VERIFICATION AND VALIDATION TESTING OF THE PILOT'S ASSOCIATE

Norman D. Geddes

Applied Systems Intelligence, Inc.
3453 Point View Circle
Gainesville, GA 30506

ABSTRACT

The DARPA/USAF/Lockheed Pilot's Associate is one of the most comprehensive and complex real-time artificial intelligence decision aids ever attempted. In the course of its development, a number of major issues in testing this system have emerged, both as a result of the rapid prototyping methodology used to develop the Pilot's Associate, and due to the fundamental nature of AI-based decision aids. The issues of verification and validation have been addressed as an integral part of the design and development process over the course of the project.

This paper discusses how the Lockheed team is conducting the verification and validation of the Pilot's Associate as an integrated methodology that brings design and test into a closer relationship. The methodology encompasses digital combat simulations, engineering development testing of individual software units and subsystems, prototype testing, manned functional testing, real-time performance testing, and a manned operational performance evaluation.

INTRODUCTION

The operational success of a decision aid depends on its perceived value by the user. Testing the effectiveness of the integration of human and machine intelligence is an important emerging concern for fielding AI-based tactical decision aids. The validation of the Pilot's Associate system, shown in Figure 1, depends not only on measured performance gains, but also on issues such as shared situation awareness, tactical flexibility, joint planning and the cognitive demands that the Pilot's Associate places on the pilot. Testing at this level requires a careful inspection of measures of merit and the role of operational scenarios in stressing the human-machine system.

The emphasis for PA development in Phase 2 falls in two important areas, real-time performance of the PA system, and demonstration of operational utility of the aiding provided by the PA. These two facets of the project are tightly related. While it has been acceptable to trade functionality for speed in order to achieve the needed timeliness of the PA system, the real-time functionality that results must provide an operationally important gain in capability for the pilot. The Phase 2 design process must make effective trade-offs in selecting the most useful functionality to implement in the real-time system design.

The current rapid prototyping design approach permits independent development of the subsystems, with periodic integration testing at the LASC simulation facility in Marietta, Ga. Each rapid prototyping cycle is controlled through a planning process that includes selection of functional goals for the next integration, definition of test cases, and a schedule for development and integration. The results of each prototyping cycle are used to formulate the functional goals for the next cycle.

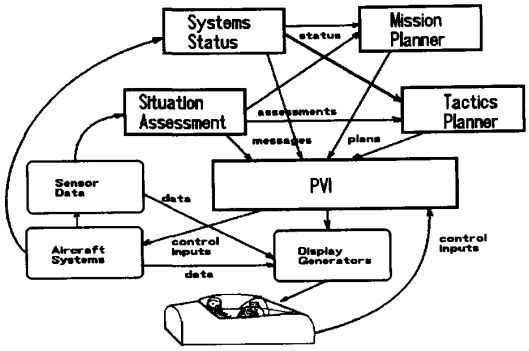

Figure 1. Pilot's Associate system

A fundamental precept of the rapid prototyping methodology used by the Lockheed design team is the evolution of requirements as the design proceeds. Testing of the design is interwoven with the design activity itself in a series of short design and test cycles, lasting typically four to six months each. The candidate requirements for the overall system are allocated to prototyping cycles as design goals. The prototype testing process must provide an accurate appraisal of the technical progress made during the prototyping cycle toward the allocated requirements. In addition, it must provide guidance on the expected utility of meeting the currently defined requirements with the current partial real-time design. This allows modification of requirements statements in cases that have shown little promise of utility or favorable cost benefit ratio for the final design. Prototype testing therefore includes elements of both design verification and requirements validation.

