Once the type of mission is determined, the weather and possibly a low level training route availability are researched. The flight lead distributes planning duties for the mission to the participating aircrew, including fuel planning, tactics planning, weaponeering, researching local airfield and airspace status information, and coordination with other flights. If multiple flights are participating in the mission together, a coordination briefing, typically lasting about thirty minutes, is conducted two-and-a-half hours before the scheduled takeoff time. The formal flight mission briefing typically begins two hours before the scheduled takeoff time, and lasts for approximately one hour. Following the flight briefing, a crew briefing is conducted to establish cockpit normal and emergency procedures. Since starting engines, checking aircraft systems, and arming ordnance can require up to thirty to forty-five minutes, there is little spare time between briefing completion and departing to the aircraft. While mission duration can vary depending on onboard fuel and A/A refueling operations, one to two hour missions are common. Postflight debriefing begins with discussing maintenance requirements for the aircraft with the aircraft crewchief and in a formal maintenance debriefing session. Next, the aircrew prepares for the multi-flight debrief (if applicable) by reviewing postflight data and mission videotapes. Multi-ship debriefs can take between thirty minutes and a couple of hours. Next is the flight debriefing where specifics of flight employment are discussed and videotapes are analyzed. Lastly, the crew debrief covers cockpit coordination and performance. All in all, the mission can easily take six to eight hours, including seven or more briefings and debriefings.

The F-15E can be employed in a number of roles. There are two major categories of employment, Offensive Counter Air (OCA) and Defensive Counter Air (DCA). In OCA, the Strike Eagle is used offensively to disrupt and destroy the enemy's ability to wage war. Commonly employed in the OCA strike role, the F-15E can carry A/A and A/G munitions to employ against the enemy. The DCA role would find the F-15E configured for A/A warfare, protecting an airbase or section of airspace. While in either role, the Strike Eagle can be employed in day or night missions, in clear visual meteorological conditions (VMC) and in instrument meteorological conditions (IMC). Additionally, the aircraft can be employed at low altitudes or medium altitudes. Low employment altitudes increase surprise and help defeat enemy radar, while medium altitudes help defeat short range SAMs and AAA and improve combat radius. Different formations can be used in each of these roles and conditions, and may vary depending on the weather and whether employing under day or night conditions.

2.4 Aircrew Duties and Responsibilities

The pilot is the aircraft commander, and has sole responsibility for basic aircraft control. This includes takeoffs and landings, departure and recovery, and mission work. Mission work can include low level flying and navigation, terrain masking and terrain following, aerial combat, A/G weapons employment and A/A refueling. Additionally, the pilot is responsible for maintaining position within the flight. As a flight lead the pilot must monitor the wingman, while as a wingman pilots must constantly monitor and control their position with respect to the flight lead's aircraft position. The pilot launches all A/A missiles, and releases A/G ordnance. The pilot has access to the aircraft weapon systems through HOTAS and display pushbuttons, and may act as the primary operator of a weapon system when the WSO engaged with other activities or systems.

The WSO manages and operates the aircraft weapon systems. In A/G operations, the WSO finds, updates, and designates ground targets for free fall and guided munitions. During A/G designations, the WSO operates the A/G modes of the radar and the TGP to define the target. Historically, the WSO has been the primary operator of the A/A radar. After performing search and sanitization functions, the WSO sorts and targets enemy aircraft. After an enemy target is designated, the pilot takes over control of the radar for missile or gun employment.

While the pilot spends much attention on aircraft control and navigation, and the WSO focuses on radar and TGP weapon systems employment duties, both aircrew have additional tasks which are critical to mission success. Despite modern electronic warfare avionics, a disciplined visual search is crucial to survival. Any spare moments away from systems operation must be spent searching the sky around the aircraft for enemy aircraft and surface-to-air weapons. Additionally, apart from employing the radar and TGP, numerous activities need to be accomplished. To name just a few, navigation systems require updating, the RWR display must be monitored, chaff and flares must be dispensed, radios must be monitored and operated, and in general both aircrew are responsible for maintaining "situational awareness" of the overall scenario.

3.0 METHODOLOGY

3.1 General

This chapter covers overall test design considerations, reviews applicable test methodology, and defines the general decision process that was used to structure the thesis test. Specifically, the test methodology has been broken down into five areas; test environment, variable selection, subjects, data collection, and data analysis.⁶ Additional details of the application of the chosen processes are delineated in chapter 4, Test Plan and Execution.

3.2 USAF Tactics Development and Evaluation.

The test methodology was primarily constrained by the requirements of the USAF tactics development and evaluation (TD&E) test. TD&Es are conducted by the USAF operational test and evaluation (OT&E) community to evaluate the combat effectiveness of weapons systems, and to develop and document tactical applications of systems for the worldwide Combat Air Forces (CAF). This project documents a portion of the methodology which was accomplished to develop and document tactics applicable to the use of the TGP in the combat environment. The TD&E had numerous objectives in addition to evaluating sensor assignment options. TD&E objectives included assessing TGP systems operating performance and utility, as well as developing tactics for the F-15E in various phases of its tactical roles as described in chapter 2. No special concessions were accomplished in order to facilitate the accomplishment of the sensor assignment determination objective. It was completed within the constraints of the TD&E. As will be noted, working within a test environment that is dictated by an overriding test can place difficult constraints on the test methodology. Tradeoffs had to be made between strict testing protocol and the need to answer the objective with the overall test constraints, some of which are described in following chapters. Of course this is not unusual in the testing community. Rarely can one accomplish each portion of a large test in the most desirable manner.

3.3 Test Environment

3.3.1 Simulators

As aircraft simulators become more capable, they are gaining acceptance and utility in combat aircraft operational training and test communities. Replication difficulties are not so much in producing a satisfactory representation of a the particular aircraft, but in creating the operational

environment and in linking more than a single simulator system. Fighters, electronic jamming aircraft, anti-radiation surface-to-air missile (SAM) killers, bombers, and other airborne assets all participate in the operational environment.⁴ The simulation community is actively pursuing improvements in creating a more realistic operational environment through linking simulators across various locations and services. A significant benefit of simulator use is that simulators are not constrained by actual flight airspace, training, and safety restrictions. These benefits and improvements, coupled with the high cost of operational training and testing, will lead to significantly increased simulator use in the near future.

The decision of the extent to which simulators would be used in this test was never in question since there were none available at the test site. All missions were flown in operationally representative F-15E fighter aircraft. If a simulator had been available, it would have been used in TGP systems instruction and for achieving a basic level of proficiency in the system. Additionally, aircrews could have evaluated sensor assignment options and determined initial workload considerations under some of the tactical scenarios under consideration in the test, reducing the need for some test mission scenarios.

3.3.2 Airspace

The test was flown primarily over southwestern desert, mountainous terrain. The humidity was low throughout the test, and cloud formations limited few VMC engagements. Both day and night missions were flown, with a majority of the missions being flown in the day. While airspace size is often a heavy constraint in fighter aircraft tactics evaluations, the airspace available to this test was not a significant limitation. Low level avoidance areas and limitations to chaff/flare use, such as minimum employment altitudes, did exist but were not a notable factor.

3.3.3 Combat Scenario

A major requirement for meaningful tactics and sensor operation evaluation is the closeness of the test environment to the anticipated actual combat scenario. One of the major concerns of an A/A TD&E is having aircraft to act as threat replicators. This test was fortunate to have F-15C and F-16 aircraft available to follow published threat replication standards. On some missions these aircraft utilized electronic combat (EC) equipment, adding further to the overall combat scenario. Both friendly ("blue") and enemy ("red", or "adversary") forces periodically utilized ground controlled intercept (GCI) techniques. And while no "live" missiles or bombs were used in

the test, simulation techniques were utilized, and chaff/flares were available and employed. Blue missions were flown self escort, as well as with air cover supplied by F-15C and F-16 aircraft. Was the combat environment realistic? While no test environment will instill the stress of actual combat, this test environment was suitable for addressing the test objectives.

3.3.4 Constraints

No test is without numerous constraints, and they played a major part in determining the overall methodology of this test. Budget constraints played a role in the ability to attain and utilize more aircraft to perform in an adversary role, and on the total number of sorties in the test. Airspace usage, data recording devices, and travel costs impacted the test budget as well. Scheduling constraints also structured the test. This testing organization is tasked with a multitude of tactical and systems evaluation every year. Each test is allocated a given period of time to accomplish the test, based on the overall test objectives, criticality of the results, and the amount of other testing that needs to be accomplished. In addition to test missions, other training and upgrade missions must take place on a regular basis in order to maintain aircrew certifications and proficiency. The budget and timeline constraints directly impact the total number of sorties available to collect data. Additionally, flight missions are subject to fuel limitations, jet availability for maintenance work, aircrew availability, and safety constraints.

