United States General Accounting Office 133874

GAO

Report to the Chairman, Subcommittee on Seapower and Strategic and Critical Materials, Committee on Armed Services, House of Representatives

August 1987

LIVE FIRE TESTING

Evaluating DOD's Programs

GAO/PEMD-87-17

GAO

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

Program Evaluation and
Methodology Division

B-207947

August 17, 1987

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

Dear Mr. Chairman:

This report responds to your May 12, 1986 letter. You asked us to review the Joint Live Fire test program of the Department of Defense.

We note that this is an unclassified version of the classified GAO report C-PEMD-87-1. Classified sections of the original report have been modified or deleted to permit the issuance in unclassified form.

As we agreed with your office, unless you publicly announce the contents of this report earlier, we plan no further distribution of it until 30 days from the date of the report. At that time, we will send copies to those who are interested and will make copies available to other authorized parties upon request.

Sincerely,

Eleanor Chelimsky
Director

Purpose

U.S. weapon systems are often procured with only computer-based vulnerability and lethality estimates, and little or no testing to determine effects against actual threats. The Joint Live Fire Test program (JLF) was initiated in 1984 to perform such testing on current U.S. systems. The Chairman, Seapower Subcommittee, House Armed Services Committee, asked GAO to evaluate JLF and related live fire programs in DOD. GAO's evaluation was organized around 4 questions: 1) What is the status of each system and munition originally scheduled for live fire testing?; 2) What has been the methodological quality of the test and evaluation process?; 3) What are the advantages and limitations of full-up live fire testing, and how do other methods complement full-up testing?; and 4) How can live fire testing be improved?

Background

JLF is a vulnerability/lethality (V/L) assessment program in which Soviet munitions are fired at combat-loaded U.S. systems, and [material deleted]. Testing began in FY 1985, and is scheduled to run through FY 1990. The program has an armor/anti-armor component (JLF/Armor) and an aircraft component (JLF/Aircraft). The focus is on first line, fielded systems. After GAO received this request, Congress mandated that developmental systems also undergo realistic live fire testing before proceeding beyond low rate initial production.

Results in Brief

There have been major slippages in the JLF Armor schedule, largely due to prolonged controversy between the Office of the Secretary of Defense and testers over objectives and methodology. In contrast, JLF/Aircraft was planned and implemented without major conflict or interruption. Lack of targets has been a problem for both components.

Although there is little completed testing to evaluate, it is apparent that the technical capabilty to do full-up testing (that is, testing with combustibles on board) is not well developed. This seems to be a consequence of the historically low emphasis on live fire testing in the U.S.

The main controversies in live fire test methodology reflect differences between the interests of proponents of full-up testing and advocates of computer modeling, resulting in largely incompatible approaches which cannot be reconciled by technical solutions alone.

It is doubtful that JLF or any future live fire testing will produce the kind or quantity of data needed to validate the sophisticated V/L models currently in use. However, the accumulated data should enable checking whether model revisions improve predictions.

Full-up live fire testing is the only V/L assessment method providing direct visual observation of the damage caused by a weapon/target interaction under realistic conditions. As such, it offers a unique, important advantage over other methods.

GAO's Analysis

Status of Tests

The slippages in the JLF/Armor test schedule have meant that the first of the originally scheduled tests began almost two years behind schedule. The JLF/Aircraft program has also been delayed (principally due to lack of targets), but less severely than JLF/Armor.

Methodological Quality of Tests

JLF's program objectives were not well enough defined to give test designers a clear direction. There have been conflicting statements of the objectives of live fire testing in general, reflecting underlying differences in the interests of the individuals and organizations involved.

The long-term planning failure of OSD and JLF/Armor officials to agree on a testing approach has caused implementation delays and the waste of resources in repeated plan revision. The approach of the most recent draft JLF/Armor master plan is similar to the first version, rejected by OSD in 1984 because of inconsistency with the objectives of JLF. JLF/Aircraft planning was generally well organized, thorough, and consistent with objectives. However, JLF/Aircraft test plans omit key information, contain inconsistencies, and specify targets which are not available.

The principal constraint faced by all JLF officials is a lack of targets. Budgetary responsibilities were never clearly designated; the services were not responsible for supplying targets, nor was this covered under JLF's budget. Consequently, test officials have had to spend much of their time marketing the program, and competing with other interests for targets. The target systems and components that JLF does receive are frequently in poor condition, yet JLF provides no funds for restoration.

In general, the sample sizes of JLF and related live fire tests have been too small to produce statistically reliable results. The most common V/L indicator—probability of a kill given a hit—is primarily based on subjective engineering judgment, and has not been shown to be statistically reliable or valid. Users of output from V/L analysis are often unaware of the subjective nature of this indicator. (See pp. 69-75; 114-15.)

Controversy over how to select test shots is largely a conflict between the two objectives of sampling efficiency and unbiasedness. Ultimately, it appears impossible to agree on how to select shots without first deciding on the relative importance of these objectives. (See pp. 75-86.)

The scientific capability to estimate human effects with confidence has not yet been achieved. This, and the fact that JLF plans have paid little attention to human effects, make it unlikely that JLF will produce precise estimates of casualties. (See pp. 86-89.)

Overall, the state of the art of live fire testing has improved since previous live fire programs, but some potentially solvable problems raised earlier have not been solved. For example, little progress has been made in the empirical validation of V/L estimates.

Advantages and Limitations of Full-Up Testing

Full-up live fire testing is the only V/L assessment method providing direct visual observation of the damage caused by a weapon/target interaction under realistic combat conditions. These observations are regarded as highly beneficial by users. Full-up testing has already produced several "surprises", i.e., results that were not predicted, and might not have been detected by other means of testing or analysis. The primary limitation of full-up, full-scale testing is cost, mainly due to the high cost and limited availability of targets. Nonetheless, live fire testing costs are a very small percentage of total program costs. Other limitations are limited information yield, limited generalizability of results, and limited redesign opportunities. (See pp. 101-04.)

Other Methods

Subscale and inert testing have some distinct advantages over full-up testing, but provide only indirect evidence of effects on realistic targets. Analysis of combat data has other advantages, but has less scientific control and is limited to systems that have been in combat. All these methods can supplement full-up, full-scale testing but not substitute for it. (See pp. 105-07.)

Models are potentially useful in extrapolating beyond test results, and have a unique advantage over testing in their applicability to systems not yet built. Still, current V L models are inadequately validated and share many limitations. Key mechanisms for producing casualties are poorly modeled if at all, limiting the models' usefulness in predicting casualties or providing insights into casualty reduction. Claims that models predict well "on the average" can be misleading, yet claims that vulnerability models predict poorly have been somewhat overstated. There are no clear criteria for success and failure in model prediction, and proponents and opponents of modeling have both claimed support from the same data. (See pp. 107-28.)

Improving Live Fire Testing

Opportunities were identified for technical improvements in the design, conduct, and interpretation of live fire tests (e.g., DOD could test whether departures from realism that reduce the cost or difficulty of conducting live fire tests do nonetheless preserve the generalizability of test results to realistic conditions), and general improvements to facilitate realistic live fire testing and the usefulness of its results (e.g., DOD could consider target costs in light of total program costs, including the concept of a percentage set-aside for live fire testing). (See pp. 132-34.)

Recommendations

In addition to the improvements noted above, there is a need to resolve current conflicts about the purpose of live fire tests and to make clear that the objective of reducing vulnerability and increasing lethality of U.S. systems is the primary emphasis of testing. Accordingly, GAO recommends that the Secretary of Defense conduct full-up tests of developing systems, first at the subscale level as subscale systems are developed, and later at the full-scale level mandated in the legislation; establish guidelines on the role live fire testing will play in procurement; establish guidelines on the objectives and conduct of live fire testing of new systems; and ensure that the primary users' priorities drive the objectives of live fire tests.

The live fire legislation requires the services to provide targets for testing new systems, but there is no similar requirement for the fielded systems in JLF. Accordingly, GAO recommends that the Secretary of Defense provide more support to JLF for obtaining targets.

Agency Comments

DOD provided oral comments on the report. DOD concurred with all recommendations and most findings, and made several suggestions to improve technical accuracy. GAO made changes based on these suggestions where appropriate.

Contents

Chapter 3 What Has Been the Methodological Quality of the Test and Evaluation Process?

Chapter 4 What Are the Advantages and Limitations of Full-Up Live Fire Testing, and How Do Other Methods Complement Full-Up Testing?

Chapter 5 How Can Live Fire Testing Be Improved?

Chapter 6 Recommendations and Agency Comments

Bibliography

Glossary

Appendixes

Tables

Figures

Abbreviations

AFATL	Air Force Armament Laboratory
AMSAA	Army Materiel Systems Analysis Activity
AP	Armor piercing
APC	Armored personnel carrier
API	Armor-piercing, incendiary
BAST	Board on Army Science and Technology (of NRC)
BRL	Ballistics Research Laboratory
CARDE	Canadian Armament Research and Development Establishment
DDTE	Director, Defense Testing and Evaluation
DOD	U. S. Department of Defense
DTP	Detailed test plan
DUSDRE(TE)	Deputy Under Secretary of Defense for Research and Engineering (Testing and Evaluation)
GAO	U. S. General Accounting Office
HASC	U. S. House of Representatives Armed Services Committee
HEAT	High explosive anti-tank
HEI	High explosive incendiary
IDA	Institute for Defense Analyses
JCS	Joint Chiefs of Staff
JLC	Joint Logistics Commanders
JLF	Joint Live Fire
JT&E	Joint test and evaluation
JTCG AS	Joint Technical Coordinating Group for Aircraft Survivability
JTCG ME	Joint Technical Coordinating Group for Munitions Effectiveness
LANL	Los Alamos National Laboratory
LAVP	Lot Acceptance Verification Program
LRIP	Low rate initial production
MEXPO	Materiel exploitation program
NRC	National Research Council
NAVAIR	Naval Air Command
OSD	Office of the Secretary of Defense
P_k	Probability of kill
$P_{k\ H}$	Probability of kill, given a hit
RPG	Rocket propelled grenade
SPC	Systems Planning Corporation
SPO	System program office
SDAL	Standard damage assessment list
SURVIAC	Survivability/Vulnerability Information Analysis Center
T&E	Testing and evaluation
TEAS	Testing and Evaluation of Aircraft Survivability
TEMP	Test and evaluation master plan
V L	Vulnerability/lethality

Chapter 1

Background

The Joint Live Fire Test (JLF) is a vulnerability/lethality (V/L) assessment program in which Soviet munitions are fired at combat loaded U.S. systems, and U.S. munitions are fired at combat loaded Soviet systems. At least four groups have been identified as potential users of JLF test results: designers, tacticians, force level planners, and procurement authorities.

The program is divided into an aircraft component and an armor/anti-armor component, which we will refer to as JLF/Aircraft and JLF/Armor, respectively. The focus is on first line, fielded systems; systems still under development are not included. According to one estimate, the program affects the lives of over 300,000 servicemen who may have to use this equipment in combat. Testing began in FY 1985, and is scheduled to run through FY 1990.

The Chairman, Seapower Subcommittee, House Armed Services Committee, asked us to evaluate the JLF program and related live fire test programs in DOD, specifically, tests of systems removed from JLF to be tested by the Army. We have organized our evaluation around 4 questions:

1) What is the status of each system and munition originally scheduled for live fire testing?

2) What has been the methodological quality of the test and evaluation process?

3) What are the advantages and limitations of full-up live fire testing, and how do other methods complement full-up testing?

4) How can live fire testing be improved?

Prior GAO Reports

We have produced two reports on the Bradley Infantry Fighting Vehicle survivability testing. The first (February, 1986) focused on the Bradley Phase I live fire testing.[1] It concluded that the Bradley as presently configured is highly vulnerable to anti-armor weapons, and noted problems with the tests already completed. (We discuss these in Appendix II). The second report (November, 1986) focused more on the Bradley's mission

[1] U.S. General Accounting Office Bradley Vehicle: Concerns About the Army's Vulnerability Testing. GAO/NSIAD-86-87, Washington, DC: February, 1986

requirements and the proposed operational tests.[2] This report concluded that new operational tests should be conducted with particular emphasis on how well the tactics devised for the Bradley will offset its vulnerability, and at the same time, permit it to retain its combat effectiveness.

What Is Live Fire Testing?

The term "live fire test" is used in several ways within DOD, but not all of these involve the vulnerability or lethality of weapon systems. For example:

- A 1985 missile firing from an F-16 aircraft was called a live fire test; its purpose was to verify the missile's compatibility with the F-16's avionics system and the performance of the missile's active radar guidance capabilities. In this sense of the term, live fire is distinguished from captive carry, i.e., using the missile's guidance system to allow the aircraft to carry the missile along the path to the target.
- A 1975 weapons proficiency training experiment was also called a live fire test; its purpose was to compare the performance of troops doing actual firing with troops simulating firing. In this sense of the term, live fire is distinguished from dry fire.
- Tests of fire suppression systems are also called live fire tests.

In this report, the term live fire will apply only to V L testing.

Lack of General Definition

None of the planning documents, briefings, or testimony we reviewed contained a general definition of live fire testing, even limited to V-L testing. The range of testing in JLF is so broad that no single definition is likely to cover all cases. Some JLF tests, such as the aircraft engine fuel ingestion tests, involve no live firing of munitions at all, relying instead on mechanically punched holes.

Types of Live Fire Testing

Live fire testing can be roughly classified by the status of the target. Targets can be full scale or subscale, and full-up or inert:

- Full-scale targets are complete aircraft or armored systems, while subscale targets are components, subcomponents, structures, etc.

[2] U S General Accounting Office. Bradley Vehicle: Army's Efforts to Make It More Survivable. GAO NSIAD-87-40

- Full-up targets contain all appropriate combustibles—typically fuel, ammunition, and hydraulic fluid—while inert targets do not contain these.

The 2 X 2 matrix in Table 1.1 illustrates these distinctions, with examples. JLF and related live fire testing currently being conducted falls into all four types.

Table 1.1: Types of Live Fire Testing

	Loading	
Scale	**Full-Up**	**Inert**[a]
Full scale	Complete system with combustibles (e.g., Bradley Phase II tests, aircraft "proof" tests)	Complete system, no combustibles (e.g. tests of new armor on actual tanks, aircraft flight control tests)
Sub-scale	Components, subcomponents with combustibles (e.g., fuel cell tests behind armor mock-up aircraft engine fire tests)	Components, subcomponents, structures, terminal ballistics, munitions performance, behind-armor tests, warhead characterization (e.g., armor/warhead interaction tests, aircraft component structural tests)

[a] In some cases, targets are "semi-inert" meaning some combustibles are on board but not all (Example tests of complete tanks with fuel and hydraulic fluid but dummy ammunition.)

Full-scale, full-up testing is generally considered the most realistic variety of live fire testing currently practiced, and is the type mandated by the authorization legislation described below. Though munitions are real, they are not generally fired from operational weapons systems such as tanks and aircraft. Rather, they are most frequently fired from stationary mockups designed to deliver them at a pre-specified realistic velocity.

Usually, the term full-up testing implies full-scale targets. We will follow this practice here; references to full-up testing imply full-scale targets unless otherwise indicated.

Basic Concepts

Regardless of type, live fire testing as currently practiced addresses questions of vulnerability and/or lethality. It does not, as is sometimes mistakenly assumed, address the larger concepts of survivability or effectiveness, or the related concept of susceptibility. Survivability and effectiveness refer to the probability of a kill, while susceptibility refers to the probability of engagement. Vulnerability and lethality, by contrast, refer to the probability of a kill given a hit. That is, for vulnerability and lethality the hit is assumed (i.e., its probability is 1 0).

Technically, this is called a conditional probability, because the probability of a kill is conditional on a hit having occurred. This means that system capabilities which reduce the probability of getting hit (e.g., maneuverability, low detection signature) have no effect on vulnerability and lethality as assessed through live fire testing (Table 1.2 illustrates the relationship among all these concepts).

Table 1.2: Relationships Between Key Concepts

		Point of view	
Symbol	Meaning	Offensive[a]	Defensive[a]
P_E	Probability of engagement	—	Susceptibility
$P_{H,E}$	Probability of a hit, given engagement	—	—
$P_{K,H}$	Probability of a kill, given a hit	Lethality	Vulnerability
P_K[b]	Probability of a kill	Effectiveness	Survivability

[a] Offensive focus on the attacking munition Defensive focus on the defending target

[b] $P_K = (P_E)(P_{H E})(P_{K H})$, for example where $P_E = 4$ $P_{H E} = 7$ and $P_{K H} = 5$ $P_K = (4)(7)(5) = 14$

Source Adapted from G Smith et al The Joint Live Fire (JLF) Test Background and Exploratory Testing, (DRAFT) Alexandria, Va Institute for Defense Analyses, March 1986

Therefore, it is the view of vulnerability experts that JLF and related live fire tests will not provide "stand alone" data from which survivability and effectiveness conclusions can be reached. Other factors, such as susceptibility, must be integrated with the V L data and then all appropriate trade-offs carefully evaluated before arriving at any conclusions about required design changes.

