United States General Accounting Office

GAO

Briefing Report to Congressional Requesters

July 1988

WEAPONS TESTING

Quality of DOD Operational Testing and Reporting

GAO/PEMD-88-32BR

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

B-222886

July 26, 1988

The Honorable Charles E. Bennett
Chairman, Subcommittee on Seapower and
Strategic and Critical Materials
Committee on Armed Services
House of Representatives

The Honorable Denny Smith
The Honorable Thomas J. Ridge
The Honorable Barbara Boxer
House of Representatives

As you know, in 1983 the Congress established the office of the Director of Operational Test and Evaluation (DOT&E) to effect several reforms concerning operational testing. Prominent among the reform objectives were independent oversight and coordination of the military services' planning and execution of operational tests, independent evaluation of the results of operational tests, and objective reporting of those results to decisionmakers in the Department of Defense (DOD) and the Congress. Fundamental concerns were that weapons were not being tested thoroughly or realistically and that complete and accurate information was not being disseminated.

This report is in response to your June 5, 1987, letter requesting that we address two evaluation questions: 1) What is the methodological adequacy of operational test and evaluation under DOT&E oversight, and 2) what is the quality of DOT&E dissemination of information to the Congress? In answering question 1, we also made an effort to determine the impact of DOT&E on the operational test and evaluation process.

To address the evaluation questions, we reviewed relevant documentation on the operational test and evaluation of six major, conventional weapon systems that had reached the full production milestone by the end of fiscal year 1987, as well as congressional testimony, DOD regulations, and outside literature on the conduct and reporting of test and evaluation in general. (The six cases were systematically selected from a universe of ten eligible cases; the specific selection criteria are described in the report.) We also interviewed DOD officials and outside experts in operational testing. We developed a standardized assessment framework to evaluate each case, after which we synthesized the information across cases to yield overall findings and conclusions. The results are

not generalizable to the test and evaluation of strategic systems or to systems that have not yet reached the full production milestone.

The DOT&E statute established the director as the principal operational test and evaluation official within the senior management of DOD, and specified that he 1) prescribe policies and procedures for the conduct o operational test and evaluation in DOD, 2) provide guidance to and consult with the secretary of defense and the service secretaries on operational test and evaluation, and 3) monitor and review all operational testing in DOD. The statute also imposed two principal congressional reporting responsibilities: 1) a report when a major defense acquisition program is to proceed beyond low-rate initial production (known as the B-LRIP report) stating whether operational test and evaluation was adequate and whether test results confirm the item or components to be effective and suitable for combat, and 2) an annual report summarizing the operational test and evaluation activities of DOD for the preceding fiscal year with comments and recommendations that the director considers appropriate. In addition, the statute requires the director to respond to requests from the Congress for information regarding operational testing. In 1984, DOD established DOT&E within the Office of the Secretary of Defense. DOT&E was without a permanent director until April 1985 and was generally understaffed during its first years of operation. However, the staffing situation improved considerably during fiscal year 1987-88; DOT&E currently has over 40 staff members.

With regard to the methodological adequacy of operational test and evaluation under DOT&E oversight, we found significant problems and limitations in the planning, execution, realism, analysis, and reporting by the service test agencies for the six systems we reviewed. Some of these problems and limitations were unavoidable due to time, resource, or safety constraints, although numerous others were not. Our conclusion is that for major, conventional systems that reached the full production milestone by the end of fiscal year 1987, the operational test and evaluation being conducted under DOT&E oversight was not methodologically adequate to assess the operational effectiveness and suitability of weapon systems. Instead, operational test and evaluation findings have tended to show more favorable assessments than are likely to be found when the weapons are employed in combat. The danger here is that this can lead to the funding of weapon systems whose operational effectiveness and suitability have not been demonstrated. In sum, operational test and evaluation under DOT&E oversight has fallen short of the objectives sought by the Congress when it established the office.

Our ability to evaluate the impact of DOT&E on the test and evaluation process was limited because much of the communication between DOT&E and other DOD components is informal and undocumented. This made it difficult to accurately determine how effectively DOT&E carries out some of its functions. As a consequence, our assessment of DOT&E impact on the test and evaluation process is inconclusive.

The interviews we conducted and the documentation we obtained show that DOT&E has had at least some impact on that process as well as on the production decisions that flow from it. However, with regard to major production decisions, we found no evidence of DOT&E's impact in three of those decisions (other than to support the decision), no opportunity for impact in one, and in the other two, impact that was either indistinguishable from that of other DOD units or that was more apparent than real.

With regard to the quality of DOT&E dissemination of information to the Congress, each of the official DOT&E reports to the Congress that we reviewed contained incomplete or inaccurate statements, and most contained both. In addition, the majority of DOT&E's favorable overall assessments of testing adequacy and of system effectiveness and suitability were not supported by the evidence. The omissions, inaccuracies, and overall assessments consistently resulted in a more favorable presentation to the Congress of test adequacy and system performance than was warranted by the facts. Our conclusion, therefore, is that for major, conventional systems that reached the full production milestone by the end of fiscal year 1987, DOT&E's dissemination of information to the Congress has not provided the complete and accurate picture of weapon performance that the Congress needs to make weapon funding decisions. As such, it has fallen short of the objectives sought by the Congress when it established DOT&E.

As noted earlier, some problems and limitations in operational test and evaluation cannot be avoided. However, we know of no reason why those problems and limitations should not be reported completely and accurately.

We believe that the law that established DOT&E and DOD's own directives together provide adequate organizational structure and guidance for the conduct and reporting of operational test and evaluation. Therefore, we offer no recommendations for changing the law or the associated directives. However, we believe that there is a need for greater management emphasis on improving the implementation of those directives, in order

to more effectively realize the intent of the law. For example, we think that it is important for DOD to improve the quality of operational test and evaluation performed under DOT&E oversight in order to remove methodological biases and to correct the tendency we found toward overly favorable assessment of weapon performance. DOD must find ways to address the many significant problems and limitations in the planning, execution, realism, analysis, and service test agency reportin of operational test and evaluation. Similarly, DOT&E must improve the quality of the information it disseminates to the Congress and avoid pr viding a more favorable presentation of test adequacy and system performance than is warranted by the facts. Specifically, actions should b taken to improve the completeness and accuracy of DOT&E reports to th Congress.

This report is divided into a summary (in four sections) followed by a set of appendixes. The appendixes provide detailed support for the fin ings and conclusions in the summary. In this version of the report, classified passages have been deleted and replaced with the phrase "[material deleted];" we are also publishing a classified version (GAO/C-PEMD-88-2BR).

As requested by your representatives, we did not seek formal agency comments on this report. We did receive informal comments from DOT&E officials on an earlier draft, and made changes where appropriate. As we arranged with your office, copies of the report will be sent to the Department of Defense. At that time, we will make copies available to interested organizations, as appropriate, and to others upon request. If you have any questions regarding the contents of this report, please cal me (275-1854) or Mr. Kwai Chan, Group Director (275-6161).

Eleanor Chelimsky
Director

Contents

Abbreviations

AHIP	Army Helicopter Improvement Program
AAST	Army Aerial Scout Test
AAW	Antiair warfare
ADATS	Army Development and Acquisition Threat Simulators
AFOTEC	U.S. Air Force Operational Test and Evaluation Center
AV	Air vehicle
BAI	Battlefield air interdiction
B-LRIP	Beyond low rate initial production
CAS	Close air support
DAG	Data Authentication Group
DCP	Decision Coordinating Paper
DOD	Department of Defense
DOT&E	Director, Operational Test and Evaluation
DSARC	Defense System Acquisition Review Council
DSMAC	Digital scene-matching area correlator
ECM	Electronic countermeasures
ELOSS	Electronic line-of-sight system
EOCM	Electro-optical countermeasure
FAA	Federal Aviation Administration
FAAO	Field artillery aerial observer
FLIR	Forward-looking infrared receiver
FOT&E	Follow-on operational test and evaluation
GCS	Ground control station
IER	Independent Evaluation Report
IG	Inspector General
IOT&E	Initial operational test and evaluation
IR	Infrared
JLF	Joint Live Fire
JT&E	Joint Test and Evaluation
LANTIRN	Low-Altitude Navigation and Targeting Infrared System for Night

LGB	Laser guided bomb
LRIP	Low rate initial production
NATO	North Atlantic Treaty Organization
OPEVAL	Operational evaluation
OPTEVFOR	U.S. Navy Operational Test and Evaluation Force
OSD	Office of the Secretary of Defense
OT	Operational test
OT/DT	Operational and developmental test
OT&E	Operational test and evaluation
OTEA	U.S. Army Operational Test and Evaluation Agency
PA&E	Program Analysis and Evaluation
RAM	Reliability, availability, and maintainability
RPV	Remotely Piloted Vehicle
SDDM	Secretary of Defense Decision Memorandum
TEMP	Test and Evaluation Master Plan
TERCOM	Terrain contour matching
TLAM/C	Conventional Tomahawk Land Attack Missile
USDRE	Undersecretary of Defense for Research and Engineering
WFOV	Wide field-of-view

Objectives, Scope, and Methodology

In 1983, the Congress established the office of the Director of Operational Test and Evaluation (DOT&E) to effect several reforms concernin operational testing.[1] Prominent among the reform objectives were inde pendent oversight and coordination of the military services' planning and execution of operational tests, independent evaluation of the resu of operational tests, and objective reporting of those results to decisionmakers in the Department of Defense (DOD) and the Congress. The DOT&E statute established the director as the principal operational test and evaluation (OT&E) official within the senior management of DOD, ar specified that he 1) prescribe policies and procedures for the conduct c OT&E in DOD, 2) provide guidance to and consult with the secretary of defense and the service secretaries on OT&E, and 3) monitor and review all OT&E in DOD. It also imposed two principal congressional reporting responsibilities: 1) a report when a major defense acquisition program to proceed beyond-low rate initial production (known as the B-LRIP report) stating whether OT&E was adequate and whether OT&E results confirm the item or components to be effective and suitable for combat and 2) an annual report summarizing the OT&E activities of DOD for the preceding fiscal year with comments and recommendations that the director considers appropriate. In addition, the statute requires the director to respond to requests from the Congress for information regarding OT&E. A fundamental congressional concern was that weapon were not being tested thoroughly or realistically and that complete and accurate information was not being disseminated.

The Chairman, Subcommittee on Seapower and Strategic and Critical Materials, House Armed Services Committee, and three other Members of Congress, asked us to address two evaluation questions: 1) What is the methodological adequacy of OT&E under DOT&E oversight, and 2) wh is the quality of DOT&E dissemination of information to the Congress? In answering question 1, we also made an effort to determine the impact c DOT&E on the OT&E process.

To address the questions, we reviewed relevant documentation on the OT&E of six weapon systems, as well as congressional testimony, DOD reg ulations, and outside literature on the conduct and reporting of testing and evaluation in general. We also interviewed DOD officials and outside experts in OT&E. Certain documents were not obtained due to lack of time; however, we believe that the effect on our overall findings and conclusions was negligible.

[1] In practice, the acronym DOT&E is used to denote both the director and the office under his direction. To avoid confusion, we refer to the director as the director and to the office as DOT&E.

All field work was conducted between September 1987 and March 1988 in accordance with generally accepted government auditing standards.

Case Selection

To select the six weapon systems, we developed case-selection criteria. These criteria and their rationales are shown in table 1.1. The use of these criteria yielded 10 eligible candidates. (See table 1.2.) The final six were selected on the combined basis of recency and number of common missions.[2] The latter criterion was important to facilitate greater comparability across systems. Final selections were: for the Army, Army Helicopter Improvement Program (AHIP) and Aquila Remotely Piloted Vehicle (RPV); for the Navy, Conventional Tomahawk Land Attack Missile (TLAM/C) and DDG-51 Destroyer (Aegis Anti-Air Warfare system only); and for the Air Force, Imaging Infrared (IR) Maverick and Low-Altitude Navigation and Targeting Infrared System for Night (LANTIRN). Common missions across the six selected systems are navigation (four systems), target acquisition (six systems), target designation (four systems), and target engagement (four systems).

Table 1.1: GAO's Initial Case Selection Criteria and Rationales

Criterion	Description	Rationale
1.	B-LRIP report filed or scheduled for FY 1987	The Beyond-Low Rate Initial Production (B-LRIP) reports is DOT&E's system specific reporting requirement under 10 USC 138, and therefore is necessary to fully address the second evaluation question. Also, for cases which met this criteria, DOD testing to justify production will be complete.
2.	Entry into B-LRIP after director swear-in	There was no permanent DOT&E director until April 1985. Cases that entered B-LRIP after that date were included.
3.	Must be a major system	The Congress is primarily interested in major systems, those for which a Selected Acquisition Report is required (that is, those over $200 million in research and development or $1 billion in production).
4.	Must be a non-strategic system	Congressional requesters expressed primary interest in conventional tactical systems (as opposed to strategic nuclear systems).
5.	Must have tri-service representation	DOT&E oversees testing across all of DOD. Sampling cases from each service allows DOD-wide conclusions.
6.	Study to include six systems maximum	Time and staff available limit the number of OT&Es and reports to the Congress that GAO can adequately evaluate.

[2]Only two Air Force candidates were eligible, so no final selection of Air Force systems was necessary.

Table 1.2: Eligible Weapon Systems and Final Selection

Service	Weapon system	
	Eligible	Selected
Army	Sgt. York (DIVAD)	No
	AHIP	Yes
	RPV Aquila	Yes
	M2 Bradley	No
Navy	AV-8B	No
	Tomahawk TLAM/C	Yes
	DDG-51	Yes
	LCAC	No
Air Force	IR Maverick	Yes
	LANTIRN	Yes

As is evident from the selection criteria, this review focuses on OT&E of major, conventional systems that reached the B-LRIP milestone by the end of fiscal year 1987. Therefore, the results are not generalizable to the entire universe of OT&E being conducted under DOT&E oversight; specifically, they do not generalize to operational testing of strategic systems or systems that have not yet reached the B-LRIP milestone. The latter limitation may mean that any effects of recent DOT&E initiatives on OT&E planning, such as the issuance of new guidelines for preparing Test and Evaluation Master Plans (TEMPS), are underrepresented. In addition, the results do not permit a direct assessment of change attributable to the legislative establishment of DOT&E. Such a study would require 1) a comparison base of OT&Es conducted and reported prior to the establishment of DOT&E, 2) sufficient numbers of cases to support a statistically valid comparison, and 3) elimination of rival explanations for observed changes (for example, increased congressional attention to OT&E). Such an assessment would have required time, resources, and data that were beyond the scope of the present study.

Assessment Framework

During a 1983 evaluation of the Joint Test and Evaluation (JT&E) program, we developed a multiple case study method to assess the quality of the tests. We later refined the method in our 1986 evaluation of the Joint Live Fire Test (JLF) program and used it again for the present evaluation of DOT&E. First, a standardized assessment framework was developed to evaluate the cases (see appendix I). Next, information on each case was analyzed and coded in terms of the assessment framework. To ensure appropriate and consistent interpretation of the framework, all coding was continually monitored and validated across cases. Lastly, the

information from each case was synthesized across cases to yield overall findings and conclusions.

Sources used to develop the framework included 1) DOD regulations on the conduct and reporting of OT&E (DOD Directives 5000.3 and 5000.3-M-1), 2) statements made by the DOT&E director during congressional testimony, 3) the legislation that established DOT&E, 4) prior studies on OT&E, and 5) the JT&E and JLF assessment frameworks. The assessment framework covered seven categories: planning, execution, realism, analysis, reporting by the service operational test agencies, DOT&E impact, and DOT&E reporting. Each category contained a set of assessment questions or items. We stress that the items in our framework and their interpretation were based on established DOD guidance. For example, the importance of a realistic portrayal of threat forces is noted in Directive 5000.3, in DOT&E's own statements, and in prior studies on OT&E. And, in each case we compared the threat as portrayed in the OT&E to the threat as portrayed in DOD-approved threat assessments.

The intent of the assessment framework was to ensure the comprehensiveness and comparability of the rating process across systems, and to support statements on the prevalence of various types of problems and limitations. It was not intended to support direct comparisons of OT&E technical adequacy across systems based simply on the number of "boxes checked." Comparisons of "boxes checked" may be unfair and misleading because they cannot account for the substantive evaluation issues that must be understood via the completeness of documentation, reporting, and interviews. Instead, box-checking comparisons favor those cases where documentation is missing or incomplete, reporting is not thorough, and officials are not informative.

Prior GAO Reports

We have issued numerous reports on or involving OT&E. The following are the most directly relevant to the present effort.

In March 1987, we reported that DOT&E had made contributions to OT&E activities, especially in test planning, but that three areas needed attention: 1) DOT&E appeared to be making only a limited number of actual on-site observations of operational tests; 2) DOT&E's analysis of operational tests was primarily based on service test reports, with little assessment of actual test data; and 3) DOT&E had not provided policy and procedural

guidance or maintained reliable records on some of its principal activities.[3] (For example, no uniform policies or procedures existed to provide guidance to action officers on how to perform their functions or document their efforts.) DOT&E officials acknowledged that these problems needed additional emphasis and attributed them partly to a lack of staff. Our March 1987 report focused on the processes by which DOT&E performs its function; the present report focuses instead on the outputs of DOT&E—specifically, on the adequacy of testing and reporting.

In September 1987, we reported on both test quality and system performance issues raised by the IR Maverick Follow-on Operational Test and Evaluation (FOT&E).[4] (The specific findings are classified.)

In October 1987, we reported that the Aquila operational test identified major problems that should be corrected prior to a production decision, including frequent inability to launch the drone, difficulty in detecting targets, and survivability concerns.[5] We also noted that certain deficiencies in the OT&E made it difficult to project the Aquila's eventual performance when fielded.

The present report draws on these prior reports where appropriate.

[3]Testing Oversight: Operational Test and Evaluation Oversight: Improving but More is Needed. GAO/NSIAD-87-108BR (Washington, D.C.: March 1987).

[4]Missile Procurement: Infrared Maverick Testing and Performance, GAO/C-NSIAD-87-21 (Washington, D.C.: September 1987).

[5]Aquila Remotely Piloted Vehicle: Its Potential Battlefield Contribution Still in Doubt, GAO/NSIAD-88-19 (Washington, D.C.: October 1987).

Methodological Adequacy of OT&E Under DOT&E Oversight

Congressional concern that the operational testing of weapons under realistic, combat-like conditions was inadequate was a principal reason for the establishment of DOT&E. In his confirmation hearing, the director expressed his intention to ensure that the weapon systems being procured by DOD are thoroughly tested and are operationally effective and suitable for combat. Members of the Senate reiterated their concern about the adequacy of operational testing during the confirmation process.

Our findings on methodological adequacy are summarized in the tables in this section. We make only one distinction in these tables, and that is whether significant problems or limitations were found in the OT&E. We define a significant problem or limitation as one that potentially affects conclusions regarding the operational effectiveness or suitability of the weapon system.[6] We do not report unimportant problems and limitations, those that in our judgment do not meet this criterion. Moreover, it is important to recognize that many problems and limitations in OT&E are unavoidable. Due to time, resource, and safety constraints, not everything can be tested or tested well. Further, it is not our intention to hold either DOT&E or the service test agencies responsible for events they cannot control.

Support for the findings in the tables and accompanying text can be found in this report's classified appendixes II through IV, sections 1 through 5. (To keep the text unclassified, system identifiers are not included in this section of the report.)

Planning and Execution

Findings on planning and execution are summarized in table 2.1. In three of the four cases where we had Test and Evaluation Master Plans to evaluate, the TEMP included a complete statement of the system's requirements. In each of these cases, however, the test plan did not then address all system requirements and critical operational issues identified in the TEMP. Requirements or critical issues that fell out included testing in all geographic, environmental, or mission conditions, testing at the edges of the performance envelope, and testing the complete, operational system.

[6]DOD Directive 5000.3-M-1 defines operational effectiveness as "the overall degree of mission accomplishment of a system when used by representative personnel in the environment planned or expected for operational employment of the system considering organization, doctrine, tactics, survivability, vulnerability, and threat." It defines operational suitability as "the degree to which a system can be satisfactorily placed in field use, with consideration given to availability, compatibility, transportability, interoperability, reliability, wartime usage rates, maintainability, safety, human factors, manpower supportability, logistic supportability, documentation, and training requirements."

cases. In three of these cases, the result was a reduction in realism that favored the system being tested.

Realism

Findings on realism are summarized in table 2.2. Directive 5000.3 states that typical users should operate and maintain the system, and prior OT&E studies as well as the law that established DOT&E also stress the importance of typical users. Directive 5000.3 provided no definition of the word "typical," but the DOT&E director shed some light on this issue at the fiscal year 1987 defense authorization hearings when he testified that adequacy of testing includes ensuring that the user participant represents what the user will be like when the system is fielded. Adopting this as a definition of typical, we found that in four of six cases the system operators were not typical, and that in four of five cases where the question was applicable, the support personnel were not all typical. The prevailing problem for operational users was that they were selected from an operator pool that was atypically high in skill or experience level—that is, so-called "golden crews." The prevailing problem for support personnel was some level of contractor involvement in the support of the system, principally in the maintenance function, although that contractor support would not be available in the field. Contractor involvement in OT&E is prohibited by the Fiscal Year 1987 National Defense Authorization Act (PL 99-661, 10 USC 2366) passed in November 1986. The law states that in the case of a major defense acquisition program, no person employed by the contractor for the system being tested may be involved in the conduct of the OT&E. In addition to being organizationally different, contractor maintenance personnel are usually better trained and more experienced on the system than military personnel would be. Consequently, their performance does not reflect what can realistically be expected when military personnel assume the maintenance burden.