3.4 Variable Selection

3.4.1 Sensor Assignment Options

With the radar and TGP as assignable sensors, and the pilot and WSO as operators, there were a total of four assignment combination options to consider:

- a) Pilot operating the TGP and the WSO operating the radar
- b) Pilot operating the radar and the WSO operating the TGP
- c) WSO operating both systems
- d) Pilot operating both systems.

Since the pilot is responsible for basic aircraft control, and the WSO is primarily designated as a systems operator, the last option was not considered a useful option to investigate. Because the

most demanding systems include the radar and TGP, the WSO would have a significantly greater spare capacity with the pilot operating both of these sensors.

3.4.2 Independent Variables Classification

The large number of independent variables (IVs) in this test should be emphasized. Most studies do not include more than a few IVs. For example, 141 reviewed experimental papers published in Human Factors from 1958 to 1972 showed that sixty percent of the experiments investigated the effects of only one or two IVs and less than three percent investigated five or more IVs.⁷ In the variable determination process in this test, three categories of independent variables were identified. Task related variables included radar and TGP options, equipment, formation types, employment altitudes, and so forth. Environmental variables included time of day (day/night), inflight visibility, and weather. Lastly, subject related variables included overall tactical experience, training, time in aircraft type, and recent flying hours.⁶ Table 1 identifies the final determination of variables and conditions which were used at the start of the test.

Table 1. Tactical Scenario Variables and Conditions.

Variables	Conditions
Relative Elevation/Background	F-15E: Low Med/High Unrestricted Adversary: Med/High Low Unrestricted
F-15E Role / Ordnance	OCA-1 DCA OCA-2 F-15E Load A Load B Load C
A/A Threat Type/ Ordnance	Threat Type A Threat Type B Loadout A Loadout B
Force-on-Force Size	< 3v2 ≥ 3v2
F-15E Formation	Day Night Tactical Terrain Following (simulated in day)
ROE	ROE-A ROE-B
GCI	Used Not Used
Weather	Clear Cloudy
Legend:	ROE - rules of engagement

The following variables had potential impact on the dependent variables, but were considered to be confounding because of an inability to satisfactorily regulate them; humidity (low/high), adversary proficiency (low, high), blue air support (available, not available), and terrain (flat, mountainous). Their impact on reaching conclusions is discussed in chapter 5. It was determined that mission success performance measures alone were insufficient to satisfy objective conclusions, due to the large number and dynamic nature of combat success measures. Despite this, the following dependent (criterion) variables were documented as described in chapter 4; threats detected/not detected, A/A missile shots defeated and taken, enemy kills and blue kills, defensive reactions, and ordnance delivered on target.

3.5 Subjects

3.5.1 Subject Operational Experience

While many types of OT&E testing strive for and use operators from operational units to help ensure the subjects are representative of those who will ultimately use the system, using operational units is not always feasible.⁸ Particularly in the USAF aircraft testing, where aircraft systems can be extremely complex and take years of training in order to become operationally proficient, it is not always possible. Because the specific system under test may take extensive training and familiarization, and the test itself may take from weeks to many months to accomplish, operational aircrew can not be frequently used. Operational aircrews and squadrons have commitments such as regular overseas deployments which prevent them from being available for a test of this duration. For this reason, aircrews are specifically assigned to OT&E testing units.

OT&E test aircrew are typically experienced aviators in their weapon system, and have commonly completed at least one three year tour in the assigned aircraft. In general, they have 500 to 1000 flying hours in the aircraft model, and are instructors, as well as being flight leads and mission commanders. In some cases, they are USAF Fighter Weapons School (USAF FWS) graduates, and military test pilot school (TPS) graduates, but the latter are exceptions in OT&E. By comparison, most developmental testing aircrew are graduates of a military test pilot school.

Aircrew selection for this project was dictated by the overriding TD&E. A total of 8 pilots and 9 WSOs participated in this test. Two were graduates of the USAF FWS and one was a test pilot school graduate (the author).

The test aircrew had significantly more experience than operationally representative squadron aircrew. This experience can "bias" test data, and must be factored into test conclusions. But while the aircrew are very experienced, OT&E aircrew spend significantly more time performing non-flying duties, including extensive traveling on test related activities. This reduction on flying time can decrease flight performance since aerial skills are very perishable despite the significant aircrew experience. It is the author's belief that subject stratification concerns did not pose a significant threat to results and conclusion validity.^{8,9}

3.5.2 Subject Test Experience

With the exception of the author, no aircrews involved in this test were graduates of a military test pilot school. And while subjective rating systems like those used in this test are fairly common in the DT&E community, no aircrew had previously used the subjective rating scales used in this project, and did not have any knowledge of them prior to this test.

3.5.3 Subject Availability

As described in chapter 2, there are significant time demands placed on flight missions. When testing is involved, the demands increase. In addition to the mission schedule described, special debriefings must be conducted on test articles and procedures. In this particular test, aircrew spare time for additional briefing and debriefing for test related documentation was minimal. Additional military duties, low manning in the unit, and other testing required test procedures to be as streamlined as possible

3.6 Data Collection

While it might have been advantageous to determine sensor assignment options using mission success performance measures from tactical scenarios, this was statistically impractical. Even after reducing the IVs as much as possible, there were still too many variables (and/or too few sorties), including confounding variables, to allow a decision based totally on tactical performance measures. This is not to say that performance data is not useful. The test team analysts still wanted this information, but in order not to be overwhelmed in data and data taking, full mission performance data was reported only during one of three test phases.

To review, data was needed to answer two major issues in determining sensor assignment. The first issue was to determine which assignment options were viable under various IVs. The second

issue was to determine how the options compared to each other, and which were considered best by the aircrew. The first issue was considered to be primarily a question of workload, i.e., to determine if the operator was able to accomplish the given sensor operation under the stated IVs. A method of measuring operator mental workload was needed. For the second issue, a method of comparing up to three options to each other was needed. Since mission success performance measures would not give conclusive results for either issue, some form of subjective preference method was needed for each.

3.6.1 Mental Workload Measurement

3.6.1.1 Workload Measurement

There are three major categories of empirical workload measurement techniques to answer the workload issue. Subjective measures, performance based techniques, and physiological techniques all have applicability.¹⁰ Subjective techniques use operator judgment of the workload associated with the task being evaluated. Performance based techniques assess workload levels based on the operator's ability to complete the task, and have long had a central place in workload assessment.¹⁰ But primary task performance measures do not always reliably vary with other workload measures since, for example, workload levels may rise from low levels without any degradation in performance (performance may actually increase for easy tasks). To counter weaknesses such as this, secondary task measurement and embedded task measurements are sometimes included, but this additional methodology has limitations as well.¹¹ For this project, these workload measurement through performance accounting methods are academic. If suitable performance measures could be readily measured there might not be a need for measuring the workload since we could determine viable sensor assignment and optimum options based on performance scores.

The most frequently used physiological workload measurement techniques include measures of heart rate and heart rate variability, brain activity, and eye activity.¹⁰ These approaches have long been of interest in the history of workload measurement since they offer the potential for objective measurement without the reliance on self-reports and questionnaires. Despite some recent advances, these measures have not lived up to their potential, and they were determined to be beyond the scope of this test.¹¹ Implementation and data reduction costs, as well as manpower and other factors eliminated these methods. With performance based and physiological measurement

tools either unsuitable or impractical, the test team was left with using subjective workload measures.

3.6.1.2 Subjective Rating Scales

Subjective rating scales, along with interviews and questionnaires, make up the two major categories of subjective measurement techniques. There are many types of workload rating scales, and they range from simple unidimensional scales which result in a single workload measure, to multidimensional scales that measure different components of workload. Three scales predominate the field, though, and include the Cooper-Harper (CH) and its derivations, the Subjective Workload Assessment Technique (SWAT), and the National Aeronautics and Space Administration (NASA) Task Load Index (TLX).

The Cooper-Harper scale, originally developed for the assessment of aircraft handling characteristics, was one of the earliest scales used for workload measurement.¹² This well known scale has a long tradition of use in the T&E community and its use is still common today because of its ease of use and simplicity. Additionally, CH rating results correlate well to more recent, increasingly sophisticated scales. In workload ratings associated with aircraft control, CH was found to have correlation values from $r=.75$ to $.79$ with scales such as SWAT and TLX.¹³ The scale involves a decision tree that leads the pilot to a rating value from 1 to 10. The CH scale is depicted in Figure 2.

Since its initial development, the CH scale has been adapted to allow application to a wider range of tasks other than pilot ratings of aircraft handling characteristics. This has been accomplished by a number of researchers, and versions include the Honeywell Cooper-Harper, the Modified Cooper-Harper (MCH), and the Bedford Scale. The Honeywell CH changes focused on overall task workload rather than handling qualities, and restructured wording from compensation and deficiencies to workload and effort.¹⁴ This shifted the emphasis from hardware orientation to task-related orientation, but maintained aircraft specific terminology. The MCH was developed in 1983 for the purpose of workload assessments in systems where perceptual, mediational and communications activity is present.^{15,16} Wierwille and Casali also kept the basic tree structure, but changed the scale descriptions to capture a wide variety of tasks, finally separating the CH scale wording from its original aircraft applications. The developers have shown that the resulting MCH ratings are valid and statistically reliable measures of overall workload and show a good level of sensitivity.