Integrated Test Objectives

Since its inception, the PA project has employed an integrated approach to testing that encompasses development, verification, validation, and performance testing. The objectives of the integrated testing process are:

CH3030-4/91/0000-0426 $1.00 © 1991 IEEE

a. Provide useful design feedback to the PA design team. The PA is primarily an exploratory development project. As a result, emphasis should be placed on investigating the validity and robustness of the design approach, rather than on merely meeting the stated PA requirements.

b. Measure the incremental design success of the PA in meeting the requirements that were defined for it at each prototyping cycle and at the conclusion of the project. The test program has an important role in providing management with an accurate technical assessment of the project's progress.

c. Construct and refine a set of measures of merit useful for evaluating pilot aiding systems. Selection of appropriate and sensitive measures for assessing the behavior of the PA is a vital importance, both in assessing the Phase 2 artifact and in extrapolating its potential into the future as a full scale engineering development project.

d. Measure the operational utility and technical success of the PA aiding concepts as implemented in the Phase 2 design. Even though it is expected that the system implementation will be less than full-mission capable, it is important to establish the operational benefits of the functionality that has been implemented and its real-time performance in the chosen hardware and operating system environment.

OVERVIEW OF THE TEST PROGRAM

The Phase 2 testing program consists of five major components, as shown in Figure 2.

a. Digital simulations. The digital simulation component utilizes Monte Carlo unmanned digital simulation to rapidly explore the potential benefits of PA-like system functionality over a broad range of mission profiles. The output of this testing component is the identification of high-leverage testing situations for planned PA functions and candidate measures of merit for PA evaluation. This test activity supports both engineering development test and the formal assessment of PA utility. Furthermore, the digital simulations are expected to support knowledge base development by providing a consistent view of future combat scenarios.

b. Engineering development testing. During each prototyping cycle, subsystem functionality is extended to meet the goals for that cycle. Engineering development testing provides local design feedback and verification for the subsystem design teams prior to and during system integration. The development tests use a common set of test cases defined for each prototyping cycle.

c. Prototype testing. At the conclusion of the prototyping cycle, the subsystem functionality is tested as an integrated system at the LASC simulation facility. These tests are based on an extended test case library that includes cases used for engineering development tests. This testing provides the primary design feedback for setting the goals for the next prototyping cycle. It is also the primary source of technical progress assessment.

d. Functional testing. Two test periods for contractor manned system testing in a mission context are planned. These testing periods provide mission-level feedback to the subsystem designers and refine test procedures and metrics for use in the final Real-Time Performance testing and the Manned System Evaluation.

e. Real-Time Performance (RTP) testing and Manned System Evaluation (MSE). When the Phase 2 development process is concluded, a formal evaluation of the PA aiding functions is scheduled. The testing will investigate both the real-time performance and operational utility of the PA aiding across a broad range of task-level and mission-level measures of merit. The RTP tests will utilize a library of test cases developed to stress the performance of the PA along certain critical paths of data flow through the system. The stressing test cases used for RTP testing may not be completely representative of actual mission conditions.

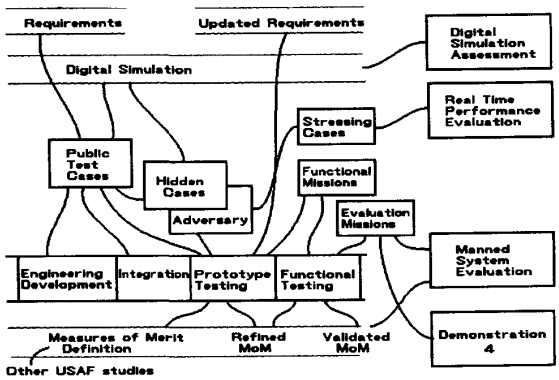

Figure 2. Integrated testing process

The MSE will provide an assessment of the PA utility using full-mission, manned simulation. The MSE provides a manned simulation with a direct comparison between a baseline aircraft configuration based on an advanced fighter, and a PA-configured version of the same simulated aircraft.

These five components provide a testing framework that supports all of the integrated test objectives. Furthermore, the integration of this framework over the design and development cycles of the PA provides an efficient and thorough test environment at each point in the development cycle.