Table 2.2: Significant Problems and Limitations in Test Realism

	Army system		Navy system		Air Force system	
Assessment questions	**AHIP**	**Aquila**	**TLAM/C**	**DDG-51**	**IR Maverick**	**LANTIRN**
Realism						
Operated by typical operational units?		X				X
Operated by typical operational personnel?	X	X	X		X	
Supported by typical support units?	X	X		X	b	X
Supported by typical support personnel?	X	X		X	b	X
Equipment put under realistic stress?	X	X	X	X	X	X
Personnel put under realistic stress?	X	X		X	X	X
Realistic combat tactics employed?	X	X	X	X	X	X
Physical environment approximates intended ranges?	X	X	X		X	X
Target systems approximate actual targets, realistically employed?	X			X	X	X
Threat systems approximate actual threat, realistically employed?	X	X	X	X	X	a
Tested system production representative and prepared for test in a realistic manner?				X		X

Note: empty cells signify "no significant problems or limitations found." X signifies "one or more significant problems or limitations found." a signifies "insufficient information to evaluate." b signifies "not applicable."

One case deserves special mention because it featured contractor involvement in operations as well as support. Two of the contractor's data collection technicians involved themselves in the conduct of the test on multiple occasions, despite warnings from service test officials. In at least one instance, the contractor technician entered the crew area, unauthorized and unsolicited, and advised the crew while a mission was under way. Evidence of these actions, along with evidence of similar contractor involvement in maintenance functions, led the DOD inspector general to conclude that 10 USC 2366 had been violated in the Aquila operational test.[7] In the fiscal year 1987 authorization hearings, when the DOT&E director was presented with the case of an earlier test in which the system contractor actively participated in the test operations, he testified that DOT&E would ensure that this would not happen again. However, it did happen again in this instance, despite the fact that DOT&E personnel and their consultants conducted on-site monitoring of the test.

[7]The issue had initially been raised by GAO officials observing the test, and reported in Aquila Remotely Piloted Vehicle: Its Potential Battlefield Contribution Still in Doubt, GAO/NSIAD-88-19 (Washington, D.C.:, October 1987).

Directive 5000.3 states that testing should be conducted under conditions simulating combat stress. The DOT&E director has also testified that adequacy of testing includes ensuring that tests are challenging. We recognize that there are safety and resource constraints that make some limitations inevitable and that most tests stressed equipment and personnel to at least some degree. Nevertheless, in all six cases, we found significant problems or limitations in the degree to which equipment was realistically stressed and, in five of six cases, in the degree to which personnel were realistically stressed. Specific details are classified but, in general, instances of insufficient stress on equipment included: 1) the absence or significant underrepresentation of countermeasures (communication jamming, radar jamming, electro-optical countermeasures); 2) the use of tactics that facilitate performance during the test but are incompatible with system survivability in a realistic threat environment as defined by DOD; 3) the use of targets that are hotter, slower, higher, more plentiful, less maneuverable, more likely to be stationary, or less likely to be camouflaged than DOD sources indicate would frequently be the case in combat, and 4) other instances where the outer edges of the specified performance envelope were not tested (although assets to test them existed). Instances of insufficient stress on personnel included crew familiarity with the test range or target area, assumptions that intelligence on enemy locations and other matters was readily available and accurate (in one case, needed meteorological data were obtained by a telephone call to the target area, an implausible method of data collection in wartime), various forms of cueing that reduced or eliminated the element of surprise, and failure to stress crew endurance commensurate with stated mission requirements. As a result, estimates of performance from the OT&Es tend to be biased upward, and performance under more realistic stress conditions remains unknown.

Directive 5000.3 states that operational testing shall be accomplished in an environment as operationally realistic as possible. Yet, in five of six cases we found significant problems or limitations concerning the extent to which the physical test environment approximated the intended range of environments as stated in TEMPs and other requirements documents. The most prevalent limitation was that systems intended for deployment in Europe were not tested in a European-like environment—that is, with terrain and weather typical of Europe. The result is that operational effectiveness in a North Atlantic Treaty Organization (NATO) environment is unknown. Other limitations included terrain that constrained operating space or maneuverability, lack of unintended countermeasures such as naturally occurring thermal clutter, and the inability (for safety reasons) to test in darkness or adverse weather.

Finally, Directive 5000.3 also states that operational testing should include threat-representative hostile forces. In 1986 testimony, the DOT&E director stated that adequacy of testing includes ensuring that the proper threat is being looked at, and he stressed that the proper threat is the one that will be faced when the weapon system is fielded, not a "vintage-type" threat that is easily overcome. In at least some of the cases we reviewed, testers went to considerable lengths to portray the threat realistically, including validation and monitoring by service threat agencies. Still, in the five cases for which we had sufficient information to compare the threat portrayed in the test to the threat portrayed in DOD threat documents, all five revealed significant problems or limitations. Again, the specific details are classified, but threat forces were in some cases less capable technologically than the Soviet forces that new U.S. systems would actually face, were numerically underrepresented, only partially portrayed (for example, ground threats present but air threats absent), otherwise not adequately depicted, or absent altogether for all or part of the test. As with stress and performance estimates, estimates of survivability from the OT&Es also tend to be biased upward, and survivability in a more realistic threat environment remains unknown.

Analysis and Service Test Agency Reporting

Our findings on analysis and service test agency reporting are summarized in table 2.3. Directive 5000.3 states that testing shall be planned and conducted to provide quantitative data and to minimize the need for subjective interpretation of system performance. We found reliance on qualitative measures to be a significant problem in only two of six cases; however, the reliability and validity of quantitative measures was a significant problem in four of six. The most prevalent problem was that system reliability data was itself not reliably or validly measured. For example, there were quality control problems during data collection and contractor participation in "scoring" the data points after collection (such as the downgrading of an "operational mission failure," which reduces system reliability, to an "essential maintenance action," which does not). There were reliability and validity problems with operational data as well. Some of the measurement problems clearly biased the results in favor of the system being tested; others simply made them uninterpretable. These biases were over and above any biases due to a lack of realism as discussed above (such as performance of maintenance by contractors).

Table 2.3: Significant Problems and Limitations in Test Analysis and Reporting

	Army system		Navy system		Air Force system	
Assessment questions	**AHIP**	**Aquila**	**TLAM/C**	**DDG-51**	**IR Maverick**	**LANTIRN**
Analysis						
Measures quantitative and non-subjective?		X				X
Quantitative measures reliable and valid?	X	X	X			X
Analytic assumptions explicit and approriate?	X	X		X	X	X
Sample size adequate to support statistically valid results?	X		X	X	X	X
Comparisons with other systems valid?	X	a	a	a	X	X
Service reporting						
Findings, conclusions, recommendatiaons consistent with the evidence and appropriately qualified?	X	X	X	X	X	X
Reporting clear and comprehensive?			X	X	X	X

Note: empty cells signify "no significant problems or limitations found." X signifies "one or more significant problems or limitations found." a signifies "not applicable."

In five of six cases, we found problems with the various assumptions underlying the analyses of the test data. In at least two cases, such assumptions led to overly optimistic estimates of the system's capability, one of which was contradicted by available data. Other questionable analysis practices included combining data from disparate sources to yield overall performance estimates of unknown meaningfulness, removing valid data from performance computations, and lowering performance criteria after data collection. The impact of these practices was significant; they frequently allowed a system to appear to meet its performance requirements. Another problem was incomplete analyses—that is, analyses that did not integrate performance across all components of the total system or did not consider the limitations of other systems necessary for mission success.

Within recent years, the Congress has indicated an interest in operational testing information that permits a comparison between the new system and the older system it is replacing. In three of the six cases we reviewed, the system was tested comparatively against one or more older systems. However, comparisons were at times not tightly controlled, were less challenging than comparisons with the new system's own user criteria would have been, or were lacking meaningful criteria altogether. In addition, some measures on which the older system would have performed better were either not included in the test or not assessed comparatively, and limitations of particular test scenarios or departures from realism were condoned on the assumption that they

affected all systems equally, which was not always the case. In at least two cases, the limitation or departure clearly favored the new system.

There were also significant problems with each service test agency's reporting of results. In all six cases, the service test agencies stated findings, conclusions, or recommendations that were not consistent with the evidence or were not sufficiently qualified. In four of these cases, one or more requirements were reported to have been met when they were not, and in one case the service test agency recommended full production despite numerous unresolved problems, one of which it had itself previously termed "urgent." One system was reported as showing "vast superiority" over its competitors in overall mission effectiveness when in fact it had demonstrated superiority in only two of the five mission areas being compared.

In four of six cases, reporting was not consistently clear or comprehensive. (There were service differences here—for example, Army reports were highly comprehensive and detailed, while Air Force reports were less so.) There were two common problems: the omission of information or assumptions important in evaluating the results, and the omission or obfuscation of key realism limitations that favored the system being tested (such as crew familiarization or cueing).

Conclusions

We found significant problems and limitations in the planning, execution, realism, analysis, and service test agency reporting of the six OT&Es. Some of these problems and limitations were unavoidable due to time, resource, or safety constraints, although numerous others were not. We therefore conclude that for major, conventional systems that reached the B-LRIP milestone by the end fiscal year 1987, the OT&E being conducted under DOT&E oversight was not methodologically adequate for assessing the operational effectiveness and suitability of weapon systems. OT&E has tended to yield more favorable assessments than are likely to be found when the weapons are employed in combat, which can lead to the funding of weapon systems whose operational effectiveness and suitability have not been demonstrated. In sum, OT&E under DOT&E oversight has fallen short of the objectives sought by the Congress when it established the office.

We believe that there is a need for greater management emphasis on improving the implementation of DOD's OT&E directives in order to improve the conduct of OT&E and more effectively realize the intent of the law. DOD should improve the quality of OT&E performed under DOT&E

oversight to remove methodological biases and to correct the tendency we found toward overly favorable assessment of weapon performance. The following actions should be taken to address significant problems and limitations in the planning, execution, realism, analysis, and service test agency reporting of OT&E.

Regarding test planning, DOD should ensure that

- TEMPs include a complete statement of the system's requirements, and that test plans address all system requirements and critical operational issues identified in the TEMP, including (where appropriate) testing representative of all geographic, environmental, or mission conditions, testing at the edges of the performance envelope, and testing the complete, operational system; and, in addition, that
- test plans provide a clear correlation between critical issues and test objectives through test-verifiable criteria, and that test criteria reflect the performance and limitations of other components that support the mission, as specified by DOD directive.

Regarding test execution, DOD should ensure that

- requirements and critical issues are tested as specified in the test plan, and to the extent possible, that significant problems and limitations are anticipated, and that contingency plans are specified.

Regarding test realism, DOD should ensure that

- typical users (user participants representative of what the user will be like when the system is fielded) operate and maintain the system as specified by DOD directive, and that tests are in compliance with 10 USC 2366, which prohibits contractor involvement in OT&E where such contractor support would not be available in the field; and that
- testing is conducted under conditions simulating combat stress as specified by DOD directive, both on equipment and personnel; that to sufficiently stress equipment, tests should include—where applicable, technically and economically feasible, and within safety constraints—representative countermeasures (communication jamming, radar jamming, electro-optical countermeasures), tactics that are compatible with system survivability in a realistic threat environment, and operationally realistic portrayals of representative targets; and that to sufficiently stress personnel, tests should—where applicable, technically and economically feasible, and within safety constraints—ensure that crews are unfamiliar with the test range or target area prior to the test,

include scenarios where intelligence on enemy locations and other matters is not always readily available or accurate, eliminate any cueing that unrealistically reduces or removes the element of surprise, and stress crew endurance commensurate with stated mission requirements; and further that

- operational testing is accomplished in an environment as operationally realistic as possible as specified by DOD directive; that systems should be tested in a representative geographic environment (for example, test systems intended for deployment in Europe in a European-like environment), in terrain affording sufficient operating space and maneuverability, with unintended or naturally occurring countermeasures such as realistic thermal clutter, and in darkness or adverse weather within safety constraints; and finally that
- operational testing include threat-representative hostile forces as specified by DOD directive, including—to the extent technically and economically feasible—threat forces that are deployed in realistic threat formations and density, are portrayed as completely as possible, and are as capable technologically as the Soviet forces that new U.S. systems would actually face.

Regarding analysis, DOD should ensure that

- testing is planned and conducted to provide quantitative data and to minimize the need for subjective interpretation of system performance as specified by DOD directive, and that quantitative measures are reliable and valid; and that
- assumptions underlying the analyses of the test data are explicit and appropriate, and do not lead to methodologically biased estimates of system capability; that aggregations of data from multiple sources are meaningful and appropriate, that valid data are not removed from performance computations, and that analyses integrate performance across all components of the total system and consider the limitations of other systems necessary for mission success; and, in addition, that
- when new systems are tested comparatively against older systems with similar missions, comparisons are designed, implemented, and analyzed in a manner that preserves the comparison's validity, are no less challenging than comparisons against the new system's own operational performance specifications, include comparative assessment on all relevant measures (including those likely to favor the older system), and do not assume that limitations of test scenarios or departures from realism affect all tested systems equally.

Regarding service test agency reporting, DOD should ensure that

- service test agency findings, conclusions, and recommendations from operational tests are consistent with the evidence and sufficiently qualified, including the avoidance of statements that requirements were met when they were not, and language that exaggerates system performance; and that
- reporting be consistently clear and comprehensive, with all information and assumptions important to evaluating the results, including limitations, presented clearly.

DOT&E Impact on the OT&E Process

In order to better specify the unique contribution of DOT&E, we assessed the DOT&E impact on the OT&E process for the six cases. Support for these findings are found in section 6 of appendixes II through IV.

Our ability to evaluate DOT&E's impact on the operational testing of the six systems we reviewed was limited because much of the communication between DOT&E and other DOD components is informal and undocumented. As our March 1987 report noted, this lack of documentation makes it difficult to determine accurately how effectively DOT&E carries out some of its functions. In addition, we did not receive all the relevant documentation we requested. And, since other sources gave us relevant DOT&E documents that we could not obtain from DOT&E, there may well be additional documents that no source provided. Consequently, our assessment regarding DOT&E influence on the OT&E process is inconclusive.

Successful Attempts to Influence the OT Process

Though IR Maverick entered into B-LRIP after the swearing in of a permanent DOT&E director (hence its inclusion in our sample), the tests had already been completed; consequently, in the case of IR Maverick the director had no opportunity to influence the tests. In all of the other five cases we reviewed, we found evidence of DOT&E influence. In at least four (AHIP, TLAM/C, DDG-51, and LANTIRN), the influence affected, or will affect, the actual conduct of the tests. In the fifth case (Aquila), the only documented DOT&E influence consisted of additions and organizational improvements to the test plan that had no apparent impact on the test itself.

Several DOT&E action officers declined to enumerate all instances of impact to us on the grounds that revealing DOT&E's influence outside DOD would impede DOT&E's future influence inside DOD. Some action officers also told us that instances of successful DOT&E influence were too numerous to state. Consequently, it may be that other instances of successful DOT&E influence on these cases have occurred, but we were unable to substantiate them.

Unsuccessful Attempts to Influence the OT Process

In the documents we obtained, we found evidence of unsuccessful DOT&E attempts to influence operational testing in two of the five applicable cases. In both cases (Aquila and LANTIRN), 1) the service convinced DOT&E that the latter's concerns were less important than other considerations (such as safety), and 2) DOT&E recommendations were simply not implemented. We do not view these incidents as serious, however, or indicative of any general problems in this area. In addition, the working

relationship between DOT&E and the service test agencies was reportedly very good.

Impact on the B-LRIP Milestone

At the beyond-low rate initial production milestone, DOT&E reports to the secretary of defense and the Congress on the adequacy of operational testing and the effectiveness and suitability of the system. Therefore, the B-LRIP milestone represents a major opportunity for DOT&E impact on the program.

In three of the six cases (TLAM/C, DDG-51 and IR Maverick), we found no evidence that DOT&E influenced the B-LRIP milestone, other than to in effect support the production decision to the secretary of defense and the Congress. In the case of IR Maverick, several other DOD units raised significant concerns about the adequacy of testing and the operational effectiveness of Maverick. They presented these concerns to DOT&E before the B-LRIP decision meeting and at the meeting itself. DOT&E's B-LRIP report nevertheless stated that testing was adequate and effectiveness was satisfactory. In addition, we could find no evidence that DOT&E attempted either to defend its position or to respond to the concerns raised at the B-LRIP milestone meeting.

The fourth case (Aquila) was proposed for termination by the Army before the B-LRIP milestone and without consultation with DOT&E.

In the fifth case (AHIP), DOT&E took the position at B-LRIP that, as tested, AHIP demonstrated an operationally effective capability in only one of the three roles planned for it. Based primarily on DOT&E's assessment, the decision was made to procure AHIP for that role only. This decision had meaningful consequences; it meant that 179 AHIPs would be procured rather than the 578 the Army had requested. However, three other DOD units were also critical of AHIP's performance; one had already recommended "only a conditional approval of limited production" based on the test results, and the other two told us that they would have objected had DOT&E assessed AHIP as effective in more than one role. Because four different offices delivered essentially the same message, the unique impact of DOT&E's position is unclear in this case.

In the sixth case (LANTIRN), we found no evidence of DOT&E impact at the B-LRIP milestone for the navigation pod other than to in effect support the production decision to the secretary of defense and the Congress. Concerning the targeting pod, DOT&E advised the Air Force that full production was not justified by the operational tests and that if the Air

Force would defer full production, a B-LRIP report to the Congress would not be required. Instead, DOT&E would report to the Congress whenever the Air Force proposed to exceed a rate of 81 pods per year. However, 81 pods was the Air Force's intended purchase for the first year of full production. Essentially, DOT&E offered the Air Force a choice between a negative B-LRIP report to the Congress and a redefinition of the B-LRIP rate to delay the report. The Air Force chose the latter and thus was able to adhere to its planned first-year, full-scale production schedule.

Conclusions

Due to the limitations stated earlier, our assessment of DOT&E impact is inconclusive.

The interviews we conducted and the documentation we did obtain show that DOT&E has had at least some impact on the OT&E process and the production decisions that flow out of it. In sum, we found successful influence on the testing of four systems, and unsuccessful DOT&E influence in two cases that we do not view as serious. We also found no evidence of DOT&E impact in three major production decisions other than to in effect support the decision, no opportunity for impact in one, and in the other two, impact that was either indistinguishable from that of other DOD units or that was more apparent than real.

Quality of DOT&E Dissemination of Information to the Congress

As noted earlier, the statute establishing DOT&E imposed two principal congressional reporting responsibilities: 1) a B-LRIP report stating whether OT&E was adequate and whether OT&E results confirm the item or components to be effective and suitable for combat, and 2) an annual report. The statute also requires the director to respond to requests from the Congress for information regarding OT&E. In addition, DOT&E has initiated on its own the publication of monthly highlight reports that provide summary information on office activities and the progress of OT&E for specific programs. These are the sources used in our evaluation of DOT&E dissemination of information to the Congress.

DOT&E issued B-LRIP reports to Congress for the following five systems from our sample of six: AHIP, TLAM/C, DDG-51, IR Maverick, and LANTIRN (navigation pod only). In general, the reports stated that, despite limitations, OT&E was adequate and the system was effective and suitable as tested, with the following exceptions: AHIP was assessed as effective in only one of its three roles, IR Maverick suitability was assessed as marginal, and LANTIRN navigation pod suitability was not evaluated on the grounds that the testing did not provide all the necessary information. B-LRIP reports for Aquila and LANTIRN targeting pod have not been written, but official DOT&E statements on the adequacy of Aquila testing and the effectiveness and suitability of the LANTIRN targeting pod (to date) were available from other sources. The May 1987 DOT&E monthly highlights report stated that the Aquila OT supplied sufficient data to address all issues adequately, and the fiscal year 1987 annual report described LANTIRN targeting pod performance as satisfactory in regard to some of its effectiveness and suitability objectives, with the rest requiring further testing. There were no official DOT&E statements on the effectiveness and suitability of Aquila or the adequacy of the LANTIRN targeting pod operational test.

Congressional concern about obtaining complete and accurate information on OT&E was a major reason for the provisions concerning dissemination of information to the Congress in the DOT&E legislation. In his confirmation hearing, the director similarly stressed the importance of candid communication with the Congress, and the importance of providing complete and accurate information was reiterated by Members of the Senate during the confirmation process. To determine the completeness and accuracy of DOT&E statements to the Congress, we compared the facts as stated in DOT&E reports to those identified during our evaluation and reported in sections 1 though 5 of appendixes II through IV. Results are summarized in table 4.1. Support for the findings in the table and

accompanying text can be found in section 7 of appendixes II through IV.

Table 4.1: Significant Problems in Completeness and Accuracy of DOT&E Reporting

	Army system		Navy system		Air Force system	
DOT&E reporting	**AHIP**	**Aquila**	**TLAM/C**	**DDG-51**	**IR Maverick**	**LANTIRN**
OT&E adequacy						
Statements complete?	X	a	X	X	X	X
Statements accurate?	X	X		X	X	X
System effectiveness and suitability						
Statements complete?		a	X	X	X	X
Statements accurate?	X	a	X	X		X

Note: empty cells signify "no significant problems or limitations found." X signifies "one or more significant problems or limitations found." a signifies "insufficient information to evaluate."