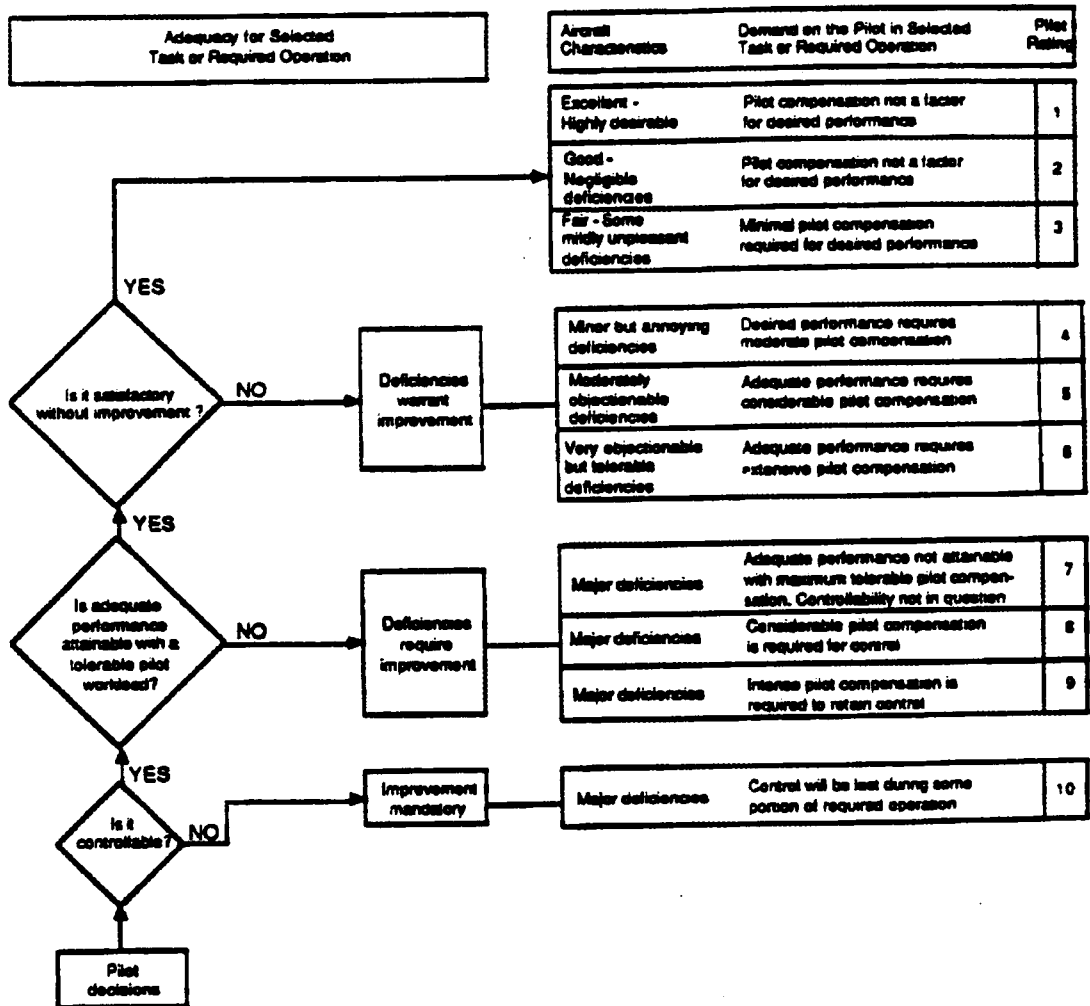


Figure 2. Cooper-Harper Rating Scale¹²

One of the more recent CH adaptations is the Bedford Scale, which was developed by Roscoe and Ellis at the Royal Aircraft Establishment, Bedford, England.¹⁷ The Bedford scale, also called the Pilot Workload Rating Scale, was designed for workload assessment in the military aviation community. This scale continues the decision tree format, but uses terminology based on ability to complete tasks and the amount of spare capacity available. The Bedford Scale application has been documented in several applied setting workload evaluations, ranging from civilian transport aircraft to advanced combat aircraft in low level operations.¹⁵ The Bedford scale is shown in Figure 3.

The Subjective Workload Assessment Technique was developed by the USAF Armstrong Aeromedical Research Laboratory (AAMRL) to provide a quantitative workload measurement tool with interval scale properties. SWAT presents a multidimensional view of mental workload broken up into time load, mental effort, and psychological stress. SWAT uses conjoint measurement and scaling procedures which give a normalized, overall workload score based on the values reported in the three areas.¹¹

The SWAT ratings for a task are not stand-alone figures. Before the three SWAT values can be converted into a workload rating, the SWAT scale must be normalized for each subject. In this scale development, the subject ranks 27 possible combinations of three different levels of time load, mental effort, and stress by means of a card sorting exercise. The results are then used to create a rule for combining the three dimensions for the particular subject. The individual's three SWAT scores are combined with the personalized normalization (using conjoint scaling) to produce a unidimensional workload score ranging from 0 (no workload) to 100 (highest possible workload).¹¹ SWAT is widely used in T&E since it produces a quantitative result with which to compare different systems, if desired. But, SWAT requires extensive preparation and pretraining of subjects before it can be used. Also, it requires significant data reduction after data collection which limits immediate results feedback and increases manpower requirements.¹¹ The SWAT rating scale is depicted in Figure 4.

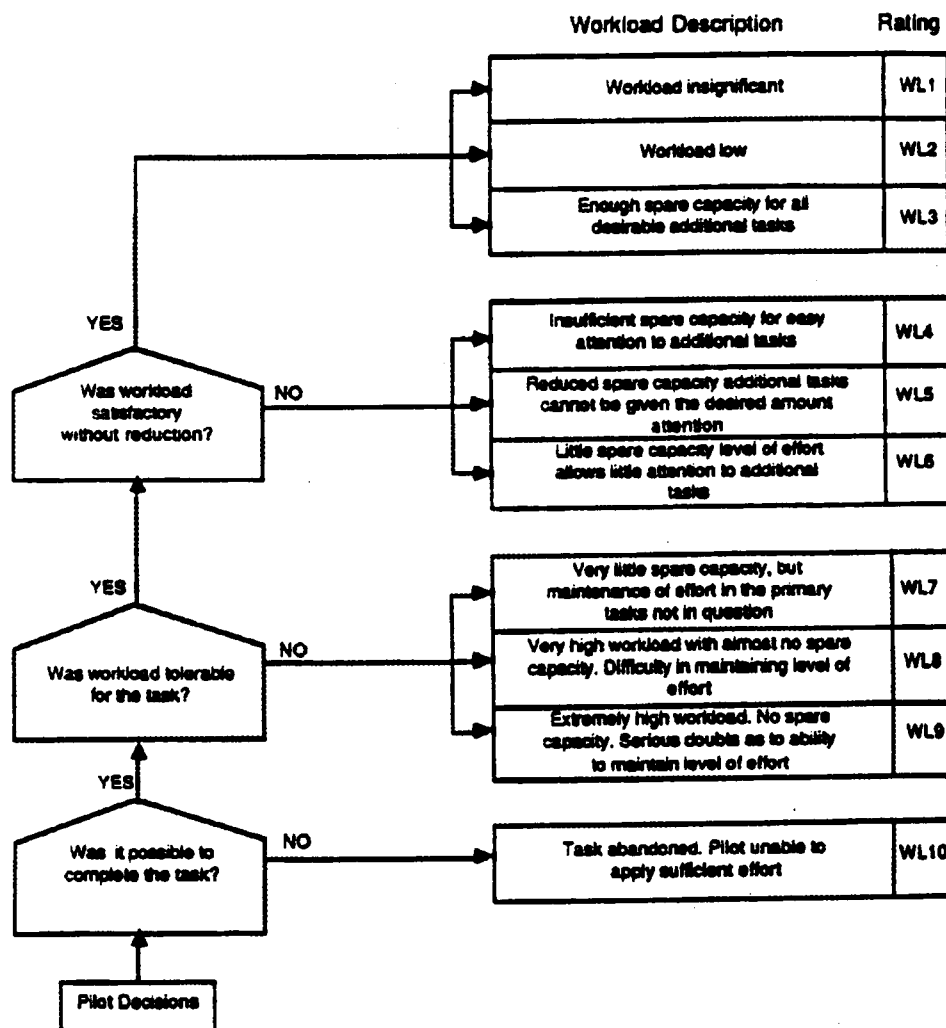


Figure 3. Bedford Rating Scale¹⁷

SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE	
NAME	DATE AND TIME
TASK BEING RATED	
(MARK AN X IN ONE CHOICE FOR EACH OF THE THREE AREAS BELOW THAT BEST DESCRIBES WHAT YOU BELIEVE THE TASK WORKLOAD TO BE.)	
I. 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.
II. MENTAL EFFORT	
☐ 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.
III. PSYCHOLOGICAL STRESS	
☐ 1	Little confusion, 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.