TEST COORDINATION ISSUES

Because this is a rapid prototyping development process in which both PA requirements and actual implemented PA functionality change over the course of a prototyping cycle, close coordination of the test planning process and the engineering development process are necessary. The test coordination involves mission environment coordination, requirements synthesis, test planning and test environment coordination.

Mission Environment

The the Fighter Sweep mission environment has been chosen for the PA-equipped aircraft. This mission type is well-established in tactical air combat doctrine and are expected to be a major role responsibility for advanced tactical fighter aircraft into the foreseeable future. Preliminary analysis of this mission environment to verify that the expected mission complexity is achieved has been conducted using the digital

simulation environment described later in this section. As a result of the digital simulation results, several forms of definition of the mission environment have been formed:

a. Public test cases. These test cases are short segments of missions that have been shown to be particularly interesting during digital simulation. The public cases support requirements definition, engineering development tests, prototype integration and prototype testing. Approximately 15 public cases are planned, each representing 6 to 8 minutes of activity.

b. Hidden test cases. The hidden cases are variations on the public cases to be used during prototype testing to assess the robustness of design. Approximately 50 hidden cases are planned.

c. Functional test missions. The functional test missions are full-mission situations used in manned testing of the PA system to validate its required performance. At least 20 functional test missions will be used.

d. System Evaluation missions. These full-mission situations will be selected from the functional missions, allowing for minor modifications in content.

Two other types of test cases, the adversarial cases and the real-time performance test cases will be related to the mission environment, but are not necessarily representative of expected real world mission conditions.

PA Behavioral Requirements Synthesis

Because requirements are fluid under the rapid prototyping development methodology being pursued, close coordination is needed to ensure that the scope and content of testing remains aligned with the requirements. In addition the test process clarifies the engineering interpretation of operational behavioral requirements.

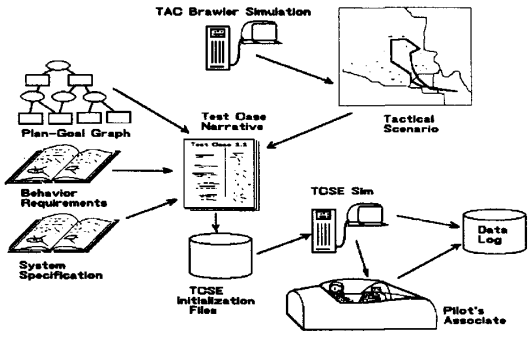

Figure 3. Test case development process

The requirements synthesis process includes:

a. requirements definition. In this stage, system and subsystem behavioral requirements are formed that define the functional relationships between the PA and the pilot, and between the subsystems of the PA. These requirements are stated in operational terms, not yet constrained by design approaches.

b. requirements allocation. Following definition, requirements are allocated to prototyping cycles for design and development. The allocation must consider the availability of input data to enable the functional requirement to be performed and the role of each requirement in shaping the integrated behavior of the PA. The test cases developed to support engineering development test, integration, and prototype testing serve as specific instances of the requirements, thus refining the definition of the requirement as the design evolves.

c. requirements revisions. As the design progresses, requirements and their interpretation as test cases may need revision, both in terms of their content and their allocation to prototype cycles. Furthermore, the requirements may need pruned in order to meet real-time processing goals.

The requirements synthesis process produces changes that must propagate to the formal evaluations at the conclusion of the development phase. The impact of this is that the content of the final missions and test cases to be used in the formal evaluation are updated and remain incomplete until the test procedures are published at the end of the prototyping process. At any point in time during development, a candidate set of missions is informally maintained that represents the current estimate of PA functionality planned for the conclusion of the development phase. These candidate missions are maintained as a set of test cases for each prototyping cycle that are directly linked to the allocated requirements for each prototyping cycle.