DOT&E Statements on Adequacy of OT&E

We found one or more individual DOT&E statements on OT&E adequacy to be incomplete in five of six cases and inaccurate in five of six cases. By incomplete statement, we mean a statement that omitted information relevant to an assessment of adequacy. Typically, such omissions consisted of the failure to report test limitations such as those discussed above. In some cases, the limitations identified by the service test agency were reported, but additional limitations were not (IR Maverick, TLAM/C); and in others, the limitations identified by the service test agency and additional limitations were not reported (AHIP, DDG-51, and LANTIRN). Inaccurate statements included the following: tests were described as more challenging and realistic than they actually were (DDG-51 and LANTIRN), certain test assets were reported not to exist when in fact they did exist (IR Maverick), and the sufficiency of the test data was overstated (AHIP and Aquila).

We further assessed whether DOT&E's assessments of overall OT&E adequacy in the B-LRIP reports and other sources of information disseminated to the Congress were supported by the evidence. Of the six favorable adequacy assessments, we found that five were not supported by the evidence (AHIP, Aquila, TLAM/C, DDG-51, and IR Maverick). In the sixth (LANTIRN navigation pod), we found no evidence inconsistent with DOT&E's assessment of overall adequacy. DOT&E made no overall adequacy statement for the LANTIRN targeting pod testing; however, the fiscal year 1987 DOT&E annual report made clear that before a favorable B-LRIP report can be written further tests are required. We concur with that assessment.

DOT&E Statements on System Effectiveness and Suitability

We found one or more individual DOT&E statements on system effectiveness and suitability to be incomplete in four of five cases and inaccurate in four of five cases. (There was no official statement on system effectiveness and suitability for Aquila.) Incomplete statements included the following: failure to mention as "urgent" a problem so characterized by the service test agency (TLAM/C), omitting key factors from an analysis, resulting in an unrealistically favorable performance assessment (IR Maverick), and omitting unfavorable test results (DDG-51 and LANTIRN). Inaccurate statements primarily consisted of overstatements of performance (AHIP, TLAM/C, DDG-51, and LANTIRN); in each of these cases we found statements in which specific aspects of system performance were reported as more successful than the test results demonstrated. We found no statements that were inaccurate because they underrated performance.

We further assessed whether DOT&E's assessments of overall system effectiveness and suitability in the B-LRIP reports and other sources of information disseminated to Congress were supported by the evidence. Of the five favorable assessments of system performance, we found that four were not supported by the evidence (AHIP, TLAM/C, DDG-51, and IR Maverick). In the fifth (LANTIRN navigation pod), we found no evidence inconsistent with DOT&E's assessment of system effectiveness and suitability. In the case of the LANTIRN targeting pod, we concur with DOT&E's statement that more testing is needed to assess effectiveness and suitability.

Conclusions

Each of the official DOT&E reports to the Congress that we reviewed contained incomplete or inaccurate statements, and most contained both. In addition, the majority of favorable overall assessments of OT&E adequacy and of system effectiveness and suitability were not supported by the evidence. As noted earlier, some problems and limitations in operational testing are unavoidable; however, we know of no reason why those problems and limitations cannot be reported completely and accurately. The omissions, inaccuracies, and overall assessments consistently presented a more favorable presentation to the Congress of test adequacy and system performance than was warranted by the facts. We therefore conclude that for major, conventional systems that reached the B-LRIP milestone by the end of fiscal year 1987, the quality of DOT&E dissemination of information to the Congress has not provided the complete and accurate picture of weapon performance that the Congress needs to make weapon funding decisions. As such, it has fallen short of the objectives sought by the Congress when it established DOT&E.

We believe that there is a need for greater management emphasis on improving the implementation of DOD's OT&E directives in order to improve the reporting of OT&E and more effectively realize the intent of the law. DOT&E should improve the quality of the information it disseminates to the Congress and avoid providing a more favorable presentation of test adequacy and system performance than is warranted by the facts. To improve the completeness and accuracy of DOT&E reports to the Congress, DOT&E should

- ensure that its assessments of overall OT&E adequacy and of system effectiveness and suitability are supported by the evidence;
- state all of the significant limitations reported by the service test agency, as well as any additional significant limitations not reported by the service test agency;
- avoid presenting tests as more challenging and realistic than they actually were, or overstating the sufficiency of the test data;
- characterize performance problems more completely, and report all relevant test results, both favorable and unfavorable; and finally,
- ensure that statements on system effectiveness and suitability are commensurate with test results.

GAO OT&E Assessment Framework

1. Planning

1.1 Did the TEMP include a complete statement of the system's requirements?

1.2. Did the test plan address all system requirements and critical issues identified in the TEMP?

1.3. Was there a clear relationship in the test plan between required system characteristics/critical issues and test objectives/missions through operationally meaningful test-verifiable criteria?

2. Execution

2.1. Was each system requirement and critical issue identified in the test plan tested for as planned?

2.2. Were there limitations in implementation that had not been anticipated in the test plan?

3. Realism

3.1. Was the system operated by typical operational units? organizational level?

e.g.Originally specified

Units themselves typical?

Contractor involvement?

3.2. Was the system operated by typical operational personnel?

e.g.Use of representative troops rather than "golden crews"?

Unfamiliarity with range/scenario?

Training representative of overall force?

Training commensurate with a wartime training schedule?

Contractor involvement?

3.3. Was the system supported by typical support units?

e.g.As with operational units.

3.4. Was the system supported by typical support personnel?

e.g.As with operational personnel.

3.5. Was the equipment put under realistic stress by design?

e.g.Outer envelope of performance requirements tested?

3.6. Were personnel put under realistic stress by design?

e.g.Element of surprise where appropriate?

Operating tempo representative of combat?

Duration of test pushes endurance?

3.7. Were realistic combat tactics employed?

3.8. Did the physical environment approximate the intended range of environments?

e.g.Terrain?

Temperature?

Weather/sea state?

Time of day?

Clutter?

Unintended countermeasures?

IFF?

3.9. Did target systems approximate actual targets, realistically employed?

3.10. Did threat systems approximate actual threat, realistically employed?

3.11. Was the tested system production representative and prepared for test in a realistic manner?

4. Analysis

4.1. Were measures quantitative and non-subjective?

4.2. Were quantitative measures reliable and valid?

4.3. Were analytic assumptions explicit and appropriate?

e.g.Data aggregations/disaggregations appropriate?

Causes of failures appropriately attributed?

Analytic assumptions supported by data?

All valid test data included?

Rationale for "no-tests" appropriate and consistent?

Other?

4.4. Was sample size adequate to support statistically valid results?

4.5. Were comparisons with other systems valid?

5. Servie operational test agency reporting

5.1. Were findings, conclusions, and recommendations consistent with the evidence and appropriately qualified?

5.2. Was reporting clear and comprehensive?

6. Evidence of DOT&E Impact

6.1. Were there successful attempts to influence the OT&E process?

(continued)

6.2. Were there unsuccessful attempts to influence the OT&E process?

6.3. What was DOT&E impact on the BLRIP milestone?

7. DOT&E Reporting

7.1. What statements did DOT&E make to the Congress regarding the adequacy of OT&E and system effectiveness and suitability?

7.2. What was the completeness and accuracy of DOT&E's statements regarding adequacy of OT&E?

e.g. All significant OTA identified limitations and problems identified and explained?

All significant limitations and problems not identified by OTA identified and explained?

7.3. What was the completeness and accuracy of DOT&E's statements regarding system effectiveness and suitability?

e.g. Unresolved system performance issues identified?

Unmet criteria identified?

Army OT&E

The agency responsible for conducting and reporting Army operational testing is the Army's Operational Test and Evaluation Agency (OTEA). OTEA produces the test plan, test report, and an independent evaluation report (IER) for each OT&E.

AHIP

System Description

The Army Helicopter Improvement Program (AHIP) OH-58D aeroscout helicopter is an enhanced, upgraded version of the Army's current OH-58C observation helicopter. Its most prominent feature is a mast-mounted sight system which protrudes above the rotor hub. The mast-mounted sight was designed to acquire, locate, and laser-designate targets day or night in obscured atmospheric conditions from stand-off ranges while the airframe of the helicopter remains below the terrain mask. This minimizes exposure of the helicopter to enemy radar and electro-optical detection devices, and therefore is expected to enhance survivability. AHIP's navigation system was designed to provide accurate, autonomous, and fully integrated positioning to meet target and self-location requirements of its aeroscout mission. Communication equipment was designed to provide simultaneous voice and digital secure communication with other helicopters by means of an automatic target handover system. AHIP has a crew of 2, a commissioned or warrant officer pilot and an enlisted aerial observer.

The AHIP was designed to fulfill 3 battlefield roles: attack, air cavalry, and field artillery aerial observer (FAAO). In the attack helicopter role, the scout and attack helicopters operate in close harmony as "hunter-killer" teams with the scout locating and designating targets for the attack helicopter's laser guided Hellfire missiles. The aeroscout sees and prepares the battlefield for attack helicopters, and the aeroscout's primary efforts are directed toward controlling and assisting attack helicopters while they destroy threat targets.

In the air cavalry role, aeroscouts use the general concepts established for the ground cavalry, but with some important differences. Air cavalry units provide an increased capability to rapidly reconnoiter and maintain surveillance over wide areas of the battlefield. They may operate independently, may work in conjunction with ground cavalry, or may be part of a combined arms team. The most frequent missions given to air cavalry units are reconnaissance and screening.

In the field artillery aerial observer (FAAO) role, the aeroscout helicopter provides an aerial platform from which to adjust the firing of conventional and precision-guided munitions. The FAAO conducts battlefield reconnaissance to gather target information to request and adjust indirect fires (i.e., fires where the firing unit does not see the target, i.e., artillery), laser-designates for precision-guided munitions such as Copperhead, and coordinates with the supported ground commander to ensure accurate and timely fire support.

Program Status

578 AHIPs were originally planned for production. The Army obtained approval to buy only 179 after an October 1985 Secretary of Defense Decision Memo (SDDM) which approved production for only 1 of the 3 roles (FAAO). 135 had been bought when the Army attempted to terminate the program in its FY88 budget submission, reportedly due to budgetary considerations. Congress voted to restore funds to buy 36 more aircraft in FY88.

OT&E History

There have been 2 operational tests of AHIP. The first—Operational II (OT II)—compared the AHIP to the OH-58C. The objective was to test AHIP in all 3 aeroscout roles discussed above. The test was conducted at Ft. Hunter-Liggett, CA, from September 1984 through February 1985, prior to the swearing in of a permanent DOT&E director. However, the B-LRIP report based on those tests was written under the director and issued in September 1985. The report concluded that as tested, AHIP demonstrated an operationally effective capability in only 1 role (FAAO). As noted above, the SDDM production decision was in accordance with this conclusion. It further directed the Army to conduct a second operational test to resolve issues not fully answered by the previous testing.

Plans were developed for the second test—the AHIP Follow-on Test and Evaluation (FOT&E)—but before the test was conducted the Army deleted production funds for any new AHIPs. The test was not cancelled, however, but redesignated the Army Aerial Scout Test (AAST). Its objective was to compare alternative candidate systems to the baseline AHIP. These included the OH-58C, OH-58C+ (OH-58C with infrared sensor), AH-1S Cobra (modernized), and AH-64 Apache. Initially, both the air cavalry and attack roles were to be tested, but only the air cavalry phase was conducted. AAST was conducted from March to May 1987, and like OT II, was held at Ft. Hunter-Liggett.

Assessment of Evaluation Questions for AHIP OT&E

1. Planning

1.1. Did the TEMP include a complete statement of the system's requirements? We cannot address this question because requested TEMPs were not provided by DOD.

1.2. Did the test plan address all system requirements and critical issues identified in the TEMP? Again, we cannot address this question because requested TEMPs were not provided by DOD.

1.3. Was there a clear relationship in the test plan between required system characteristics/critical issues and test objectives/missions through operationally meaningful test-verifiable criteria? We found no significant problems or limitations.

2. Execution

2.1. Was each system requirement and critical issue identified in the test plan tested for as planned? The OT II test plan was significantly changed after approval. At the suggestion of Program Analysis and Evaluation (PA&E), it was modified to include an investigation of the aeroscout contribution to the attack helicopter, including a sensitivity experiment on the scout/attack mix. This change required deleting the AH-1S/AHIP trials (leaving only the AH-64/AHIP trials) and restructuring of the test matrix. The modification of the OT II test plan after its approval to include a scout/attack mix experiment broadened the scope of the test but significantly diluted the originally planned testing. According to the DOT&E action officer, DOT&E argued against the change but at that time lacked the influence to prevent it (as noted earlier, this test took place prior to the swearing in of a permanent director). OTEA officials characterized the change as losing sight of what the test was to accomplish so that a lot was done but not done well. This was further exacerbated by pressure to meet a production milestone review date.

2.2. Were there limitations in implementation that had not been anticipated in the test plan? When OT II record trials began, 44 trials were planned. As testing proceeded, some trials had to be terminated for various reasons. Due to time lost from delays in the training phase, the added task of conducting scout/attack mix sensitivity trials, and the inflexibility of the established dates for the Milestone III review, these lost trials were never rescheduled. Consequently, the originally scheduled 44 trials were reduced to 24. The time constraints were also partly

responsible for measurement validity problems (see 4.2) and the poor performance of threat air defenses (see 3.10).

The sensitivity trials were themselves diluted by 1) the nonavailability of a sufficient number of attack helicopters and 2) the invalidation or cancellation of trials due to instrumentation problems and adverse weather conditions. Due to limited AHIP availability, OH-58Cs were used in some FAAO trials in their place, i.e., the wrong aircraft was used. This considerably complicated the data reduction process and reduced the validity of the AHIP versus OH-58C comparison. According to OTEA officials, the accelerated schedule combined with inadequate resources caused them to lose control of the test. For example, the number of OTEA personnel on the ground was not sufficient to ensure proper execution of the test. Ultimately, morale was also affected; the prevailing attitude among testers as well as player personnel was that it was more important to complete the test quickly than to do it correctly.

AAST ran much more smoothly but still had unanticipated limitations. To compensate for reliability problems in the electronic line-of-sight system (ELOSS), which records engagement opportunities, ELOSS was to be augmented in AAST with a scanning laser system. However, during exploratory trials it was discovered that the beam from the scanning lasers was visible to aircrews through both day video optics and night vision goggles, creating a test artificiality. Consequently, it could not be used. The consequence was that the number of engagement opportunities was unknown (for implications, see 4.2). In addition, the use of flash simulators was discontinued during night trials after it was determined that the flash was interfering with the aircrew's use of night vision goggles, which potentially affected safety. This reduced the operational realism of the test since flashes would occur on the battlefield.

3. Realism

3.1. Was the system operated by typical operational units? We found no significant problems or limitations.

3.2. Was the system operated by typical operational personnel? After the personnel selection and training in OT II had failed to produce pilots and observers capable of operating AHIP as intended (see 3.7), a provisional attack helicopter battalion (TF1-112) was specially established as the test unit for AAST (all aircraft). The pilots selected were to be a cross-section of new, medium, and highly experienced pilots. However, these crews were not typical, particularly those flying AHIPs: 1) there were 2

instructor pilots in the AHIP crews, none in the AH-64 crews (the AH-64 was AHIP's closest competitor in the scout role); and 2) TF1-112 was formed a full year in advance of record trials to train for the test. OTEA officials stated that crews would not experience this level of training given a wartime scenario, but noted that they would also not be charged with developing tactics and doctrine, one reason why initial training takes longer. However, the fact that the crews used in the operational test had been used to develop tactics and doctrine only strengthens the argument that they were atypical. Therefore, the capability of typical aircrews to perform AHIP's mission in an operational environment has still not been demonstrated.

3.3. Was the system supported by typical support units? Some AHIP maintenance was performed by contractors in OT II, potentially resulting in better AHIP reliability, availability, and maintainability (RAM) performance than could be expected in an operational environment. In addition, the quality of the RAM data was poor, and contractors participated in RAM scoring conferences and assessments (see 4.2). This is particularly significant because RAM was not assessed in AAST, where contractors performed all intermediate and depot level maintenance. Consequently, the supportability of AHIP by typical units and troops in an operational environment has still not been demonstrated.

Early in AAST, AHIP crews complained of difficulty in detecting and recognizing targets using the forward looking infrared receiver (FLIR) thermal imaging system, which they believed was lowering their detection and recognition rates. Approximately 3 weeks into record trials, a manufacturer's representative was brought in to present a class on how to tune the system to improve performance. On the first trial after the class, the crew detected and reported targets in all 5 of the presented target arrays, whereas their best previous effort was 2 arrays. Crew interviews confirmed that the initial training was incorrect, and only corrected by the manufacturer's class. Though clearly not a realistic event, the class was defended on the grounds that it would have been inappropriate to continue the test knowing that the crews had not been sufficiently trained to use the system correctly. However, the AH-64 crews also had difficulty in detecting and recognizing targets and complained about the lack of definition in the FLIR imagery and the lack of a thermal signature when viewing a target area. They also stated, as did the AHIP crews, that their training on how to use the FLIR had been inadequate. Yet AH-64 crews did not get a class from the contractor to help them as AHIP pilots did. Providing the class to AHIP crews and not AH-64 crews biased the results in favor of AHIP.

3.4. Was the system supported by typical support personnel? See 3.3.

3.5. Was the equipment put under realistic stress by design? In OT II, testers attempted to employ smoke as a countermeasure, but it was used in only 5 of the 24 trials because smoke generators were not generally available. In AAST, with the exception of some limited communication jamming, no countermeasures of any type were used during the operational trials. Smoke and camouflage were employed in a non-operational subtest, on the grounds that employing them during regular reconnaissance trials would have degraded the performance of instrumentation. Other infrared (IR) countermeasures (e.g., flares, lasers) were not used at all. Effects of IR jamming were accounted for in the data, i.e., assessed through models. The absence of actual IR jamming is significant because AHIP relies on IR technology. Its capability in an operational countermeasure environment remains unknown.

3.6. Were personnel put under realistic stress by design? OT II players were extremely familiar with the test area, to the point where their navigation task was negligible. Cockpit workloads, therefore, were less than they normally would be during combat operations. In AAST, TF1-112 crews were trained at a separate location prior to arriving at Ft. Hunter-Liggett, in part to diminish terrain familiarity problems. However, TF1-112 still spent a full 6 weeks conducting training and practice trials at the test site prior to record trials, and crews were therefore familiar with key terrain features before the test began.

In OT II, pilots knew the enemy's location. In AAST, 8 different target laydowns were used, and aircrews were not told which laydown they would face. However, they did know there would be 5 arrays, one in each reconnaissance zone. Therefore they would know that once an array was detected no further arrays could be in that zone, which might not be the case in an actual battle. Also in AAST, AHIP crews learned how to insure during mission planning that they would have line-of-sight into a suspected target location area. This is only possible when repeating trials over known terrain.

Also in AAST, crews initially had difficulty with enemy communications jamming, but they rapidly learned to work through it by using shortened and rapid bursts of voice communication. This appears to have been possible only because the jamming was kept periodic and predictable. The jamming technique used was only one of many that might have been used to represent Soviet jamming. OTEA officials admitted that 1) the Soviets have more effective jamming techniques available, and 2) it

might have been more realistic to change the jamming once it was broken through. However, this was not done because it would have "shut down the test," i.e., the aircraft would not have been able to perform.

Requirements specify that AHIP crews must be capable of performing all flight and mission operations in chemical protective equipment. OT II included a limited test of conditions which would prevail in a chemical environment, but safety restrictions limited the degree of realism. For example, night tests were conducted on the ground with only the observer in protective equipment. There were consistent complaints about the difficulty of operating AHIP's small, multifunction keyboard while wearing protective gloves, and pilots said they would not be able to fly the aircraft safely wearing all the protective equipment. In AAST, no further attempt was made to test operations in a chemical environment. Therefore, the capability remains undemonstrated.

3.7. Were realistic combat tactics employed? The principal improvement of AHIP over OH-58C is its mast-mounted sight, which allows AHIP to minimize its exposure to the enemy by hovering behind the terrain mask. During OT II, however, AHIP exposed itself to threat units at a 60 percent greater rate than did OH-58C. AHIP pilots who had been trained to fly earlier scouts (e.g., OH-58C) had marginal confidence in their observer and insisted on seeing the battlefield directly. In so doing, they climbed above the terrain mask and exposed the entire aircraft. The failure of AHIP pilots to follow intended tactics biased downward its likely survivability, but also may have biased upward its target acquisition performance.

3.8. Did the physical environment approximate intended range of environments? Both OT II and AAST were conducted at Ft. Hunter-Liggett, where the test area consists of 3 long, narrow valleys bordered on both sides by mountains. It was primarily selected for its instrumentation and laser safety clearance; Army officials stated that it is not typical of a European environment. While there are mountainous regions in Central Europe, the prevailing terrain is one of rolling hills with heavier forestation than we observed in the target areas (valleys) of Ft. Hunter-Liggett. The amount of forestation is significant because it affects how well ground targets can hide from scout aircraft.

Ft. Hunter-Liggett's terrain limits maneuverability and thus limits the types and sizes of trials that could be run. In both tests, test details were "tailored" to fit the terrain, and in AAST the terrain restrictions greatly influenced both the choice of scenario and the tailoring and sizing of the

threat forces (see 3.9), significantly limiting the generalizability of the test. AHIP proponents claim that the Hunter-Liggett test area works against AHIP because 1) it is too small to adequately showcase AHIP's capabilities, and 2) the long lines-of-sight over the mountains enable the AH-64 to scout for itself and survive. Others in DOD have said that the boundaries were too restrictive to stress the system, and that AHIP crews could simply fly up the mountain outside the valley and know they'd see targets when they looked in (see 3.6). In either case, the operational capability of AHIP in a European-like environment remains unknown.