AFOTEC FORM
JUN 83 81

Figure 4. SWAT Rating Scale¹¹

The NASA Task Load Index is another multidimensional scaling system which adjusts the subject's initial rating in order to normalize it. But the TLX method is significantly different from SWAT in a number of ways. Instead of three workload components, TLX uses six: mental demand, physical demand, temporal demand, performance, effort, and frustration. And while SWAT raters chose subscale values of 1, 2, or 3, TLX raters use a subscale broken up into twenty levels, as depicted in Figure 5.

Additionally, TLX methodology for converting the subject's scores into a normalized value is a bit simpler than SWAT. The subjects weight the relevancy of each scale by making paired comparisons of each subscale to each other. The least weighted scale is dropped out by giving it a weight of zero. After the data is collected, each subscale is multiplied by the weighting factor then added together. Lastly, the total score is divided by the number of paired comparisons (15) to obtain workload scores from 0 to 100. The multidimensionality theoretically allows overall workload score use and comparison, as well as subscale use for diagnostic purposes. The logistics and training, administering, and data reduction of TLX make it less amenable to the T&E environment.¹¹

3.6.1.3 Rating Scale Selection

There are a number of guidelines for selecting workload measures. Six key features of good workload metrics include:

- a) sensitivity - the ability to detect changes in workload for a given task
- b) intrusiveness - the degree of interference with the operator task
- c) diagnosticity - the ability to discriminate the source of the workload
- d) convenience - time, money, and instrumentation required for measurement
- e) relevance/transferability - the applicability to a variety of different operator tasks
- f) operator acceptance - operators' willingness to employ the measure

A good measure should be sensitive, diagnostic, convenient to use, relevant to different tasks, and possess good operator acceptance.¹¹

NAME	TASK	DATE
------	------	------

TASK LOADING INDEX

Mental Demand How mentally demanding was the task?

Very Low
Very High

Physical Demand How physically demanding was the task?

Very Low
Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low
Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect
Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low
Very High

Frustration How insecure, discouraged, irritated and annoyed were you?

Very Low
Very High

AFOTEC FORM
JUL 84 SAH-10

Figure 5. TLX Rating Scale¹¹

All of the rating scale methods presented, the CH and derivatives, SWAT, and TLX have the favorable characteristics described, in varying degrees. For this project, the Bedford scale, a CH derivative, was chosen as the method which best satisfied the above guidelines. The Bedford scale was preferred over the SWAT and TLX methods for operator acceptance, implementation, and data reduction requirements. Additional comments of Bedford scale test implementation is continued in chapter 4, Test Plan and Execution.

3.6.2 Subjective Comparisons

The second issue in determining sensor assignment concerned how the sensor assignment options compared to each other, and which were considered the most preferable among the three. Statistically, the large number of IVs and DVs make traditional paired comparison data of little utility. In the test we had few subjects, low sample sizes, and lots of variables.

It was decided to evaluate the use of the Analytic Hierarchy Process (AHP), which is a methodology for developing ratio scales from paired comparison data. AHP is stated to have several advantages over traditional psychophysical approaches for data presentation. First, is its ability to quantify consistency in human judgments. Second, is its ability to provide empirical results with only small sample sizes, where traditional statistical procedures offer insufficient value. Additionally, AHP requires no statistical assumptions with respect to the distributions of human judgments.¹⁸

The AHP process begins with the user providing paired comparison data on the alternatives being studied. Each alternative is compared to each other, for a total of $[n(n-1)/2]$ comparisons. For this project, with 3 alternative sensor assignment options, a total of three paired comparisons existed. A suggested format is provided in Figure 6. This "nine-point" scale gives subjects the format to document their judgments of one alternative to another. The AHP comparison indicates not only that one alternative dominates a second, but also shows the degree to which one alternative dominates the second. The data analysis involves placing the data into a reciprocal matrix reflecting the dominance of the alternatives to each other. In the AHP, the weights assigned to the ratio scale are defined by the principal eigenvector that is calculated for the matrix. Commercial computer programs are available to perform the matrix manipulations. The eigenvector is typically normalized such that its components add up to one. This allows the final presentation of the data in a visual format such as a pie chart, displaying the alternatives and their degree of dominance.

For example, in our test if all three options were rated equally, with normalized values of 0.333, 0.333, 0.333, then a pie chart would show three equal areas. This methodology is further detailed in a number of references.^{18,19,20}

I	Absolute	Very Strong	Strong	Weak	Equal	Weak	Strong	Very Strong	Absolute	J
a_1	—	—	—	—	—	—	—	—	—	a_2
a_1	—	—	—	—	—	—	—	—	—	a_3
.										.
.										.
.										.
a_1	—	—	—	—	—	—	—	—	—	a_4
.										.
.										.
a_{k-1}	—	—	—	—	—	—	—	—	—	a_k

Figure 6. Recommended AHP Rating Scale¹⁸

3.6.3 Interviews and Questionnaires

In addition to rating scales, interviews and questionnaires are a second major category of subjective assessment techniques. While rating scales use psychometric methods to derive quantitative estimates of behavior, questionnaire and interview data tends to be qualitative.¹⁵ Open questions and open remark/comment sections on questionnaires allow the subject to document areas unforeseen by the questionnaire designer. These sections can also allow clarification to be made to answers in other sections of the questionnaire. While the author has found that the more general the open remark section is made the less responses are attained, their inclusion sometimes illicit useful responses. Significant data can also be attained from interviews, both one-on-one and group discussions. One-on-one interviews give the tester an environment to attain detail data that is often not captured on debrief questionnaires, for a variety of reasons, including time constraints and user acceptance of questionnaire formatting. Group discussions, or test team meetings also have their utility. Although data contamination is a danger from some personal observations

influencing others' opinions, the ability to bring together numerous ideas in a group format often stimulates new concepts and ideas.

3.7 Data Analysis

The data analysis methodology needed to be able to support the needs of the Bedford scaling, AHP operations, and mission performance measures. Numerous inflight operations are captured by aircraft computers and can be downloaded after the flight. A data transfer module (DTM) stores the information until downloading is accomplished using the debriefing section of the Mission Support System (MSS) computer. This data includes missile shot details such as weapon used, time fired, time of flight, and other pertinent information.

Inflight video recording was also accomplished during each mission. The airborne video tape recorder (AVTR) is capable of recording up to two hours of mission video. The system can record HUD video as well as MPD video, recording up to two displays at once. Aircrew and radio audio is recorded on AVTR tapes as well.

A Red Flag Measurement and Debriefing System (RFMDS) was also available for the test. With RFMDS, participating aircraft carry an aircraft instrumentation system (AIS) pylon which sends aircraft position and other data to ground receiving stations. Data analysis and mission viewing can be accomplished both real-time and postflight from a variety of orientations with the recorded data.

Operations Analysts (OAs) are also part of the test team. Specializing in data analysis, they have access to a variety of analytical tools including data reduction computer programs.

4.0 TEST PLAN AND EXECUTION

4.1 General

This chapter details the specific test methodology which was documented in the TD&E test plan. After test plan working groups, test team reviews, and a final review board, the plan was approved by the test center commander. Determining considerations for sensor assignment was an objective within the overall TD&E test. Next, this chapter covers the highlights of the test execution, focusing on the inevitable changes to the test plan which occurred due to circumstances beyond the control of the author.

4.2 Data Collection Plan

After a general review of applicable subjective test methodologies, the following five part data collection plan was adopted:

a) Workload scale. As described in the previous chapter, the Bedford scale was chosen for the test. Two changes were made to the original Bedford scale. First, the initial scale entry block was changed from "Pilot Decisions" to "Aircrew Decisions" since both pilots and WSOs would be using the scale. Second, the decision block between workload ratings scores (WL) 1-2-3 and WL 4-5-6 was modified. The test team management wanted to keep the emphasis on task completion instead of workload reduction because workload reduction is typically not an option in a combat scenario. Additionally, the original wording is most applicable for evaluation of a system under development, while in this case the system was already fielded. It was the workload and the ability to complete mission tasking that was under question in this test. It should be emphasized that the original Bedford workload descriptions that were used to assign the ordinal WL ratings were not modified in any way. These descriptors already use primary and secondary task completion for the assignment, the only effect of the change was to not require the aircrew to make a decision on the need for system modification. Lastly, as required for scale application, the evaluation and additional tasks were detailed at the bottom of the inflight kneeboard reference card. The test version on the Bedford workload scale is depicted in Figure A-1.