Test planning coordination

To obtain maximum efficiency of the testing efforts in the PA program, many elements of the engineering development test will be shared with the formal evaluations at the conclusion of the development process. The role of the library of test cases in shaping the evaluation missions has already been discussed. Other major areas of planning coordination are:

a. Functional testing. The two functional test series provide trial runs for the procedures to be used in the MSE. This permits validation of the training of subjects, the content of the missions and the data collection and analysis methods to be used in the MSE. The statistical design of the functional tests will permit the results from the functional tests to be analyzed both separately and in concert with the MSE results.

b. Digital simulation assessment. The on-going digital simulation assessment provides not only the possible contents of missions, but also feeds directly into the knowledge engineering process. This assures that the missions used for test purposes are covered appropriately in the knowledge bases of the PA System. In addition, the same measures of merit and analysis methods used in functional testing and the System Evaluation are planned for use in the digital simulation assessment to provide convergent validity for the results observed in manned simulation.

Test environment coordination

The test environment for the integrated test framework involves several important areas for coordination across test and development. The development of the PA target hardware environment is a parallel activity with the development of the PA system software. The hardware environment includes the processors and memory architecture, the operating system, and the compiling/linking/loading system. The overall features of this activity are:

a. progressive build up of the hardware environment. The initial architecture started with multiple processor cards using the 68030 processor and migrated to processor cards using the R3000 RISC processor. Both of these configurations were integrated in a VME bus chassis and connected to the rest of the development environment using ethernet interfaces. The 68030 processor configuration was used through prototype K.

b. development of an enhanced version of the VxWorks operating system for use on the target hardware. The extensions to the operating system provided specific functions for collection of real-time performance data and PA behavioral data from any level within the PA software.

c. development of a hardware event signal manager to replace software event signalling in VxWorks. This is a single VME bus card, known as the GEF board.

Because of the phased introduction of hardware and the need to validate the development portion of the environment, these activities required coordination with the integrated test framework. The critical constraint from the hardware is that no actual real-time testing could be performed until following the integration of prototype L, the first to be executed in the R3000 configuration.

The aircraft and external environment for hosting and testing the PA system is the LASC Tactical Combat Simulation Environment (TCSE). This environment is being developed in parallel with the PA system as a multi-purpose tactical simulation system that can host a variety of aircraft and threat models at several levels of fidelity.

For the functional tests and the System Evaluation, the TCSE is used to represent a baseline aircraft without the PA System and an identical aircraft equipped with the PA System. The aircraft modeled is a low-observable advanced fighter aircraft with performance and avionics similar to that expected for the planned LASC F-22 program. Issues that must be addressed in testing are:

a. functional integration with the PA system. It is extremely important to identify those aspects of aircraft or avionics performance that may be affected by PA system integration. In order to conduct a valid test, it is necessary that no differences in performance or capability between the baseline aircraft configuration and the PA configuration exist except for those that are a natural and intended part of the PA system itself.

b. data collection. Approximately 50 percent of all data to be collected for performance analysis of the PA is ground truth data from the TCSE. The capabilities of the TCSE to capture and record data for test purposes have been coordinated with the planned data collection and measures of merit development and validation process.

c. cockpit configuration. The cockpit environment for the phase 2 development program closely resembles the space constraints and geometry of an actual fighter cockpit. This cockpit is integrated with the TCSE simulation environment to provide the manned interface to the simulated aircraft systems, including control inputs and data display. A major effort to limit the cockpit to operationally viable displays has been made. This properly shifts the burden of evaluating PA subsystem behavior away from the cockpit displays and instead to the planned data collection and analysis process described above.

A final concern for consistency between the baseline and PA configured aircraft and cockpits is the limited resources available for training the pilots to be used in the manned evaluation. The statistical design of the MSE requires that each subject be trained in both configurations. Maintaining reasonable similarity in the displays and mechanization reduces the potential effects of incomplete training of the pilots across the two configurations.

Digital simulation environment

Unmanned digital simulation of the PA aircraft and baseline aircraft in a variety of mission situations is an important supporting activity to the PA test and evaluation. Three issues of concern can be effectively addressed by use of digital simulation.