As noted earlier (see 2.2), flash simulators were not used during night trials after it was determined that the flash interfered with aircrew use of night vision goggles, which potentially affected safety. This reduced the operational realism of the nighttime test environment since flashes would occur on the battlefield.

3.9. Did target systems approximate actual targets, realistically employed? As described in 3.8, the terrain restrictions greatly influenced the choice of scenario and the tailoring and sizing of threat forces used in AAST. The reconnaissance trials were specified to be limited counterattacks against threat motorized rifle forces in hasty defensive positions. In accordance with this scenario's storyline, targets were stationary and non-camouflaged. OTEA officials admitted that holding all targets stationary may not have been realistic, but believed it permitted a more challenging test, moving targets are generally considered to be easier to detect. However, AHIP would encounter moving as well as stationary targets in the course of the air cavalry mission, and camouflaged as well as non-camouflaged targets. While the targets as portrayed may have realistically portrayed one possible scenario, the relative likelihood of that scenario occurring was not made clear, and the results do not permit generalization to other air cavalry scenarios.

3.10. Did threat systems approximate actual threat, realistically employed? In OT II, the Army Development and Acquisition Threat Simulators (ADATS) did not provide an adequate simulation of threat air defenses. Despite many verified engagement opportunities, very few engagements were attempted. The reasons were: 1) contractors had difficulty maintaining the several different types of ADATS equipment and meeting the test's 7-day-a-week schedule; and 2) ADATS players simply did not perform to expectations, frequently commenting that they could not detect blue aircraft. In fact, the test director personally chided the ADATS representatives to motivate players to do better. Ineffectiveness

of ADATS decreased the operational realism of the test, and likely led to an overestimate of AHIP survivability.

ADATS was more successful in AAST, but portrayal of threat capabilities was still limited by the availability of forces and equipment and the sophistication of instrumentation and methodologies used. Specifically, 1) ADATS systems were available to simulate only one Soviet air defense system, with all others portrayed by U.S. surrogates, 2) instrumentation limited the ability of gunners to obtain valid engagements, and 3) machine guns and other small arms weapons (commonly on the battlefield in large numbers) were not instrumented. Additionally, ADATS crews were fresh (i.e., not fatigued) and totally focused on the aircraft with no distractions from ground or air based air defense suppression. Finally, no red air threat was played in either test, other than for the purpose of determining blue helicopters' ability to detect and recognize counterair targets; no air-to-air maneuver or engagements were conducted.

3.11. Was the tested system production representative and prepared for test in a realistic manner? We found no significant problems or limitations.

4. Analysis

4.1. Were measures quantitative and non-subjective? We found no significant problems or limitations.

4.2. Were quantitative measures reliable and valid? The method of determining detections and recognitions in OT II was to score a detection whenever a target array was observed and recognitions for each individual target identified within the array. Consequently, there were more target recognitions than detections. The result was a data base without a uniform basis for relating detection to recognition, recognition to hand-off, and hand-off to engagement. Moreover, the collection of data to address detection, recognition, and hand-off was described by testers as sometimes difficult and sometimes impossible. Detection and recognition events could only be captured accurately if players commented verbally, e.g., "there's a tank," which is not a normal operating procedure to the aircrews. Therefore, data were either not collected or collected with degraded accuracy.

There were reliability problems with ELOSS in both tests, exacerbated in OT II by the accelerated trial schedule which precluded validation of the ELOSS data between trials. The reliability in AAST was not expected to

exceed 60 percent, and in the test it was around 50 percent. Attempts to augment ELOSS with a scanning laser system were not successful (see 2.2). The consequence was that the number of detection opportunities was unknown. This limited the ability to explain some performance differences between candidates, such as why AHIP detected targets at much greater ranges than AH-64 when technically their FLIR systems are the same. Any limitations to explaining performance differences are significant since the purpose of the test was to compare candidate aircraft.

In OT II, there was no accountability for RAM data flow and quality assurance. Recommended procedures for entering, correcting, and uploading data to the data base were not followed in many cases. As a result the quality of the RAM data was described as "extremely poor" by OTEA officials. Also in OT II, the AHIP prime contractor participated in Data Analysis Group (DAG) meetings. (The DAG validates test data and resolves problems associated with it.) By virtue of its participation, the contractor could influence the evaluation of its own system. Similarly, there was extensive contractor participation in the RAM scoring conferences. Out of 31 total attendees at the first scoring conference, 17 were contractor personnel. In accordance with the 1986 law prohibiting contractor involvement in operational testing, contractors were not permitted to participate in the AAST DAG, although they had originally been scheduled to.

4.3. Were analytic assumptions explicit and appropriate? In OT II, AHIP's day detection range was initially calculated based on 120 detections achieved during all AHIP attack, air cavalry, and FAAO day trials. However the IER used only the 88 detections achieved during attack trials. The use of this subset favored AHIP. This was significant because the mean detection range based on the attack subset met the user criterion, whereas based on the full data set it did not. Consequently, the IER reported that the criterion was met. The user criterion as stated was not limited to performance in the attack role, and using more data points provides greater statistical confidence in the results. Yet no justification was provided in the IER for using only the attack subset.

The AAST IER's statistical analysis of the effect of the FLIR tuning class and other mid-test interventions showed only a small effect on AHIP performance. This was surprising in that the class had been given specifically to solve problems in the use of the FLIR that were degrading AHIP's performance. On closer look, we found that several factors contributed to statistically underpowering this analysis (that is, precluding it from demonstrating a larger effect). They were: 1) the use of separate

pairwise comparisons between each of 9 "event groups" rather than a before-after comparison using all the data; 2) the use of a 95 percent confidence interval when other analyses in the IER had used 80 percent; and 3) the use of a highly conservative statistical test. As a result, some confidence intervals were so broad that only extremely large differences could have been statistically significant. Thus the analysis obscured the fact that the FLIR tuning class in the middle of the test - obviously an unrealistic event - did influence the results in AHIP's favor (see 3.3).

4.4. Was sample size adequate to support statistically valid results? Changes in the OT II test plan combined with some additional unanticipated problems (see 2.1 and 2.2) brought about reduced statistical validity. For example, there were only 2 air cavalry missions for AHIP and 2 for the OH-58C, and neither of the latter was conducted at night. A principal objective of the test was to compare AHIP to the OH-58C, yet numerous intended comparisons could not be made due to small sample sizes.

4.5. Were comparisons with other systems valid? Both OT II and AAST were comparative tests. OT II tested AHIP against the existing scout OH-58C, whereas AAST tested various scout alternatives against the AHIP as a baseline. The validity of the AHIP comparisons can be questioned on 4 grounds. First, they permitted a less challenging test than would tests against pre-established user requirements. In AAST, for example, the IER concluded that AHIP was "dominant" in locating enemy targets over all other scout candidates. [material deleted] Second, measures on which the comparison system would perform better were either not included or not tested comparatively. The OT II test report and IER present AHIP RAM performance data, but neither presents RAM performance data for the OH-58C. Since AHIP has a more complex, technologically advanced mission payload than the OH-58C, it is also more expensive, requires more training, and has a greater maintenance burden (e.g., AHIP's laser boresight had to be verified before and after each trial). Yet, none of these issues were examined comparatively in either test. Third, comparisons were at times invalid because they were not well controlled. The AHIP crews receiving a contractor class during AAST on the thermal imaging system while the AH-64 crews did not (see 3.3) is one example; in addition, AH-64 pilots were less experienced than AHIP pilots in terms of flight hours in their assigned helicopter. Both factors favored AHIP for reasons that were extraneous to the aircraft's capability. Finally, departures from realism were condoned by testers on the grounds that they affect all systems equally, when in fact the assumption may be incorrect due to unmeasured interaction effects. For example, the approval of the threat

force laydowns for AAST was based on AAST being a comparative test, and therefore not having to accurately represent the threat in all respects. Because of the mast-mounted sight, however, AHIP's survivability tactics are different from the other candidates; it is therefore likely that observed differences in survivability will depend on how the threat is portrayed (e.g., modern radars that can "see" behind the terrain mask might be relatively more effective against AHIP). Other limitations cited in the test report were effectively downgraded in importance because they would apply to all the systems tested.

5. Service Test Agency Reporting

5.1. Were findings, conclusions, and recommendations consistent with the evidence and appropriately qualified? The operational requirement for AHIP target acquisition performance specified that target acquisition be achieved at specified ranges with [material deleted] As noted in 4.3, the OT II IER based its statement that the day range criterion was met on a subset of data; had the entire data set been used the criterion would not have been met. As to the probability requirement, probability of detection and recognition were never presented in the test report or IER, as they were judged by the test evaluator to be of little value for performance comparisons. In sum, neither the acquisition range nor probability requirements were demonstrated. The IER nonetheless stated that the target detection and recognition requirements were met or exceeded.

The target location criterion was that targets would be handed off with operational accuracy within [material deleted] under simulated battlefield conditions. Data on target location accuracy were collected during OT II but not published in the test report. The report did state, however, that target handover location radial error values were grossly higher than the requirement (greater than 1 km in some cases). In addition, data from the live fire trials indicated that the accuracy requirement was not met, but the IER states that the same requirement was met or exceeded. This latter statement is not consistent with the evidence.

Some human factors conclusions have not been consistent with the evidence. The OT II IER concluded that human factors aspects of AHIP do not detract from system performance, but made no mention of a 1985 human factors analysis which concluded that AHIP should not be approved for production unless adequate cooling and ventilation were provided to crewmembers. (The problem had not been addressed in OT II, which was conducted in winter.) In AAST (conducted in spring),

crewmembers unanimously reported that the cockpit temperatures were too high in hot weather and that the high cockpit temperatures degraded their performance. Once again, the IER made no mention of the problem. Also in AAST, AHIP crews were surveyed on the effectiveness and suitability of 8 cockpit modifications made as a result of developmental testing II. Crewmembers registered complaints on 7 of 8 of the modifications. The test report concluded that aircrews indicated that the modifications were, for the most part, effective.

We believe the AAST IER's conclusions were overstated. It reported that AHIP had demonstrated it was "overwhelmingly" the most effective scout for detection, recognition, and location of enemy targets. In fact, AHIP statistically outperformed the other aircraft on only 7 of 16 measures of performance. On 3 other measures, it outperformed some of the aircraft but not all. Similarly, AHIP was reported to show "vast superiority" over all alternative aircraft in overall mission effectiveness. In fact, AHIP did outperform the other scouts in 2 of the 5 functional areas evaluated (navigation and target acquisition) but no scout demonstrated clear superiority in the other 3 (survivability, target handover, and reporting of information).

5.2. Was reporting clear and comprehensive? We found no significant problems or limitations.

6. Evidence of DOT&E Impact

Our ability to assess DOT&E's impact on the operational testing of AHIP was limited because 1) much of the communication between DOT&E and the Army was informal and undocumented, and 2) we may not have been provided all the documentation that exists. Since AHIP OT II trials were completed prior to the swearing in of a permanent director, we will not assess attempts by DOT&E to influence the conduct of that test. However, the B-LRIP report on OT II appeared several months after the swearing in, so we will discuss that (see 6.4).

6.1. Were there successful attempts to influence the OT&E process? In their memo approving the AAST test plan, DOT&E stated that while not included in the plan, it was their understanding that unit level maintenance of AHIP would be performed by soldiers. This was incorporated in the final test plan and was implemented in the test. (Intermediate and depot level maintenance was still performed by contractors.) OTEA officials told us they would not have had soldiers do the maintenance had

they not been directed to by DOT&E because AAST was not intended to address RAM issues.

We found no other evidence of successful attempts to influence AAST. The DOT&E action officer told us that due to the unusual nature of the test—it was being held to compare alternative aircraft to AHIP, rather than to support an AHIP production decision, because the AHIP program was technically dead at the time—DOT&E was hesitant to recommend major design changes. DOT&E's concern was that AHIP could be removed from the test by the Army, which would then effectively remove any requirements for DOT&E oversight. However, this should not be construed as indicating that DOT&E had no opportunity to influence the test. They approved the test plan and, according to OTEA officials, closely monitored the implementation. It therefore seems likely that there were opportunities for influence below the level of major design changes.

6.2. Were there unsuccessful attempts to influence the OT&E process? We found no evidence of unsuccessful attempts to influence the AAST.

6.3. What was DOT&E's impact on the B-LRIP milestone? In their B-LRIP report and corresponding DSARC memo, DOT&E stated that as tested in OT II, AHIP demonstrated an operationally effective capability in only 1 of the 3 roles planned for it (FAAO). Based primarily on DOT&E's assessment, the decision was made to procure AHIP only for the FAAO. This decision had meaningful consequences; it meant that only 179 AHIPs would be procured rather than the 578 the Army had requested.

The DOT&E conclusion needs to be placed in context. The DOD Inspector General's (IG) office had done its own audit of the AHIP program prior to the 1985 Defense Systems Acquisition Review Council (DSARC) meeting. The IG concluded that the operational test data was only sufficient to show that AHIP exceeded the technical threshold ranges required for target recognition and laser designation; it did not show that AHIP could increase the kill rate of attack helicopters or provide greater survivability than existing helicopters. The IG advised giving only a conditional approval of a limited production quantity until the Army provided more conclusive test data. Officials from PA&E and USDRE, who also participated in the AHIP DSARC, both told us that they would have objected had DOT&E concluded that AHIP had been proven ready for production in the other 2 roles. In sum, we determined that at least 3 other OSD offices were delivering essentially the same message as DOT&E. The unique impact of DOT&E is therefore unclear.

7. DOT&E Reporting

7.1. What statements did DOT&E make to the Congress regarding the adequacy of OT&E and system effectiveness and suitability? DOT&E issued a B-LRIP report following OT II, stating that 1) the OT was adequate to assess the operational effectiveness and suitability of the aircraft in performing the 3 mission roles planned for it, and 2) as tested, AHIP demonstrated an operationally effective capability in the FAAO role. DOT&E has not issued a B-LRIP report following the AAST because additional production funds have not been requested for the air cavalry or attack roles. The DOT&E action officer told us that if they were asked to issue another B-LRIP report, they would certify that: 1) AAST was adequate for testing the air cavalry role, and 2) they would certify AHIP for that role. They would not certify it for the attack role, which has never been demonstrated.

7.2. What was the completeness and accuracy of DOT&E's statements regarding adequacy of OT&E? The B-LRIP cited the following test limitations in OT II: 1) terrain was used repeatedly due to fixed instrumentation and physical size of the test site; 2) safety considerations precluded full mission oriented protective posture operations (i.e., chemical warfare defense); 3) threat simulators were unable to fully replicate all aspects of the postulated threat; and 4) the quantity of AHIP and AH-64 aircraft limited the scope and flexibility of the test. These were the same 4 test limitations described in the IER as acknowledged in the DOT&E-approved test design plan. No attempt was made in the B-LRIP report to explain these limitations or their implications for the test results. For example, the severity of the third limitation was not apparent (see 3.10).

The B-LRIP report did not include the additional limitations described in the IER as disclosed by the test. These were: 1) instrumentation/simulation software precluded valid force-on-force data concerning FAAO performance; 2) ELOSS captured only 60-80 percent of exposure data; 3) there were several software changes to test and support system test program set; and 4) the fault detection/location system was in an immature state of development. The B-LRIP report also did not include a further set of limitations described in the original test report but not in the IER, e.g., that the accelerated trial schedule hampered attempts to establish an on-going assessment of the data to determine if sufficient valid data were being collected to adequately answer each test issue. Finally, it did not mention that OTEA had lost control of the test (2.2), or that the quality of the RAM data was poor (4.2).

The B-LRIP report also included statements on test adequacy that were inaccurate. First, it stated that test sample sizes and numbers of aircraft were small but adequate. In fact, they were small and inadequate. As noted in 4.4, numerous comparisons between the AHIP and OH-58C—a principal objective of the test—simply could not be made due to small sample size. Second, it stated that the threat force employed simulated chemical agents. In fact, the test included simulated chemical warfare, but this was limited to the performance of only parts of missions in protective clothing, and was conducted on the ground.

Apart from individual statements that were incomplete and inaccurate, DOT&E's overall statement that the operational test reviewed (AHIP OT II) was adequate to assess the operational effectiveness and suitability of AHIP in performing its 3 roles was not, in our view, supported by the evidence.

7.3. What was the completeness and accuracy of DOT&E's statements regarding system effectiveness and suitability? As noted in 7.1, the B-LRIP report stated that AHIP had demonstrated operational effectiveness in the FAAO role. In this role, AHIP is required to acquire, locate, and designate targets. As previously discussed in 4.2, 4.3, and 5.1, the OT II test results did not demonstrate that AHIP could meet its acquisition and location requirements (location accuracy is particularly critical in the FAAO role). Moreover, much of the data that bear on these issues are highly questionable (see 2.2, 4.2). In sum, AHIP did not attain the performance thresholds required for the FAAO mission in OT II. Therefore, DOT&E's statement that AHIP had demonstrated operational effectiveness in the FAAO role was not supported by the evidence.

Aquila

System Description

The Aquila Remotely Piloted Vehicle (RPV) system is a target acquisition and reconnaissance system operated by a 94-man battery. The battery's equipment consists of 10 unmanned air vehicles (AV), 2 hydraulic launch vehicles, 2 net recovery vehicles, 5 ground controls stations (GCS) with remote ground terminals, 2 AV handlers and mobile maintenance support facilities, all of which are mounted on 5-ton trucks (no air strip is required to launch or recover the AV). The AV is flown by computer on a preprogrammed course that can be modified in flight. The sensor video and AV telemetry are sent to the GCS for real-time transmission to the

supported headquarters. The current system has daylight-only capability; a follow-on FLIR sensor was planned to provide 24-hour near all-weather capability. The system is mobile and designed for up to 3 AVs to be operating simultaneously on 3-hour missions, with ranges of up to 45 km.

Aquila was designed to provide real-time battlefield information to the ground commander by detecting, recognizing, identifying, and locating stationary and moving enemy forces that are beyond the line of sight of ground-based target acquisition and sensor systems. It is also intended to adjust artillery fire and laser-designate for precision-guided munitions such as the Copperhead. The Aquila battery is organized with a battery headquarters section, 5 operational sections—2 central launch and recovery sections and 3 forward control sections—1 ground support maintenance section, and 1 AV maintenance section. The operational sections are manned and equipped to provide AV mission support to Army divisions.

Program Status

At the time of the operational test, 9 batteries were planned for fielding in the 1990s. After the test, however, the Army postponed a production decision review pending the completion of additional testing to develop better aerial reconnaissance techniques. With the program coming under increasing congressional criticism, DOD has proposed to terminate Aquila in its FY89 budget request.

OT&E History

There has been only one operational test of Aquila: OT II. OT II was conducted at Ft. Hood, TX, from November 1986 through March 1987. There were 3 primary objectives: determine whether Aquila could 1) successfully conduct flight operations in an operational environment, including launch on command, flight, and recovery, 2) detect, recognize, and locate tactical target arrays, and 3) adjust conventional artillery fire and laser-designate for the Copperhead round. All 3 objectives were termed critical issues. Other OT&E issues included survivability, RAM, training, and human factors. These were addressed only to the extent that they affected the ability of Aquila to meet the criteria on the 3 critical issues.

Assessment of Evaluation Questions for Aquila OT&E

1. Planning

1.1. Did the TEMP include a complete statement of the system's requirements? In some cases, the test criteria differed from or could not be directly traced to the previously specified user requirement. For example, the requirement specified a 50 percent probability of detecting a single stationary or moving target. However, the TEMP based the criterion on detecting, recognizing, and locating a target array (defined as 3 or more targets) rather than a single target. It also lowered the required probability of detecting stationary targets (now target arrays) to 30 percent, and eliminated the need to identify the target. DOD officials said the user requirement specified only technical performance thresholds which were not intended as operational test criteria; however, the requirements document stated that it represented both technical and operational requirements.

1.2. Did the test plan address all system requirements and critical issues identified in the TEMP? We found no significant problems or limitations.

1.3. Was there a clear relationship in the test plan between required system characteristics/critical issues and test objectives/missions through operationally meaningful test-verifiable criteria? There were no established test criteria for suitability issues. These include communication, survivability, human factors, RAM, mobility, logistics, training, and safety. Instead, these factors were treated as "associated data requirements" to be assessed in conjunction with their contribution to the system's operational mission performance. OTEA officials told us that Army regulations require an assessment of all suitability issues, but only require specific criteria for those which are designated critical issues. Aquila had started out with 17 critical issues (including the suitability issues), but these were reduced to 3 by the time the revised TEMP was approved. Therefore, only flight operations, target acquisition, and fire support for artillery were required to have criteria. As one OTEA official put it, "The more hurdles, the more problems ... This way you don't hang them up on a banner."

Where there were criteria, their operational meaningfulness was not always clear. There were quantitative criteria to evaluate launch, flight, target detection, directing artillery fire to targets, and recovery of the AV. However, no analysis was made to determine whether meeting these criteria, either individually or cumulatively, would produce an operationally effective system. For example, an 80 percent probability of successful launch (the launch reliability criterion), combined with a 30

percent probability of detecting, recognizing, and locating a stationary target array given a launch (the stationary target criterion), combined with a 50 percent probability of directing Copperhead hits on a stationary tank-sized target given acquisition (the Copperhead designation criterion), suggests a low probability of overall operational success in the Copperhead designation mission when these probabilities are multiplied together. Admittedly, performing multiple missions during a single flight could raise the probability of having at least one successful mission, but none of this was laid out by the Army.