Workload rating statistics were planned as the primary method for illuminating sensor assignment options that were not viable for a given set of tactical conditions. As an initial planning tool, the TD&E test plan set a WL evaluation criteria above 7 (modal statistic) for identifying

sensor assignments which should be considered as non-viable options. Workload statistics would not be used totally in isolation. This data would form a baseline to assignment option viability decisions. Because of the large number of tactical scenarios under consideration, the test sortie matrix would be truncated as sensor assignment options were deemed not viable. This would allow remaining test missions to focus on viable alternatives.

b) Analytical Hierarchy Process. The AHP process would be used to illuminate aircrew preferences between the options, under changing variables and conditions. The recommended AHP terminology was used, as well as the recommended ratio scale.¹⁹ The questionnaire depiction of the AHP scale is shown in Figure A-2. AHP statistics were planned as the primary method for illuminating operator preferences for sensor assignment. AHP statistics would be used in conjunction with other test data to make test report recommendations.

c) Open Comments Section. As depicted in Figure A-2, an open comments section was made available following the workload and AHP assignment sections. This section could be used by the evaluating aircrew to supplement WL and AHP rating data.

d) One-on-one Interviews. One-on-one interviews were to be conducted with test team members periodically throughout the test period. These interviews would allow detailing information on sensor assignment as well as for data pertaining to other TD&E test objectives. Because of the wide topics available for discussion and limited aircrew availability, sensor assignment option discussions did not follow an additional questionnaire, but rather focused on areas of higher workload ratings and strong sensor assignment preferences. This forum also provided an additional area for test methodology and conduction feedback.

e) Test Team Discussions/Meetings. Group test team meetings were to be held by the author, as the TD&E test team project manager. During these forums, results from WL, AHP, and interviews would be presented for discussion. Test mission matrix modifications would be covered in these bi-monthly meetings.

4.3 Detailed Test Methodology

The data collected from the above data collection plan would be used by the test team to formulate the recommended sensor assignments for various tactical conditions. As described, three sensor assignment options would be considered for a given set of tactical conditions. After flying under the tactical conditions, aircrews would assign WL ratings. Workload statistics would then

be computed for each of the three sensor assignment options. If the WL rating for a sensor assignment was above 7 (modal statistic), then it would be eliminated as a viable option. The test mission matrix would be truncated to reflect this data.

Next, AHP comparisons would be considered for each of the viable sensor assignment options. After test matrix missions were completed, AHP data would be reduced to show aircrew preferences. This data would be used by the test team management in the drafting of test report recommendations. Additionally, interview data, comments from questionnaires, discussion from previous test team meetings, and WL data would be considered in developing the recommendations. Preliminary recommendations would then be presented before the test team for their review. After the test team review, comments would be incorporated, and the sensor assignment recommendations for specific tactical conditions would be documented in the test report. Figure 7 highlights the detailed methodology application for a single tactical condition.

4.4 Aircrew Test Preparation

Prior to active testing, it is typical to get the test team aircrew prepared for the test methodology, as well as to achieve an operational level of proficiency in the given system. A test team meeting was held before the first test mission to brief the team on the chosen methodology and involved systems. The team briefing covered mission procedures, mission folder contents, and questionnaire procedures. The WL scale and the AHP scale were covered in detail since they had not been used during previous testing in the unit. Examples were presented on application of these systems. The author made instructional comments on ratings accomplished during training missions described below. Additionally, a TD&E read file was created for covering the methodology in detail, as well as for providing additional feedback and informational updates to the aircrews throughout the test.

The test involved two of the most complicated systems in the aircraft, the radar and the TGP. Because both systems had recently had operational flight program (OFP) updates, "spin up" was accomplished in order to attain what would be an expected operational level of proficiency. In OT&E testing it is not uncommon to have longer periods of time between accomplishing a given type of mission than operational units. But, once in a test the particular mission is often accomplished with greater frequency and in greater depth than compared to an operational unit. While the OFP had been loaded in the test aircraft for approximately six months, these spin-up

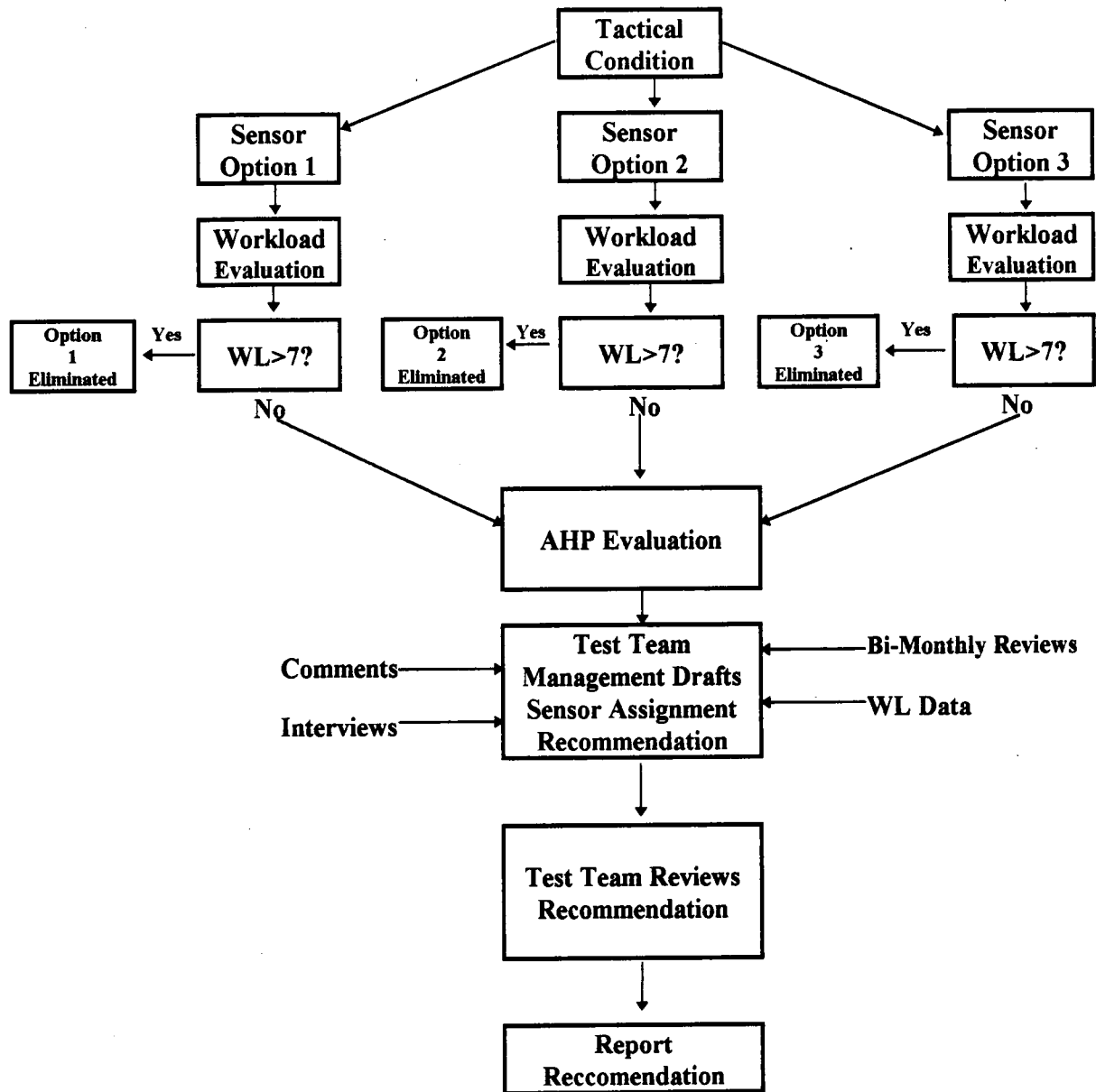


Figure 7. Tactical Condition Sensor Assignment Methodology

missions focused on TGP systems, and consisted of system briefings and readings, and included dedicated inflight drills to practice with the new systems. Typically, inflight drills were followed with benign tactical intercepts and finally included full up tactical scenarios. For the "spin-up" flights, pilots were limited to a minimum of 500 feet above ground level flight while operating the TGP.

4.5 TD&E Mission Phase Plan

Test events in DT&E and OT&E testing often use different terminology. This test used the term "mission" to annotate an entire test formation, typically a two, three, or four-ship flight. Within the flight, 1 of 3, 2 of 3, etc., was used to identify the specific aircraft and crew. Mission numbers began with the first mission and continued in chronological order. Within a particular mission, "data passes" were used to annotate fully scripted maneuvers accomplished for system performance characteristics, and the term "trial" was used to describe a complete ingress, target strike, and egress. "Engagements" were not specifically annotated, but referred to particular occurrences within a trial.