Relationship to Developmental and Operational Testing

Traditionally, DOD testing falls under one of two categories: developmental and operational. Live fire testing, however, does not fall neatly under either category. Organizationally, it has been conducted under the Deputy Undersecretary of Defense for Research and Engineering (Test and Evaluation), which has oversight for developmental, but not operational, testing. Though live fire testing sometimes has been part of developmental testing, this is not generally the case. The JLF and related tests (Bradley vehicle, M1 tank) are all being conducted post-development. Yet live fire tests are not operational tests either, because there is no attempt to simulate an operational environment.[3] Organizationally as well, live fire testing has been kept separate from operational testing.

[3] A notable exception was the GAU-8 lethality test (described in Chapter 3), where the targets (tanks) were arranged in operational formation and fired at from actual flying aircraft

History of Live Fire Testing

Live fire testing in the U.S. goes at least as far back as early WW II, when live fire tests demonstrated that the U.S. M2-series light tanks could be defeated by .50-cal. armor piercing (AP) machine gun fire. It continued through the 1950s, culminating in the Canadian Armament Research and Development Establishment (CARDE) trials in 1959. CARDE—the last comprehensive series of live fire tests on armored targets—looked at a number of generic shaped charge warheads in an attempt to assess their lethality against enemy targets. Although the trials were conducted under a number of handicaps, such as non-functioning test vehicles, limited weapon classes, and old tactics, CARDE data still form the empirical foundation for the computer models used by weapons effects and vulnerability analysts. In the 25 years between CARDE and JLF, there were only isolated instances of live fire testing on armored vehicles (most notably, the GAU-8 lethality tests, described in Chapter 3).

On the aircraft side, the only systematic live fire testing was the Test and Evaluation of Aircraft Survivability (TEAS) in the early 1970s. TEAS grew out of the Southeast Asia conflict, in which the large number of aircraft losses made it clear that survivability had not been given sufficient emphasis during design (at least 60 percent of the 5,000 U.S. aircraft lost in Viet Nam were downed by fire and explosion). TEAS was a tri-service program to 1) evaluate the vulnerability of the F-4, A-7, and AH-1 aircraft, 2) develop vulnerability reduction concepts for those aircraft, and 3) apply the knowledge gained to future aircraft. Following TEAS, funding emphasis moved from evaluation by full-scale live fire testing toward evaluation by analysis (i.e., computer modeling).

Thus both the armor and aircraft V L communities took a general turn away from live fire testing and towards modeling. Test and evaluation funding for vulnerability reduction has been limited, despite a recognition by experts that the analytical models utilized in some areas are inadequate or lack validation.

Recent Legislation

At the time of the Chairman's request, there were no laws requiring live fire testing; JLF and related live fire programs of existing systems were DOD initiatives. Congress has since mandated live fire testing of certain weapon systems in the National Defense Authorization Act for Fiscal Year 1987. There are two live fire sections in this legislation: one on the testing of new systems and one on the testing of the Bradley vehicle.

Live Fire Testing of New Systems

As stated in Section 910, the Secretary of Defense shall provide that:

1) a covered system may not proceed beyond low rate initial production until realistic survivability testing is completed.[1]

2) a major munition or missile program may not proceed beyond low rate initial production until realistic lethality testing is completed.

Survivability and lethality tests are to be carried out sufficiently early in the development phase of the system or program to allow any design deficiency demonstrated by the testing to be corrected before proceeding beyond low rate initial production. Testing costs will be paid from funds available for the system being tested. The Secretary of Defense may waive such testing if he certifies to Congress that live fire testing of that system would be unreasonably expensive and impractical. The President may waive it in time of war or mobilization.

The section's definitions emphasize the full-up, live fire nature of the testing requirement:

> "The term 'realistic survivability testing' means . . testing for vulnerability and survivability of the system in combat by firing munitions likely to be encountered in combat . . . at the system configured for combat, with the primary emphasis on testing vulnerability with respect to potential user casualties and taking into equal consideration the operational requirements and combat performance of the system."

> "The term 'realistic lethality testing' means . testing for lethality by firing the munition or missile concerned at appropriate targets configured for combat."

> "The term 'configured for combat' . . means loaded or equipped with all dangerous materials (including all flammables and explosives) that would normally be on board in combat."

Testing of the Bradley Vehicle

As stated in Section 121, the Secretary of Defense shall:

- require both live fire testing and testing of operational combat performance.
- develop a plan for said testing and evaluation.

[1] A proposed amendment to the legislation substitutes the term "vulnerability" for "survivability," so as to be more consistent with general DOD practice.

The plan is to include both the Army's "high survivability" Bradley and the "minimum casualty vehicle." The latter will be a specially configured vehicle previously encouraged by the Office of the Secretary of Defense (OSD). The two vehicles will then be compared.

The live fire tests were to be developed in consultation with the OSD Director of Defense Research and Engineering and the National Academy of Science. This has already been done. The operational performance aspects are to be developed in consultation with the Director of Operational Testing and Evaluation.

Objectives, Scope, and Methodology

In a letter of May 12, 1986, the Chairman of the Seapower Subcommittee, House Armed Services Committee asked us to collect and analyze information on the Joint Live Fire Test program (JLF) and related live fire testing in DOD. From the letter and subsequent meeting with staff, we derived the following evaluation questions and adopted tasks to answer each:

Question 1. What is the status of each system and munition originally scheduled for live fire testing?

Tasks:

- Determine the current scheduling status of each system or component in the JLF program from appropriate testing officials and documents, including systems formally removed from JLF (Bradley, M1, and M113).
- Compare current schedules with original schedules and determine principal reasons for slippages.

Question 2. Over a variety of tests, what has been the methodological quality of the test and evaluation process?

Tasks:

- Assess the JLF methodology in terms of setting test objectives, planning and implementing tests, and analyzing and reporting results
- Elaborate and clarify the controversy over the objectives of the JLF program.
- Determine the obstacles and proposed solutions to obtaining sufficient numbers of targets.
- Determine if and how testers are maximizing information yield from small samples through design efficiency.

- Elaborate and clarify the controversy over shot selection methodology for the Bradley vehicle, and draw implications for live fire testing in general.
- Compare JLF objectives and approaches with previous U.S. live fire testing programs.
- Compare JLF objectives and approaches with foreign live fire testing programs, if information is available.

Question 3. What are the advantages and limitations of live fire testing, and how do other means complement full-up testing?

Tasks:

- Review the relationship of subscale and inert testing to full-up tests.
- Review the advantages and limitations of using combat data in vulnerability assessment.
- Identify and evaluate claims about the advantages and limitations of modeling and full-up testing.
- Determine how models are used in live fire tests.
- Determine how models can be validated in live fire testing.
- Determine how well current models predict vulnerability.

Question 4. How can live fire testing be improved?

Tasks:

- Identify potential technical improvements to live fire testing.
- Identify potential general improvements to live fire testing.

The scope covered JLF and other live fire testing not currently part of JLF. We conducted the work in Washington, DC, Aberdeen Proving Ground, MD, Wright-Patterson Air Force Base, OH, and China Lake, CA. All data were gathered between June and December, 1986.

To answer our questions, we:

- observed live fire tests;
- interviewed DOD officials and outside experts in V/L testing and analysis;
- reviewed JLF and related live fire testing documentation; and
- reviewed literature on V/L model validation and other literature as applicable.

Assessing Test Quality

During a prior review of the Joint Test-and-Evaluation Program (JT&E), we developed a case-study method to assess the quality of the tests.[5] The method included an examination of the seven steps in the test process from understanding the context and defining the objectives through planning and implementation, data analysis and reporting, to using the results. For each of the steps, we identified threats to the quality of the test and assessed the probable effect of each threat. The JT&E review also considered test features such as realism in the selection of test objectives; whether there were unjustified departures from the test plan during implementation; whether the data analysis used explicit and justifiable criteria for excluding data and appropriate statistical controls for threats such as attrition; and whether the reporting of results was clear and comprehensive, with appropriately qualified conclusions and recommendations congruent with the findings.

Because the JLF review was similar in many respects to the JT&E study, we used the JT&E assessment methods wherever possible in the case-study analysis of the completed Joint Live Fire Tests. Many of the standards are relevant to judging the quality of live fire tests, but there are some points at which the nature of live fire testing dictated the use of different standards and emphases and, correspondingly, the consideration of different methodological issues. For example:

- The operational tests reviewed in JT&E often managed to obtain fairly large sample sizes in spite of the high costs of testing, through the use of repeated trials with simulated firing. The extremely limited availability of test targets and the limited number of shots possible on each target for destructive live fire testing make issues related to small sample size more important.
- Tactical realism and training of participants are important in operational tests, but not important in live fire testing as currently practiced.

Since JLF was in the second year of its funding, and little actual testing had taken place, we assessed the methodological quality of plans for tests not yet conducted, in addition to completed tests. Our goal was to identify methodological issues that could be expected to arise in the course of JLF tests and live fire testing in general. Therefore, we reviewed:

[5] U.S General Accounting Office. How Well Do the Military Services Perform Jointly in Combat? DOD's Joint Test-and-Evaluation Program Provides Few Credible Answers GAO PEMD-84-3 Washington, DC: February 22, 1984

- all available master plans and detailed test plans that had been prepared by the end of our review, and
- all completed draft JLF reports.

We examined the process of overall planning, setting test objectives, test planning, implementation, and the analysis and reporting of results. We were able to capitalize on the existing variation in the test cycle across systems to review different stages of the testing process occurring simultaneously in our time window. Wherever reports were complete at least in draft form we proceeded with a version of the JT&E case study method.

By agreement with our requester, we did not conduct a detailed case study of the Army's tests of the Bradley Fighting Vehicle. There are three reasons for this:

- We have issued two reports on Bradley testing and will issue another after Phase II testing is completed.
- The House Armed Services Committee, our requester's parent committee, has already conducted their own investigation and issued a report on the Bradley tests.
- All Bradley testing was suspended during the time of our review pending approval of a new test plan, which we obtained only when our review was nearly complete.

However, we did examine the methodological positions of the testers, critics, and outside reviewers to derive general lessons to be learned for the design and conduct of live fire tests. We reviewed and elaborated the controversy over Bradley shot selection methodology, which we believe has important implications for live fire testing in general. Several of the armor test plans are being or will be revised in light of the Bradley controversy.

External Comparisons

In order to provide a comparative context for methodological judgments of current U.S. live fire test programs, we also examined past programs and foreign programs. Past programs included the CARDE trials and the GAU-8 tests on the armor side, and TEAS and live fire qualification testing on the aircraft side. In each case we obtained and reviewed reports and interviewed knowledgeable experts and, when possible, testers who had been involved in the original tests. Foreign programs included activities in Israel and the U.K. (we found no evidence that other allies conduct live fire tests). We were not able within the time of our review to obtain documentation of foreign test programs or to interview officials actually conducting tests. Our understanding of these tests is based on interviews with U. S. test officials and others who have knowledge of the programs, and in one case, videotaped interviews with foreign testers.

Models

Answering the third question of the request letter required that we expand what was a minor issue for the JT&E work—the balance between modeling and testing—into a major part of our review. In order to assess claims about vulnerability modeling we focused on three questions about the vulnerability models that play a role in live fire tests: 1. How well do the models currently predict vulnerability? 2. What role will the models play in live fire tests? and 3. How will live fire tests be used to validate or calibrate the models (one of the JLF objectives)?

During the course of reviewing the cases and the test plans, we assembled information relevant to the role of vulnerability/lethality computer models in live fire tests that have been conducted or proposed. We interviewed vulnerability modelers at the Ballistics Research Laboratory (BRL) and elsewhere to obtain their reaction to the model validation literature and to learn what procedures or decision rules they follow in updating or revising their models in the light of live fire test data. We were briefed by them about their general modeling procedures and views of live fire testing and obtained documentation of the modeling process and vulnerability methodology in general. We assessed the realism of model assumptions and the quality of input data, but did not attempt to evaluate the accuracy of the models' computer code or the physical theory supporting it.

We then reviewed the literature that attempts to assess the predictive validity of vulnerability models with respect to test and combat data. The goal of this review was to assess conflicting claims that were made during the planning and implementation of JLF about the ability of models to predict live fire results. We were able to obtain the main studies and analyses identified by both proponents and critics of modeling. It was not possible in our time frame, though, to conduct a complete synthesis of this literature.

Suggestions for Improvement

One goal of the design outlined above is to move beyond judgments of the quality of the few cases of live fire tests examined in the most detail, and to isolate the distinctive problems of live fire testing as they emerge from both the case studies and the review of test plans and modeling. Because the Joint Live Fire program as a whole is still at an early stage of implementation, and recent legislation has mandated live fire testing of major new weapon systems, there is an important opportunity now for any general lessons learned to be fed back into the live fire test design and implementation process. We assembled suggestions made by testers and experts we interviewed for resolving some of the difficulties and added others that reflect accepted standards of applied research practice. Our recommendations and suggestions are set forth in Chapter 5.

Organization of the Report

The report is organized around the four evaluation questions noted above, with one chapter devoted to each question:

- Chapter 2 traces the development of the JLF program, and presents the status of the systems and munitions originally scheduled for live fire testing.
- Chapter 3 assesses the methodological quality of the test and evaluation process to date, delineates key general issues in live fire testing, and compares current U.S. efforts with past and foreign live fire testing programs.
- Chapter 4 reviews the advantages and limitations of full-up live fire testing, and assesses the capability of other methods to complement live fire tests, with particular attention to computerized models.
- Chapter 5 identifies opportunities for technical and general improvements in live fire testing.
- Chapter 6 contains recommendations for the Secretary of Defense and agency comments.

What Is the Status of Each System and Munition Originally Scheduled for Live Fire Testing?

To understand the current status of specific live fire tests, it is first necessary to review how the Joint Live Fire (JLF) program developed.

Development of the JLF Program

In the summer of 1983, the OSD Director, Defense Testing and Evaluation (DDT&E) nominated to the services a joint test and evaluation (JT&E) initiative involving live fire of munitions against full-scale operational targets. As with previous JT&E's, a joint test force would be established to plan and conduct the tests. The tests were to include ships as well as aircraft and combat vehicles.

Service Response

According to a memo from the Principal Deputy Undersecretary of Defense for Research and Engineering (USDRE) describing the services' response to the proposal, the need for the program was "recognized by all." In fact, however, official responses from Joint Chiefs of Staff (JCS) and services do not confirm this:

- The Marine Corps recommended against the nomination.
- The JCS, Army, and Navy recommended varying degrees of further study before reconsidering the nomination.
- The Air Force alone agreed to support the proposal pending resolution of its concerns, one of which was participation by the other services.

General concerns were:

- The work would duplicate ongoing efforts.
- Necessary Soviet targets would not be available.
- Financial responsibility for providing test assets was unclear.
- Materiel costs would be high, with uncertain return.
- The purpose and/or scope was unclear.

Additionally, the JCS noted that there was no indication of how the test tied in with joint operational requirements, how the data would be used, and who would benefit from the results; they also questioned the need for the test on the grounds that simulation in this area had been generally successful. The Navy stated that they constantly review ship design considerations with respect to survivability and vulnerability, implying that inclusion of ships in the test was unnecessary. The Marine Corps cited concern with the manpower requirement of assigning one or more officers to a joint test directorate.

Eventually, all the services did agree to participate, on the following conditions:

- A joint test force would not be established. Instead, the program would be planned and conducted by the Joint Technical Coordinating Groups (JTCG's) for Aircraft Survivability (aircraft tests) and Munitions Effectiveness (armor/anti-armor tests), two existing elements of the Joint Logistics Command (JLC) with expertise in V/L assessment.
- Services would bear no financial responsibility for providing test assets.
- Ships would be excluded.

The JTCG Tasking

The services were strongly against establishing a new joint test force. The OSD program manager agreed to the tasking of the JTCGs provided DDT&E retained a level of control equivalent to that of any other joint test. Reportedly, this was the first time DDT&E had directed a joint test to an existing organization. This decision had both positive and negative aspects. On the one hand, the core technical expertise was already in place, and the test personnel had longstanding professional relationships, thereby minimizing the "training" requirements and interservice coordination problems characteristic of many joint tests. On the other hand, the JTCGs had strongly held views about the objectives and methodology of V/L assessment which were frequently at odds with those of the OSD program manager.

While the aircraft JTCG was able to negotiate an acceptable compromise strategy, the armor JTCG was not. This set the stage for much subsequent conflict and delay. In the view of at least one outside expert, tasking JLF to the JTCG's was a mistake because their attachment to their vulnerability models made it difficult for them to plan and conduct objective full-up tests.