Finally, test criteria do not consider the limitations of other systems Aquila must operate with. In the conventional artillery mission, one component of the performance criterion is a mission response time of 5 minutes or less exclusive of the field artillery system processing time and the projectiles' time of flight. The Copperhead mission criterion also assumes a reliable Copperhead round. When a mission failure could be clearly attributed to a portion of the fire support system outside the Aquila subsystem, the trial was to be counted as an Aquila success. While the knowledge of Aquila's performance in isolation is necessary for purposes of diagnosis and attribution, the criteria would still need to include consideration of other systems to be operationally meaningful. (For example, in determining the likely success of a course of action, a commander must adjust for the combined limitations of each system contributing to the accomplishment of the mission.) This is consistent with DOD policy, which states that "thresholds ... must reflect the performance and limitations of other components that support that mission."

2. Execution

2.1. Was each system requirement and critical issue identified in the test plan tested for as planned? The test plan specified that electro-optical countermeasures (EOCMs)—specifically obscurants (smoke), video, and laser countermeasures—be employed during operational trials. EOCMs were tested, but with the exception of smoke, not as part of the operational test. Further, the EOCM results were not integrated with performance estimates from the operational test. Consequently, the OT&E results are only representative of Aquila performance in an EOCM-free environment (see 3.5).

The Aquila system threat assessment states that the Aquila ground components will likely come under chemical attack. Wearing chemical protective equipment has been shown to significantly degrade military

performance generally, and can be particularly problematic in a high workload system requiring fine motor skills (e.g., keying in data) such as Aquila. The test plan specified that the crew would 1) enter, verify, and/or correct some flight plans, 2) set up and break down the system, and 3) decontaminate major equipment items, all while wearing chemical protective clothing. None of this was done in the test.

2.2. Were there limitations in implementation that had not been anticipated in the test plan? OTEA did not have exclusive control over Ft. Hood training areas during the test. Consequently, vehicles not under OTEA's control frequently crossed areas under surveillance by the Aquila, and were processed and reported as targets. While the acquisition of these uncontrolled target arrays was segregated in the data base, their presence nevertheless degraded the development of clear measures of system performance, particularly detection rates and time usage. The degradation in overall mission performance caused by time spent in processing these extraneous targets cannot be conclusively measured. We believe test planners should have anticipated this problem, and made necessary adjustments prior to the test.

3. Realism

3.1. Was the system operated by typical operational units? A full Aquila battery consists of 5 operational sections: 2 central launch and recovery systems and 3 forward control systems, each with its own GCS. Due to shortages of AVs and other subsystems, the operational test used only 1 of each type of system. This limited the evaluation of the battery's capability to command, control, maintain, and sustain geographically separate elements in accordance with operational goals. It also restricted the evaluation of the flexibility inherent in having redundant assets, of how the battery operations center would perform with the full battery present, and of how the failure rate would have changed. The last point is important because more units in operation would have led to more unit failures, increasing the maintenance burden.

3.2. Was the system operated by typical operational personnel? Two of the Aquila contractor's data collection technicians involved themselves in the conduct of the test on multiple occasions, despite warnings from OTEA officials. In one instance, the contractor technician entered the GCS and advised the crew on a problem they were having with launch aborts. The advice was not solicited, and the technician had no authorization to enter the GCS. In another instance, the technician provided unsolicited advice while a mission was underway. Evidence of these

actions, along with similar contractor involvement in maintenance (see 3.3 and 3.4), led the DOD IG to confirm that there had been illegal contractor involvement in the conduct of the test. (The Army has since taken corrective actions to prevent this from recurring in future OT&Es.)

3.3. Was the system supported by typical support units? The Army waived the requirement for military personnel to perform maintenance above the unit level on the grounds that sufficient numbers of soldiers could not be trained and retained, and test equipment to maintain the system had not been developed. Rather, the contractor performed all intermediate and depot level maintenance, as well as some of the unit level maintenance. Some of this maintenance was performed at the contractor's plant.

3.4. Was the system supported by typical support personnel? In addition to being organizationally atypical (see 3.3), contractor maintenance personnel are better trained and more experienced on the system than typical soldiers will be. Consequently, their performance does not reflect what can realistically be expected when soldiers assume the maintenance burden in the field.

3.5. Was equipment put under realistic stress by design? ADATS forces were only present for the first 2 weeks of record trials, before being removed to participate in another test. The Aquila operators were aware that ADATS had left, i.e., that the AV was no longer being tracked and engaged, and they altered flight parameters to improve Aquila detection capabilities. [material deleted]

EOCMs were not employed during operational trials as specified in the test plan (see 2.1). In this EOCM-free environment, the Aquila successfully met the laser designation criterion for Copperhead employment. However, results from a separate EOCM subtest indicate that the designator can be very successfully countered by both simple and sophisticated countermeasures within the current Soviet capability.

3.6. Were personnel put under realistic stress by design? Missions did not consistently approach the full 3-hour flight endurance specification. This is significant because the Aquila is very demanding on its operators; an Army Human Engineering Laboratory study reported that operations within the GCS impose workloads that may exceed human performance capabilities, and other observers have noted that GCS operations are highly fatiguing. One problem is that operators have to concentrate for extended periods of time on a black-and-white video image

(notably degraded when in anti-jam mode). In addition, the GCS was not configured to enhance operator effectiveness and efficiency; the operators require access to information and controls that are not located at their respective consoles or are not readily accessible consistent with their importance and frequency of use. In sum, the operational test did not demonstrate whether crews could effectively continue operations for the required length of time. Additional factors which unrealistically reduced crew stress were: 1) ADATS leaving the test after 2 weeks (see 3.5); and 2) not having to perform missions in chemical protective clothing as originally planned (see 2.1).

3.7. Were realistic combat tactics employed? As described in 3.5, the search rates and altitudes employed by the Aquila operators after ADATS left the test were not the same as those they would be forced to employ were an air defense threat present. Given the deployment concept envisioned for Aquila, however, it will face a dense array of numerous air defense threats. The absence of a full Aquila battery also limited the range of tactics that could be employed (see 3.1). Therefore, the tactics employed did not fully represent those that would need to be employed in an operational environment.

3.8. Did the physical environment approximate the intended range of environments? Although the Aquila is expected to be employed in Europe, where the terrain is typically hilly and heavily forested, the operational test was limited to the rolling, sparsely forested terrain found at Ft. Hood. OTEA considered the impact of this limitation to be negligible, but the absence of trees and other vegetation would have likely affected the test results since the Aquila cannot perform tracking and laser designation when a target moves behind trees or other heavy vegetation.

Due to installation boundaries at Ft. Hood, the maximum range of operation was less than it would be in an operational environment. Test officials stated that operating at shorter distances made it easier for crews to maintain electronic line of sight to the AV, and allowed more time for reconnaissance and target acquisition because the AV spent less time traveling to the mission area. It also forced the use of some artificial flight patterns.

3.9. Did target systems approximate actual targets, realistically employed? We found no significant problems or limitations.

3.10. Did threat systems approximate actual threat, realistically employed? The Aquila system threat assessment states [material deleted] OTEA officials acknowledged that the Soviets can [material deleted] Despite this assessment, there was no attempt to include threats to the GCS during the test. According to OTEA officials, they were told that nothing was available to replicate the threat.

The air defense threat was portrayed with considerable rigor while ADATS was present (as described in 3.5, ADATS was removed after the first 2 weeks of record trials), but there still were important realism limitations, and the biases ran in both directions. Factors favoring ADATS were: 1) no threat was portrayed against the ADATS units; 2) ADATS resupply was assumed; and 3) ADATS was cued to Aquila launch times. Factors favoring Aquila were: 1) ZSU-X air defense gun simulators were not available, only the less capable ZSU-23/4 (this is significant because ZSU-X will have replaced many of the ZSU-23-4s in the forward area by Aquila's projected fielding date); 2) ADATS did not have the vehicle density of the motorized rifle regiment they were supposed to represent; 3) ADATS crews were changed without coordination, reportedly degrading performance.

3.11. Was the tested system production representative and prepared for test in a realistic manner? We found no significant problems or limitations.

4. Analysis

4.1. Were measures quantitative and non-subjective? As described in 1.3, suitability issues were treated as "associated data requirements," to be assessed in conjunction with their contribution to the system's operational mission performance. While several had quantitative components—e.g., reliability included mean time between operational mission failures—others did not. For example, the data requirements for survivability included comments by threat personnel on their ability to detect the AV by various means, but no hard data on detection rate or ranges. Regardless of whether it included quantitative components, each issue was ultimately evaluated subjectively. This led to numerous overall conclusions which were more favorable than the evidence warranted. (See 5.1 for specific examples.)

4.2. Were quantitative measures reliable and valid? The maintenance concept waiver (see 3.3) affected the validity of the maintenance measures. There are several reasons why the maintenance times were probably biased downward: 1) the collection of maintenance times was incomplete, and no OT data were collected when an item was taken to the contractor's plant; 2) the contractor performed some unit level tasks which did not show up in the unit level estimates; and 3) the contractor's data collection technicians on occasion acted as maintenance personnel, rather than waiting for the authorized maintenance team (thus underestimating statistics such as mean-time-to-repair). The size of the bias is unknown, but it could have a significant impact on the maintenance burden which will be associated with the system when fielded. Note that these biases are over and above the bias due to performance of all intermediate and depot level maintenance by contractors (see 3.3 and 3.4).

Contractor representatives participating in reliability scoring conferences exceeded their roles as technical advisors. According to the test policy, the system contractor representatives' participation was to be limited to answering questions directed to them by the members. This rule was repeatedly violated, however. If the scoring conference members disagreed with the contractor representative, he would debate and argue the issue. One incident was reportedly argued for approximately 3 hours. As a result, several incidents were rescored in a manner more favorable to the system, e.g., downgrading an operational mission failure, which reduces system reliability, to an essential maintenance action, which does not. Evidence of these actions led the DOD IG to confirm that there had been illegal contractor involvement in the scoring of the Aquila operational test.

4.3. Were analytic assumptions explicit and appropriate? Analyses conducted after the data were collected showed that the accuracy criterion for conventional artillery adjustment - mean point of impact within [material deleted] of the time. The revised criterion was reportedly based on the operational standard for field artillery accuracy which uses a complex computational procedure involving the number of firing tubes, type of artillery weapon, and other factors. While the logic behind the new criterion may have been sound, it is not clear why the problem with the original criterion was not apparent during planning. The revised criterion was met [material deleted] whereas the original criterion would not have been.

As described in 1.3, the criteria did not consider the limitations of other systems Aquila must operate with; consequently neither did the analyses. For example, artillery adjustment trials in which failure was attributable to a non-Aquila system were scored as no-tests. Over half of such no-tests represented operationally realistic problems in interfacing with other systems, and would have resulted in mission failure. Given that Aquila's conventional artillery criterion was barely met (see above paragraph), the inclusion of even a small proportion of these no-tests would have changed the outcome.

4.4. Was sample size adequate to support statistically valid results? We found no significant problems or limitations.

4.5. Were comparisons with other systems valid? There were no comparisons with other systems.

5. Service Test Agency Reporting

5.1. Were findings, conclusions, recommendations consistent with the evidence and appropriately qualified? The IER's ratings of operational suitability issues were not consistent with the evidence. For example, RAM was rated "overall satisfactory." Yet, the system failed to meet its total system operational reliability requirements, largely due to problems with the launch phase (less than half of all attempted launches were successful). There also were repeated incidents in which the AV was difficult or impossible to fuel, despite repeated filter changes and fuel unit repairs and replacements during the test. Availability estimates were based on unrealistically favorable maintenance statistics that nonetheless failed to meet requirements. The IER did in fact conclude that there were several significant problems identified in the RAM area, and recommended that some of these problems be corrected prior to fielding the system. This makes the "overall satisfactory" rating all the more inconsistent. Human factors was also rated "overall satisfactory" and concluded to be "generally adequate." Yet high noise levels during launch operations interfered with the ability of crewmen to communicate between the GCS and the launcher, causing errors resulting in mission aborts and lengthy delays. Initialization of the data link during darkness was observed to be difficult, which adversely impacts the input and verification of necessary data, and the fuel service unit is very cumbersome to use and inefficient (requiring approximately 8 manual crankturns per pound of fuel), meaning that crews may not be able to get AVs ready for launch when needed. Finally, safety was also rated "overall satisfactory" and concluded to be "generally adequate."

Yet, safety hazards with the potential for seriously impairing the continuous operations of both the launch and recovery systems were reported.

OTEA did a detailed analysis of detection performance, in which they calculated 3 separate indicators of detection rate. All 3 use the same number of target arrays detected for the numerator. The difference is in the population of target arrays available for detection used as the denominator. The first and most conservative rate is "in mission area," which includes target arrays in all search areas assigned by the mission commander whether or not the AV entered the search area. The second is "in mission area searched," which includes only target arrays in search areas entered by the AV. The third is "field of view," which includes only target arrays that had actually been within the Aquila's field of view during the trial. For moving targets, the percentages from the 3 indicators were 23 percent, 31 percent, and 42 percent, respectively. In forming their overall conclusion on mission performance, OTEA used the "field of view" percentage and further modified it on the basis of target geometry; they concluded that the Aquila can adequately detect, recognize, and locate targets which appear in the operator's field of view when the AV search geometry is correct. Selecting this choice to describe operational mission performance is highly questionable.

5.2. Was reporting clear and comprehensive? We found no significant problems or limitations.

6. Evidence of DOT&E Impact

Our ability to assess DOT&E's impact on the Aquila OT was again—as with AHIP—limited because 1) much of the communication between DOT&E and the Army was informal and undocumented, and 2) we may not have been provided all the documentation that exists. The DOT&E action officer defended informality, saying it is important to keep DOT&E's influence low-level and invisible to be effective.

6.1. Were there successful attempts to influence the OT&E process? DOT&E reviewed a draft test plan and provided written comments to the Army. The following changes were requested by DOT&E and at least partially supplied in the final plan: 1) provide a statement of scope for each of the 3 performance issues; 2) update the plan to accommodate events occurring since its initial preparation; and 3) improve the organizational consistency.

6.2. Were there unsuccessful attempts to influence the OT&E process? The following changes to the draft test plan were requested by DOT&E but not supplied in the final plan: 1) provide suitability issues and their criteria; 2) define the planned reconnaissance area to be searched during a mission; 3) reorganize the test plan around the 6 test objectives presented in the TEMP, rather than having 2 levels of data requirements ("primary" and "associated"); and 4) clarify relationships between measures, data elements, and evaluation processes. The latter included concern by DOT&E over the absence of any criteria for vulnerability/survivability, and the absence of intent to record quantitative data on detection and acquisition of the Aquila. None of these changes was made. Despite this, the revised test plan was approved by DOT&E.

DOT&E's consultant also reviewed a November 1985 draft TEMP, and DOT&E forwarded the consultant's written comments to the Army. The TEMP was later revised, but none of the suggestions had been incorporated. Finally, at a test concept briefing, DOT&E personally raised the concern that a 50 percent probability of detection, identification, and location of a target array coupled with only a 50 percent chance of successful engagement with a Copperhead round leads to a low probability of destroying tank-size targets in the threat array. After a discussion, according to an Army memo, DOT&E accepted the fact that although a low overall probability of destroying tanks existed, the two performance standards appeared realistic.

6.3. What was DOT&E's impact on the B-LRIP milestone? The program was proposed for termination prior to the Milestone III decision point. DOT&E was not consulted and had no impact in the decision to terminate Aquila; it was an Army budgeting decision.

7. DOT&E Reporting

7.1. What statements did DOT&E make to the Congress regarding the adequacy of OT&E and system effectiveness and suitability? DOT&E had been scheduled to write a B-LRIP report in FY87, but because no request was made to move forward with the Aquila program, no report was written. Internal DOT&E documents and discussions with the DOT&E action officer for Aquila indicate that the DOT&E action officer has concluded that 1) the Aquila OT II was a well designed and conducted operational test that supplied sufficient data to address all issues adequately; and 2) the Aquila has met the parameters against which it should be judged and has proven it can do its job. He also told us that were it up to him, he would "put it in the field tomorrow."

7.2. What was the completeness and accuracy of DOT&E's statements regarding adequacy of OT&E? We cannot fully evaluate this question because no B-LRIP report was written, and the overall adequacy of Aquila OT&E was not addressed in the FY87 annual report. The only statement to the Congress appeared in the May 1987 DOT&E monthly highlights report, which stated that the test supplied sufficient data to address all issues adequately. In light of the numerous problems and limitations cited in sections 1 through 4, we do not believe the statement is supported by the evidence.

7.3. What was the completeness and accuracy of DOT&E's statements regarding system effectiveness and suitability? We cannot evaluate this question because DOT&E has made no such statements.

Navy OT&E

The agency responsible for planning, conducting, and reporting Navy OT&E is the Operational Test and Evaluation Force (OPTEVFOR).

Tomahawk TLAM/C

System Description

Tomahawk cruise missiles are small, subsonic, low altitude guided missiles. The Navy has developed four Tomahawk variants, all with common airframes and engines but differing guidance or warheads. [material deleted]

TLAM/C navigates itself over a series of geographic waypoints. At each waypoint a [material deleted] measures ground contour elevations and the distance between them. The missile's Terrain Contour Matching (TERCOM) system correlates the detected terrain profile with a reference map stored in a computer to determine navigational accuracy. A Digital Scene Matching Area Correlator (DSMAC) further refines TLAM/C's navigational accuracy near the target. To do this, DSMAC uses an [material deleted] of ground scenes with prestored digitized scenes. [material deleted].

U.S. Navy Theater Mission Planning Centers store mission flight data and route plans (incorporating TERCOM maps, DSMAC scenes, terrain measurement data, and threat data). Mission plans are digitized and provided to TLAM/C launch platforms to enable it to navigate to the target. The maps, scenes, terrain data, and threat data employ various sensors and intelligence assets. The data are collected and prepared by various DOD components. Mission planning failures can result from inaccurate data collection for any of the elements of flight data and route plans or from terrain contour (for TERCOM) or scene contrast (for DSMAC) that is insufficiently unique or that changes sufficiently between data collection and missile flight due to a variety of factors.

Program Status

Between FY 1980 and FY 88, 608 TLAM/C were authorized for production; 1486 are planned by the end of FY 93.

OT&E History

TLAM/C operational flight testing started in 1981. Subsequent flight test failures resulted in suspension of TLAM/C testing. After modifications, operational flight testing resumed in January, 1985. This later TLAM/C

operational testing, which is the subject of our analysis, included a total of eight flight tests: four in 1985 and four in 1987. OPTEVFOR conducted its Operational Evaluation (OPEVAL) in January - March, 1985. It consisted of one OT/DT and three OT TLAM/C test flights. This testing provided the basis for a determination of operational effectiveness and suitability. After these 1985 tests, a permanent DOT&E director was sworn in; the Navy approved TLAM/C to progress from LRIP to "limited" production, and DOT&E wrote its B-LRIP report. To resolve various issues, FOT&E consisted of four test flights ending on September 24, 1987. On September 1, the Navy approved full TLAM/C production. Thus, the full production decision occurred before all FOT&E tests were completed and before OPTEVFOR could provide its final FOT&E report. Consequently, approval for the TLAM/C full production decision was granted provided there were no subsequent adverse recommendations from OPTEVFOR. On November 2, the commander of OPTEVFOR stated in writing that FOT&E testing did not yield any results that would change the full production decision for TLAM/C.

Assessment of Evaluation Questions for Tomahawk TLAM/C OT&E

1. Planning

1.1. Did the TEMP include a complete statement of the system's requirements? We found no significant problems or limitations.

1.2. Did the test plan address all system requirements and critical issues identified in the TEMP? The capability to conduct successful Tomahawk missions in [material deleted]. After OPEVAL, OPTEVFOR called for follow-on operational tests in these conditions before TLAM/C moved into full fleet introduction, but the operational test plans were not written for the tests to occur before full fleet introduction was approved.

Employment at [material deleted] which was not tested in OPEVAL due to test range limitations.

Although TLAM/C is described as a [material deleted] were conducted in OPEVAL or FOT&E. This capability was tested in 1) a developmental test, in 1981 when a TLAM/C was captive-carried to a test range and flown over a test course [material deleted] and 2) a contractor test in 1987 employing DSMAC equipment but not on a TLAM/C missile.

The ability of the missile to remain reliable over its storage requirement is an element of TLAM/C's overall RAM critical issue. [material deleted] The responsibility to collect storage reliability data has been given to the

program management office, which will collect data from depot and contractor facilities as missiles are returned from operational deployment for dismantling, checking and refurbishment. Consequently, FOT&E also did not address the storage reliability issue. Storage reliability data will include some degree of contractor self-reporting, unmonitored by OPTEVFOR. Data from this program will be made available in 1988. [material deleted]

1.3. Was there a clear relationship in the test plan between required system characteristics/critical issues and test objectives/missions through operationally meaningful test verifiable criteria? [material deleted] Thus, it is not clear how meaningful the test results will be to decision-makers.