First, there is no set of well-established measures of merit for evaluating the benefits of intelligent systems in air combat. While hypotheses abound as to the probable effects of intelligent systems, the appropriate operationalization of these hypotheses in terms of specific measures of merit remains problematic. Exploration of sensitive and stable measures of merit using manned simulation would be prohibitively expensive. Furthermore, an established body of measures of merit that has been validated for decision-aiding systems like the PA System is not directly available from previous studies.

Secondly, despite the structuring of air combat into mission types, and phases within mission, and so forth, an immense diversity of situations can potentially occur. The specific situations in which intelligent systems have potentially high payoff are not always intuitively obvious. Without some clear guidance on the types of situations in which intelligent systems may help, a manned test series may fail to investigate the most appropriate cases.

Finally, even the most ambitious manned tests can only evaluate a very small set of the potential situations that make up air combat. A comprehensive evaluation of the potential of the PA system needs to look beyond the specific mission events in the manned test series to include a broader sampling of potential situations.

To support the digital simulation process, a specially adapted set of programs has been assembled, which includes proprietary LASC models and a modified version of TAC Brawler prepared by DSC, the original creators of the program. Four important areas of coordination are necessary to make effective use of the digital simulation process as a part of the integrated test approach:

a. Baseline aircraft representation as a reflection of unaided manned simulation. TAC Brawler represents aircraft capabilities and tactics in a specialized format that must be coordinated with the actual baseline aircraft and tactics used in manned simulation.

b. PA capabilities representation as reflections of manned simulation and PA design. The original versions of TAC Brawler do not consider the direct effects of decision aiding in its representation of the combat pilot decision cycle. To enable a better depiction of the common cognitive problems of pilots while engaged in combat, and to represent the possible effects of the PA system on those cognitive problems, a modification to TAC Brawler has been initiated. This modification, along with changes in the tactics descriptions used in TAC Brawler must be coordinated with the functional design of the PA system to ensure that the digital simulation of

the PA system is a reasonable reflection of the manned PA simulation. This will include enhanced representation of PA subsystem design and knowledge engineering as cognitive effects and tactics in TAC Brawler.

c. Mission environment in terms of threat positions and behaviors, mission goals, and rules of engagement. In addition to coordination of PA rules of engagement, the digital simulation requires coordination with the representation of the mission environment in TCSE.

d. Data collection and analyses, measures of merit. In order to attain the most consistent comparisons between what was observed in manned simulation and the digital simulations, a common set of data collection and measures of merit will be used for both studies. This requires coordination in defining the exact collection conditions and computational algorithms for use in the digital simulation assessment.

TEST APPROACH

The approach to testing the PA integrates the five main aspects of testing discussed earlier.

Unmanned digital simulation testing

The digital simulation environment consists of two major simulation programs, supported by data collection, analysis, and visualization programs. The first stage is to evaluate platform detections for the PA-equipped aircraft from both surveillance and fire-control radar systems in the enemy Electronic Order of Battle. This is performed by a LASC proprietary classified program that fully supports current sensors, weapons and low-observability technology. The output provides the locations and geometries of both surface-to-air missile engagements and vectored air intercepts from the enemy Air Order of Battle.