Limited Service Responsibility

A March, 1984, OSD memo stated explicitly that "no unique service support will be required" for JLF. The interpretation of this by the JLF program managers was that the services were under no obligation to provide targets or related support. JLF's budget was to cover the cost of replicas and munitions, but <u>not</u> costs of:

- actual or surrogate armor and aircraft systems.
- transport of actual or surrogate systems.
- target restoration.
- modification of vulnerability methodologies.

As such, neither OSD nor the services were assigned responsibility for these costs. In other words, no one was.

Exclusion of Ships

The Navy claimed that they had already conducted live fire tests in 1969 and 1970, with WW II hulks and Harpoon missiles. The OSD program manager did not believe those tests obviated the need for more testing on ships, but told us he did not have sufficient time to work with the ship community, in addition to the armor and aircraft communities, to bring them on board by a spring, 1984, deadline. Consequently, ships were excluded from the program.

Program Organization

Figure 2.1 is a JLF organization chart. Additionally, OSD contracted with the Institute for Defense Analyses (IDA) to monitor JLF and provide technical support. IDA issued their first report on JLF in March, 1986.

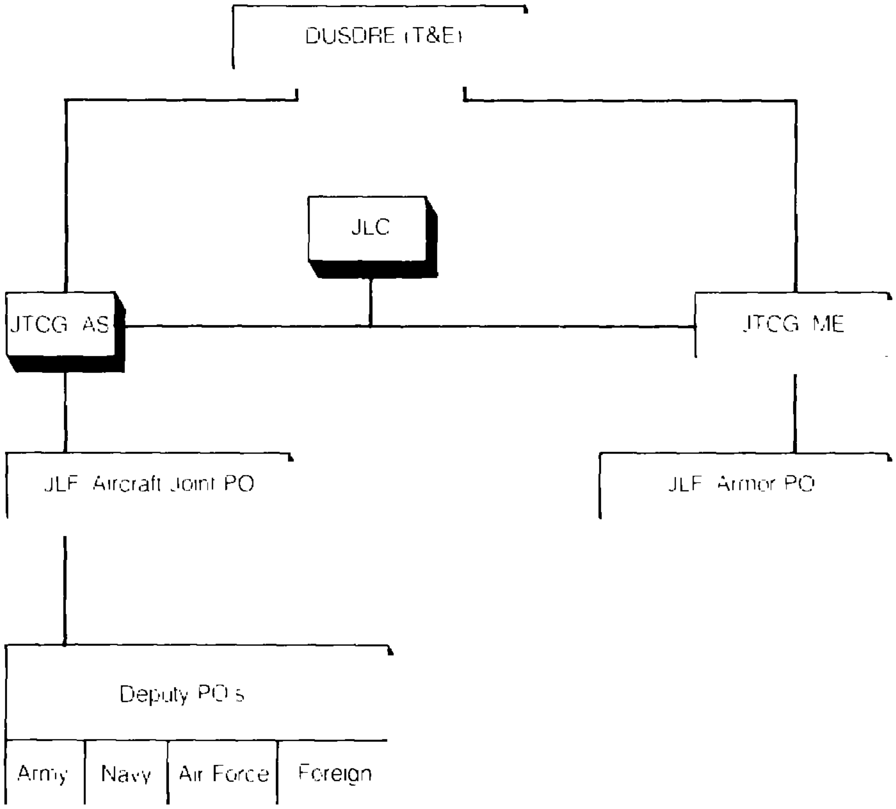
Current Status of Tests

JLF and related live fire testing in DOD has been marked by delays, interruptions, and removals of systems from the program. We describe the status of tests separately for JLF/Armor and JLF/Aircraft.

JLF/Armor

There have been substantial delays and changes in the JLF/Armor schedule. Table 2.1 shows the schedule as of January, 1985, and the revised schedule as of October, 1986.

Figure 2.1: JLF Organization Chart



DUSDRE (T&E) Deputy Undersecretary of Defense for Research and Engineering Test and Evaluation (formerly, Director Defense Test and Evaluation)

JLC Joint Logistics Command

JTCG AS Joint Technical Coordinating Group Aircraft Survivability

JTCG ME Joint Technical Coordinating Group Munitions Effectiveness

PO Program Office

Table 2.1: Initial Schedule, Revised Schedule, and Reported Reasons for Changes, for JLF/Armor

Test	Initial reporting date (1/85)	Revised reporting date	Reported reasons for changes
[material deleted]	4QFY85	2QFY90	Target unavailable test design controversy
[material deleted]	2QFY87	4QFY90	Split into 4 series by munition type
U.S. M113 APC	4QFY86	[a]	Removed by Army
U.S. M60 Tank	4QFY86	4QFY90	Test design controversy
[material deleted]	3QFY86	4QFY89	Test design controversy
U.S. M1 Tank	2QFY88	[a]	Removed by Army test design controversy
U.S. M1E1 Tank	2QFY89	[a]	Removed by Army; test design controversy
[material deleted]	4QFY88	[b]	Unanticipated expense of live fire tests
[material deleted]	1QFY88	[a]	Target unavailable
[material deleted]	2QFY89	[b]	Unanticipated expense of live fire tests
[material deleted]	2QFY89	4QFY90	Target unavailable, test design controversy
[material deleted]	4QFY89	4QFY90	Target unavailable: test design controversy
U S. LVTP-7 Amphibious Assault Vehicle	4QFY89	4QFY89	n/a
U S LAV Light Armored Vehicle	4QFY89	3QFY89	Removal of other tests

[a] Not yet scheduled

[b] Dropped from program; will not be tested

Sources JLF/Armor January 1985 Plan; JLF/Armor October 1986 Draft Revised Plan Interviews with test officials

Schedule Changes

- FY 86 was to have been the second year of JLF, the first for a full schedule of tests after modest beginnings in FY85. In fact, no shots were fired at armored vehicles in FY86 within JLF.
- No detailed test plans were approved by OSD during FY86.
- The Army removed some systems from JLF in order to conduct the tests themselves, and others have been dropped altogether.
- The type of tests (see Table 1.1) proposed has changed repeatedly, from inert tests including surrogates for threat vehicles, to exclusively full-up tests, and back, as conflicts between JTCG ME and OSD over testing philosophy caused delays and uneasy compromises, and as direction from OSD changed with changes in personnel.
- Soviet target vehicles originally thought to be among the easiest to obtain were not obtained in time to meet the original schedule.
- The Army Chief of Staff suspended the Bradley testing in April, 1986; it was not resumed until October.
- Testing of the M1 tank was put on hold, awaiting completion of the Bradley tests, so as to reflect changes in the Bradley test methods.

- The current JLF schedule is much reduced from the one proposed as recently as January 1985. Only six distinct vehicles of the twelve originally scheduled will be tested within JLF before the scheduled end of the program in FY 90, and although a number of munitions have been added, the total number of munition types to be tested has been reduced.

When JLF was initiated in 1984, the Bradley vehicle, the M1 Abrams tank, and the M113 armored personnel carrier were part of the program. Initial JLF/Armor efforts were focused on planning the Bradley tests. Shortly thereafter, however, the Bradley was removed from JLF, by agreement between the Army and OSD. The reasons given were that the Army could test the Bradley earlier than the JLF program could and on a larger scale. The Army could also provide target vehicles. Subsequently, the Army also pulled the M1 and M1E1 (now M1A1) versions of the Abrams tank, as well as the M113, out of the program, giving the same reasons. The JLF/Armor program manager predicted that the Army will eventually pull out the M60A3 tank as well. OSD retained oversight responsibility for these tests.

Tests conducted by the Army on the Bradley have included comparison tests between the M113 and [material deleted] vehicle, but these were not the full test series contemplated by the original JLF planners. The only test series completed within JLF was primarily a training exercise for damage assessors conducted in September of 1985 on a single training version of the M-48 tank and a previously tested [material deleted] hulk. This test does not appear in any JLF schedules and is not mentioned in the draft of the revised plan. The first of the originally scheduled JLF/Armor tests began in January, 1987.

Reasons for Delays

We have identified three reasons for the delays in the JLF/Armor testing.

1) Controversy over testing methodology. The Bradley tests were to serve as a methodological model for JLF/Armor tests. When questions were raised by the former OSD JLF program manager about the Army's Phase I Bradley live fire results reported to Congress in December 1985 and the conduct of Phase II Bradley tests in the Spring of 1986, the Army stopped all testing on the Bradley. OSD and the Army at first appeared to reach a compromise on methods for selecting shots and test conditions, but in May of 1986 directed that two independent panels of experts provide guidance to the Army on methodology for live fire testing. While the Army waited for the recommendations of the advisory

panels, which were to be incorporated in the revised Phase II test plan, no shots were fired.

The M1/M1A1 tank test plan was frozen in draft form, to be revised to reflect changes in the Bradley test methods. Similarly, plans for testing the [material deleted] were suspended pending resolution of disagreements between OSD and JTCG ME.

2) Unanticipated problems in obtaining targets. A major cooperative arrangement with a foreign nation was to have been completed during the first months of JLF. The JLF schedule had depended on this agreement to provide a large number of [material deleted] for testing in JLF. When the agreement collapsed the OSD Program manager sought to locate replacements, without immediate success.

The test conducted on the M-48 tank and single [material deleted] hulk in the Fall of 1985 was a stop-gap measure. It exploited targets that were available, though not the ones originally scheduled for JLF testing.

3) Funding cuts. In addition, FY 86 funds were delayed and reduced by 20 percent by OSD as of March 86.

JLF/Aircraft

Unlike JLF/Armor, JLF/Aircraft has planned and implemented their test program without major interruption, yet not without delays. Table 2.2 shows the initially scheduled reporting date, the revised reporting date, and the reasons for slippage, for all FY85 and FY86 tests.

Table 2.2: Initial Reporting Date, Revised Reporting Date, and Reported Reasons for Changes, for JLF/Aircraft FY85 and FY86 Tests

Test	Initial reporting date (10/84)	Revised reporting date	Reported reasons for change
F-15/16 Engine Steady State Fuel Ingestion	6/85	1/86[a]	Data reduction facility overloaded hardware modification required
UH-60 Tailboom Hydraulic System	8/85	4/87	Hydraulic tubing unavailable; testing and reporting resources diverted to non-JLF test, reporting time underestimated
F-15/16 Hydraulic Fluids	2/86	5/87	AF requested trade-in of airflow engines replacements were missing parts; 2 fluids added, facility fire
F-15/16 Engine Quick Dump Fuel Ingestion	4/86	4/87	Test required invention of new injection device AF requested capacity increase, test personnel diverted to non-JLF test, reporting time underestimated
F/A-18 Engine Rotating Core	5/86	12/87	Non-operational engine unavailable, test deferred
UH-60 Main Rotor Blade	7/86	6/87	Reporting time underestimated
UH-60 Engine Controlled Damage	8/86	4/87	1st attempt to create hydraulic load unsuccessful, subsequent use of a load absorbing pump delayed by design problems; reporting time underestimated
UH-60 Main Rotor Flight Controls	9/86	6/87	Prototype servos discovered to be inadequate; new servos arrived disassembled with parts missing; complete set of flight controls still unavailable
F-16 Emergency Power System Hydrazine Tan	9/86 k	7/87	Facility fire
AV-8B Flight Controls Mechanical Component	11/86 s	5/87	Aircraft arrived 6/86, support equipment arrived 10/86 test personnel diverted to non-JLF test, reporting time underestimated
AV-8B Flight Controls Reactive Control System	12/86	5/87	Aircraft arrived 6/86, support equipment arrived 10/86, test personnel diverted to non-JLF test reporting time underestimated
UH-60 Engine Compartment Fires	3/88	9/88	Moved to FY87 due to unavailability of engine compartments fuselage, and flight deck

[a] The report was reviewed and revised for an additional 6 months before reaching final form This was attributed to additional coordination required because a Navy test agency had tested an Air Force engine The additional coordination time was not considered in the initial planning

Sources JLF/Aircraft October, 1984 Master Plan JLF/Aircraft FY86 Detailed Test Plans Interviews with test officials

Schedule Changes

FY85 and FY86 testing was initially scheduled to result in 9 reports by the end of FY86; however, only 1 was completed and it only in draft form. If the revised schedule is met, the average delay will have been almost 11 months.

Reasons for Delay

The principal delaying factor has been lack of test targets. Some other factors, such as diversion of test personnel and facilities, are in part secondary effects of the lack of targets; the personnel and facilities are reallocated to other tests, and may not be immediately available when the awaited test targets finally arrive. JLF/Aircraft test officials anticipate further delays in FY87 testing and beyond for the same reasons. An FY87 funding cut of 33 percent and anticipated Gramm-Rudman-Hollings cuts are also expected to cause delays in FY87-FY90 testing. Several test reports are behind schedule because the time required for report writing was underestimated.

Differential Progress of JLF/Armor and JLF/Aircraft

As is evident from the above discussion, JLF/Armor has had considerably more difficulty implementing their program than JLF/Aircraft:

- The JLF/Aircraft plan was approved by the former OSD Program manager in 1984. A 1984 JLF/Armor plan was rejected.
- A revised JLF/Armor plan, produced in January, 1985, was approved by OSD but approval was later rescinded.
- A new revised plan was tentatively approved in Nov., 1986 by the current OSD program manager pending revisions, but final approval had not been obtained as of March, 1987.

According to a memo from the Principal Deputy USDRE, the original armor plan was inconsistent with the objectives of JLF. The primary objection was that major contributors to armored vehicle vulnerability—specifically fuel, ammunition, and hydraulic fluid—were not present in the majority of tests. Instead, the plan emphasized tests on inert targets to "characterize" warheads and assess behind armor effects, so as to perform final vulnerability assessments by computer modeling.

By contrast, the aircraft plan was described as responsive to the program objectives, because its planned replica and component tests included fuel, ammunition, and hydraulic fluid, to be followed by "sufficient" full-up, full-scale aircraft tests to validate component and replica results.

A January, 1985 revision of the JLF/Armor plan was approved by OSD. This plan was in effect for over a year, after which the incoming DUSDRE(T&E) (formerly, DDT&E) required another revision, at the urging of the former OSD program manager. He also required a revision of the [material deleted], the one JLF/Armor test for which a formal plan had been completed. Meanwhile, JLF/Aircraft has proceeded, essentially implementing their original 1984 proposal.

The JLF/Armor program manager told us he was "dictated to almost daily" by the former OSD program manager, while JLF/Aircraft was left alone. In his view, their initial proposals were essentially no different:

- Both had similar phasing logic, with a controlled progression up to full-scale firings.
- Both emphasized component, or "off-line" tests.
- Both relied on replicas and surrogates.

In fact, both armor and aircraft officials had serious differences of opinion with the former OSD program manager as JLF was being planned. According to JLF/Aircraft officials, he had initially wanted to forego the component, replica, and surrogate testing and concentrate on full-up, full-scale targets from the start. They report persuading him that this would not be feasible from a cost and availability standpoint, and that little would be learned by destroying aircraft without first studying the components. However, this is essentially the same argument made—unsuccessfully—by JLF/Armor officials.

The former OSD program manager admitted he had been harder on JLF/Armor but with justification. Essentially, he trusted JLF/Aircraft and did not trust JLF/Armor. His reasons were:

- JLF/Aircraft's proposal focused on fire and explosions; JLF/Armor's did not.
- JLF/Aircraft component tests were primarily full-up; JLF/Armor's were not.
- JLF/Armor's program logic was dominated by models; JLF/Aircraft's was not.
- JLF/Aircraft had a track record with TEAS; JLF/Armor did not.

As a result, JLF/Aircraft was able to negotiate an acceptable approach with the former OSD program manager while JLF/Armor was not. The consequences are apparent in the differential progress made by the two components. The mistrust problem appears to have been solved by replacing the OSD program manager, but the differential progress persists.

Conclusions

In this chapter, we addressed the evaluation question, "What is the status of each system and munition originally scheduled for live fire testing?" Our conclusions follow.

Development of JLF

- The services' original response to the proposal for JLF was unenthusiastic.
- The arrangement worked out by OSD for conducting JLF with the JTCGs avoided some problems characteristic of joint tests, but also contributed to dissension over the objectives of JLF and led to many implementation difficulties.
- The arrangement did not adequately designate budgetary responsibilities. The services were not responsible for supplying targets or related support, nor were these covered under JLF's budget.

Current Status of Tests

JLF/Armor

- There have been major slippages in the JLF/Armor test schedule.
- Of the Army's four vehicle types initially in JLF, three have been removed (Bradley vehicle, M1/M1A1 tanks, and M113 APC). Only the M60A3 tank remains, and it too may be pulled out.
- The Army's Bradley testing resumed in October, 1986, after a six-month suspension.
- Prolonged controversy between OSD and the armor testers over the purposes of JLF and appropriate methods for conducting and analyzing tests has pushed back the overall JLF/armor schedule. The first of the originally scheduled JLF/Armor tests began in January, 1987, almost two years behind schedule.
- JLF/armor has been hampered by greater difficulty than was anticipated in acquiring target vehicles, especially [material deleted].
- Even with no further schedule slippages or problems in obtaining targets, only half (six of twelve) of the originally scheduled armor/anti-armor tests would be completed during the term of JLF.