2. Execution

2.1. Was each system requirement and critical issue identified in the test plan tested for as planned? [material deleted]

2.2. Were there limitations in the implementation that had not been anticipated in the test plan? [material deleted]

3. Realism

3.1. Was the system operated by typical operational units? We found no significant problems or limitations.

3.2. Was the system operated by typical operational personnel? The destroyer USS Merrill was employed for surface launched TLAM/C tests in OPEVAL. The same ship was employed in previous tests of other variants of Tomahawk, and the crew had an exceptional skill level in Tomahawk shipboard preparation and employment. Testers told us they have no choice in what ship is made available for testing, and that the Merrill was the only ship available. We found no similar problems in the submarine launched tests in OPEVAL or in the ship or submarine launch platforms in FOT&E. Nonetheless, it cannot be stated that the system was operated by personnel with a typical skill level for the ship launched portions of OPEVAL.

3.3. Was the system supported by typical support units? We found no significant problems or limitations.

3.4. Was the system supported by typical support personnel? We found no significant problems or limitations.

3.5. Was equipment put under realistic stress by design? [material deleted] Before TLAM/C was tested at these ranges, other variants of Tomahawk were flight tested there. This means that earlier tests of the nuclear variant of Tomahawk (which employs TERCOM but not DSMAC) provided TERCOM pre-testing for TLAM/C. Safety concerns also restricted the selection of DSMAC scenes in both OPEVAL and FOT&E. DSMAC scenes in OPEVAL were limited to a selection of six scenes that were often reused in DSMAC correlations. Thus, earlier TLAM/C test flights validated the effectiveness of DSMAC scenes for subsequent flights. Consequently, for the entire TERCOM portions and for parts of the DSMAC portions of TLAM/C flight tests in OPEVAL, the test occurred only with mission plans that were pre-tested and known to work effectively. This potentially biased performance upward.

Additional TERCOM and DSMAC pre-testing occurred in the form of overflight of the test range TERCOM routes by contractor personnel in a [material deleted] Initially, OPTEVFOR told us that DSMAC scenes in operational tests are not pre-tested with King Air. However, OPTEVFOR's FOT&E report and contractors made it clear that [material deleted] All TLAM/C route plans were even further pre-tested on computers for safety reasons by a contractor to insure that test missiles would maintain proper course and altitude when flown in tests. [material deleted] As noted above, operational tests took place using only pre-tested mission plans; the opportunity for navigational errors or failures during flight tests was substantially reduced and did not represent realistic combat conditions.

[material deleted]

In sum, OPEVAL and FOT&E TLAM/C flight tests were conducted only with TERCOM maps and DSMAC scenes known to work effectively. All this very probably biased test results upwards.

[material deleted]

3.6. Were personnel put under realistic stress by design? We found no significant problems or limitations.

3.7. Were realistic combat tactics employed? To ensure effective DSMAC operation, real-time information is important to determine weather over the target area. In OPEVAL this information was collected through a "special weather report", a telephone call to the test range to determine target area weather. This is an implausible method of data collection in

wartime. Thus, the possibility of test failure resulting from adverse weather was averted.

3.8. Did the physical environment approximate the intended range of environments? FAA safety regulations have prevented adverse weather and night operational tests. The FAA requires cruise missiles flying over FAA controlled airspace to be accompanied by two chase aircraft with the ability to visually observe the test missile and to control it, if necessary. Any adverse weather or darkness that prevents the chase aircraft from taking off before a test or from being able to observe the test missile during a test causes the cancellation of the Tomahawk flight test. [material deleted]

3.9. Did target systems approximate actual targets, realistically employed? We found no significant problems or limitations.

3.10. Did threat systems approximate actual threat, realistically employed?Although test plans called for survivability tests to be conducted, the tests as planned and as executed demonstrated sufficient limitations that OPTEVFOR termed Tomahawk survivability tests "inconclusive" just before OPEVAL and TLAM/C survivability as "unknown" after TLAM/C's full production decision in September, 1987.

[material deleted]

3.11. Was the tested system production-representative and prepared for test in a realistic manner? We found no significant problems or limitations.

4. Analysis

4.1. Were measures quantitative and non-subjective? We found no significant problems or limitations.

4.2. Were quantitative measures reliable and valid? [material deleted]

4.3. Were analytic assumptions explicit and appropriate? We found no significant problems or limitations.

4.4. Was sample size adequate or findings properly qualified/interpreted? OPTEVFOR reported the sample of four test flights in OPEVAL to be insufficient to provide meaningful statistical results. To rectify this, they called for sufficient test assets to determine meaningful results

before full production. However, FOT&E consisted of only four test flights, again an insufficient number to generate statistically valid results before the September 1987 full production decision. Referring to the lack of statistical validity, OPTEVFOR reported that the four FOT&E test flights provided insufficient data for comparison to three thresholds. In sum, the sample size remained inadequate. This is important because it means that, although the absence of statistical validity was appropriately qualified, the statistically significant test sample sought earlier was not generated in time for the full production decision.

4.5. Were comparisons with other systems valid? There were no comparisons with other systems.

5. Service Test Agency Reporting

5.1. Were findings, conclusions, recommendations consistent with the evidence and appropriately qualified? At the time of TLAM/C's full production decision in September 1987, OPTEVFOR was permitted time to provide only brief "quicklook" reports and a special five page interim assessment on most FOT&E testing. This reporting documentation gave no overall evaluation of TLAM/C performance in FOT&E. [material deleted]

5.2. Was reporting clear and comprehensive? Some significant information was not explicitly presented in OPTEVFOR's OPEVAL report. We learned of the King Air pre-testing of TERCOM routes and the detailed nature of various safety restrictions not from OPTEVFOR's OPEVAL report, but during interviews of OPTEVFOR personnel. We learned of [material deleted] from Defense Mapping Agency personnel, not OPTEVFOR. We found certain OPTEVFOR report passages cryptic and uninformative: a "special weather report" turned out to be a telephone call to the target area (see 3.7). Certain significant information, such as the use of only one TERCOM route being available at a test range, only became apparent after we compared data presented in report appendices. The FOT&E report did disclose [material deleted] or whether low altitude terrain following was employed in FOT&E. In sum, neither report fully addressed whether the level of performance achieved in operational tests could be expected in realistic conditions.

6. Evidence of DOT&E Impact

6.1. Were there successful attempts to influence the OT&E process? DOT&E initiatives included a successful effort to adopt common, inter-service

measures of effectiveness for cruise missiles used by more than one service. This initiative included the incorporation of height-of-burst miss distances in calculations of Tomahawk accuracy. However, because the initiative applied only to inter-service cruise missiles, it does not now apply to TLAM/C. Additional successful initiatives included increased funding requests initiated by DOT&E to the Congress for surrogate threat assets that would be useful in TLAM/C survivability testing, and for a special aircraft to reduce the number of test related aircraft now needed to accompany cruise missile flight tests to collect data and satisfy FAA safety regulations. Both of these initiatives have been funded by the Congress and are expected to be available after 1989. DOT&E action officers stated that there are many additional examples of DOT&E impact on TLAM/C OT&E: for example, an assessment of whether the contractors' developmental [material deleted] which should be available in 1988. They also stated that much of DOT&E's work is conducted informally—in person and over the telephone. This operating style has resulted in scarce documentation to demonstrate evidence of DOT&E impact on the OT&E process beyond that identified above.

6.2. Were there unsuccessful attempts to influence the OT&E process. We found no evidence of unsuccessful DOT&E efforts to influence the TLAM/C OT&E process.

6.3. What was DOT&E impact on the B-LRIP milestone? We found no evidence of any DOT&E impact on TLAM/C's production milestones, other than to in effect support the production decision to the secretary of defense and the Congress, nor any objection to the limited or full production decisions or the schedule allowing certain operational testing after production decisions were made. Those production decisions were Navy decisions that DOT&E did not directly participate in. Had DOT&E desired to affect the decision, it had the opportunity to inform the Congress of any dissent in DOT&E B-LRIP or annual reports.

7. DOT&E Reporting

7.1. What statements did DOT&E make to the Congress regarding adequacy of testing and system effectiveness and suitability? DOT&E's November 27, 1985, B-LRIP report to the Congress stated TLAM/C operational testing 1) was adequate to assess the operational effectiveness and suitability of the TLAM/C, and 2) demonstrated the system to be effective and suitable.

7.2. What was the completeness, clarity and candor of DOT&E's statements regarding adequacy of OT&E? DOT&E's B-LRIP report stated only those test limitations that were already in OPTEVFOR's OPEVAL report. [material deleted]

OPTEVFOR's recommendation for additional testing was reiterated by DOT&E's B-LRIP report. However, OPTEVFOR's recommendation that those tests be held, and the results verified, before full fleet introduction was not reiterated by DOT&E's B-LRIP report.

In sum, we believe that the operational testing held was not adequate to assess TLAM/C's overall effectiveness and suitability, and that DOT&E's favorable assessment of OT&E adequacy was not supported by the evidence.

7.3. What was the completeness, clarity and candor of DOT&E's statements regarding system effectiveness and suitability? DOT&E stated that four developmental flight tests, in combination with the OPEVAL results, provided results adequate to assess the expected combat performance of the production TLAM/C missile. We believe the addition of these developmental test results to TLAM/C's OPEVAL results in DOT&E's analysis neither provides a sufficient statistical sample to overcome the statistical validity problem in OPEVAL (as noted by OPTEVFOR), nor does it address [material deleted]

The B-LRIP report stated that thirteen separate missions incorporating 40 TERCOM maps and ten DSMAC scenes were planned for OPEVAL. The report did not point out that 1) the TERCOM route plans were repetitions of each other, 2) only one TERCOM route exists at the OPEVAL test range, and 3) of the ten DSMAC scenes only six were used and were used repeatedly.

DOT&E stated that OPEVAL route plans were "highly satisfactory products" for OPEVAL but that two deficiencies in the mission planning system (manning and source imagery quantity, quality and timeliness problems) were experienced. However, the details and significance of these problems, including OPTEVFOR's characterization of them as urgent, was omitted (see 5.1).

[material deleted]

In sum, we believe that DOT&E's B-LRIP report statements could lead the reader to overly favorable conclusions about TLAM/C's performance in

those tests, and did not find DOT&E's overall assessment of system performance to be supported by the evidence.

DDG-51 Aegis AAW System

System Description

DDG-51 destroyers will replace obsolete guided missile destroyers in the 1990s. As with previous surface combatants, DDG-51 will be equipped with missile launching systems for a mix of Tomahawk, Standard surface-to-air, Harpoon anti-ship, and ASROC anti-submarine missiles; a 5" rapid fire gun, a Phalanx close-in-weapon system, torpedoes, sonar systems, a multi-purpose helicopter pad, sensors, and command and control systems. The DDG-51 is designed to operate in high and medium threat areas.

A prime DDG-51 feature is the Aegis Anti-Air Warfare (AAW) System with a SPY-1D phased array radar and three Mk 99 guided missile illuminating radars. [material deleted] The SPY-1 radar comes in three variants. DDG-51's SPY 1-D radar is a product improvement of the SPY 1-A currently deployed on CG-47 cruisers and a derivative of the SPY 1-B under development for more recent CG-47 cruisers.

Program Status

The first DDG-51 began construction in 1985. Twenty-nine ships are planned. For FY 1988, Congress disapproved funding for three DDG-51s for budget and scheduling reasons but did approve initial funding for future DDG-51s.

OT&E History

Aegis SPY-1D AAW operational testing was conducted in 1986 at a land based facility in Moorestown, New Jersey, where operational test realism is limited, for certain objectives, by the on-land location, the absence of actual missile firing capability, and other equipment and operating constraints. These constraints are discussed in detail below (see 3.5 and 3.11). Operational testing at sea of the SPY 1-D AAW with Standard missiles and unmanned target drones will occur in 1990 after DDG-51 is deployed.

Because of similarities in the SPY 1-A, 1-B, and 1-D systems, the Aegis AAW OT&E is interrelated with SPY 1-A and 1-B operational testing for CG-47 cruisers. DOT&E and other DOD components assessed that DDG-51 AAW performance had been demonstrated in prior operational tests of SPY 1-A at sea with live missiles and unmanned targets. Such tests took place in 1983 and 1984—before a permanent DOT&E director was sworn in—and again in 1986. These operational tests were of Aegis CG-47 cruisers and of the SM-2 Block II Standard missile on Aegis cruisers. For successful engagement of targets, the performance levels demonstrated in these operational tests were generally consistent, except for one set of tests in April 1984 when the performance was [material deleted]

In addition, SPY 1-A and 1-B operational testing has taken place at the land based facility at Moorestown. This testing started in 1979 for SPY 1-A and in 1986 for SPY 1-B. The level of success for engagement of targets in these Moorestown tests, which was done by simulation only, was also consistent; however, meaningful differences between the results of this simulation testing and operational testing at sea against aerial targets did occur. We evaluated each of these operational tests employing the Aegis AAW, and their mutual relationship; all of these tests were relevant to DDG-51 procurement.

Assessment of Evaluation Questions for DDG-51 Aegis AAW OT&E

1. Planning

1.1. Did the TEMP include a complete statement of the system's requirements? We found no significant problems or limitations.

1.2. Did the test plan address all system requirements and critical issues identified in the TEMP? DDG-51 TEMPs stated as a critical operational issue that DDG-51 should fully support simultaneous action against air, submarine and surface threats in the 1990's during operations with and without support from friendly aircraft. Action against air threats includes the search, detection, tracking and engagement (i.e. successful intercept with missiles) functions. However, no capability was developed to test the DDG-51 SPY-1D AAW system with actual missile firings against aerial targets until the first DDG-51 goes to sea in 1990. SPY 1-D search, detection, and tracking tests were held at Moorestown (see 3.5 and 3.11), but computer simulations of SPY 1-D target engagements are not scheduled for Moorestown until the future. Computer simulations of target engagements were held with the SPY 1-B radar at Moorestown in 1986. SPY 1-D engagement operational tests at Moorestown and the tests at sea will occur well after DDG-51's B-LRIP milestone.

In 1983, the former director of Defense Test and Evaluation (a predecessor office to DOT&E), who was a Navy Admiral, criticized this OT&E approach. Shortly before he left the office, he rejected a draft DDG-51 TEMP, and termed the Moorestown facility not sufficiently realistic for tests prior to a production milestone. He called for the Moorestown facility to be augmented or replaced to permit SPY 1-D tests to include live missile engagements against threat representative aerial targets before a major production milestone. We found no such features in any test plan.

1.3. Was there a clear relationship in the test plan between required system characteristics/critical issues and test objectives/missions through operationally meaningful test-verifiable criteria? [material deleted] TEMPs for DDG-51 stated only "qualitative assessment" for AAW effectiveness thresholds. OPTEVFOR contended that there were not enough missiles and targets available to be used in all the scenarios Aegis is likely to encounter in combat to yield statistically valid results; therefore, no point was seen in establishing quantitative thresholds. While this argument explains why results may not be statistically valid, it does not explain why thresholds indicating desired levels of effectiveness should not be set. As a result, testers and decision makers were provided no specific criteria for evaluating Aegis' engagement performance in an environment where the system will be realistically stressed.

2. Execution

2.1. Was each system requirement and critical issue identified in the test plan tested for as planned? We found no significant problems or limitations.

2.2. Were there limitations that had not been anticipated in the test plan? [material deleted]

3. Realism

3.1. Was the system operated by typical operational units? We found no significant problems or limitations.

3.2. Was the system operated by typical operational personnel? We found no significant problems or limitations.

3.3. Was the system supported by typical support units? Before Ticonderoga's April 1984 OT III C, flaws were found and repaired by a special Navy engineering unit. A component that had previously caused failures

in operational tests was repaired. During the middle of Ticonderoga's September 1983 OT III B, the radar was found to be operating improperly. The test was interrupted, and the ship returned to port for repairs provided by "outside assistance" and then resumed testing. Support services not representative of the ship's own crew were used to repair flaws that had degraded performance. Test officials stated that this in-port maintenance was in fact typical; however, in combat a ship may not have the opportunity to disengage at will to obtain outside assistance.

3.4. Was the system supported by typical support personnel? See 3.3.

3.5. Was equipment put under realistic stress by design? [material deleted]

In sum, the absence of stress biased the results in favor of Aegis and left actual performance in a more realistic and stressful environment unknown.

3.6. Were personnel put under realistic stress by design? Some forms of warning of attacks in tests have already been discussed (see 3.5). Other forms of warning also occurred. First, for safety in operational tests at sea, aircraft dropped chaff and left the immediate test area in advance of each ECM test event. This activity provides the crew with warning of the time and sector of tests events. Only in the 1983 Ticonderoga tests was a restriction applied to prevent the SPY radar being on while pre-target presentation activities (laying of chaff, launching of drones) were being conducted. Second, intelligence notices were provided to CG-47 crews before tests. Although no sample of these notices was provided for our review to determine their specificity, the provision of any intelligence before tests eliminated the possibility that specific test events would occur as a complete surprise to the crew. We found no examples of tests where intelligence was either unavailable or inaccurate (which would increase stress) as might be the case in wartime. Third, during tests, OPTEVFOR listed notices of test events that were available to the crew. These notices would have some information blacked out, but they would state the time of a test within six to eight hours and the assets being used for tests, such as types of drone launcher aircraft, from which the type of drone to be presented and, thus, target characteristics, could be deduced. Fourth, safety regulations and the geometry of the test ranges required ships to be located in the test range such that drones would come only from a predictable sector of the test range. All this enabled CG-47 crews to deduce the general direction, timing and type of the test threats.

OPTEVFOR personnel stated verbally that it took no "mental magician" for the crew in CG-47 tests to know the direction from which drones would come in tests, [material deleted] Target specific knowledge was available to the crew; stress from surprise was therefore reduced or eliminated.

OPTEVFOR stated that the restrictions imposed by the physical and safety constraints of test ranges could not be overcome because tests in the open sea could not be properly instrumented. However, a target drone control system does exist for tests far out at sea, and the use of live warheads, as done at test ranges, could sharply reduce instrumentation needs to determine engagement success. Such tests in the open ocean would remove the capacity for crew to determine the most likely sector of attack in tests from the geometry of test ranges.

OPTEVFOR personnel contended that the crew warning we found in tests was irrelevant because Aegis AAW can operate as a fully automatic system without human intervention. [material deleted]

3.7. Were realistic combat tactics employed? See sections 3.5 and 3.6.

3.8. Did the physical environment approximate intended range of environments? We found no significant problems or limitations.

3.9. Did target systems approximate actual targets, realistically employed? [material deleted]

3.10. Did threat systems approximate actual threat, realistically employed? For Aegis, the targets are the threats; see 3.9.

3.11. Was the tested system production-representative and prepared for test in a realistic manner? As constituted for land-based tests at Moorestown in June 1986, SPY-1D was not production-representative. Differences existed between the system as tested and a production version in the following categories: 1) one radar array, not four, was available; 2) the one array used was a pre-production model with a commercial power supply; 3) some components were not installed or were replaced with non-production surrogates; 4) computers and computer programs employed were pre-production, and 5) interfaces between the radar and computer control systems were incomplete. SPY 1-D tests of a more fully integrated, more production representative system are scheduled for November 1988. (Also, see 3.3.)

4. Analysis

4.1. Were measures quantitative and non-subjective? We found no significant problems or limitations.

4.2. Were quantitative measures reliable and valid? We found no significant problems or limitations.

4.3. Were analytic assumptions explicit and appropriate? Because DDG-51's SPY 1-D is a derivative of the SPY 1-A in CG-47 class cruisers, TEMPS, and other program documents stated that DDG-51 components experienced extensive operational testing in CG-47 operational tests at sea. DDG-51 documentation also stated that SPY 1-A test results at Moorestown were validated by subsequent SPY 1-A tests at sea; therefore, DDG-51 SPY 1-D test results at Moorestown are anticipated to be confirmed by tests yet to be held at sea. This linkage formed a basis for DDG-51's B-LRIP production decision. Although SPY 1-D tests at Moorestown in June 1986 were for detection and tracking only—not engagements—Moorestown simulation engagements were conducted with SPY 1-B, to which the DDG-51 SPY 1-D is most similar. [material deleted]

The linkage used in forecasting SPY 1-D operational test results at sea from SPY 1-B results at Moorestown would be supportable if SPY 1-A Moorestown results forecast SPY 1-A results at sea for the critical portion of the tests—the successful engagement of targets. We did not find this to be the case. Because OPTEVFOR reported no overall aggregation of Aegis' success rate in all operational tests at sea, we did our own analysis of Aegis operational tests at sea. We compared these results to the SPY 1-A Moorestown results, as reported by OPTEVFOR. The same tests analyzed in section 3.5 were used in our calculations. [material deleted]

4.4. Was sample size adequate or findings properly qualified/interpreted? SPY 1-A operational tests covered a range of tactical scenarios against several types of targets with and without ECM, but not enough operational tests were conducted in any single scenario to yield statistically valid results. OPTEVFOR adopted a policy of simply reporting the results of individual sets of tests without referring to the lack of statistical validity.

4.5. Were comparisons with other systems valid? There were no comparisons with other systems.

5. Service Test Agency Reporting

5.1. Were findings, conclusions, and recommendations consistent with the evidence and appropriately qualified? [material deleted]

5.2. Was reporting clear and comprehensive? OPTEVFOR's report on the April 1984 CG-47 operational tests at sea contained several statements that were insufficiently clear to be fully comprehensible. OPTEVFOR did state that the Ticonderoga's commander was alerted just before tests of the threat sector and geometry of target presentations, but the report did not address the rest of the crew's ability to deduce warning from various aspects of target presentations. [material deleted] The report's appendix revealed only the range and altitude of the target at detection and that it was continuously tracked, without making any reference to the existence of unique atmospheric conditions, i.e. ducting.