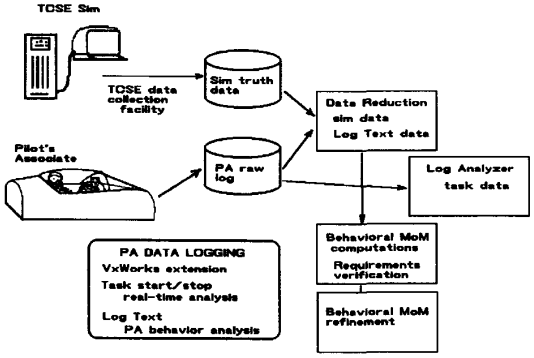

Figure 4. PA data collection process

Mission outcome files are progressively built up from the detections and TAC Brawler runs. The engagement geometries are used in a specially modified version of TAC Brawler to evaluate engagement outcomes. By selecting a set of representative engagement outcomes from the TAC Brawler data and using them as starting points for the route continuation, a full mission scenario can be constructed. This iterative process permits analysis of the cumulative effects of both detections and engagements over the course of the full mission. The Monte Carlo nature of TAC Brawler enables a large number of engagements to be played out and statistics collected that provide estimates of the long-term consequences of the engagement situation.

The output files from TAC Brawler will be analyzed for interesting changes in success as a result of engagement situations and PA aiding. The format of the output files is shared with the manned simulations to permit the same data analysis to be conducted on both TAC Brawler output and manned simulation output.

The digital simulation efforts are organized as a set of three cycles of parametric exploration, each cycle keyed to one of the three prototyping cycles of Phase 2. Each cycle is approximately 4 months in duration. Variations developed in one cycle may be re-used in succeeding cycles if appropriate.

A final Digital Simulation Assessment Study will be performed in conjunction with the MSE. This study will use the best representation of PA functionality to explore the mission outcomes for a larger number and variety of missions than can be tested directly in manned simulations.

Prototype cycle test

Development testing is conducted at the module development site in a stand-alone fashion to evaluate the readiness of the module to undergo integration testing. Only partial testing of a subsystem during development test is practical due to the interdependencies among subsystems and the logistical burden of attempting to maintain accurate emulations of the other subsystems under development. Successful completion of the planned development testing is a prerequisite to beginning integration of the subsystems for the prototyping cycle.

The second part of each prototyping cycle involves the integration and testing of the entire PA subsystem over a broader set of test cases than used in development test. It is expected that the PA will pass all public test cases and most of the hidden test cases for each prototyping cycle during integration testing. Successful performance of adversarial test cases is not required. Both operational measures and real-time performance measures are collected and analyzed during integration testing.

Functional testing

Functional Tests L and M comprise two functional test periods. The functional test periods are manned simulation exercises that are conducted as formal designed experiments using contractor team pilots. The content of the functional testing is determined by the public and hidden test cases performed to date. A detailed set of informal test procedures is prepared prior to the start of the functional tests. Each Functional test is scheduled immediately after the completion of the respective prototype integration testing.

The basic design of the functional tests is a one-way factorial experiment, with two aircraft conditions (PA, BASELINE). A full-mission context is desired to explore and refine measures of merit issues as well as to evaluate cumulative mission effects with man in the loop. Each subject will perform several fighter missions under both aircraft conditions, allowing a "within-subjects" design. The mission instances will be balanced across aircraft condition and order. Subjects for the experiment are drawn from the LASC team, totalling 7 experienced former fighter pilots. Five subjects (4 plus 1 spare) are used for each experiment, selected to as similar in experience and knowledge of the PA as feasible.

Training is expected to be reduced as a result of using team pilots already familiar with the PA and the mission context. All subjects are provided with 4 hours of briefing on

the PA and 4 hours briefing on the baseline aircraft, written material describing each system, a simulated 1 hour air intelligence briefing, and two practice flight sessions in the simulator (one PA, one BASELINE). A written quiz is used to provide a pre-test measure of knowledge about both the baseline aircraft and the PA.

System test and evaluation

The final phase of testing in the PA integrated test framework is the formal system test, supported by USAF pilots. This test phase consists of two major parts, the RTP test and MSE using a full-mission simulator.

The RTP tests are conducted to measure the response times of the PA to mission situations and to evaluate system architecture considerations, such as operating system overhead, scheduling discipline, multi-processing scale-up, bus bandwidth, and memory contention. The detailed data to be collected and evaluated is defined in the formal test procedures for real-time testing.