JLF/Aircraft

- In contrast to JLF/Armor. JLF/Aircraft has planned and implemented their program without major conflict or interruption.
- The schedule has been delayed, but less severely than JLF/Armor.
- Target availability has been the principal problem.

Differential Progress of JLF/ Armor and JLF/Aircraft

- JLF/Aircraft's initial proposals (high emphasis on fire and explosion and full-up shots, low emphasis on computerized V/L modeling) were compatible with OSD's interpretation of JLF program objectives, while JLF/ Armor's proposals (low emphasis on fire and explosion and full-up shots, high emphasis on modeling) were not. The differences between OSD and JLF/Armor were fundamental and never satisfactorily resolved, contributing to a relationship of mutual mistrust between JTCG ME officials and the former OSD program manager. The mistrust problem appears to have been solved by replacing the OSD program manager, but the disparity in progress between the two components in implementing their programs has continued.

General

- Target availability is expected to remain a problem for both JLF/Armor and JLF/Aircraft. Recent live fire legislation requires the services to provide targets for testing new systems, but this has no impact on the fielded systems in JLF.
- Both JLF/Armor and JLF/Aircraft have suffered budget cuts from OSD. These ranged from 20 percent to 33 percent.

What Has Been the Methodological Quality of the Test and Evaluation Process?

In this chapter, we review the overall JLF planning and the few detailed test plans (DTPS) and draft reports that exist. Because JLF is in an early stage, we could make only a limited assessment of the methodological quality and realism of tests at this time. We realize that testing to date may not be representative of subsequent testing; our emphasis is on identifying methodological issues of potential concern for the remainder of JLF and for future live fire testing programs.

The reasons that limited information is available for assessing JLF test quality are as follows:

For JLF/Armor,

- None of the originally scheduled JLF/Armor tests have yet been completed within the formal JLF structure. The only armor tests completed within JLF are: 1) an unscheduled training exercise and methodological demonstration, in which four shots were fired at an old U.S. M-48 tank and three at a [material deleted] hulk; and 2) a series of 10 shots fired against the [material deleted] vehicle. The [material deleted] shots were intended only for comparison with the full-up Phase I Bradley shots conducted by the Army. They were not the complete series of tests listed in the initial JLF schedule.
- The detailed plans and JLF reports for these tests exist only in preliminary draft form.
- The only completed DTP for a JLF/Armor test in the initial schedule was for the [material deleted]. This plan exists only in several draft versions that had not been given final acceptance by OSD during our review.
- A draft DTP for the M1 Abrams tank was written, but is being revised to reflect the approach taken in the Bradley Phase II plan.

For JLF/Aircraft,

- No full-up full-scale testing is scheduled before FY1988. Consequently, no full-up testing was conducted during our time frame. Since JLF/Aircraft does not publish a DTP prior to the fiscal year in which the test will be conducted, DTPs for full-up tests were also unavailable.
- Though FY85 and FY86 testing has proceeded, only one draft report was produced as of December, 1986. Consequently, we have only limited knowledge—primarily from discussions with test officials—of implementation, analyses, and results.

After discussing the overall program objectives covering both components of the program, we review JLF/Armor and JLF/Aircraft separatel*y*

These sections are followed by a more in-depth discussion of the key general issues identified during our evaluation, and comparisons with past and foreign live fire testing programs.

Program Objectives

The objectives of JLF and live fire testing in general have not been consistently stated by all the involved participants. (We discuss this in detail later in the chapter). At the working level, however, the JTCG's for JLF/Armor and JLF/Aircraft attempted to unify the program by specifying a common set of objectives. These were to:

1) Gather empirical data on the vulnerability of U.S. systems to [material deleted] weapons and the lethality of U.S. systems against [material deleted] targets.

2) Develop insights into design changes necessary to reduce vulnerabilities and increase lethalities.

3) Enhance the data base available for battle damage assessment and repair.

4) Use test data and results to validate (calibrate) lethality and vulnerability models.

The only differences between JLF/Armor and JLF/Aircraft were:

- In objective 4, "validate" was used by JLF/Aircraft while "calibrate" was used by JLF/Armor. We attach no particular significance to this difference because the terms appear to be used interchangeably in the V L community
- The objectives were described as in order of priority by JLF/Aircraft; no priority was assigned to them by JLF/Armor.

For our purposes, the first three objectives are not stated in an evaluable way. "Gather empirical data" on vulnerability and lethality, "Develop insights" into design changes, and "Enhance the data base" for battle damage assessment and repair, will all be accomplished regardless of the methodological quality, realism, cost-effectiveness, or usefulness of the program. There are no specified comparisons to be made or criteria to be met, only a statement that the state of knowledge on the vulnerability or lethality of weapon systems will somehow be improved. In contrast, the fourth objective—validate (or calibrate) lethality and vulnerability models—has at least an implied criterion to be met (validation by test data).

JTCG Objective 3 exploits a side benefit of live fire testing in that damaged systems are produced as a by-product of the testing process. While battle damage assessment and repair is clearly an important function for DOD, it is outside our main evaluation focus, which is the methodological quality of the test and evaluation process. Therefore, we do not address Objective 3 in this review.

JLF/Armor

For JLF/Armor, we address overall planning, setting test objectives, test planning, implementation, and analysis and reporting. Our review included any planning and related information on all proposed tests, as well as all available reports on completed tests. Table 3.1 identifies the individual tests we reviewed and Table 3.2 identifies the principal documents we used. Individual tests are treated in more detail in Appendix II.

Table 3.1: Live Fire Tests Individually Reviewed

Test	Munitions	Status
Armor		
[material deleted] (March 1986 Plan)	AT-4 Antitank Weapon (12)[a] LAW Light Antitank Weapon (12) TOW II Missile (14) Basic TOW (6) BLU-97 Bomblet (10) M42 Grenade (9) M718 Mine (8) M833 105mm Projectile (12) M392 105mm (4)	Plan not approved by OSD
U.S. M-48 Tank	Steel Long-Rod Penetrator (Foreign simulant of Soviet 115mm projectile) (4) TOW Missiles (3)	Testing completed, preliminary draft report prepared
[material deleted]	LAWs (6) 120 mm HEAT (1)	Testing completed: preliminary draft report prepared
U.S Bradley Vehicle Phase I	[material deleted] TOW Missiles (14) 120mm HEAT (2) Rockeye Bomblets(7) M718 Mine (6) 30mm (UK) (5) 3.2" HEAT (15)	Testing completed by army, reported to Congress
U.S Bradley Vehicle Phase II	[material deleted] TOW2 Missiles (7) TOW (2) 30mm (UK) (8) 120mm HEAT (4) Mine (1)	Testing interrupted: resumed after preparation of detailed test plan
Aircraft		
U.S. F-15/F16 Engine Steady-State Fuel Ingestion	Munitions not used	Testing completed, draft report prepared

[a] ( ) = Number of shots with each munition type

Table 3.2: Main Documents Used in Reviewing the Quality of Live Fire Tests

Program component	Documents
Overall JLF Program	JLF Charter, March 27, 1984
	IDA March 1986 JLF Report
JLF/Armor	January 1985 Plan
	October 1986 Revised Plan (Draft)
JLF/Aircraft	February 1984 Preliminary Plan
	October 1984 Master Plan
	FY86 Detailed Test Plans
[material deleted]	Detailed test plan (in January 1985 JLF/Armor Test Plan)
	March 1986 Detailed Test Plan
	Outline Test Plan (in October 1986 Revised JLF/Armor Plan)
	IDA March 1986 JLF Report
U S M-48 Training Test	Detailed Test Plan
	Preliminary Draft JLF Report
	IDA March 1986 JLF Report
[material deleted]	Detailed Test Plan
	Preliminary Draft JLF Report
	Bradley Phase I December 1985 BRL Report
U S Bradley Fighting Vehicle Phase I Test	December 1985 BRL Report
	OSD Program Manager s December 1985 Report to Congress
	GAO Report of March 1986
	HASC Staff Report
	BAST June 1986 Report
U S Bradley Fighting Vehicle Phase II Test	BAST October 1986 Report
	October 1986 Detailed Test Plan
U S F-15/F-16 Engine Steady-State Fuel Ingestion Test	Detailed Test Plan
	January 1986 Draft JLF Report

Overall Planning

The first JLF/Armor plan (submitted in 1984) was rejected by OSD. According to a memo from the Principal Deputy USDRE, it was inconsistent with the objectives of JLF. The primary objection was that major contributors to armored vehicle vulnerability—specifically fuel, ammunition, and hydraulic fluid—were not present in the majority of tests. Instead, the plan emphasized tests on inert targets to "characterize warheads and assess behind-armor effects," so as to perform final vulnerability assessments by computer modeling.

The second plan, published in January, 1985, was accepted by OSD. This plan focused more on vehicle tests, with only occasional indications that warhead characterization and behind-armor effects tests would also have to be conducted (the one DTP contained in the plan did not propose any such tests). There were to be twelve series of tests, organized around eleven distinct armored vehicles, [material deleted] (In addition, [material deleted] were to serve as surrogates for [material deleted].) Thirty-six different munition types were proposed for testing. It was estimated that there would be between 1,370 and 1,870 total shots fired in the program. In addition to the four formally stated JLF objectives, the early tests were also to result in new damage criteria and an updated version of the Standard Damage Assessment List (SDAL) used to quantify observed damage to armored vehicles. The plan contained the first version of a DTP for the [material deleted], which was to be the first JLF/Armor test.

An October 1986 draft revision of the JLF/Armor plan retreats from the position that JLF will accomplish all four JLF objectives.[1] Acknowledging that the varied needs of the users of live fire test data prevent any one test series from fully addressing all concerns, it proposes that the JLF/Armor tests therefore focus on the empirical representation of the lethality and vulnerability of U.S. weapons (Objective 1). It includes outline test plans organized around Soviet and U.S. armored vehicles. These specify the number of each type of vehicle and munition required, and the approximate time span required for planning, implementing, analyzing, and reporting each test. Actual shotlines—that is, the projected path of each shot—are not specified, but the plan indicates that each test will employ a mixture of systematically selected and randomly selected shots. There are descriptions of test setups and instrumentation and discussions of the uses to be made of the data.

The October 1986 draft plan acknowledges that experimental validation of sophisticated vulnerability models would be prohibitively expensive, and will not be accomplished by JLF "in a rigorous mathematical or statistical sense." Instead, the test data will be used to increase confidence in models or to suggest improvements in them.

In important respects, the October 1986 draft revised JLF/Armor plan resembles the first plan produced by JTCG ME for JLF in 1984, which had been rejected by OSD. Specifically:

[1] As of March 1987, this plan had still not been accepted by the OSD program manager. He cited two reasons: lack of an evaluation plan and inadequate procedures for assessing crew effects He cited no disagreement with the general direction of the plan

- Approximately two-thirds of the shots will be "warhead characterization" or "behind-armor debris" studies involving the interaction of munitions with armor plate targets rather than shots at vehicles.
- Target condition is mostly inert or semi-inert; only one-third of the shots will be on vehicles, and only 20 percent of these will be full-up; in 80 percent of the shots the vehicles will be loaded with dummy major-caliber ammunition or without fuel in their tanks.

The draft plan is also more similar than the earlier plan to the JLF/Aircraft DTPs in its use of components and replicas, e.g., a test of a "fuel cell in a steel box, the equivalent of a surrogate vehicle."

Two additional topics fall under overall planning rather than individual test planning: selection of target vehicles and selection of munitions for inclusion in JLF testing.

Selection of Vehicles

The January, 1985, plan states that the selection of armored vehicles to be tested, while subjective, was based on military value, critical data deficiencies, target cost, test cost, and the projected worth of the data. Without more detail about the process, it is not possible to assess the strengths or limitations of vehicle selection. The plan acknowledged that availability of equipment might prove to be the limiting factor in implementing the tests. The October 1986 draft revised plan contains no overall rationale for the selection of vehicles, but each of the outline test plans discusses the interest in the particular target vehicle that is to be tested. Some problems with target availability are still anticipated. Overall, only 6 of the 30 vehicles that will be required for the tests outlined in the plan were on hand as of October, 1986.

Selection of Munitions

The January 1985 JLF/Armor plan states that the anti-armor munitions selected were as exhaustive a list as time and money would allow. They were required to be in the current inventory and to have been designed with the objective of defeating armor. The plan included a matrix which displayed the munitions proposed for testing versus the target vehicles in the program. The range of munitions chosen appears to be realistic.

The draft revised plan of October, 1986, introduces a criterion for excluding munitions such as the Hellfire missile (originally included) because they are large "overmatches" for currently fielded armored vehicles. It also excludes obvious "undermatches", which would have little chance of producing significant damage to a particular vehicle. The stated rationale is that testing a munition that cannot be prevented from penetrating or will clearly not penetrate a vehicle is a waste of scarce and costly resources. We were told that more detail on the rationale for including particular munitions would require plans to be classified.

Setting Test Objectives

The outline test plans in the October 1986 draft revision of the JLF/Armor plan each list several of the overall JLF objectives as major test objectives. Each test outline lists a specific objective as well. Some of these may be infeasible goals for the tests as described. For example: one test proposes to determine the likelihood that an uncontrollable fire will result from penetration into the diesel fuel stowage cells of armored vehicles. But only one vehicle with replaceable fuel cells and armor panels will be tested. Although the setup may produce reliable results for the particular configuration tested, the generalizability of the results to other armored vehicles is questionable. The objectives of the Bradley Phase II plan are more specific.

The JLF/Armor outline test plans are primarily focused on the objective of quantifying the vulnerability of armored vehicles and the lethality of anti-armor munitions (that is, JTCG Objective 1). They propose to accomplish this by using the results of numerous "off-line," or subscale tests of armor-warhead interaction, behind-armor debris, and component tests as input to computerized vulnerability models. Concern about target availability and loss of information in catastrophic events has led planners to propose that only 20 percent of the shots on vehicles involve full-up targets. With respect to model validation (JTCG Objective 4), the results of the tests will be used informally to improve or update models, but the plan acknowledges that statistically rigorous model validation will not be pursued in JLF/Armor because it is not economically feasible.

The objective of developing design insights for vulnerability reduction and lethality improvement (JTCG Objective 2) is not addressed by the plans in the most direct way possible: There were no plans in JLF as of December 1986 for comparative tests of proposed vulnerability fixes on full-scale U.S. systems, like the Army's Bradley Phase II tests. There are plans, though, to conduct comparative tests of the effects of radiation liners and applique armors on the vulnerability of [material deleted] and of the applique armor to be added to the M60A3. Because only fielded anti-armor munitions will be tested in JLF, the use of the tests for the improvement of lethality is toward future munitions.

Rather than directly reflecting the JTCG statement of program objectives, draft versions of other plans contained objectives that reflect methodological questions whose answers could aid in the design or conduct of live fire tests:

- One version of the [material deleted] plan proposed to test the differences between static firings (munitions fixed to the target) and dynamic firings (munitions fired from a distance) of shaped charges.
- The M-48 test was intended to train damage assessors.
- Inert M-48 shots were compared to full-up shots on a [material deleted] hulk and these were treated as a methodological comparison.

A main objective of one test conducted, that of training damage assessors during the M-48 test, was unrealistic in the time available. The brief course conducted at the test site did not succeed in training the damage assessment teams to fill out forms with acceptable accuracy or consistency across assessors. The training consisted of only a few classroom sessions supplemented by discussions with an instructor between the shots. One tester suggested that a full year's experience might be necessary to produce an acceptable level of competence in damage assessment.

Test Planning

Test designs in the outline plans are generally congruent with test objectives. Newer plans such as the Bradley Phase II place more emphasis on estimation of casualties and the effects of fire and explosion than previous Bradley or JLF/Armor plans.

The outline plans in the October 1986 draft revision of the JLF/Armor plan are more realistic in providing lead time for the acquisition of targets. However, some have still specified targets that may not become available. For example, the planned test of the Marine Corps' Light Armored Vehicle (LAV) is designed to require a minimum of two prototypes and a ballistic hull, but only one vehicle has been obtained, and the JLF/Armor program manager described the prospect of obtaining others as "dim."

Design Efficiency

Tests are designed to be efficient with respect to conservation of target resources. For example:

- Inert testing of vehicles will precede full-up testing.

- Effects on volatile components will be determined prior to placing those components in vehicles.
- Shots will be sequenced to stave off catastrophic damage.

The newer plans pay considerable attention to target preservation and the sequence of shots, because target availability has been such a persistent problem.