"Phases" were used to divide and annotate major categories of mission types. The initial plan was to begin with phase one missions, which would consist of 2-ship or 3-ship missions which primarily focused on data passes, and to increase proficiency and identify concepts of employment. WL and AHP ratings would not be assigned in phase one. Phase two would develop tactics and sensor assignments, under various conditions, during tactical intercepts involving up to 4-ship versus 4-ship formations. The trials in this phase would involve both "scripted" and unrestricted adversary tactics. Three trials would be accomplished on a phase two mission, with each of the three sensor assignment options being accomplished (assigned option order randomized). This would allow full completion of the AHP chart each phase two mission. Phase three was planned as the "graduation" exercise, or validation phase, with multiple support and adversary aircraft and unrestricted tactics. Phase three missions would have the largest airspace and ground controlled intercept (GCI) support. Workload ratings would be assigned on each mission, but since only one full trial was expected on a mission, AHP ratings could not be assigned. Full mission performance data would be documented only during phase three missions. A preflight mission tasking sheet is depicted in the Figure A-3.

4.6 Mission Scheduling

Unforeseen circumstances at the beginning of the test eliminated the support and adversary aircraft that were to be used for phase three missions at the end of the test. This would drive the test team to make tactical validation decisions based on smaller combat scenarios, which is less than desirable. But, a separate TD&E test that was being completed at the beginning of this test requested F-15E test support. This would allow the test team to get phase three type data, but at earlier point in the test. The test team opted to fly these missions, thereby accomplishing "combined" TD&E testing. During this portion of the test the test team management did not have total control over the scenarios, and did not get all the type of setups that were planned. Additionally, employing in phase three missions was difficult since aircrew proficiencies were not at their highest level and tactics had not been fully developed. On the positive side, the early difficult missions served as a baseline against any later phase three missions which would be flown, illuminating beneficial aspects to training with the system and developed tactics. This would show how much improvement could be expected with dedicated training. Additionally, the level of proficiency at this point in the test was not unrealistic for an operational unit, since test aircrews had been flying with this particular OFP for six months, and had accomplished a dedicated spin-up phase.

Phase two missions also did not go completely as planned. The length of the trials, combined with the lack external fuel or air-to-air refueling prevented three full trials from being accomplished on most phase two missions. This limited AHP comparisons of all sensor assignment options on a single mission. Sensor assignment options and test conditions were varied in order to give the best opportunity for like comparisons.

4.7 Data Compilation

In general, the quality and quantity of the data that was obtained in the test increased as the test progressed, for a number of reasons. Time allotted for proper questionnaire completion was not adequate in the beginning of the test. Due to short manning, some aircrews in a mission were required to "double-turn" into a second flight after landing from the first flight. When this occurs, there is little time between landing from one mission and beginning the brief for the second mission. This drives debriefing details, tape review, and paperwork from both missions to be completed after the second mission. If only one of the two crewmembers is flying twice, typically the

crewmember flying once will complete all postflight questionnaires, possibly losing some data from the missing crewmember.

There were some initial misunderstandings as to the proper way to complete the Bedford and AHP scales during spin-up and some early missions. These problems were addressed in one-on-one interviews, test team meetings, and through e-mail. For WL ratings, it was observed that during spin-up and early missions that pilots were occasionally assigning WL ratings when the WSO was assigned both sensors. Also, it was noted that WL ratings were only infrequently assigned while inflight, due to the dynamic nature of the environment, as well as the fact that it was felt it could adequately be addressed after the mission. In some Cooper-Harper handling qualities rating applications, numerous ratings are made throughout a single mission, and these perishable ratings are difficult to remember and assess postflight. In this particular application of the WL scale, typically only a total of one or two assignments were made for the entire mission, and postflight video and other data documentation aides facilitated the rating assignment. For these reasons postflight documentation was considered acceptable.¹¹

For AHP, it was observed that in some early cases all three option comparisons were being ranked when three trials had not been completed in that mission. It appeared that the comparisons were being annotated on a general level of how the assignments compared, as opposed to only ranking the comparison based on the mission just flown, for that given set of conditions. Additionally, there were occasions where the pilot and WSO did not agree on the AHP ratings. The test team decided in this scenario that each aircrew could document their given preference, separately, with a "P" or a "W". The open comment section of the questionnaire generated almost no aircrew response with respect to sensor assignment. This section was available for test comments pertinent to anything in the entire TD&E test, and it is felt that had the priority of this objective been higher, an open comment section dedicated to this issue might have generated more comments.

Interviews and test team meetings generated significant data from a quantity and quality standpoint. This format allowed greater probing of aircrew insight, and also drove the subjects to put additional time into formulating their comments and opinions of test data and the developed test procedures.

5.0 RESULTS, CONCLUSIONS, AND RECOMMENDATIONS

5.1 General

It should be kept in mind that the results, conclusions, and recommendations are made on the test methodology, not on the specific problem of actual sensor assignment options. While some generalized test data is presented for illustrative purposes, attention is not intended to be taken away from the methodology.

It should also be reiterated that the conclusions apply specifically for the subjective measures as applied to the unique type of military T&E testing accomplished in this test. Some of the test characteristics were that this application was part of a larger T&E, included numerous types of variables, lacked mission performance measures for substantiation of subjective measures, and had significant time, cost, and manpower limitations.

5.2 Data Analysis and Reduction Results

As delineated in the test plan, TD&E limitations would not permit a large number of missions to be flown with every assignment option for every variation of the factors and conditions. This was taken into account in the planned mission test matrix by reducing the number of planned missions for sensor assignment options under conditions which it was felt that the workload would not be acceptable. As these missions were flown, confirmation was attained, using the described methodology and results delineated in the following paragraphs, that these variations were not viable. For these cases where limited missions were flown, it was felt by the test team that full statistical analysis (AHP computations) were not needed. Essentially, the raw data from these data points were evaluated to see if the expected results had occurred.

The following paragraphs highlight data analysis and the results which were observed in each of the subjective methodology areas.

5.2.1 Workload Ratings

Workload rating data reduction discussion first needs to address the different levels of measurement, including nominal, ordinal, interval, and ratio.²¹ Nominal measures only place objects in classes, while ordinal measures rate objects in an order of magnitude. But the distance between the objects in ordinal measures is not defined. Interval measures have equal scaling

intervals but do not have a fixed zero point. Ratio measures have equal intervals and a known zero point, as in a ruler.¹¹ Bedford's scale, along with other Cooper-Harper derivatives, are ordinal in nature. The implication of this, as an example, is that a ranking of a WL 2 and a WL 6 does not translate into an average WL 4. The intervals between WL ratings are not treated as being equal.

For measures of central tendency of workload ratings, the sample mean or average is inappropriate. Instead, the mode, the value that occurs most often, or the median, the middle value of a series, statistic should be used to summarize the data. While not specified in the test plan, modal statistics were used to aide in workload viability decisionmaking.¹¹

WL data was used in a number of ways. First, when limited missions were flown to show that certain sensor assignments were not viable, WL statistics provided this confirmation. This allowed the truncation of the mission test matrix and saved valuable resources. For example, only six data points were used to show that the WSO should be operating the sensors when the pilot is flying as a wingman at night while performing low altitude terrain following.

WL data reduction showed definitive increases in WL ratings under the following conditions for the pilot; operating at low altitudes with either sensor, with wingman formation duties with either sensor, and with TGP control during terrain following at night. Combinations of the above conditions resulted in WL ratings above or near the established WL threshold. For the WSO, higher WL ratings were noted when the WSO was assigned both sensors and the conditions included adverse TGP backgrounds or multiple adversary groups. WL ratings approached or exceeded thresholds with these two conditions combined. As specified in the test plan, sensor assignments which resulted in modes above seven were not considered as viable options.

While WL statistics were used to determine option viability, they were also used in forming recommended sensor assignments, as a supplement to AHP ratings. For example, sensor assignments under conditions where pilot modal statistics were seven or above were not recommended in the TD&E test report.

Data reduction analysis showed that there was a tendency, under some tactical conditions, for pilot WL ratings to be lower than anticipated as compared to interview and test team results. This might be attributable to a lack of previous experience using WL rating scales in earlier TD&Es, especially considering that the participants were not graduates of a military test pilot school (with the exception of the author). Additionally, it is possible that aircrews may have been reluctant to annotate assignment options as unaccomplishable. Pre-test cautions should be given to minimize

this potential effect.¹⁰ For the WL ratings annotated during postflight work, the delay in making the assignment may have weakened some of the perceived difficulty of the task. As discussed earlier, the requirement for typically only one or two WL ratings and sophisticated postflight debriefing aids should have limited the impact of this procedure.¹¹

5.2.2 AHP Ratings

AHP data reductions, which are generally accepted as ratio measurements, were in fair agreement with WL measurements. For example, under conditions with lower pilot WL rankings (medium altitude flight as flight lead) and higher WSO WL rankings (numerous threat groups and adverse background concerns), paired comparison data showed preferences against WSOs being assigned both sensors to pilot control of the radar or TGP.

Paired comparison results were used in the formulation of sensor assignment recommendations made to the operational community in the TD&E test report. As with WL ratings, AHP data reduction was not accomplished under sensor assignment options which were accomplished during limited test scenarios. But, raw paired comparison data could be used in these instances to confirm that certain sensor assignments were not favorable for the tactical scenario.