6. Evidence of DOT&E Impact

6.1. Were there successful attempts to influence the OT&E process? DOT&E has sought the incorporation of tests of DDG-51's 5" gun in future AAW operational tests and obtained funding from the Congress to support additional and improved targets that would be useful in Aegis test engagements at sea. DOT&E action officers stated that other instances of DOT&E impact on DDG-51 testing are too numerous to describe. Because of DOT&E's practice of working informally without producing documentation to support claims of impact, we can confirm only the two items described above.

6.2. Were there unsuccessful attempts to influence the OT&E process? We found no evidence of unsuccessful DOT&E attempts to influence the DDG-51 OT&E process.

6.3 What was the DOT&E impact on the B-LRIP milestone? The DDG-51 B-LRIP milestone was a Navy Program Decision Meeting in which DOT&E was not a direct participant. We found no evidence of DOT&E impact on the DDG-51 B-LRIP milestone, other than to in effect support the production decision to the secretary of defense and the Congress.

7. DOT&E Reporting

7.1 What statements did DOT&E make to the Congress regarding adequacy of OT&E and system effectiveness and suitability? DOT&E supported the assessment that the similarity between SPY 1-A on CG-47 and SPY 1-D on DDG-51 enabled CG-47 operational test results to support DDG-51 milestones. Accordingly, we reviewed DOT&E's statements for both DDG-

51 and CG-47. DOT&E reported to the Congress regarding DDG-51 operational testing in their September 30, 1986 B-LRIP Report. They reported on some CG-47 tests in a March 19, 1986 letter to Congressman Denny Smith.

DOT&E stated that Aegis testing was extensive on CG-47 and overall was adequate to determine DDG-51 AAW effectiveness and suitability and that DDG-51's combat systems were sufficiently effective and suitable to support DDG-51 procurement. Although CG-47 operational tests were conducted in 1983 and 1984 and evaluated by the former acting director of DOT&E, the permanent director stated those tests were adequate.

7.2. What was the completeness and accuracy of DOT&E statements regarding adequacy of OT&E? DOT&E stated that testing of the integration of all the various DDG-51 systems would not be possible until 1990 when the ship first goes to sea and that SPY 1-D testing at Moorestown was as realistic as safety and equipment restraints permitted. The long list of testing limitations existing at Moorestown, and set forth in OPTEVFOR reports, was not presented in the DOT&E B-LRIP report. In the March 19, 1986 letter to Congressman Smith, DOT&E also stated that CG-47 testing was as realistic as safety and equipment restraints permitted. He either did not identify or make explicit the following limitations, some of which were not disclosed by OPTEVFOR: [material deleted]

DOT&E's description of CG-47 operational tests also did not reveal other limitations that may have been unavoidable but that nonetheless may have affected test results. The description of the April 1984 Ticonderoga tests in the letter to Congressman Smith did not make explicit the existence of naturally occurring ducting conditions that assisted the tracking of two low altitude targets. [material deleted]

In sum, we found that DOT&E made numerous incomplete or inaccurate statements, and we found that the overall assessment of test adequacy was not supported by the evidence.

7.3. What was the completeness and accuracy of DOT&E's statements regarding system effectiveness and suitability? [material deleted]

The letter to Congressman Smith stated that SM-2 Block II testing on the Ticonderoga in September 1984 had target tracking results consistent with the highly successful April 1984 Ticonderoga Aegis operational tests. [material deleted]

The DDG-51 B-LRIP report stated that testing at Moorestown, which was highly successful, was extensive and adequate to demonstrate satisfactory performance. The DOT&E FY 1985 Annual Report stated Moorestown tests had an "excellent correlation" with Aegis operational tests at sea. We did not find these statements to be accurate for the engagement of targets (see 4.3).

In sum, DOT&E's statements regarding Aegis effectiveness present a picture of a high level of effectiveness, which we found to be unsupported by the evidence.

Air Force OT&E

The Air Force Operational Test and Evaluation Center (AFOTEC) plans and conducts Air Force operational testing. AFOTEC produces both test plans and test reports for Initial Operational Test and Evaluation (IOT&E), and for FOT&E.

IR Maverick

System Description

The Imaging Infrared (IR) Maverick is a rocket-propelled, air-to-surface guided missile that develops tracking signals that differentiate between the heat of a target and its surroundings. It can be used on a variety of aircraft and is intended to destroy small, hard, tactical targets, both fixed and moving, such as tanks, armored personnel carriers, or aircraft shelters. It is designed to be used during the day or night and in limited adverse weather conditions in interdiction and in close air support operations by the Tactical Air Force. The missile is used with navigational and targeting aids to find or acquire targets. The IR guided Maverick is a follow-on improvement to a television (TV) guided version currently in the Air Force's inventory.

Program Status

In September 1982, the IR Maverick was approved for low-rate production. In March 1986, the IR Maverick was approved for full rate production, with a total buy of 60,697 missiles. Recently, the intended procurement of the IR Maverick has been stretched out; for the present FY the Air Force intends to buy around 2000 missiles.

Operational Test and Evaluation History

AFOTEC completed IOT&E testing of the IR Maverick in 1982. The Office of the Secretary of Defense (OSD) criticized the validity of IOT&E results because of crew familiarity with the test range and lack of realism, and directed AFOTEC to conduct FOT&E(1); phase two of FOT&E was conducted later and was not part of this set of tests. FOT&E was conducted from May 1984 through December 1985 primarily in two areas, Volk Field and Eglin Air Force Base. The Volk Field tests specifically addressed the OSD concerns, such as crew familiarity and lack of realism, by insuring that the pilots did not know the test area and by using National Guard tank units trained in Soviet tactics to simulate the threat. Survivability was assessed by taking the flight profiles and by applying them to a theoretical model. The Volk Field tests assessed IR Maverick only in the interdiction role. The Eglin tests addressed only proposed engineering

changes, and have been characterized as essentially developmental tests. The full rate production decision followed FOT&E.

Assessment of Evaluation Questions for IR Maverick OT&E

1. Planning

1.1. Did the TEMP include a complete statement of the system's requirements? We cannot address this issue because requested TEMPs were not provided by DOD.

1.2. Did the test plan address all system requirements and critical issues identified in the TEMP? We cannot address this issue because requested TEMPs were not provided by DOD.

1.3. Was there a clear relationship in the test plan between required system characteristics/critical issues and test objectives/missions through operationally meaningful test-verifiable criteria? Neither absolute nor relative criteria were given for aircraft survivability, a critical issue in the delivery of IR Maverick.

2. Execution

2.1. Was each system requirement and critical issue identified in the test plan tested for as planned? We found no significant problems or limitations.

2.2. Were there limitations in implementation that had not been anticipated in the test plan? We found no significant problems or limitations.

3. Realism

3.1. Was the system operated by typical operational units? We found no significant problems or limitations.

3.2. Was the system operated by typical operational personnel? The use of highly trained pilots for FOT&E was a major concern of the DOD IG. The DOD IG stated that aircrews used for the operational testing were not representative of the aircrews in operational units. The aircrews used were highly experienced, particularly in working with IR Mavericks, senior captains and majors considered among the Air Force's most proficient aircrews. Such crews do not represent typical operational personnel and may have biased IR Maverick performance estimates upward.

3.3. Was the system supported by typical support units? This evaluation question does not apply to IR Maverick because the missile does not need support units or personnel in the field. Support units may load or unload the missile, but they do not perform maintenance per se.

3.4. Was the system supported by typical support personnel? See 3.3.

3.5. Was equipment put under realistic stress by design? Since IR Maverick is an infrared based system, it is vulnerable to intentional and unintentional countermeasures which diffuse the IR signature or present competing IR signatures, such as other thermal images present in the battlefield (burning jeeps and tanks, warm shell holes, sun-warmed roads, dust, etc.). Very few of these CMs were included in FOT&E. (See 3.8 for unintentional countermeasures.)

3.6. Were personnel put under realistic stress by design? Pilots were not stressed during the FOT&E test. First, they were not required to react to onboard sensors which warned them if they were being tracked, acquired, or locked-on by threat air defenses. If the pilots had been forced to react realistically, this might have substantially lowered IR Maverick's overall effectiveness because acquiring and locking-on to targets requires undivided pilot attention compressed into a very short time period; any evasive response by the pilot, such as jinking, will take the targets out of the missile's field of view and consume time. As a result, the pilot may not have enough time to reacquire the target, thereby lowering first pass success rates. Second, the pilots were audibly warned if they spent [material deleted] however, this warning was not consistently exercised, and the test did not require pilots to react to it. That is, they could continue to attempt target lock-ons after the warning. AFOTEC did not report how many lock-ons were achieved after this warning, but it did report that [material deleted] USDRE did not [material deleted] to be a realistic option because the aircraft would be too vulnerable. Similarly, the Dutch government concluded that this time over the terrain mask, coupled with a [material deleted] Third, the pilots received high quality intelligence briefings which informed them of the location (accurate to within a few kilometers) of a moving tank column 60 kilometers behind the front. An assumption of high quality intelligence may not always be warranted.

3.7. Were realistic combat tactics employed? As stated in 3.6, pilots did not have to jink or perform any sort of evasive action even when they were informed they had been acquired by threat air defenses. This is not

a typical tactic, as pilots normally jink and perform evasive actions when entering a high threat area.

3.8. Did the physical environment approximate the intended range of environments? Volk Field was picked as a test site on the grounds that the selected pilots were not familiar with it and it closely represented NATO's environment. Although the overall area replicates some geography that can be found in NATO, the target area is flat, clear, and generally quite distinct from the rest of the area. Furthermore, the initial point used during the test is a very prominent rock outcropping that is higher than the surrounding terrain. Its prominence makes it easier for pilots to find, but it could also be the site of a radar and surface-to-air missile site.

Clutter and unintentional countermeasures were not fully assessed during the Volk Field tests. For half the sorties, after the aircraft's second pass, burning oil drums were placed in the target area to simulate clutter. However, the amount of this clutter was limited and it does not replicate hot craters, empty shells, or other items which would often be found on a battlefield.

3.9. Did target systems approximate actual targets, and were they realistically employed? During the test, the armored vehicles moved both in column and on-line, that is, abreast. Testing against vehicles on-line was unrealistic because at 60 kilometers from the battle area, tanks do not go on-line. Air Force intelligence states that enemy forces go to on-line formation when they are about to attack.

The scenario AFOTEC used had a high proportion of armored vehicles; 14 out a total of 17 vehicles were armored. While this scenario may be realistic, Air Force intelligence officials have stated that other scenarios, which have more trucks than tanks, could equally be encountered. A fuller operational test would have included a diverse range of the scenarios which Air Force intelligence believes would normally occur.

The live fire and captive carry tests at Eglin used M-47 tanks. The thermal signature of these vehicles was estimated to be 200 degrees celsius hotter than operational M-60s. AFOTEC did not report how the thermal signature of the M-47 or of the M-60s compared to actual threat tanks, but threat tanks have diesel engines similar to the M-60s, so they could likewise be cooler than M-47s. If so, the hotter M-47 tanks provide stronger signatures for the IR sensors to detect, acquire, and lock-on.

3.10. Did threat systems approximate actual threat and were they realistically employed? The Air Force said that simulated threat systems could not be employed because mobile air defenses did not exist. In fact simulated mobile air defense simulators did exist and the Air Force was planning to use these systems in later exercises.

3.11. Was the tested system production-representative and prepared for test in a realistic manner? We found no significant problems or limitations.

4. Analysis

4.1. Were measures quantitative and non-subjective? We found no significant problems or limitations.

4.2. Were quantitative measures reliable and valid? We found no significant problems or limitations.

4.3. Were analytic assumptions explicit and appropriate? The FOT&E reported [material deleted] however, numerous other DOD agencies and GAO questioned this. GAO, PA&E, USDRE, and the office of the assistant secretary of the Air Force for acquisition noted that probability of mission success should be calculated, rather than hit probability, and that it should be based upon the product of seven individual steps. They are the probability of: 1) locating the target area, 2) acquiring the target array, 3) maintaining a lock-on, 4) being a vehicular target, 5) being an armored target, 6) launching and hitting the target, and 7) killing the target. [material deleted] However, the above analysis also makes several favorable assumptions which raise the calculated probability of mission success. [material deleted] As stated, this depends on very high quality real time intelligence on moving vehicles, which may not actually exist in a typical operational environment. Second, a disproportionately high number of vehicles used in the tests were armored (14 of 17), thereby increasing the probability of success in acquiring an armored target. Third, effects of unintentional and intentional countermeasures are not included. [material deleted]

4.4. Was the sample size adequate to support statistically valid results? Based on a comparison of percentages calculated from a model, AFOTEC concluded that aircraft using IR Maverick were more survivable than aircraft using other weapons such as TV Maverick during the day and unguided conventional bombs during the night. However, the sample size was too small for the difference in the night trials of 6 percent to be

statistically significant; that is, the results do not permit a conclusion that a true difference in capability exists at the 95 percent level of confidence. AFOTEC did not report the lack of statistical significance, and thereby gave the impression that aircraft using IR Maverick are more survivable at night when in fact this was undemonstrated, not only because the results were not significant but also and especially because, for 2 of the 5 threats, the comparison actually did not favor IR Maverick.

4.5. Were comparisons with other systems valid? OSD criticized the IOT&E tests because they failed to answer several questions, among them one on survivability. OSD had directed AFOTEC to address survivability as a critical operational issue. AFOTEC's objective was to evaluate the survivability of delivery aircraft during IR Maverick missions by comparing IR Maverick with TV Maverick in the day and overflight weapons at night. The survivability analysis that AFOTEC performed is questionable, not because it is model-based, but also because of the lack of absolute and relative criteria. Because there were no criteria on absolute survivability, it cannot be inferred whether IR Maverick is sufficiently survivable regardless of how it did compared to other weapon systems. Because there were no relative criteria versus the other systems, it cannot be inferred whether just doing "better" achieves a level of survivability which is operationally meaningful.

A comparative test such as this one is vulnerable to unmeasured interaction effects. For example, the report stated that the conditions of the comparison represented a worst case scenario. IR Maverick may show higher survivability than TV Maverick in a worst case scenario, while that difference could disappear in a typical scenario. Similarly, the AFOTEC test director acknowledged there may be unmeasured interaction effects between pilot skill and type of Maverick. The test did not examine how low or average skill pilots would perform on IR versus TV Maverick. An AFOTEC official stated that pilot skill did not affect the comparison because the pilots were not versed in IR imagery interpretation. However, the pilots were not required to interpret IR images in FOT&E, so other factors such as pilot skill may have been relatively more important.

The type of Maverick can also interact with how the targets are located. AFOTEC's analysis assumed that all air defenses would be co-located with the target, some 60 kilometers behind the front. Under this assumption, IR Maverick's increased standoff range directly increases survivability over TV Maverick. However, both IR and TV Mavericks are over the terrain mask and exposed to threat air defenses [material deleted]

5. Service Test Agency Reporting

5.1. Were findings, conclusions and recommendations consistent with the evidence and were they appropriately qualified? As stated in 4.3, AFOTEC reported high IR Maverick hit probabilities without acknowledging several favorable assumptions and without giving an overall mission success criterion. As stated in 3.7 and 4.5, [material deleted] but the results were not qualified by acknowledging either one of these issues.

5.2. Was reporting clear and comprehensive? The survivability section omits key data, such as lock-on time, sample size, etc., which are critical in order to fully assess the system. Furthermore, a model was used in the comparative survivability tests, but AFOTEC neither reported all its assumptions nor stated how the model's assumptions might have influenced the outcome. Without knowing all the model's assumptions, we cannot estimate the extent to which the survivability data may have been biased. Finally, AFOTEC did not break out the results from the two types of target areas which were presented, so we cannot tell if acquisition and lock-on ranges were higher against the online scenario.

6. Evidence of DOT&E Impact

6.1. Were there successful attempts to influence the OT&E process? Since the FOT&E trials were completed prior to the swearing in of a permanent DOT&E director, we will not discuss attempts by DOT&E to influence the conduct of that test. However, the B-LRIP report followed the swearing in, so we will discuss that (see 6.3).

6.2. Were there unsuccessful attempts to influence the OT&E process? See 6.1.

6.3. What was DOT&E's impact on the Beyond-LRIP milestone? In their B-LRIP report and corresponding DSARC memo, DOT&E stated that as tested in FOT&E, the IR Maverick was operationally effective. DOT&E's memo to the DSARC-IIIB meeting basically concurred with AFOTEC's position that aircraft delivering the IR Maverick were more survivable than aircraft delivering TV Maverick. DOT&E also agreed with AFOTEC's findings on IR Maverick's hit probabilities. However, several other OSD offices, such as the DOD IG, PA&E, and USDRE, raised very strong concerns at the DSARC-IIIB meeting regarding testing adequacy, test reporting, survivability, the overall realism of the operational test, and how mission success was defined. According to those present, DOT&E knew about these various concerns before the meeting and offered no objection to them either

before or during the meeting. We could find no evidence that DOT&E attempted to defend its position or respond to the concerns raised at this meeting.

7. DOT&E Reporting

7.1. What statements did DOT&E make to the Congress regarding the adequacy of OT&E and system effectiveness? In its B-LRIP report, DOT&E states that, first, the IR Maverick FOT&E(1) testing was adequate to provide the information necessary to reach a full-scale production decision concerning the IR Maverick, and second, the overall effectiveness of the IR Maverick system on the F-16 in the interdiction role was satisfactory (as noted earlier, this was the only role that was tested). DOT&E also stated overall suitability was marginal due to problems with the guided missile test set, a diagnostic tool for analyzing potential problems in the missiles.

7.2. What was the completeness and accuracy of DOT&E's statements regarding adequacy of OT&E? In its B-LRIP report, DOT&E contended, as had the Air Force, that absolute survivability could not be tested because mobile threat defense simulators did not exist. As already pointed out by the DOD IG, however, mobile threat defense simulators did exist and the Air Force had planned to use them in later exercises. The DOT&E officials told us that mobile air defenses did exist at the time, but that using them would have been too expensive, delayed the test, and in their view not added sufficient value to warrant the cost. He did not explain the reason for both AFOTEC's and DOT&E's contrary report statements.

The B-LRIP report also states that the limited survivability analysis was adequate to answer the survivability issue compared to other current generation weapons by virtue of the use of simulation. These limitations which DOT&E mentioned are the same limitations that AFOTEC reported. Several other DOD agencies, such as PA&E and the DOD IG, questioned the survivability analysis—specifically, the assumption that all air defenses would be co-located with the target and that pilots do not have to perform evasive action such as jinking in combat (see 6.3). [material deleted]

7.3. What was the completeness and accuracy of DOT&E's statements regarding system effectiveness and suitability? The B-LRIP report stated that only a limited survivability comparison of IR Maverick to TV Maverick and to overflight weapons could be performed and that these limitations denigrated the answers to the issue of IR Maverick's

survivability. The B-LRIP report did not report that in AFOTEC's modeled comparative survivability analysis, aircraft using IR Maverick appeared to do worse against some threats than aircraft using overflight weapons during the night, or that IR Maverick advantages at night were not statistically significant. Finally, the report did not point out that IR Maverick would be equally as vulnerable to ground-to-air threats as TV Maverick when those threats are within range (see 4.5). In sum, because survivability results were based on a theoretical model, relative rather than absolute, and because there are so many caveats to the relative results obtained, we found IR Maverick's survivability to be undemonstrated.

DOT&E concluded that [material deleted] however, numerous other DOD agencies questioned this assessment. They argued that overall mission success should be calculated, rather than hit probability. Given their calculations, [material deleted] (see 4.3). DOT&E's limited analysis produced a more favorable effectiveness result than did other, more realistic and comprehensive calculations.

DOT&E's statements regarding both survivability and mission effectiveness were incomplete. Based on the more realistic, comprehensive analyses described above, we conclude that DOT&E's assessment of overall effectiveness in the interdiction role was not supported by the evidence.

LANTIRN

System Description

LANTIRN is a two pod set that is to be placed under the F-16 or F-15E aircraft. The LANTIRN system consists of three main components. First, there is a wide field-of-view (WFOV) heads up display (HUD) in front of the pilot in the cockpit. It takes infrared imagery and displays it, to give the pilot a "night window" of what the terrain in front of the aircraft looks like. Second, there is a navigational pod which utilizes a forward-looking infrared receiver (FLIR) and a terrain avoidance radar. The navigational pod aids the pilot in navigating and avoiding terrain during the night, and can also be used for acquiring and attacking targets. Third, there is a targeting pod which also utilizes a FLIR, but the targeting pod FLIR is used for target acquisition and weapons delivery. The targeting pod has two fields-of-view which provide the pilot with magnified display images for standoff target acquisition and precise target aiming. When coupled with the targeting pod's laser and target autotracker, the

system will provide a laser designation capability for delivering precision laser guided bombs (LGB). The targeting pod contains a missile boresight correlator for automatic lock-on of IR Maverick missiles. In sum, the WFOV HUD displays night-time images to the pilot that the navigational pod "sees," and the targeting pod's main function is to help the pilot deliver LGBs and IR Mavericks at night.

Program Status

IOT&E ended in April, 1986. In November 1986, the navigation pod and the WFOV HUD received a Beyond-LRIP report from DOT&E; 143 navigation pods were approved for production in FY87 and 169 were approved in FY88.

The FOT&E testing which took place in FY87 was to support a B-LRIP decision for the targeting pod. The targeting pod did not receive a B-LRIP report, but its 81 pod production for FY88, originally characterized as full-scale production, remained unchanged. The Air Force plans to buy approximately 700 of each pod.