Shot Selection

Shot selection has been a point of controversy in the planning of JLF/ Armor tests as well as in the Army's Bradley tests. The concern was that judgmental selection of shots could bias the results, directing a disproportionate number of shots at less vulnerable areas, and thereby make the vehicle appear less vulnerable than it actually was. Earlier draft JLF plans also relied on judgment for the selection of shots. The March 1986 [material deleted] DTP, for example, proposed the selection of shotlines with azimuths at 30 degree intervals around the vehicle and impact points based on engineering judgments about particular components. We found no evidence of intentional bias in this proposal. For example:

- In the March 1986 [material deleted] DTP, no shots with the LAW or AT-4 light infantry antiarmor weapons were planned to impact the lower frontal glacis plate or the front of the turret because previous tests indicated that an infantryman would have little or no chance of killing the [material deleted] with shots in these locations. Firing most test shots at the more vulnerable sides of the tank would constitute bias only if the results were generalized to the vulnerability of the tank as a whole, and
- the October 1986 draft plan clearly indicates that any randomly selected shots not fired because their effects are claimed to be known will nonetheless be incorporated into the analysis. Shots of certain munitions into main gun ammunition, for example, will not be fired, but still scored as catastrophic kills.

The newer plans, including the outline plans in the October 1986 draft plan, incorporate some of the changes introduced by the Bradley controversy:

- The use of a method of random shot selection from combat distributions—that is, data showing where vehicles were hit in combat—proposed by the Board on Army Science and Technology (BAST). This is a change from the judgmental shot selection proposed in some earlier drafts of DTPs. It will improve the combat representativeness of the shot selections and reduce the opportunity for biases to enter the shot selection process.

- There are proposals to combine center-of-mass aimpoints with dynamic firings from combat realistic ranges. The resulting dispersion of impact points is expected to resemble combat distributions, but this is open to question.

However, the Bradley Phase II plan misstates the position of the BAST report on shot selection methodology as recommending the minimum number of shots required per munition type to establish with reasonable confidence that observed vulnerability differences between specified test targets are true differences. In fact BAST explicitly stated that the number of live fire shots required for reliable vulnerability assessment is an open question, and their selection of 20 shots was dictated by the OSD request, not by statistical considerations.[2] In addition, the plan states that BAST selected the specific shotlines to be used in the test, but the BAST report and its cover letter make it clear that BAST was only suggesting an interim method for selecting shotlines. BAST specifically stated that the shotlines appended to their report were the results of a "trial use" of the method, and did not constitute recommendations as to which shotlines should be used for the Bradley. They stated that the responsibility for choosing shotlines was the Army's. The Bradley Phase II plan stated that the BAST selections sufficed for the Army's goals, and the 20 BAST shots were therefore adopted without modification.

Omitted Information

The 1986 draft outline test plans omit some important information, such as the rationale for specific shotlines. However, more complete DTPS are to be produced prior to each test. It remains to be seen whether the JLF DTPS, when written, will adequately address these topics. An earlier draft DTP for the [material deleted] did contain detailed rationales for shotline selection based on engineering judgment.

The Army's Phase II Bradley plan is the most detailed and thoroughly specified of the live fire test plans we have reviewed. It omits little information about test procedures that could be required of a DTP. Its six volumes include:

- detailed descriptions of the procedures to be followed in all the subtests,
- predictions generated by vulnerability models for all proposed impacts,

[2] OSD had stated that 13 shots were available. BAST felt that 20 shots, distributed over four munitions, would be "more representative."

- detailed diagrams of the vehicle configuration and stowage plans and
- a detailed evaluation or analysis plan.

It removes test implementation decisions from the informal judgments of testers, specifying contingency plans for departures from planned procedures during the tests, and requiring that they be approved by OSD.

Statistical Validity

The statistical validity of live fire results has been problematic from the beginning of JLF. Recent plans are more explicit about the extent of the problem and contain at least some attempts to cope with it. For example,

- The January 1985 JLF Armor plan argues that costs and other constraints on the testing of complex full-scale systems would prohibit the collection of data sufficient to produce stable, reliable estimates of kill probabilities directly from live fire data. The sample sizes would be too small. This is the rationale for using live fire results to calibrate models, instead, and then to use the models to produce estimates of overall kill probability. There is, however, no discussion of the sample sizes that would be required to validate or calibrate V L models in the January 1985 plan.
- The October 1986 plan acknowledges the impossibility of experimentally validating models in a rigorous statistical sense with the numbers of shots possible in live fire tests.
- The Bradley Phase II plan specifies a statistical analysis of matched shots on two versions of the vehicle, even though its appropriateness is questionable.

Implementation

The JLF Armor tests have encountered difficulties in implementation resulting from disagreements over test design and target availability Four tests were originally scheduled for completion by the end of FY86. Two tests were completed, but these were not among those originally scheduled, and they were made possible by the existence of the Army's Bradley testing program and by using available but less-than-realistic targets.

Not all departures from test plans were justified in the draft test reports. For example, U.S. surrogates were substituted for the actual Soviet munitions that were to be loaded into the [material deleted] for the shots comparing its vulnerability with the Bradley's. Because the surrogate munitions may react more violently than the actual Soviet munitions the comparison may be misleading, although test officials argued that only one of the ten comparison shots could have been strongly affected by the differences.

In sum, the implementation problems of JLF Armor have been very serious, since of the twelve tests originally scheduled, none has been conducted within JLF.

Analysis and Reporting

The JLF reports on the [material deleted] and M-48 were still preliminary drafts as of December, 1986, although a summary report of the [material deleted] results was included in the Army's Bradley Phase I report that was sent to Congress in December, 1985. The JLF drafts each consist of a few pages of rough copy and more than a hundred pages of raw data in the form of photographs of damage and damage assessment forms. BRL personnel assigned $P_{k/h}$s to the four M-48 shots, but the results were not analyzed further. It is not possible to assess the way JLF data will be analyzed and reported on the basis of these preliminary drafts. We note that the results for the first two tests were still in this preliminary form more than one year after the tests were completed.

Although the draft JLF report on the [material deleted] tests acknowledged that the surrogate munitions may react more violently than the actual Soviet munitions, potentially biasing the Bradley [material deleted] comparison in favor of the Bradley, this is not mentioned in the report presented to Congress. Because the [material deleted] shots were intended to provide context for the perception of the Bradley's vulnerability, we believe that questions about the equivalence of the surrogates should have been reported.

The three shots against a burned-out [material deleted] hulk conducted at the time of the M-48 test have not been reported in a separate JLF report, but only as part of the IDA report on the first year of JLF. IDA concluded that there are tradeoffs between inert and full-up testing (recall that the M-48 was tested inert while the [material deleted] was tested full-up). Inert testing provides detailed information on behind-armor effects, while full-up testing provides unambiguous information on catastrophic kills. The IDA conclusion reflects the fact that all three of the [material deleted] shots led to catastrophic fire or explosion resulting from impacts on main gun ammunition.

We believe that empirical demonstrations of the consequences of adopting different methodologies are desirable, but this test was a questionable basis for a general conclusion about inert versus full-up testing:

- Although the two tanks are of the same general class, they were in very different states of repair and combat-readiness, apart from whether they were loaded with fuel and ammunition.
- The [material deleted] was missing most of its internal components. The presence of internal components can prevent some impacts on ammunition that cause catastrophic kills, and provide data on component damage by debris when catastrophic kills do not result. Although some components were simulated with sheet metal after the first shot, it is not known whether their masking effects were equivalent to actual components, and unambiguous assessment of damage to components would not have been possible even if catastrophic kills had been avoided.

The proper comparison for conclusions about the differences between inert and full-up tests is between vehicles of the same model differing only in whether they are loaded with fuel and ammunition.

JLF/Aircraft

For JLF/Aircraft, we again address overall planning, setting test objectives, test planning, implementation, and analysis and reporting. Our review includes any planning and related information on the complete universe of proposed tests, as well as all available reports on completed tests. Table 3.1 identifies the individual tests we reviewed and Table 3.2 identifies the principal documents we used. Individual tests are treated in more detail in Appendix II.

Overall Planning

The overall plans were published in a Preliminary Plan in February, 1984, and a Master Plan in October, 1984. The Preliminary Plan presented the general test concept and program logic, while the Master Plan documented the funding requirements, objectives, test approach, hardware and facility requirements, and schedule for each test, organized by aircraft system. In all, this encompassed 97 tests. The Master Plan closely followed the concept and logic of the Preliminary Plan. Both documents were clear and well organized.

The program was conceived in 6 phases:

- tri-service test plan development
- test preparation
- component testing
- replica/surrogate tests
- full-scale aircraft tests
- vulnerability reduction technology tests

There is some overlap between phases. but in general they were intended to proceed sequentially. As a consequence. full-up, full-scale testing comes late in the program.

Selection of Aircraft

Aircraft were selected based on the JLF objective to test the vulnerability of first line tactical aircraft. The selections were made to ensure a representative, tri-service cross section of currently employed aircraft, i.e., swing/swept wing, fixed and rotary wing, single and multi-engine, turbofan and turbojet. thrust vectoring, and metal and composite construction. The probability of obtaining a particular aircraft was also considered, but JLF officials report that no aircraft were actually excluded on the grounds that they could not be obtained. The selected aircraft are listed in Table 3.3.

Table 3.3: JLF/Aircraft Target Systems Selected for Testing

Source	System
United States	
Navy	F/A-18 AV-8B A-6E/F F-14
Air Force	F-16 F-15
Army	UH-60 AH-64
Foreign	[material deleted] [material deleted]

Source JLF/Aircraft February 1984 Preliminary Plan

All aircraft listed in the Preliminary Plan are still in the program except for the F-14. According to the deputy program manager (Navy), the F-14 was removed because JLF/Aircraft was over budget, and the Navy had four aircraft in the program compared to two each for the other services. We could not obtain a clear statement of the reason the F-14 in particular was removed; it does not appear to have been a lack of availability, as the F-14 is older and more plentiful than the F/A-18.

Non-tactical aircraft were never seriously considered. In part, this reflected the limited testing budget, and in part, the historically tactical focus of the aircraft survivability community as a whole. This focus would not normally involve vulnerability issues in, say, strategic bombers.

Selection of Subsystems

The testing program is organized around nine critical subsystems: fuel and dry bay, propulsion, flight control, hydraulics, structures, armament, crew station, rotor/drive train, and miscellaneous unique. A formal process was used to designate test priorities:

- 1st, the 10 target aircraft were crossed with the 9 subsystems to form a target/subsystem matrix.
- 2nd, officials rated their confidence in current vulnerability estimates based on "as installed" configuration testing, for each cell in the matrix.
- 3rd, they rated test priority, again for each cell in the matrix, in part based on the confidence ratings. Both confidence and priority were rated high, medium, or low.

We overlaid the matrix of FY85 and FY86 test selections on the confidence and test priority matrices, and saw no discernible relationship between the tests selected and either confidence or priority. When questioned about this, the JLF/Aircraft program manager stated that FY85 and FY86 test selections were actually driven by:

- availability of hardware (e.g., F100 engines were already in hand)
- the need to ensure tri-service interest and cooperation, and related bureaucratic concerns (e.g., the need to start testing as quickly as possible to show JLF/Armor that JLF/Aircraft was contributing to the program).

That is, it appears that these practical concerns took on greater importance than confidence levels and associated priorities.

Selection of Munitions

The Preliminary Plan states that specific threats applicable to both the U.S. and foreign systems were selected by a tri-service review. No further rationale was provided. They are listed in Table 3.4.

Table 3.4: JLF/Aircraft Munitions

Munition type	Size
Foreign threats to U.S. Aircraft	
Projectiles	12.7mm API 14.5mm API 23mm HEI and API 30mm HEI and API
Warhead Fragments	45 Grains 70 Grains 110 Grains 200 Grains
U.S. Threats to Foreign Aircraft	
Projectiles	7.62mm API 12.7mm API 20mm AP, API and HEI 25mm API and HEI 30mm HEI 40mm HEI
Warhead Fragments[a]	2-1000 Grains Steel 3.5-30 Grains Tungsten

[a] Nominal fragments representative of those produced by current surface-to-air and air-to-air missile warheads. Specific warheads represented by these grain sizes are classified.

Source: JLF/Aircraft February 1984 Preliminary Plan

Constraints on Realism

Live fire testing of aircraft is conducted on the ground. Consequently, there are inherent limitations to the realism of tests. Although JLF/Aircraft program planners devoted considerable attention to realism issues, both in terms of targets and test conditions (e.g., assuring the presence of appropriate combustibles), there are nonetheless several technical constraints on realism. Of the four discussed here, the first three reflect the difficulty of simulating flight conditions on the ground.

1) Limitations of airflow. Two test ranges used by JLF—China Lake and Wright-Patterson—have the capability to simulate the airflow conditions of a plane in flight. High speed airflow is considered essential for the realism of aircraft tests involving fire, whether component-level or full scale. It also affects the probability of sustaining a fire once one starts, and causes fires to take unexpected paths through the aircraft. While some JLF/Aircraft tests for which airflow would be warranted will not have it, in general there is an attempt to use airflow whenever possible.

However, current airflow facilities are limited in that they cannot blow air over an entire aircraft. Coverage is about 5 ft. in diameter. In a wing test, for example, the airflow does not cover enough of the wing to generate the appropriate lift, or the interplay between the loaded wing and the fuselage. Maximum velocity is also limited to about Mach .8, which is considerably below the top speed of fixed-wing tactical aircraft. A larger airflow facility with higher maximum velocity has been proposed for China Lake, but even if funded, it may not be built in time for the full-scale testing phase of JLF/Aircraft.

2) Other environmental limitations. While airspeed is believed to be the most critical, other environmental factors affect the probability of fire. However, these will not be simulated in JLF. They are:

- altitude (affects fuel vapor pressure)
- altitude history (affects fuel vapor composition and subsequent volatility)
- maneuver load (affects effective fuel weight and subsequent leak rate)
- slosh (affects vapor distribution).

The JLF/Aircraft Preliminary Plan stated that altitude, altitude changes, and slosh would be considered in setting test conditions, but according to JLF technical staff, there currently is no satisfactory capability to simulate these factors. The JLF/Aircraft program manager considers altitude simulation unnecessary because all tests are for air-to-ground missions, in which typical combat altitudes are low enough to be little different than sea level, practically speaking.

3) Restricted attack angles. The focus of JLF/Aircraft is ground-to-air fire. Though the aircraft are slightly elevated (on pads), shot angles greater than 45 degrees (typical in a ground-to-air scenario) are not possible.

As noted above, all three of these constraints stem from the difficulty of simulating flight conditions on the ground. More realistic flight conditions could be obtained with drone targets. However, this was not seriously considered. The stated reasons were:

- Costs would be prohibitive.
- Specified hit points would frequently be missed.
- Combat realism would still not be achieved.

4) Inability to use actual warheads. A fourth constraint on realism stems from a different problem, the inability to use actual missile warheads. All threats are actual munitions with the exception of missiles, the fragments of which are being simulated by metal cubes launched at the target.

Test officials justify not using actual warheads on grounds of:

- safety and security (only extremely remote test ranges can support the detonation of actual missile warheads)
- controllability (impact location and fragment size are harder to control, potentially leading to accidental target loss)
- cost (testing costs as well as materiel costs would be several times greater if actual warheads were used)
- availability (particularly for foreign warheads).

Nonetheless, IDA analysts and other experts have raised concerns over the realism of simulating warhead fragments with metal cubes. The JLF Aircraft program manager claims that the simulators give realistic results because the mass, velocity, impact orientation, and shape are all reasonably close to an actual warhead fragment. However, he was not aware of any direct comparison studies or other equivalence tests that would support the claim. As the use of cubes has become standard practice, and will likely be continued in live fire testing of aircraft throughout and beyond JLF, we believe it should first be validated in controlled comparison studies with actual warheads.

Setting Test Objectives

For the FY85 and FY86 DTPS,

- All thirteen specified test objectives.
- All were congruent with the program objectives as stated by JTCG.
- Most test objectives were feasible. The principal exceptions were objectives related to determining probabilities (e.g., $P_{k/h}$s) or validating vulnerability models. The available sample sizes simply do not permit this level of quantification of results. Furthermore, test officials generally designed the test matrices to maximize the range of threat target interactions. In order to generate credible probabilities or validate models, they would need to trade off the range of shots for more replications of shots.

The principal focus of the JLF Aircraft tests is gathering empirical data on the vulnerability of U.S. systems to Soviet weapons. The objective of developing design insights is not typically made explicit, but the F100 engine draft report suggests that developing design insights will also be important. None of the FY85 or FY86 tests include empirical comparisons of proposed vulnerability fixes for U.S. systems; however, vulnerability reduction technology tests are not scheduled until after the full-scale aircraft tests of FY89. If designed to do so, these tests could provide such comparisons.