AHP ratio values computed for large categories, such as comparing the three sensor assignments over the various factors, presented the preferred sensor assignments in a generalized format used in summarizing the test results for the TD&E test report. Operations analysts reduced data using computer methodology discussed in paragraph 3.6.2.

5.2.3 Comments, Interviews, and Test Team Meetings

While open comments, interviews, and test team meetings did not have statistical data reduction requirements, their qualitative information was compared to and used in conjunction with WL and AHP results. Comparison of these methodologies is detailed in the following sections.

5.3 Individual Methodologies Conclusions

5.3.1 WL Ratings

a) While no general threshold value is established for WL rankings, values above seven were felt to have identified conditions where the option being evaluated was not viable. Sensor assignments with modes above this value were used to reduce the TD&E test matrix, allowing more attention on tactically feasible divisions of duty.

b) WL ratings can help with making sensor assignment recommendations. For example, options which generated pilot WL rating modes of seven were not considered to be unaccomplishable, but they were not recommended. Combined with data from other measures, WL values helped to formulate recommended sensor assignments.

c) Pre-test preparation and operator instruction is minimal. Depending on the operators' previous experience, ground instruction and application examples, combined with a limited amount of postflight instruction might be adequate.

d) WL ratings had utility from the standpoint of generating discussion of the required elements required under various assignment options, even without statistical analysis.

e) WL scale usage added to the overall methodology. But, it was felt that this method alone would not have supplied sufficient information needed for sensor assignment conclusions.

5.3.2 AHP Comparisons

a) While WL ratings helped outline which options were viable, AHP forced direct comparisons between them. Since AHP values are derived with ratio properties in mind, the reduced data allowed this type of result presentation. For example, results could indicate that one option was preferred over another by a three to one margin.

b) Missions which accomplished all three options mirrored the overall workload rankings of the mission in that the options with the higher workload received lower rankings. Additionally, an amount by which options were preferred was present in the AHP results. This information was valuable from the standpoint of making conclusions regarding various sensor assignments.

c) A major limitation of getting useful information from the AHP comparisons was that few missions were able to accomplish all three alternatives. Comparisons of the assignment options between different missions were not felt to be similar enough to allow AHP rankings.

d) AHP data reduction required advanced computer processing, which reduced the utility of the method in this application.

e) Completion of AHP ratings instigated productive discussion of the utility of the various sensor assignment options, and was most valuable when used in conjunction with the other methods. Stand-alone use of AHP in this application would have not been sufficient for answering the objective.

5.3.3 Open Comments

Due to the additional requirements of the TD&E, an open comments section was not dedicated to workload evaluation. Few workload responses were solicited under this forum. If this objective was a higher priority in the overriding TD&E, a dedicated open remarks section with leading questions might have produced more data.

5.3.4 Interviews

One-on-one interviews were probably the most beneficial subjective measures for reaching conclusions to this project objective. Focused subject attention, the ability of the subject and the examiner to detail responses, and the ability to probe into subject preferences contributed to this conclusion.

5.3.5 Test Team meetings

Test team meetings were productive for helping to draw conclusions. While these group discussions can sometimes bias subject responses, the ability for an innovative thought process to occur is increased because of the wide variety of perceptions of the subjects. Used sparingly and with proper managing, they were a useful tool.

5.4 General Conclusions

The overall test methodology was successful in developing working considerations for sensor assignment for various critical tactical conditions. Neither WL or AHP ratings alone would have been capable of reaching the conclusions drawn from combining both methodologies. Interviews and test team group discussions were the major source of data used in reaching sensor assignment conclusions. Because of the large number of variables in this T&E environment, WL and AHP ratings had reduced effectiveness and increased data reduction requirements. This limited their impact on conclusions to a secondary role. Even with limited statistical effectiveness, their inclusion raised aircrew awareness, and forced sensor assignment preference decisions. Overall, their inclusion in the TD&E was helpful but not critical to reaching objective conclusions.

5.5 Recommendations

While the combined application of the chosen subjective measures was developed in order to attack a relatively unique problem, the methodology could have further application. In military

tactics development, where qualitative tools are the primary focus of reaching conclusions, the combination of WL and AHP can provide a simple, effective means of adding statistical basis for judging various options.

LIST OF REFERENCES

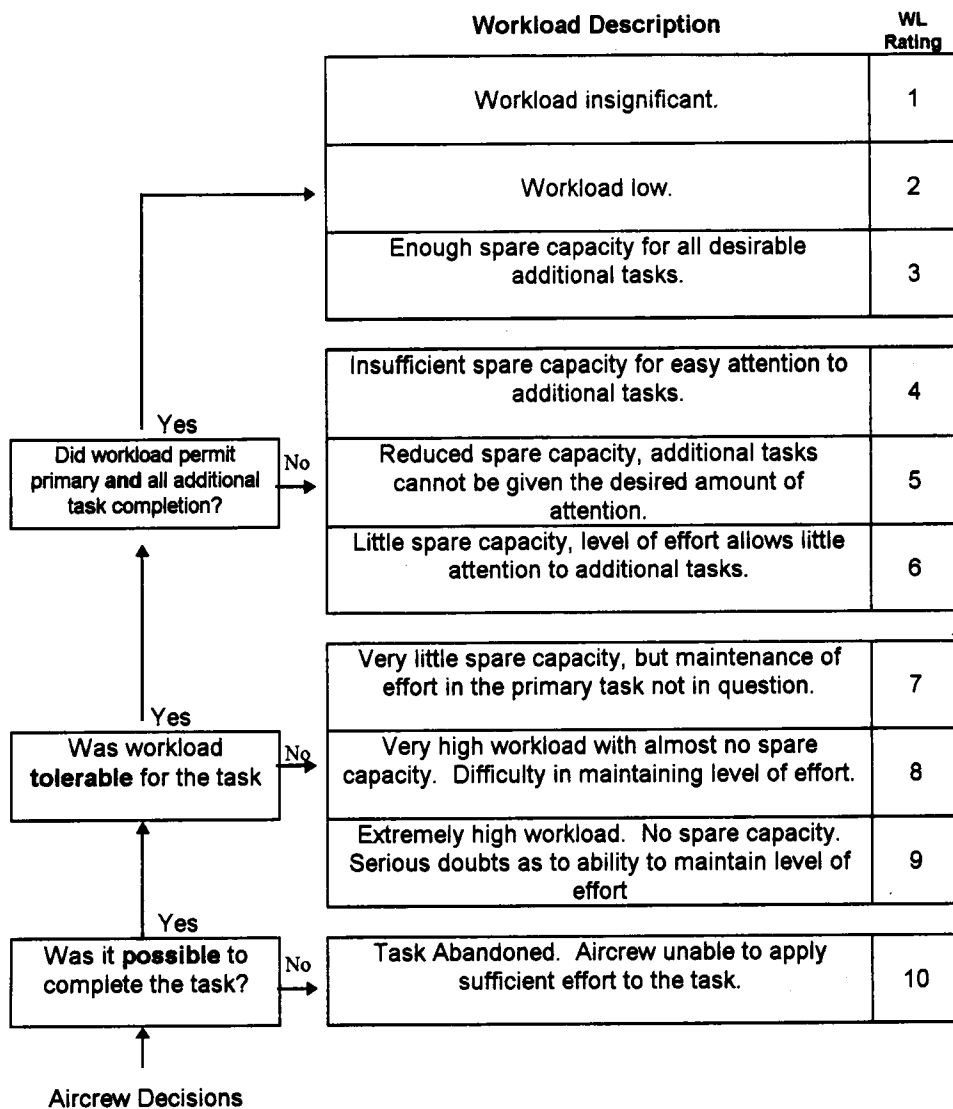
LIST OF REFERENCES

1. Stimson, George W. Introduction to Airborne Radar. El Segundo, CA: Hughes Aircraft Company, 1983.
2. Scott, William B. "LANTIRN Gives Tomcat Night Attack Role." Aviation Week & Space Technology. 10 June 1996, 40-43.
3. TO 1F-15E-1 Flight Manual, 15 April 1993, with Interim Change No. 5, 01 October 1995, 1-1 to 1-196.
4. Gunston, Bill, and Mike Spick. Modern Air Combat: The Aircraft, Tactics, and Weapons Employed in Aerial Warfare Today. New York: Crescent Books, 1983, 124-125.
5. TO 1F-15E-34-1-1 Weapons Manual, 15 January 1994, with Interim Change No. 4, 15 February 1996, 1-350 to 1-391.
6. Sanders, Mark S., and Ernest J. McCormick. Human Factors in Engineering and Design. New York: McGraw-Hill, 1993, 24-39.
7. Simon, C. "Analysis of Human Factors Engineering Experiments: Characteristics, Results, and Applications." Westlake Village, CA: Canyon Research Group. August 1976.
8. Kantowitz, B. H. "Selecting Measures for Human Factors Research." Human Factors. Vol. 34, 1992, 387-398.
9. Weimer, Jon. Research Techniques in Human Engineering. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1995, 377-383.
10. Wierwille, W. W., and F. T. Eggemeier. "Recommendations for Mental Workload Measurement in a Test and Evaluation Environment." Human Factors. Vol. 35, 1993, 263-281.
11. O'Brien, Thomas G., and Samuel G. Charlton. Handbook of Human Factors Testing and Evaluation. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 1996, 93-98 and 180-199.
12. Cooper, G. E., and R. P. Harper. "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities." NASA TN-D-5131. Moffett Field, CA: NASA Ames Research Center, 1969.
13. Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., Zaklad, A. L., Bittner, A. C., and R. J. Wherry. "Operator Workload: Comprehensive Review of Evaluation of Operator Workload Methodologies." US Army Research Institute Tech. Rep. 851. Alexandria, VA. June 1989.