OT&E History

Dedicated IOT&E for the HUD/Navigation pod was performed from September 1984 to January 1985. Additional IOT&E testing on the HUD/Navigation pod occurred from April 1985 to September 1985. The focus of this test was navigation across rough terrain in a variety of operationally representative environments. This testing addressed deficiencies found in the previous IOT&E, specifically the HUD display unit. The targeting pod underwent IOT&E from January to April 1986. The targeting pod FOT&E was conducted from May to September 1987 for the purpose of addressing issues outstanding from IOT&E—most notably, LGB deliveries and RAM.

Assessment of Evaluation Questions for LANTIRN OT&E

1. Planning

1.1. Did the TEMP include a complete statement of the system's requirements? We found no significant problems or limitations.

1.2. Did the test plan address all system requirements and critical issues identified in the TEMP? The TEMP identifies four LANTIRN missions: 1) close air support (CAS), (2) battlefield air interdiction (BAI), 3) air interdiction, and 4) counter air. There have been no operational tests for the air interdiction, the counter air, or CAS roles. AFOTEC did test LANTIRN

in a BAI scenario. However, the BAI scenarios AFOTEC used for the targeting pod's FOT&E did not include moving targets, which Air Force intelligence has said could be encountered. Furthermore, the targeting pod tests primarily focused on delivering weapons, and not on how LANTIRN will perform with those weapons in a specific operational scenario.

1.3. Was there a clear relationship in the test plan between required system characteristics/critical issues and test objectives/missions through operationally meaningful test-verifiable criteria? The TEMP and the test plan do not always make the connection between the most critical issues and some operationally measurable criteria. For example, the TEMP states the LANTIRN system must provide an effective means of ingress and egress with acceptable attrition rates. In the FOT&E(1) test plan, a critical issue is stated as the capability of a LANTIRN-equipped F-16 to deliver LGBs. But, in both these examples, the required system characteristic is stated only generally; for example, there is no specification of what acceptable attrition rates or acceptable LGB delivery rates should be. Furthermore, these criteria are not related to a clear operational need or statement.

The criteria put forth in the TEMP and the test plan do not account for or reflect the effectiveness of the weapons that LANTIRN is designed to use, such as LGB and IR Mavericks (for implications see 4.3). Utilizing criteria just for LANTIRN is inconsistent with DOD policy, which states that, "thresholds...must reflect the performance and limitations of other components that support the mission."

2. Execution

2.1. Was each system requirement and critical issue identified in the test plan tested for as planned? We found no significant problems or limitations.

2.2. Were there limitations in implementation that had not been anticipated in the test plan? We found no significant problems or limitations.

3. Realism

3.1. Was the system operated by typical operational units? LANTIRN is intended to be operated by a pilot in a single-seat aircraft, but in the FOT&E targeting pod tests there was a second pilot in the back seat. DOT&E did not want a second pilot in the back seat during the test, on the grounds that even if the pilot did not say anything, it would affect how

the first pilot flew. The second pilot was required for safety reasons; nonetheless, the major issue of how well a single-seat pilot can perform the targeting pod missions remains undemonstrated.

3.2. Was the system operated by typical operational personnel? We found no significant problems or limitations.

3.3. Was the system supported by typical support units? For all IOT&E and the FOT&E tests, there was contractor maintenance and no integrated logistics support. Since contractors maintained LANTIRN, there are questions whether the reported findings can be generalized to typical operational personnel. The contractors who developed the system are usually more experienced than typical Air Force personnel. Although future tests assessing the maintenance capability of Air Force personnel are planned, these tests will occur after the B-LRIP decision.

3.4. Was the system supported by typical support personnel? See 3.3.

3.5. Was equipment put under realistic stress by design? The LANTIRN targeting pod was not stressed during the test in several areas. First, the effects of jamming against the targeting pod, (other than lasers), was not tested, [material deleted] Second, decoy targets were not used. Decoys are important because according to the LANTIRN threat assessment, the Soviets will use them to create false targets, thereby forcing LANTIRN to discriminate between valid and invalid targets. Third, obsolete, gasoline engined M-47 tanks were used for live launches in FOT&E; these tanks are hotter than Soviet tanks with diesel engines (see 3.9). Finally, the full range of LGB delivery angles was not tested. Laser tracking performance depends in part on the angle at which it lazes the target; at a certain point tracking performance is degraded because of the delivery angle. For missions, the Air Force plans on the optimal angle—45 degrees ingress and 45 degrees egress—however, this cannot always be achieved. Since AFOTEC examined only a portion of the possible attack angles, and those the most optimal ones, LANTIRN's performance at other angles is not known.

3.6. Were personnel put under realistic stress by design? Aircrews were not stressed in several areas. First, pilots were allowed a pre-flight test during the day over the same terrain they would navigate at night during the navigation pod testing. Therefore, the test did not assess how well single-seat pilots could navigate over unfamiliar terrain. Second, pilots were not required to react to any sensors which would inform them if they were being acquired, tracked, or locked-on by threat air

defenses. Forcing pilots to react not only replicates an operational environment, but it may also lower their ability to acquire, track, and lock-on with IR Maverick. Third, in the FOT&E targeting pod tests, there was a second pilot in the back seat (see 3.1). In sum, the major issue of how well a single-seat pilot can perform a high-stress mission in unfamiliar terrain remains to be fully demonstrated.

3.7. Were realistic combat tactics employed? (See 3.6 on pilots not being required to react.)

3.8. Did the physical environment approximate intended range of environments? The TEMP and test plan state that LANTIRN is intended for worldwide deployment. However, IOT&E and FOT&E targeting pod testing was only performed at Eglin and Nellis Air Force Bases. Although Eglin's humidity levels are as high or higher than NATO's (high humidity levels stress weapon delivery), its terrain is flat and sparsely forested, while NATO's is rolling and heavily forested. On the other hand, Nellis is a desert. These two areas do not represent the range of terrain/weather possibilities which LANTIRN will encounter, and neither is representative of NATO.

During the FOT&E targeting pod tests, LGBs were not tested with any type of clutter. In particular, the absence of other buildings in the target area is not realistic, because LGB targets do not typically stand alone. Other buildings could have interfered with LGB delivery and degraded the overall performance of the LANTIRN/LGB system.

3.9. Did target systems approximate actual targets and were they realistically employed? The targets used for the FOT&E targeting pod test were limited in realism. First, obsolete gasoline engined M-47 tanks were used for live launches in FOT&E. This is understandable for reasons of cost. However M-47 tanks with gasoline engines are hotter than Soviet tanks with diesel engines. The hotter vehicles have more vivid IR signatures, so it is easier for pilots to acquire and lock-on to them with the IR Maverick. (During LANTIRN's EOCM testing, the testers banded a column of diesel powered M-60s with M-47s, because pilots could not locate the M-60s.) Second, the tanks were online and not moving—which does not accurately represent a typical BAI threat target scenario. Moving targets are harder to lock-on to, and they create EOCM problems for the IR Maverick. Since the targets as presented favored target acquisition, the results may be biased upward.

3.10. Did threat systems approximate actual threats and were they realistically employed? We cannot address this issue because all the required data were not supplied.

3.11. Was the tested system production-representative and prepared for test in a realistic manner? In all the tests for both the navigation and targeting pods, production-representative equipment could not be used. For tracker performance, WFOV resolution, etc. (items whose production configuration has not been finalized), the pre-production model may perform differently than the production model would. So, LANTIRN's performance with a production-representative model is unknown.

4. Analysis

4.1. Were measures quantitative and non-subjective? LANTIRN overall capability to deliver LGBs was subjectively rated marginal. How this was determined is unclear. The seven measures of performance were not weighted in terms of importance. Only four of the measures had criteria, one of which could not be tested because of targeting pod problems. Each of the three measures with criteria that was tested was rated marginal, but two of these (auto-tracking and F-16/LANTIRN overall missions) are questionable. First, the F-16/LANTIRN overall success data is averaged across actual and simulated deliveries; based on actual deliveries alone, it would fall to unsatisfactory. Third, neither result is statistically higher than its marginal criterion, as the lower confidence bound for each dips well into the unsatisfactory range (see 4.4). Taking all this into account, an overall rating of unsatisfactory seems equally justifiable.

4.2. Were quantitative measures reliable and valid? AFOTEC did not report that navigation and targeting pod failure rates are derived from the same data pool. Both the navigation and targeting pod have similar items; when those items fail they go to a single maintenance facility which does not determine whether the piece of equipment comes from a navigation or a targeting pod. DOT&E officials have stated that it is impossible to identify the origins of some of the failure rates because of this problem. As a result, the RAM data for both the navigation pod and, especially, the targeting pod are of doubtful utility.

4.3. Were analytic assumptions explicit and appropriate? AFOTEC has not tested the whole system—i.e. a "soup-to-nuts" test--at any one time because all the tests were segmented to address specific issues. However, the assumption that the whole system can be evaluated through

segmented tests is questionable. First, from the Air Force analysis we cannot tell from the segmented testing what LANTIRN's success rates are from take-off, through navigation, to finding the target area, to hitting a valid target for an operational scenario. Second, the Air Force analysis does not consider any problems that could emanate from the weapons LANTIRN was designed to deliver; as a result, there are many trials where LANTIRN succeeded but the target was not destroyed. This means that a mission failure does not impede LANTIRN from being scored as successful. [material deleted] As is true for all systems, LANTIRN is constrained by its weakest link. AFOTEC officials acknowledged that there has not been a whole system assessment, but stated that such an assessment is not their responsibility; rather, they test the specific items. On the other hand, DOT&E officials acknowledged that DOT&E can assess the whole system, but in LANTIRN's case, such an assessment will occur only after full-scale production is completed.

Although AFOTEC does break out the data regarding all the simulated and live LGB launches for both Eglin and Nellis, the data are aggregated to report an average over both types of launches and over both ranges. In our view, the results are not meaningful because: 1) the success rates between Eglin and Edwards/Nellis varied greatly—for example, the difference between them on actual targets being hit is a 60 percentage point difference; 2) success rates also changed with type of mission—high loft deliveries scored much better than low-loft ones; and 3) simulated launches were 23 percentage points better than live launches.

4.4. Was sample size adequate to support statistically valid results? Overall LANTIRN/LGB success was rated marginal. In order to achieve a marginal rating, 50 percent of the auto-tracking and 60 percent of the F-16/LANTIRN overall missions had to be successful. (Note that these marginal requirements are substantially lower than the 85 percent and 75 percent rates needed for a satisfactory rating.) [material deleted]

4.5. Were comparisons with other systems valid? The survivability objective in IOT&E was to evaluate the effect of the LANTIRN system on aircraft survivability during under-the-weather operations and in battlefield conditions. The objective was addressed with a comparative test between LANTIRN-equipped and non-LANTIRN aircraft. AFOTEC reported that LANTIRN enhanced aircraft survivability relative to other aircraft, but did not report on how survivable LANTIRN is in absolute terms. In fact, there was no criterion for absolute survivability. This is important because AFOTEC notes reduced radar detection range and fewer valid launches against the LANTIRN-equipped aircraft, but does not report that

LANTIRN's radar detection range still gives ample time and warning to enemy air defenses, or that the number of launches against LANTIRN, while lower, is still high enough to threaten the aircraft. Thus, the comparative analysis may have obscured the fact that LANTIRN-equipped aircraft are still not survivable. It is questionable whether comparing LANTIRN-equipped aircraft to aircraft not equipped for night attack is meaningful. From the outset, a LANTIRN-equipped aircraft could not do any worse than the aircraft it was being compared to. An alternative would have been to compare LANTIRN-equipped aircraft to other planes used for night-time missions, such as the F-111.

In addition, a comparative test such as this one is vulnerable to unmeasured interaction effects. The comparison aircraft did not have night-time under-the-weather capability. Consequently, they had to fly at least 500 feet higher than the LANTIRN-equipped aircraft. On the flat desert terrain where the test was conducted, this permitted the higher flying comparison aircraft to be acquired and tracked from much greater distances. Were the test conducted in rough or mountainous terrain, the results would change in two ways. First, the LANTIRN-equipped aircraft would fly higher, thereby lessening the altitude difference between it and the non-LANTIRN aircraft (navigation pod tests revealed that in rough terrain pilots could not get as low to the ground as in the desert and they could not consistently maintain low altitudes because of the tendency to overfly ridge lines). This tendency will be lessened when the terrain allows the pilot to fly through gaps or canyons, but this will not always be the case. Second, terrain affects radar acquisition capabilities and thereby survivability rates. In the desert, the radar's line of sight is unobstructed, while in a rougher environment the line of sight is limited by the terrain. As a result, radars could not acquire non-LANTIRN aircraft as far away as they could in the desert; therefore, the difference in survivability rates between the two aircraft would change.

5. Service Test Agency Reporting

5.1. Were findings, conclusions, and recommendations consistent with the evidence and appropriately qualified? AFOTEC reported that the EOCM critical issue was satisfactorily met, despite a lack of evidence to support this claim. The targeting pod FOT&E report states no rating was provided for this objective because IOT&E testing gave LANTIRN a satisfactory evaluation for this objective. However, the IOT&E testing was described as "limited." The IOT&E testing primarily examined digital laser threats; simulation models were used for this testing. Further testing was turned

over to another agency, which has not yet published its final report on the LANTIRN targeting pod.

5.2. Was reporting clear and comprehensive? In many cases the reports do not provide all essential data nor inform the reader of all relevant assumptions. For example, the FOT&E report does not show how many no-tests there were or what constitutes a no-test. Based on interviews, there were around 47 no-tests and 63 valid tests. The FOT&E report does not make it clear what analysis it performed to reach its criteria. Furthermore, in the LGB analysis, the report does not include the original sample size. The LGB results are also unclear. In reporting the probability success of each of the 5 steps, AFOTEC does not specify which ones are conditional probabilities, i.e., dependent upon the success of the previous step. AFOTEC reports success rates in percent, without stating if that percent is derived from the previous step or if it is a completely separate analysis. Without the sample size, we cannot tell if the results are statistically significant, and without the criteria for recording a result as a no-test, we cannot determine if their exclusion from the test was warranted.

The FOT&E report does not point out that at least one of the test aircraft used was a two seater, often with a pilot in the back seat for safety reasons, nor do they report that in the navigational pod IOT&E, the pilot flew pre-tests over the area during the day (see 3.6). Furthermore, they did not make clear that it was the same pilot who flew both the pre-testing and the test. Finally, AFOTEC does not report that both the navigation and targeting pod are maintained by the same maintenance shop, and when fixing similar items that shop does not record whether a failure is a navigation or a targeting pod one (see 4.2).

6. Evidence of DOT&E Impact

Our ability to assess DOT&E's impact on LANTIRN testing was limited because 1) much of the communication between DOT&E and the Air Force was informal and undocumented, and 2) we may not have been provided all the documentation that exists. The DOT&E action officer defended informality, saying it is important to keep DOT&E's influence low-level and invisible to be effective.

6.1. Were there successful attempts to influence the OT&E process? DOT&E asked for more live IR Maverick launches with LANTIRN, and the number of firings was increased from one to six. They also asked for more conventional weapons delivery, which is planned for summer 1988.

6.2. Were there unsuccessful attempts to influence the OT&E process? DOT&E did not want a second pilot in the back seat during the test, on the grounds that even if the pilot did not say anything, his presence would affect how the first pilot flew (see 3.1). However, the issue was resolved in favor of safety, i.e., putting the second pilot in during tests. DOT&E also has raised some concerns over how AFOTEC will test some unresolved LANTIRN issues. The Air Force and DOT&E have agreed to resolve these issues before an anticipated FY88 B-LRIP report. However, some issues will not be resolved by this time (see 6.3).

6.3. What was DOT&E impact on the Beyond-LRIP milestone? We found no significant DOT&E impact on the B-LRIP milestone decision for the LANTIRN navigation pod except to support the production decision, but DOT&E significantly impacted the decision for the LANTIRN targeting pod. In a memo to the assistant secretary of the Air Force, DOT&E stated that a decision for full production is not justified by the results of OT&E, due to: 1) the need for more integration with specific aircraft models, 2) incomplete conventional weapons testing, 3) poor LGB results, and 4) marginal targeting pod RAM data. However, the memo also stated that if the Air Force were to defer a full production decision for the targeting pod, a B-LRIP report would not be required at this time. Instead, it would be submitted only in October 1988 or when the Air Force proposed to exceed a rate of 81 pods per year. However, it is important to note that 81 pods was the Air Force's intended buy for the first year of full-scale production, so basically DOT&E offered the Air Force a choice between a negative B-LRIP report to the Congress or a redefinition of B-LRIP that would delay the report. The Air Force choose the latter, keeping its planned first year full-scale production schedule without calling it B-LRIP.

DOT&E reviewed and supported the Air Force's new test schedule and criteria for LANTIRN. The test schedule is intended to resolve DOT&E's concerns mentioned above, prior to October 1988, so that DOT&E can make a B-LRIP report. However, the agreement between DOT&E and the Air Force will not resolve many of the problems raised by DOT&E. This is because: 1) There will be no dedicated LANTIRN OT&E test sorties, 2) there will be a limited opportunity to flight test the production targeting pod on both the F-15E and Block 40 F-16 test aircraft; 3) there will be only limited data available for production targeting pod reliability growth tracking, 4) the production pod delivery schedule will not support accumulation of sufficient test data to validate improved reliability projection by October 1988; 5) many of the success milestones, from which the Air Force and DOT&E will judge the system's new performance, do not reflect any quantitative criteria; 6) much of the needed data will come from

contractors or from laboratory results; and, finally, 7) LGB data will be simulated.

7. DOT&E Reporting

7.1. What statements did DOT&E make to the Congress regarding the adequacy of OT&E and system effectiveness and suitability? DOT&E's navigation pod B-LRIP report of November 1986 stated that testing was adequate to demonstrate operational effectiveness of the navigation pod on a single-seat fighter. Four of six operational suitability areas were satisfactory, but the other two depended on contractor support. The report stated that because of this, testing did not provide all the information necessary for a complete evaluation of operational suitability.

A B-LRIP report has not yet been issued for the targeting pod; however, the 1987 annual report to the Congress does address FY87 LANTIRN targeting pod testing. In that report, DOT&E did not comment on test adequacy. On effectiveness, it stated that of the seven objectives, two were satisfactory (IR Maverick delivery and LANTIRN controls and displays) and one was marginal (LGB delivery). In the 1987 annual report, DOT&E stated that EOCM vulnerability had been judged satisfactory in previous testing. DOT&E considered three suitability issues to be satisfactory: 1) logistics support reliability, 2) mission performance reliability, and 3) availability. Overall system maintainability was rated marginal primarily because of targeting pod problems. The DOT&E FY87 annual report concluded that both the operational effectiveness and suitability of the targeting pod required further improvement and testing before it will fully meet the needs of the user.

7.2. What was the completeness and accuracy of DOT&E's statements regrading adequacy of OT&E? DOT&E's B-LRIP report on the LANTIRN navigation pod stated that pilots flew over unfamiliar, "first look" routes and terrain. We learned from an AFOTEC data document that pre-testing was performed, and from DOT&E's action officer that the same pilots flew both the pre-testing and the testing missions. We also learned from DOT&E action officers that a second pilot was in the back seat during some of the FOT&E tests for safety reasons. This is important because the tests were intended to assess single-seat effectiveness over unfamiliar terrain, something which the tests did not do.

DOT&E's FY87 annual report further states for the targeting pod, "...effective sorties were flown from two geographically and meteorologically different locations." Although the locations are different, they

have very specific environments which represent neither the scope of NATO-like environments and terrain, nor the world wide range of environments where LANTIRN is intended to be deployed (see 3.8). In its report, AFOTEC stated the limitations of each test site; however, DOT&E's annual report did not. As a result, DOT&E's statement is not complete, and it gives the misleading impression that LANTIRN environment testing was more robust than it was. DOT&E officials do not dispute the test limitation, however they told us that the annual report is not required to be as complete as the B-LRIP report, and that this issue will be addressed then.

7.3. What was the completeness and accuracy of DOT&E's statements regarding system effectiveness and suitability? We found no significant problems regarding DOT&E statements on effectiveness and suitability of the navigation pod. However, for the LANTIRN targeting pod LGB results, the annual report informed the Congress that they were marginal, whereas in earlier memos to the Air Force, DOT&E had termed them "poor." DOT&E thus described the results in less negative terms to the Congress than to the services. DOT&E stated that IR Maverick performance was satisfactory with the LANTIRN targeting pod. This satisfactory rating came from testing where: 1) the targets were stationary and presented an unrealistically vivid IR signature (see 3.5); 2) pilots were familiar with the test range and did not have to react to any threat defenses (see 3.6); and 3) there was no EOCM (see 5.1).

In its 1987 annual report, DOT&E states that for targeting pod testing, "... EOCM vulnerability had been judged satisfactory in previous testing." However, the statement was incomplete in that it did not mention that the previous testing was limited and much of it depended on simulations (see 5.1). The report also did not mention that the final conclusions regarding EOCM have not yet been reached (see 5.1).

Requests for copies of GAO reports should be sent to:

U.S. General Accounting Office
Post Office Box 6015
Gaithersburg, Maryland 20877

Telephone 202-275-6241

The first five copies of each report are free. Additional copies are $2.00 each.

There is a 25% discount on orders for 100 or more copies mailed to a single address.

Orders must be prepaid by cash or by check or money order made out to the Superintendent of Documents.

United States
General Accounting Office
Washington, D.C. 20548

Official Business
Penalty for Private Use $300

First-Class Mail
Postage & Fees Paid
GAO
Permit No. G100