Test Planning

Test designs in the FY85 and FY86 DTPs were generally congruent with test objectives. However, some DTPs specified target requirements which exceeded the availability of those targets. Some examples:

- The AV-8B flight control DTP required one AV-8B airframe or prototype. This was not then and is not now available, so an AV-8A will serve as a surrogate.
- The UH-60 engine compartment fire DTP required 8 ground-operable engines, but as of July, 1985, only two engines had been obtained. Fourteen months later (in September, 1986) two more engines had been obtained, but a fuselage and flight deck were still unavailable.
- The UH-60 flight control DTP required four sets of flight control components. As of September, 1986, less than one complete set had been obtained.

Some DTPs had contingency plans or fallback options (e.g., substituting surrogates or replicas), while in others, testing was simply postponed pending availability of the actual targets. Test officials believed the need for realism justified the postponement.

We address 5 test planning issues for JLF Aircraft: design efficiency, efforts to ensure realism, omitted information, inconsistency in selecting threat velocities, and statistical validity.

Design Efficiency

Since target availability is the principal problem facing test officials, a high priority is placed on conserving test articles through efficient test design. This is done in at least two ways: use of multiple outcome measures and control for testing effects.

An example of multiple outcome measures is the F-16 wing test (FY87), where the test set-up will be instrumented to measure 1) presence of fire, 2) structural effects, 3) amount of fuel leakage, 4) self-sealing capability, and 5) effectiveness of halon extinguishers, all from a single shot Testing effects are controlled, where possible, by replacing components between shots. For those which cannot be replaced, repair teams are used whenever possible. After a repair, stresses and other key factors are checked to ensure they have not changed. For known irreparable targets, shots are strategically sequenced to avoid premature catastrophic loss of the target, as well as to prevent testing effects. For example, it is well known for engine tests that shots into the turbine must come last, since such shots have a high likelihood of destroying the engine.

Efforts to Ensure Realism

A high priority was placed on ensuring realism on an objective-specific basis. The FY85 and FY86 tests are all component tests, and as such, their objectives are relatively limited. Consequently, the DTPs contain apparent departures from realism that test officials maintain are appropriate and efficient. For example, in the F-15/F-16 hydraulic fluids test:

- Projectiles were to be fired at realistic velocities for selected standoff distances. However, they were not to be fired from real weapons at realistic ranges, but from Mann barrels at close range to minimize hit dispersion. This was justified on the grounds that the hydraulic line was only .75 inches in diameter, and hit dispersion simply causes misses which waste time and ammunition. As the objective of the test is to learn what will happen when the line is hit, the higher level of realism is not considered appropriate.
- The plan specified pressurized hydraulic lines of the exact diameters, wall thicknesses, and material characteristics of those used in the F-15 and F-16. However, the structure to house the lines was a modified replica fuselage previously used in A-10 refueling tests. Again, as the test objective was simply to determine the incendiary effect of penetrating the line, use of the unrealistic replica was not considered to detract from the objective-specific realism of the test. All the testers needed was a "metal box" to contain the fire.

In this fashion, scarce and expensive resources are being conserved for the tests whose objective-specific realism requires them. For example, structural tests of an F-15 wing will necessarily require an actual F-15 wing.

Omitted Information

The DTPs omitted information we believe necessary for assessing how well a test plan meets its test objectives. Table 3.5 breaks this down by individual DTP. Note that:

- most plans omitted rationales for threat munitions and shotlines.
- all plans omitted data analysis plans and rationales for sample sizes.

Table 3.5: Inclusion of Selected Information in JLF/Aircraft FY85 and FY86 Detailed Test Plans

Test	Munitions specified	Rationale for munitions	Shotlines specified	Rationale for shotlines	Sample size specified	Rationale for sample size	Test matrix specified	Data analysis plan
F-15/16 Engine Steady State Fuel Ingestion	n a	n a.	X			n a.	X	
UH-60 Tailboom Hydraulic System	X		X			n a.		
F-15/16 Hydraulic Fluids	X		X		X		X	
F-15/16 Engine Quick Dump Fuel Ingestion	n. a	n a		n a		n. a	X	
F/A-18 Engine Rotating Core	X		X	X	X		X	
UH-60 Main Rotor Blade	X	X	X		X		X	
UH-60 Engine Controlled Damage		n a	X	X		n a		
UH-60 Main Rotor Flight Controls	X	X	X		X		X	
F-16 Emergency Power System Hydraulic Tank	X		X		X		X	
AV-8B Flight Controls Mechanical Components	X		X	X	X		X	
AV-8B Flight Controls Reactive Control System	X	X	X		X		X	
UH-60 Engine Compartment Fires		n a		n a	X			
F-15 Conformal Fuel Tank Tests	X	X	X	X	X		X	
Totals	**9/11**	**4/9**	**11/13**	**4/11**	**9/13**	**0/9**	**10/13**	**0/13**
	(.82)	(.44)	(.85)	(.36)	(.69)	(0)	(.77)	(0)

[a] X = Information included in the plan

The JLF/Aircraft deputy program manager (Navy) requires a more detailed revision of test plans from his test engineers prior to implementing the test. He is unique in this practice, there being no central JLF/Aircraft requirement. The JLF/Aircraft program manager told us he did not require further detail because all three deputy program managers are highly experienced and "know what they're doing." Some omission of detail was justified on grounds of keeping the DTPs unclassified (e.g., damage predictions).

Inconsistency in Selecting Threat Velocities

The DTPs lack consistency in their selection of threat velocities. Some examples:

- 110-grain fragments are projected at 5,000 ft./sec. in some tests and 6,000 ft./sec. in others, yet both represent the impact velocity 50-75 feet from the burst.
- 45-grain fragments are projected at 6,000 ft./sec. to represent the same 50-75 foot range; this similarity of velocities across fragment size is questionable since small fragments decelerate faster than larger ones.
- The 12.7, 23, and 30 mm projectiles impact at a wide range of velocities with no explanation for specific selections.

Such inconsistencies raise questions about the usefulness of the results in building a systematic data base but the Preliminary Plan, Master Plan and DTPs contained no explanation for them.

Some, though not all, of the inconsistencies may be due to incomplete documentation in the DTPs. One test official said the engineers tend to start from what was tested previously, and attempt not to duplicate those shots. However, DTPs do not list what's been tested in the past in any detail; consequently, selected calibers and velocities can appear arbitrary.

Statistical Validity

The JLF Aircraft Preliminary Plan, while implicitly acknowledging that the full-up testing phase would not produce statistically valid results due to small numbers of test targets, stated that those tests would validate the statistically significant data collected on the replica targets from earlier phases. In fact, there is no indication that statistically valid results will be produced by the replica testing conducted to date as represented in the FY85 and FY86 DTPs.

Statistical validity requires careful attention to cell sizes, i.e., the number of replications of each test condition. In JLF Aircraft, there is no formal process to determine cell sizes, which in any case are changed during implementation as engineering judgment warrants. Test matrices are "as big as you can get or afford." The apparently haphazard distribution of cell sizes is primarily driven by what is available to shoot and shoot at. Test officials freely admit that JLF will not provide statistically valid results, even in the component phases.

Implementation

Observations of several test shots and discussions with test officials indicated that tests are being implemented much the way they were specified in the DTPs. We did not attempt to formally monitor any of the tests, so we cannot state this definitively (IDA has been monitoring the tests more closely).

Changes From the DTP

Test officials do not implement the design exactly as prescribed in the DTP if circumstances arise which, in their view, warrant changing it. The rationale is increased efficiency. Some examples·

- In the F-16 hydrazine testing, two shots in a particular test condition produced no fires, so all further planned shots in that test condition were judged unnecessary and eliminated.
- A Navy decision to replace the AV-8A led to two AV-8As going to JLF, so the AV-8B flight control test design was changed to capitalize on their availability.

In the one draft report we reviewed (F100 steady state fuel ingestion), we discovered that the basic test conditions had changed since the DTP. Inlet ram pressure was to have been supplied to simulate flight at Mach .8, 3,000 ft. above sea level. In implementation, conditions were set at Mach .7, 2,230 ft. above sea level. No explanation for the change was provided.

Efforts to Maintain Realism

Test officials appear to be making reasonable efforts to maintain the realism of test conditions as specified in the DTPs. For example, in the F100 fuel ingestion test:

- The engine was "trimmed" to establish nominal relationships among engine temperatures, pressures, and rotor speeds. This was done to ensure that the engine's reaction to fuel ingestion would be representative of engines currently in use.

- Some equipment problems were observed in the early runs. (i.e., flood lamps and cameras failing due to engine vibration, difficulty establishing and maintaining the desired fuel injector pressure drop), but were eventually solved.

Analysis and Reporting

As only one JLF/Aircraft test had been completed and written up during our time frame, only one report (in draft) was available for our review. This was the FY85 test on steady state fuel ingestion in the F100 engine. which powers both the F-15 and F-16 aircraft. Fuel ingestion is a potential kill mechanism experienced by jet aircraft when a projectile (small arms, warhead fragment) penetrates a fuel cell in such a manner that fuel is injected into the engine inlet. The test's objectives were to determine the fuel ingestion tolerance of the F100 to steady state fuel leakage, and compare the results with previous vulnerability analyses for enhancement of applicable models.

For the fuel injections, the report stated that clean round holes were selected over other hole shapes, in part to meet the primary objective of "controlled" fuel ingestion. A test matrix showed the hole size and its position on the inlet duct for each test run. However, there was no mention of the size or type of ballistic threat the holes are supposed to be simulating; nor was there any explanation for the choice of hole sizes and positions.

The report states that the pressure of the inlet air successfully simulated the specified flight conditions of Mach .7, 2,230 ft. above sea level. but the temperature corresponded to a very hot day—about 113 degrees F. at sea level. To simulate Mach .7 on a cooler day would have required a different inlet pressure and temperature values. The report states that total pressure and density describe not just a single Mach/altitude flight condition. but a locus of points in the Mach/altitude map; therefore, the test data are applicable to flight conditions other than those tested, including some with cooler temperatures (i. e., Mach 1.51, 25,000 ft., 20 degrees F.). It also states that the flight conditions simulated in the test are within the flight capabilities of the F-15 and F-16.

We believe the draft report overstates the generalizability of the findings. No statement is possible on the effect of changing a single parameter—all 3 must change. So, for example, the effect of Mach .7, 2,230 ft. at a cooler temperature (e.g., one representative of a European scenario) cannot be inferred. The fact that the test conditions were within the flight envelopes is irrelevant; it does not make them generalizable. Only additional testing could do that.

The report presented a computer model for predicting damage from a steady flow of fuel into a turbofan inlet. It was derived principally from the F100 test results, although in part from "the author's intuition alone." We question the value of this model for the following reasons:

- It only addresses turbine section thermal failure; no other failure modes are included.
- It was developed after the fact.
- The user varies parameters independently, even though several of these were held constant in the test; consequently, the quantitative effect computed by the model is highly speculative.
- No justification is provided for the model's basic hypothesis, and no basis in the test results was evident.

The report's recommendations were congruent with the results, and sensitive to the likelihood of user acceptance. It concluded that given the engine designers' focus on thrust-to-weight ratios, performance, fuel consumption, and signature, little opportunity existed for engine design changes. Recommendations were therefore focused elsewhere, on airframe and fuel system design.

General Issues

We have identified six issues which bear on the methodological quality of JLF and related live fire testing, past and future. Some of these were introduced earlier, but their importance and/or complexity warrants a separate discussion. They are: conflict over objectives, availability of targets, statistical validity, shot selection methodology, characterization of human effects, and incentive structure.

Conflict Over Objectives

It was clear before the program was launched that different actors in the program process had different, potentially incompatible, agendas. For example, in their official response to OSD's 1983 proposal, the Army replied that the idea of a live fire program appeared to have merit, with the most important benefits being validation of the current vulnerability and lethality models, validation of computer programs on the penetration of armor, and assessment of the "fightability" of damaged weapon systems. The identification of areas requiring further vulnerability reduction was mentioned as a "spinoff." Reducing casualties was not mentioned at all.

The program objectives set forth by the major participants in the planning process are outlined below.

Joint Logistics Command Objectives

As noted earlier, the services agreed to participate on the condition that rather than establish a new joint test force, the existing JTCGs would plan and implement the program. In assigning the task to the JTCGs, the JLC specified the objectives as:

1) Gather empirical data on the lethality of U.S. weapons against foreign systems and the vulnerability of U.S. systems to foreign weapons

2) Use the test data and results to correct any model deficiencies and to validate foreign lethality and vulnerability models.

3) Develop insights into design changes necessary to reduce vulnerabilities and increase lethalities.

The same day, the JLC sent a memo to the Undersecretary for Research and Engineering confirming their support for the program. In it, they stressed that an integral part of the effort must be to obtain empirical data which will provide confidence that the models developed by JTCG are correctly portraying the actual effects, i.e., to validate models. No other objectives were mentioned.

JLF Charter Objectives

The JLF charter, which was not promulgated by OSD until 2 1/2 months after the JLC objectives were communicated to the JTCGs, specified the priority objectives as:

1) For the aircraft component, assessment of the survivability of first line air-to-ground attack aircraft, both U.S. and [material deleted].[3]

2) For the armor/anti-armor component, quantification of the lethality of major caliber anti-armor munitions against first line armored vehicles, both U.S. and [material deleted].

It did not mention objectives 2 and 3 from the JLC version, nor did it specify what was meant by assessment of survivability or quantification of lethality.

[3] Although the charter used the term survivability, the implicit meaning was vulnerability

Former OSD Program Manager Objectives

In his own description of program objectives, the former OSD program manager testified that the objectives of the program were to ensure that U.S. weapons platforms do not unnecessarily endanger their crews, and that the munitions U.S. servicemen fire actually stop the enemy. On other occasions he stated, more succinctly, that the purpose of live fire testing was to reduce casualties. There had been no mention of crew members or casualties in either the JLC or OSD versions.

JTCG Objectives

In their respective plans, the JTCG JLF, Armor and JLF/Aircraft program managers slightly modified the JLC objectives—the reference to correcting model deficiencies was deleted, "validate" was changed to "calibrate" in the armor version, and an additional objective was inserted:

1) Gather empirical data on the vulnerability of U.S. systems to [material deleted] weapons and the lethality of U.S. systems against [material deleted] targets.

2) Develop insights into design changes necessary to reduce vulnerabilities and increase lethalities.

3) Enhance the data base available for battle damage assessment and repair.

4) Use test data and results to validate (calibrate) lethality and vulnerability models.

These objectives were described as in order of priority in the JLF Aircraft plan; no priority was assigned to them in the JLF Armor plan.

Working Level Objectives

We found differences in statements of program objectives among JLF working level personnel. Typically, these differences reflected differences in these individuals' roles, with modelers emphasizing model-related objectives and testers downplaying them.

Current OSD Program Manager Objectives

The current OSD program manager provided his statement of the objectives in an interview with us. His version was highly similar to the JTCG version, although he did not use the terms validate or calibrate in the model objective. Rather, he phrased it as providing necessary data to model and simulate vulnerability and lethality. In his implementation of these objectives, he intends to increase the emphasis on crew effects, with crew survivability as his principal concern. In this sense, his emphasis is consistent with his predecessor's. However, they differ in their emphasis on models in implementing the objectives; the former program manager believed models should not be used unless they were validated, whereas the current program manager believes they are useful tools even when not validated. Toward this end, he has contracted with the Institute for Defense Analyses (IDA) (for $900,000) and The Analytical Sciences Corporation (TASC) for ($150,000) to review the state of V/L methodology, with the majority of the funding focused on improving the models.[4]

Our Analysis

We believe the conflicting statements of objectives reflect underlying differences in the interests of the individuals and organizations involved. The differences are rooted in the differing position of testers, OSD officials, and consulting analysts over the proper role of computerized V/L models in live fire tests and the relative value of models and live fire tests in determining vulnerability and lethality.

Within the V/L community, the use of models has become firmly established over the past twenty-five years. The assessment of the survivability and effectiveness of U.S. weapons has come to depend increasingly on their use. The output from these models is also used in a variety of other activities in DOD, including war games, other simulation models, weapons design, and logistical planning for repair times and stocks of spare parts.

The logic of vulnerability modeling is to build up a full working model from submodels, each of which is based in part on subscale test data. If the submodels are working properly the overall prediction should be accurate, but the focus is on getting the data necessary to make the parts work. Modelers therefore tend to design live fire tests to produce data on fundamental interactions.

Unconcerned with what will improve models, proponents of full-up testing consider more directly the possible areas of vulnerability of a target, and with or without consideration of sampling and generalizing shots, focus on realism as a test design criterion. To them, the tests should directly benefit the serviceman who will depend on the system in battle and the designer/developer who can improve it; not the modeler. This

[4] This is not JLF money, so does not negatively impact the testing budget

leads to an emphasis on full-up shots likely to induce catastrophic fire and explosion, which are of relatively little interest to modelers.