14. Wolf, J. D. "Crew Workload Assessment: Development of a Measure of Operator Workload." Air Force Dynamics Laboratory AFFDL-TR-78-165. Wright Patterson AFB, OH. 1978.
15. USNTPS Staff. "US Naval Test Pilot School Subjective Test Techniques." USNTPS, Patuxent River, MD. 1994, 81-90.
16. Wierwille, W. W., and J. G. Casali. "A Validated Rating Scale for Global Mental Workload Measurement Application." Proceedings of the Human Factors Society 27th Annual Meeting. 1983, 129-133.
17. Roscoe, A. H. "In-flight Assessment of Workload Using Pilot Ratings and Heart Rate." The Practical Assessment of Pilot Workload, AGARDograph No. 282. Neuilly Sur Seine, France. 1987, 78-82.
18. Mitta, Deborah A. "An Application of the Analytic Hierarchy Process: A Rank-Ordering of Computer Interfaces." Human Factors. Vol. 35, 1993, 1141-1157.
19. Saaty, Thomas L. The Analytic Hierarchy Process. New York: McGraw-Hill, 1980.
20. Saaty, Thomas L. "How to Make a Decision: The Analytic Hierarchy Process." Interfaces. November-December 1994, 19-43.
21. Stevens, S. S. "On the Theory of Scales of Measurement." Science. Vol. 103, No. 2684, 07 June 1946, 677-680.

APPENDIX

SENSOR OPERATION WORKLOAD RATING SCALE



Evaluation Task: To maintain primary operation duties of the assigned sensor (radar, TGP, or both) for the entire trial. Specifically, detecting, tracking, monitoring, and maintaining situational awareness on the target.

Additional tasks: Include checking RWR, formation positionkeeping, navigation, situational awareness building, communication, etc.

Figure A-1. Aircrew WL Rating Scale

C. SENSOR OPERATION ROLES

1. Workload ratings:

Trial 1: Sensor setup: Primary Operator: FCP: Radar Pod None RCP: Radar Pod Both

FCP Workload Rating (if applicable) WL _____. RCP Workload Rating WL _____

Trial 2: Sensor setup: Primary Operator: FCP: Radar Pod None RCP: Radar Pod Both

FCP Workload Rating (if applicable) WL _____. RCP Workload Rating WL _____

Trial 3: Sensor setup: Primary Operator: FCP: Radar Pod None RCP: Radar Pod Both

FCP Workload Rating (if applicable) WL _____. RCP Workload Rating WL _____

2. Role Comparisons.

After trial completion, annotate aircrew consensus (the applicable pilot and WSO) of preferred operator duties (mark an x in each applicable comparisons).

These roles better	Abs Better		Much Better		Better		Slightly Better		Same		Slightly Better		Better		Much Better		Abs Better	These roles better
WSO Both																		P-Rdr W-Pod
WSO Both																		P-Pod W-Rdr
P-Rdr W-Pod																		P-Pod W-Rdr

D. ADDITIONAL COMMENTS. Make any additional comments on sensor assignment. Note any strong likes or dislikes or potential improvements for Suite 2 TGT IR Pod Software or Hardware. Describe any OFP anomalies.

E. TEST CONDUCTION FEEDBACK

How can we make the test better, or run more smoothly. Include comments on gameplan, documentation, standards, etc.

Figure A-2. Posflight Aircrew Debrief Form

MISSION TASKING

Phase# _____ Mission# _____ Phase: 1 - Spinup/Systems
 2 - Tactics Development (single ship / element)
 3 - Tactics Development & Validation (Large Force Scenarios)

REQUIREMENTS FOR THIS MISSION:

Force Minimum = _____ or TD&E Piggyback Day / Night Other:

MISSION PLANNING NOTES / MISC:

Date: _____ 1st Go 2nd Go

RFMDS: Yes / No Adversaries: POC

Airspace: Type/Number

(See the TD&E Read File or the test plan for explanation of options and abbreviations below)

A. TRIAL SETUP AND VARIABLES - CONDITIONS

1. Test Role: OCA Strike DCA

2. Primary Sensor Operator:

Trial 1:	FCP: Radar	Pod	RCP: Radar	Pod	Both
Trial 2:	FCP: Radar	Pod	RCP: Radar	Pod	Both
Trial 3:	FCP: Radar	Pod	RCP: Radar	Pod	Both

3. Armament

Adversary:

OCA1	xxxxxxxxxxxxxxxxxxxx	XXXX	xxxxxxxxxxxxxxxxxxxx
OCA2	xxxxxxxxxxxxxxxxxxxx		
DCA	xxxxxxxxxxxxxxxxxxxx	XXXX	xxxxxxxxxxxxxxxxxxxx

4. ROE: ROE1 ROE2

5. Altitudes/Formation: (Phase 2 Missions)

For testers this is the primary altitude *gameplan* for Ingress - Attack - Egress

For adversaries this is the *initial look* (maintain until a minimum of 20 nm, pre/post F-15E A/G delivery).

Altitude Definition: Med/High >10,000 msl, Low < 10,000 msl

Test Aircraft:	Med/High	Low	Low
	Day-Tactical / Night-Trail	Tactical	TF (use night blocks)

Adversary: : Med/High Low Night Blocks, entire trial ≥15K'

6. Adversary Tactics Unrestricted Scripted

7. TGR IR Default Setup:

Option: TRK boxed TRK unboxed PB7: xxxxx xxxxxx

Figure A-3. Preflight Mission Tasking

VITA

Thomas Plumb was born in Washington, D.C., on March 29, 1962. He attended Oakton High School where he was a member of the National Honor Society as well as a 4-year letterman in wrestling and soccer. After graduation from high school, Mr. Plumb attended the University of Virginia, and graduated with a bachelors degree in Aerospace Engineering in May, 1984. He was a member of the USAF ROTC, attended US Army Airborne training, and was commissioned a second Lieutenant. While attending college, he was a black belt Karate instructor, member of the Sigma Pi fraternity, and a participant in many sports activities. When not attending college, Mr. Plumb worked at the Naval Air Systems Command, and the Naval Ship Research and Development Center, under a fellowship program.

Mr. Plumb attended Undergraduate Pilot Training (UPT) at Vance AFB, OK, 71st Flight Training Wing. He was selected as a T-38 first assignment instructor pilot following graduation. Mr. Plumb attended Pilot Instructor Training at the 12th Student Squadron, 12th Wing, Randolph AFB, and was a distinguished graduate. While an instructor pilot in the 25th Flight Training Squadron, he was a check pilot, and a T-38 Flight Examiner in the 71 Flight Training Wing.

Mr. Plumb was selected for the first F-15E B-course, and attended Lead In Fighter Training (LIFT) at the 435th Fighter Training Squadron, 49th Wing, Holloman AFB, NM, where he was the top graduate. Following LIFT, Mr. Plumb attended F-15E Replacement Training Unit at the 550th Fighter Training Squadron, 58th Wing, Luke AFB, AZ. Mr. Plumb was the first non-previous fighter pilot to be qualified in the F-15E.

Mr. Plumb was initial cadre in the F-15E while assigned to the 336th Fighter Squadron, 4th Wing, Seymour Johnson AFB, NC, from May 1990 to June 1993. During this time he was an F-15E instructor pilot, mission commander, flight lead, wing supervisor of flying, and Functional Check Pilot. While deployed to Southwest Asia during Desert Storm, he flew 34 combat sorties for 116 combat hours, both over Kuwait and over Southwest Iraq in Scud hunting missions.

Mr. Plumb was the first F-15E pilot to be selected for attendance at the Naval Test Pilot School, Patuxent River Naval Air Station, where he graduated in July of 1994. He flew 27 different models of aircraft while at USNTPS. Mr. Plumb has been a test pilot at Nellis AFB, Air Warfare Center, and F-15E Tactics and Test Manager from August 1994 to the present.

Mr. Plumb is married to the former Terri Ralson, and resides in Las Vegas, NV.