A set of stressing test cases assembled from past prototype test cases, including adversarial cases, will be selected as the real-time performance test battery. These cases are supported by data files to supply pilot and enemy forces behavior in a fully repeatable manner. The behavior produced by the PA and the simulator for these cases must be judged to be fully acceptable from a tactical perspective prior to their acceptance as part of the real-time performance test battery. The duration of a typical test case will be from 2 to 4 minutes in length. A minimum of 15 cases will be selected for the test battery.

The real-time tests consist of running the stressing test cases identified during prototyping and collecting data using the internal data logging system added to the operating system during Phase 2 to support real-time analysis. No further optimization of the PA task scheduling or processor assignments will be permitted once the real-time performance testing has begun. Data collection using the logging tools will allow investigation of task delays and processor loading. Analysis consists of the evaluation of the processing time of critical threads through the PA System. Although no direct measurements of degraded operation is planned, the data is expected to support discussions of the consequences of processor degradation and scheduling efficiency.

Following the RTP testing, the MSE will be a direct comparison between the baseline aircraft configuration and a PA-configured aircraft with identical systems using current USAF pilots as subjects. This manned evaluation will take place at the LASC TCSE simulation in Marietta, GA.

The MSE is a small-scale experiment requiring careful design to reduce sources of extraneous variation. Although the risk of failing to detect true differences is high (that is, the test has low power) for the objective measures, it will provide a satisfactory setting for collection of well-controlled subjective responses and may provide concrete objective indications of advantages for the PA system if the magnitude of the improvements is sufficiently large.

The basic design of the Manned System Evaluation is single factor experiment, with two aircraft conditions (PA, BASELINE). A full-mission context is desired to investigate PA subsystem interactions as well as to evaluate cumulative mission effects with man in the loop. Each subject will perform both aircraft conditions, allowing a "within-subjects" design.

The use of low-time, inexperienced pilots is highly desirable from a test perspective. The opportunity for mission experience with the PA by more highly-experienced aircrews that are more representative of the current skill levels of TAC will be conducted as a part of the demonstration aspects of the PA project rather than as a part of the testing. Training is expected to be a significant part of the subjects' time at Marietta. All subjects will be provided with 8 hours of briefing on the PA and 8 hours briefing on the baseline aircraft, written material describing each system, a simulated 1 hour air intelligence briefing, and four practice flight sessions in the simulator (two PA, two BASELINE). A written quiz will be used to provide a pre-test measure of knowledge about both the baseline aircraft and the PA. The training sessions will cover three days for each subject.

The final set of approximately 15 middle-level measures of merit and the associated high-level measures of merit will be identified for analysis as a part of the detailed test procedures prior to the start of testing. This set of measures is a validated set used in Functional Test M. It is expected that these middle-level measures will require collection of between 100 and 200 primitive measurement types.

In addition to the objective data collection, a battery of subjective questions and ratings will be administered after each simulator session as a structured interview. Approximately 50 ratings or questions are anticipated for the MSE, each question validated in prior functional test periods. The subjective questions will be aimed at corroborating and interpreting the performance oriented measures.

An advantage to the data collection is the ability to analyze the data from Functional Test M and the MSE both as two separate experiments and as one single experiment with more replications. Comparison of the Functional Test M results with the MSE results may also indicate how the results might change with more training for the subjects, since the contractor pilots used in Functional Test M are in general more experienced and more knowledgeable of the PA.

SUMMARY

The successful evaluation of intelligent decision aids will require careful, integrated testing throughout the design and development phases. The testing approach taken on the Pilot's Associate program is designed to provide as much assurance as feasible with efficient use of program resources. Beyond the development phases, it is likely that intelligent systems will require testing over their entire life cycles. As a result, the establishment of a principled approach to testing is of great importance in the introduction of this technology into the operational forces.

ACKNOWLEDGEMENTS

This work was conducted under the support of the Lockheed\USAF\DARPA Pilot's Associate program, contract F33615-85-C-3804.