The failure of the two "philosophies" to co-exist is largely a function of resources. The former OSD program manager openly expressed his distrust of both models and modelers, and viewed them as impeding what he held to be the primary objective—finding ways to reduce casualties. To him, spending the funds on model-oriented shots were a waste of the program's budget. The modelers, on the other hand, claim to have been waiting years for an opportunity like JLF to supply their data needs, which they claim are necessary for valid vulnerability assessment. To them, spending the funds on full-up shots was squandering that opportunity.

Availability of Targets

Both U.S and [material deleted] targets are in seriously short supply and represent the principal constraint faced by all JLF test officials.

- Target availability drives test schedule, test methodology, and the need for complementary approaches.
- Cost is a key factor for complete functional systems, yet crash hulks, old prototypes, and many components are also scarce ("you take what you can get").
- Both the aircraft and armor programs are affected, but the constraint is particularly acute with aircraft due to higher unit cost.
- The problem is an old one, present at least as far back as the late 1940's.

[material deleted]

The principal obstacles faced by JLF test officials in obtaining suitable targets are:

- absence of assigned responsibility for providing targets.
- competing interests within and outside DOD.
- negative attitudes toward destructive testing.
- poor condition of targets upon arrival.

Absence of Assigned Responsibility

Because of the way in which JLF was chartered, the services have no responsibility to provide test articles or bear any support costs. Additionally, JLF test officials told us that no individuals were designated within the services to assist them in obtaining targets. The former OSD program manager played a role (e.g., he helped to obtain an F-15 prototype redirected from the Air Force Academy, as described later) but most arrangements have been made directly between the JLF program managers and the services.

Since the services had no responsibility to provide targets or assistance in obtaining them, and OSD provided no funding to purchase targets, JLF test officials were essentially left to fend for themselves. In order to obtain targets, they have had to continually "sell" the program, to try to convince individual service components that they would benefit from JLF, or as expressed by one JLF program manager, "sweet talk" them out of hardware. Therefore, a substantial proportion of the time they might otherwise have spent implementing the test schedule was used giving briefings. It is important to remember that the bulk of this effort was directed not at obtaining new or operational hardware, but obsolete hardware that might have otherwise been discarded.

[material deleted]

Competing Interests

JLF generally only requires obsolete hardware for full-scale tests, i.e., crashed or prototype systems of no use to operational forces. Nonetheless, it still must frequently compete with other government interests. For example:

- In May of 1986, sixteen F A-18 prototypes stricken by the Secretary of the Navy were transferred to NASA to upgrade their fleet of tracer planes. NASA planned to keep six of the sixteen flyable, and "cannibalize" the remainder for spare parts. The JLF Aircraft deputy program manager (Navy) knew of NASA's interest in the planes, but assumed that JLF had priority; the Navy had granted JLF a formal acquisition priority and knew of its requirements for F A-18s. He learned of the decision to transfer the planes to NASA only three days before it was finalized. After JLF officials briefed the responsible flag officer, two of the sixteen planes were redirected to JLF. However, these were considered the worst two, having already been severely stripped by NASA. JLF is now dependent on NASA to supply the parts needed for restoration. The JLF deputy program manager had no explanation for the decision to give the planes to NASA rather than JLF.
- At the direction of the Chairman, Senate Armed Services Committee, an F-15 prototype that JLF wanted was to go on display at the Air Force Academy. According to the OSD program manager, the Secretary of the Air Force was unwilling to request reconsideration. Consequently, the

OSD program manager briefed the senator's staff himself, after which the aircraft was redirected to JLF. However, by the time JLF received it, it had already been gutted—the horizontal and vertical tails, the cockpit, the majority of avionics, and other components were gone. The F-15 system program manager and the manufacturer have since agreed to restore the removed parts.

There are competing non-government interests as well. If a crash causes a fatality, and the crash hulk is recovered, it will be impounded pending the outcome of private litigation brought against the manufacturer and/or the government. This can delay access to JLF for years. Currently, an AH-64 helicopter tentatively promised to JLF is in litigation (JLF has obtained no other AH-64's).

All three services have approved a force activity designator priority of 2 (1 is highest, 5 is lowest), or FAD-II, for JLF (Air Force in July, 1985; Navy in December, 1985; Army in June, 1986). The FAD-II officially gives JLF priority in obtaining assets over various non-operational interests with lesser priority. It in no way guarantees the availability of targets. For example, the Army's approval of the FAD-II request explicitly states that it will not materially affect the availability of armored vehicles or aircraft for tests. JLF test officials believe the FAD-II may have had some impact in obtaining test targets, but cannot document that it has provided them specific targets they otherwise would not have obtained, or that in general it has made them easier to obtain. For example, the FAD-II was used to reclaim the two F/A-18s described above, but finding a supportive flag officer was believed to be equally or more important. The most frequently mentioned improvement was that since obtaining the FAD-II, JLF now has priority over museums.

Negative Attitudes

According to JLF test officials, live-fire testing is alien to most DOD officials, including many flag officers, because of its destructive potential. A test official described live fire testers as the "Ralph Naders" of the testing business. Another noted an "immediate fear reaction" when mentioning JLF to system program offices (SPOs)

However, we found no evidence that SPOs for systems being tested under JLF have impeded the process. In general, test officials reported good acceptance and cooperation from all SPOs. Nonetheless, it must be noted that SPOs are not required to actually provide targets.

Poor Condition

The systems and components that JLF does receive are frequently in poor condition, particularly if they were in crashes. This applies to components as well as full-scale targets which arrive stripped. An example is the UH-60 Blackhawk helicopter servo assembly, required for the flight control tests. After a crash, the manufacturer disassembles the servos, analyzes them, and boxes the disassembled parts. Multiple disassembled units are sometimes boxed together, with no indication of which parts go with which unit, or, parts are simply missing (e.g., mixers were missing from the servo parts received by JLF).

Despite the fact that targets do not typically arrive in a condition that is suitable for testing, JLF provides no funds for restoration. The JLF program managers attempt to get restoration support from the system program offices.

What Could Have Been Done

According to the JLF test officials, the above problems might have been lessened by:

- an education program for high level military officials
- a small procurement fund to keep some baseline level of testing going while "selling" the program and waiting for crashes
- more top-level effort from OSD.

There was some disagreement on the last suggestion. There had been some high level effort—the original FAD-II request came from the DDT&E director in May, 1984—but nothing came of it. Eventually, it became apparent that the JLF program managers and deputy program managers would have to try to obtain the FAD-II priorities themselves, separately, after the services had agreed to participate with the understanding they would not have to provide acquisition priorities. According to one test official, another "bottom-up staffing drill" should not have been necessary; rather, this role should have been handled by OSD. According to another, the lower level "selling" of the program was probably unavoidable. As he saw it, the charter alone was meaningless; the credibility required to get hardware out of the services could only come from the test officials' demonstration that JLF was useful.

Target Availability for Army Programs (Non-JLF)

As noted earlier, the Army removed the Bradley vehicle, M1 tank, and M113 APC from JLF to conduct the tests themselves. In so doing, they took responsibility for supplying targets, and supplied twelve Bradleys (thus far) and four M1A1 tanks. These are not prototypes or crash hulks, but fully operational vehicles off the production line. This clearly shows that while obtaining targets for live fire testing is currently a problem, it need not be an inherent problem (at least for ground vehicles). If the services have sufficient motivation, they will supply targets.

Statistical Validity

The problem of statistical validity in destructive defense testing is not new, having been noted at least as far back as the early 1950's. Lack of statistical validity is related to the unavailability of targets. However, the numbers of targets JLF officials have specified in their requirements and tried to obtain would still be insufficient for statistical analysis. Where shot selections are potentially catastrophic, or sufficiently damaging to invalidate subsequent shots on the target, the cost of supplying statistically adequate samples could easily run several hundred million dollars.

Need for Statistical Validity

JLF officials and other experts contend that live fire results yield valuable knowledge despite the lack of statistically adequate sample sizes. A single shot can identify an excessively vulnerable component or unexpected kill mechanism, reveal model flaws, and generally provide qualitative insights into the vulnerability or lethality of the system and how to improve it. A few shots can be used to tentatively characterize these phenomena, e.g., show how vulnerability progresses with threat size. Matched comparison shots, such as those being fired at the standard version Bradley M3 and the high survivability M3, can be particularly useful because absolute measures of vulnerability are not required.

Small numbers provide, at minimum, descriptions of directly observable damage. These descriptions can be highly beneficial to users. Descriptions based on direct visual observation avoid reliance on indirect sources with unverified assumptions (e.g., an analytic estimate of what would have happened had ammunition been on board).

Small numbers would suffice if the same shot could be guaranteed to yield precisely the same damage each time it was repeated, i.e., it could be predicted deterministically. With many simpler phenomena, deterministic prediction is reasonable. For example, a 23 mm API round impacting a hydraulic line will virtually always cut the line. However, the collateral structural damage, the likelihood of fire and its effect, the impact on redundant systems, etc., are much less certain. Total system damage mechanisms are inherently complex and involve too many variables to be predicted deterministically, particularly for full-up firings.

Due to random variations in these variables, results can at best be predicted probabilistically. The same type of shaped-charge round fired at identical pieces of armor under identical controlled conditions of impact produces a variety of spall patterns and damage levels. It is impossible to predict deterministically the spall pattern and damage level produced by a specific shot.

As a result, with small numbers of shots,

- there is no way to assess whether the test results are typical or atypical.
- there is no way to assess the likely range of variation.

The problem is most readily apparent with JTCG Objective 4—the validation of V L models—because of the quantitative precision required. However, it also affects the more primary objectives of gathering V L data and developing insights into design changes. Without confidence that the observed results represent what typically would occur, both the data gathered and the design insights it provides could be misleading For example, a decision to harden a particular component on the basis of one or a few shots would be misdirected if the results were atypical, adding unnecessary cost or performance penalties to the system. The problem was also noted by the BAST group, who concluded that·

- it is extremely unlikely that a statistically credible assessment of vulnerability will result from the current state of live fire testing.
- with so few shots per weapon, many unanticipated damage mechanisms may be overlooked.

Statistical Validity and Engineering Judgment

According to the JLF Aircraft Preliminary Plan, the number of shots per test would be based on a combination of statistical probability analysis and engineering judgment. In fact, the number of shots has been primarily determined by target availability, as has the way the shots are configured into a test matrix Engineering judgment was clearly the next most important factor, with statistical probability analysis given little if any consideration. One explanation was that with such small samples, there was no hope of attaining statistical validity and therefore no point in considering it in the test design. However, it also appeared that the test engineers believe their engineering judgment alone will correctly guide their design decisions as well as provide valid interpretations of test results.

Engineering judgment is a necessary tool, particularly when data are sparse However, as the sole criteria for interpreting small sample test results, it has serious drawbacks. Each engineer will not necessarily interpret a single result in the same way. Some test engineers typically specify only one shot per condition, claiming it will be obvious that is all that is needed. If the result is uncertain, and there is still a component to shoot at, the shot can be repeated. If the results of the two shots diverge, some may try to ascertain the reason analytically, others may take the average, and still others may fire a "tie-breaker" shot. An old problem that has reappeared in JLF is the lack of upfront coordinated planning by test engineers and statisticians. While there have been some interactions between JLF test engineers and statisticians, there is no requirement that statisticians have input into test design decisions. The statistical input has been minimal and has had little effect. Our observation is that the input that is provided is not always well understood by the testers; as described by one statistician providing consultation for JLF, engineers and statisticians are still "talking past each other."

Engineering judgment parallels the "clinical" judgment of physicians and surgeons in evaluating the effectiveness of new drugs or medical and surgical procedures. Much research has shown that in fact clinical judgment is not effective for such evaluations, and controlled experimental trials have gradually come to be recognized by the medical community as the only reliable way of assessing the effectiveness of medical innovations. Just as it is precarious to entrust our lives to medical treatments whose effectiveness is assessed solely by clinical judgment, it would seem equally precarious to trust soldiers' and airmen's lives to engineering judgment if the alternative of a more credible procedure is feasible.

Statistical Validity and P_{KH}

The limitations of engineering judgment are particularly apparent in the generation of probability of kill given a hit P_{KH} values. The P_{KH} is the most common form of quantitative output from V L assessment, whether live fire or analytical. Component-level P_{KH}s are the basic input to assessing vulnerability of the full weapon system, and vulnerability is in turn input to successively higher level analyses, such as survivability, mission effectiveness, exchange ratios, and force planning. The P_{KH}s are generally not actual statistical probabilities—number of kills divided by number of hits—rather, they are products of prior test data, combat data, and models, as well as engineering judgment. For example, it is common to assign a P_{KH} based on a single shot; if the shot is a kill, it might be assigned a P_{KH} of .8.

Getting analysts to agree on $P_{K\ H}$s is reportedly very difficult, yet as far as we could determine, there has never been any attempt to assess the interrater reliability of judgmental assignment of $P_{K\ H}$s. Essentially, they have no demonstrated reliability or validity.

When $P_{K\ H}$s are based on actual probabilities (i.e., empirically generated relative frequencies) it is often from a very small number of shots; e.g., two shots with one kill and one non-kill yielding a $P_{K\ H}$ of .5. Consequently, estimates are extremely unstable. JLF officials agree that target availability generally precludes doing enough live fire shots to get empirical $P_{K\ H}$s. Probability measures, by their nature, require large samples for statistically reliable results. Figure 3.1 illustrates how confidence increases with sample size when the sample results show a $P_{K\ H}$ of .5. With only 10 shots, for example, the approximate 95 percent confidence interval runs from .19 to .81. As the number of shots gets larger, the confidence interval narrows, reflecting greater precision and reliability. Most JLF tests contain considerably fewer than ten shots per condition.

Examples of Statistical Analysis in Live Fire Tests

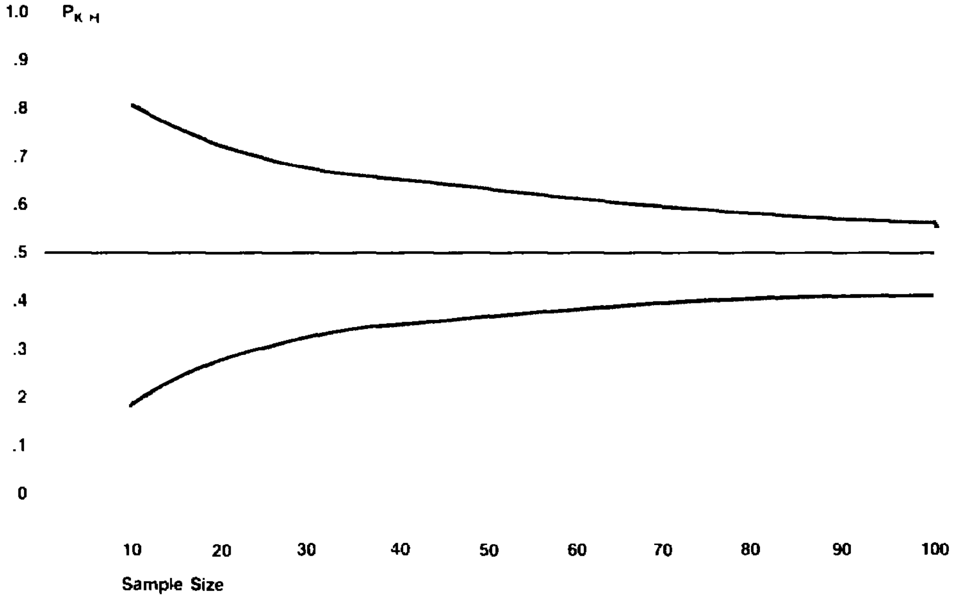
In the few instances in which we found formal statistical tests, we question their appropriateness.

1) Bradley Phase II. A key feature of the overall evaluation in Bradley Phase II is a paired comparison of twelve matched RPG-7G shots against the standard version and the high survivability version of the Bradley M3. The results of this comparison test will be used by the Army and OSD as part of the information upon which to decide whether to apply the enhancements tested to production vehicles.

The comparison will be based on a statistical procedure called the sign test. This approach has numerous problems. Most importantly, a sign test with only twelve pairs of shots will fail to detect differences between the two vehicles unless the differences are very large. We performed the test for numerous possible outcomes and found that, barring ties and noncomparable shots, the high survivability version would have to win ten of the twelve shots for the comparison to be statistically significant.[5] In the event of ties or noncomparable shots, the percent of wins needed by the high survivability version would be still greater.

[5] This is based on a 1-tailed test as specified in the plan, and a .05 significance level (as no level was specified we are using the most commonly accepted value). This means that the likelihood of a difference being due to chance is less than 1 in 20.

Figure 3.1: Distribution of 95% Confidence Intervals by Sample Size, for $P_{K/H}$ = .5



The DTP does not acknowledge that the sign test will require a difference this large for the testers to declare the high survivability version the winner. This criterion, as well as the significance level and rationale, should be made explicit in the plan. If these criteria are set after the fact, the test's credibility as a decision tool is damaged.

The sign test has the following additional problems:

- It assumes that each pair of data points is an observation on a random sample. The assumption is not supported because four of the twelve pairs are based on Phase I shots, which were selected by BRL personnel to target areas of uncertainty. As such, they were selected systematically, not randomly.
- It assumes observations are mutually independent. This assumption is not supported, again because the Phase I shots were selected systematically. Personnel selecting shots were fully cognizant of previously selected shots, hence the selections were not independent. Additionally, the BAST method specifies re-sampling shots which are duplicative, which also violates independence.

- The test is inefficient; it only utilizes direction of the difference (hence its name), ignoring any relevant information on the size of the difference.
- The comparisons will be based on casualties, penetrations, and assessed levels of firepower and mobility loss. However, there is no explanation of how these measures will be combined, or how they will be reduced to a single win, loss, or tie.

2) A-6 dry bay foam. On the aircraft side, a statistical test was applied in assessing the effectiveness of reticulated foam in preventing fires in the A-6 dry bay. This live fire test was not technically part of JLF, but was conducted by the same performing organization and personnel that are carrying out the JLF/Aircraft tests. The results were later used to persuade NAVAIR and the aircraft's manufacturer against using that particular foam.

The test design specified six shots each for the foam and baseline (no foam) conditions. The outcome measure was fire vs. no fire. Fisher's exact test was the statistical procedure selected for the hypothesis test, with a specified confidence level of 90 percent. Fisher's exact test permits computation of exact probabilities for a 2 X 2 table when, as in this case, sample size is too small to meet the assumptions of the more commonly used chi-square approximation test. The test assumes that shots are sampled at random, which was violated by the engineers' systematic selection of shots.

The test was applied to all possible outcome scenarios. The test engineers reported that with six shots per condition, no outcome provided the 90 percent level of confidence. Unfortunately, the number of shots had already been fixed by time and budget constraints. The testers concluded that the statistical analysis supported the finding suggested by direct observation, that the foam was ineffective in improving survivability. In fact, the statistical analysis supports a very different conclusion, i.e., there were not enough shots to conclude that the foam was ineffective. The testers stated they also used engineering judgment, which they described as more "critical" than statistics. We do not dispute the importance of engineering judgment nor do we assert that the conclusion of ineffectiveness was incorrect. Rather, such a conclusion simply could not be reached given the statistical basis used. We do assert that application of the statistical test did nothing to improve the decision making process, and by confusing the statistical logic, potentially muddied it

Efforts to Improve Statistical Validity of Live Fire Testing

We learned of three ongoing efforts to make live fire tests statistically interpretable. The common theme seems to be placing the problem in a stochastic, or probabilistic, context.

1) Air Force project. A Wright-Patterson analyst is attempting to develop "front end stochastic simulations," whose theoretical stochastic distribution can be used to bracket an expected effect (a stochastic distribution incorporates randomness or chance). A small-sample value from a test can be compared to the theoretical distribution. If it falls within one standard deviation of the distribution mean, it provides reasonable confidence that the test sample is not a statistical outlier. If it does not, then the distribution is assumed to be incorrect. The logic is reasonable; the problem is how to form the basis for the theoretical distribution. Prior test data is an obvious candidate, but the analyst admits he will probably start with engineering judgment, the limitations of which were described above. Even if the theoretical distribution is correct, a substantial percentage of sample results will fall outside one standard deviation by chance alone. In these instances, altering the theoretical distribution to accommodate the data would represent the wrong decision, and potentially mislead interpretations of subsequent test results.

2) Army project. A BRL analyst is doing a similar project for armor. They have added a stochastic distribution to the impact point, depth of penetration, vehicle geometry, number of spall fragments hitting components, etc. As in the Air Force project, the problem is how to form the basis for the theoretical distributions. Some are based on data (e.g., depth of penetration data has been collected from subscale testing), others on engineering judgment. Unlike the Air Force project, there is no simple interpretation rule for a small sample result falling outside the distribution.

3) BAST project. As a follow-up to their Bradley shot selection work, the BAST group is attempting to develop a valid statistical approach to live fire testing. This will include the determination of sample sizes needed for statistical validity.

Shot Selection Methodology

Many of the methodological issues raised by live fire testing surround the question of how to select shots. The immediate cause of the controversy that halted the Phase II Bradley tests in April of 1986 was a disagreement between the Army and OSD about how the shots should be selected. The House Armed Services Committee (HASC) investigation concluded that this was a basic disagreement about testing methodology. The Chairman of the Procurement Subcommittee of the House Armed Services Committee pointed out that this amounted to a failure to decide what the live fire test program was about. IDA, Los Alamos National Laboratory (LANL), and the Board on Army Science and Technology (BAST) were asked to examine the issues involved in the disagreement, and BAST was asked to recommend an interim selection method. The Bradley tests were suspended until the BAST report on shot selection was produced. But the controversy goes beyond the Bradley series to live fire tests in general. The question of how shots are to be selected is also relevant for aircraft tests, and will be more apparent when full-up aircraft tests are conducted.

The conditions that define a particular live fire shot include:

- angle of attack, or azimuth;
- point of impact, given the azimuth;
- elevation;
- velocity at impact; and
- range of firing.

The range at which shots are fired, in particular, can vary from the distance limit of the munition to the special case in which a warhead is detonated while fixed to the vehicle to ensure that a particular point is hit. This case is known as static firing

Static vs. Dynamic Firing

Traditional ballistic testing practice has relied on the static firing of warheads. A shaped charge warhead is fixed to the armor surface and detonated by remote control. The effects on a specific preselected location can thus be determined regardless of the munition's accuracy. Early in the planning of JLF, there was some question about whether the results of this kind of test differ from the results obtained when a shaped charge munition is fired from a distance as in combat:

- Some experts argue that when the kinetic energy of a missile's flight is added to the effect of the shaped charge warhead at impact, there is additional damage, especially to lightly armored vehicles. They also note that dynamic firing changes the fuzing and yaw angle. It is therefore unrealistic to fire such warheads statically.

- Other experts argue that the jet produced by shaped charges travels at such extreme speeds as it penetrates armor that any energy added by missile flight is negligible.

Apparently the differences between static and dynamic firings had not been studied systematically with modern weapons prior to JLF. The initial [material deleted] tests in the January 1985 JLF/Armor plan were intended to help resolve the issue.

The JLF/Armor test planners were sufficiently confident that there would no substantial differences between the two modes of firing that all the shots after the first few were tentatively planned to be static. Because static tests are cheaper, simpler, and more controllable, the testers wanted to confirm their ability to rely on static firing. They also pointed out that dynamic firings of some foreign munitions would pose problems in the absence of a suitable launching platform and a qualified operator.

The former OSD program manager disagreed. He pressed for dynamic firings, invoking the general principle of maximizing "combat realism" as OSD guidance for the design of JLF tests. He also asserted that additional damage was observed to result from dynamically fired TOW missiles in the Phase I Bradley tests, as compared to one static firing. There was structural damage to the armor in the vicinity of the impact and additional debris, including the body of the missile, entered the vehicle and caused further damage.

We believe that the question cannot be resolved by theoretical arguments in the absence of relevant evidence. It must be decided by the kind of comparative empirical test proposed in the January 1985 JLF/Armor plan. Only then would it be possible to formulate a methodological rule for live fire tests stating the conditions under which dynamic firings are necessary.

As part of the Bradley Phase II tests, the Army recently conducted a limited number of static/dynamic comparisons. Using two types of threat munitions and multiple measures, the analysts did not find consistent differences between static and dynamic firings. However, the small sample sizes (for some comparisons only three shots) and the large observed round-to-round variability in effects meant that true differences of moderate or smaller size would have been difficult to detect statistically in these tests. In addition, the shots were fired on configurations of Bradley armor, not on full-up or full-scale vehicles.

We were told that further tests using other munitions have begun as part of the first series of JLF tests on the [material deleted] tank.

Selecting Shotlines: Two Approaches

The other main controversy has been over how to select shotlines, i.e., the location and angle of impact. There are two basic approaches. The first, advocated by both armor and aircraft test planners, emphasizes judgment. It holds that shotlines should be chosen by using:

- the knowledge of designers, who are familiar with design features and developmental test results;
- knowledge of the way munitions are typically used against targets in combat;
- other aspects of engineering judgment; and
- the predictions of, and uncertainties identified by, vulnerability models.

The rationale for this judgmental approach is that:

- much is already known about vulnerability in some areas;
- judgment based on this knowledge and expertise can be used to select the shotlines from which the most new knowledge can be gained;
- it is a waste of scarce test resources to fire shots unlikely to yield new knowledge;
- if models are improved they can be used to generate whole-target estimates of vulnerability;
- shotlines of interest for improving models will permit extrapolation of test results.

Critics of using judgment to select shotlines point out that:

- it does not make the process of shot selection explicit enough to be easily explained or evaluated;
- it relies on fallible processes of judgment that can introduce inadvertent biases into the sample of shotlines;
- it is vulnerable to intentional biases from testers (e.g., intentionally selecting shotlines so as to underestimate system vulnerability);
- it allows the requirements of modeling to guide the design of tests;
- it does not produce results that can be directly generalized to statements about a target's overall vulnerability, because the shotlines chosen are not representative of combat hits.

The second approach to shotline selection, advocated by the former OSD program manager and several outside experts, uses random selection, most often from a distribution of hits observed in combat (combat distribution). Attack angle, impact location, and range can be chosen randomly. The rationale for some form of random selection of shots is that:

- it is the only way to ensure that individual or community biases do not enter into the shot selection processes, even inadvertently;
- it gives shots that lead to unexpected occurrences a chance to appear in the sample of shots;
- selection from combat distributions of hits theoretically permit test results to be generalized to target vulnerability as a whole.

Critics of random selection point out that:

- it can be very wasteful of test resources if shots likely to destroy targets are actually fired;
- many of the shotlines chosen will be near duplications and provide little new information;
- sample sizes will still be too small to generalize directly from tests to target vulnerabilities;
- combat distributions of hits are biased.

There is considerable disagreement over the validity of combat distributions as the principal basis for shot selection. The criticisms and rebuttals generally asserted are as follows:

Criticism 1: Combat distributions are biased because they fail to include aircraft and vehicles that are not recovered. For aircraft, a gap in the shot distribution may mean the aircraft is never hit there or it may mean that those hits are catastrophic. For vehicles, a gap may mean the vehicle is never hit there or it may mean that vehicles hit there are repaired and returned to service. In either case, the hits do not show up in combat distributions.

Rebuttal: For aircraft, it has long been known that a low hit-density spot on an aircraft signified a catastrophic hit point, rather than a spot which avoided hits (during WW II, the 8th Air Force recorded locations of bullet, shell, and fragment impacts on returning B-17's; missing locations implied vulnerable impact points). First, they are typically in places where the reason for aircraft loss is readily evident (e.g., proximity to fuel tank). Second, there are no logical reasons for a particular spot avoiding hits, given the general uniformity of aircraft hit data. For vehicles the claim is more justified, but for the purpose of designing live fire tests, practically unimportant. If the shots in the low density spot failed to do serious damage, they are not of primary interest.

Criticism 2: Combat distributions are biased by particular wars, battles, threats, ranges, etc., that do not represent future wars.

Rebuttal: In fact, distributions of shots on vehicles have been quite similar from World War II through the 1973 Middle East war. However, significant increases in missile accuracy could change this eventually. There is also considerable similarity in air to air hits from WW II through the 1982 conflict in Lebanon. The criticism may be most justified in the case of helicopters, where tactics have changed substantially (helicopters now fly lower than in the Vietnam era, so shots disperse over the entire lower hemisphere, rather than just the lower quadrant).

Criticism 3: Reporters of damage are frequently inexperienced and may incorrectly identify the munition causing the observed damage, or otherwise make errors.

Rebuttal: Some combat data sets are undoubtedly higher quality than others. However, when the Survivability/Vulnerability Information Analysis Center (SURVIAC) receives data, it is cleaned to some extent (e.g., removal of obvious outliers) and identified by collector. Thus, analysts can use only data collected by experienced professionals if they wish. On the other hand, we do not know of any systematic interrater agreement studies assessing the reliability of the damage assessment process.

Criticism 4: [material deleted]

Rebuttal: In typical anti-aircraft fire (non-missile), hits occur one in every 4,000-8,000 shots; there is no way to aim at a particular spot. For vehicles, the criticism is reasonable for close shots such as might occur in ambushes or urban warfare, but these cases do not constitute the preponderance of the data. In most battles involving tanks, the weapon is aimed at the apparent center of mass so as to maximize hit probability.

Criticism 5: Combat distributions do not distinguish between kills and non-kills on vehicles, because 1) the enemy will continue to fire on a dead vehicle until the kill is confirmed, and 2) soldiers use dead vehicles for target practice.

Rebuttal: The first point is not relevant because a combat distribution does not presume to identify cause of kill, but simply the distribution of hits during combat. The second point is relevant in as much as target practice produces non-combat hits, and therefore adds noise and possibly bias to the combat data. However, it is questionable how often soldiers will expend their ammunition that way during wartime, so the percentage of non-combat hits may be negligible.

While not all of the rebuttals are equally convincing, we believe that generally, they do refute the criticisms. More importantly, the use of combat distributions removes the potential for bias, intentional or inadvertent, introduced by systematic selection of shots.

If combat distributions are very close to the uniform distribution, sampling randomly from the uniform distribution would be a sensible solution—in effect giving all locations on the target an equal chance of being selected. It would have the advantage of avoiding objections to the use of combat data based on possible biases in the data, while at the same time preventing any personal bias from entering into the selection of shots.

Attempts to Reconcile the Two Approaches

Several attempts have been made to reconcile the claims made by proponents of these two approaches to the selection of live fire shots. These attempts have sought to use technical principles of statistics and experimental design to resolve some of the methodological issues, or to provide a shot selection method that meets the concerns of both positions. We review three of these here: center-of-mass aiming (Army proposal), the LANL proposal, and the BAST proposal.

1) The Army's center-of-mass aiming. The first proposal was an Army response to OSD guidance that live fire shots be selected for the Bradley Phase II tests using combat data. The Army procedure had three steps:

1) The most common attack azimuths in combat distributions were selected.

2) An ellipse was laid over the vehicle's apparent center of mass. Its size was determined by the dispersion of hits for the weapon in question as determined in range tests.

3) The test director was allowed to select impact points from within the ellipse.

The rationale was to simulate what happens in combat. The Army claimed that most combat hits would fall within an ellipse of the size used. because gunners are taught to aim at the target's center-of-mass and the ellipse included 68 percent of the hits expected when shots were so aimed. The approach thus combines the use of a combat distribution (of azimuths) with judgment based on training doctrine, test range data, and statistical reasoning.

The OSD program manager replied that:

- through an apparent statistical error, Army analysts had failed to note that only 39 percent of even test range hits would fall within an ellipse of the size used (technically, one standard deviation removed from the center of mass in each direction).
- the apparent center of mass of targets in combat changes as vehicles are concealed by terrain in varying degrees, and combat hits therefore occur "all over a vehicle;"
- combat data on RPG-7 hits show that only 29 percent fall within the ellipse outlined by the Army;
- a statistical test suggested that the observed combat shots are unlikely to have come from the distribution hypothesized by the Army.

The result of the center-of-mass aiming approach is to place more shots in the center of the vehicle than could be expected in combat. This was important in the case of the Bradley because of controversy about the relocation of less vulnerable components toward the center of mass of the vehicle.

We note that it is unlikely in general that test data based on center-of-mass aiming will reproduce combat distributions, because of additional features of combat likely to affect hit points. In combat:

- targets are often moving;
- gunners are being fired upon;
- smoke and fire obscure battlefields.

All of these, in addition to the changing apparent center-of-mass of a target, tend to increase the dispersion of hits.

The failure of the center-of-mass aiming procedure to reproduce combat data is an illustration of how a shot selection procedure can sometimes be exposed to the test of data. It demonstrates that a procedure for selecting shots can rely in part on data, take account of training doctrine, employ statistical reasoning—all of which suggest a striving for realism—and still produce a sample of shots that is unlike those to be found in combat data. And although there is no evidence of intent to bias the sample, it is in fact likely to be biased in its implications for vulnerability assessment.

2) LANL report. The Statistics and Operations Research Group at Los Alamos National Laboratory (LANL) was asked to conduct a statistical assessment of the two competing shot selection methods we have outlined. They placed the issues in the context of statistical sampling theory and the discipline of experimental design. Their general observations were:

- Formal experimental design can help research to be efficient by optimizing some criterion that can be measured.
- The judgmental selection criterion that live fire shots be "of interest" cannot be stated in a way that experimental design principles can maximize it.
- The former OSD program manager's determination to eliminate all possible biases in JLF shot selection would require the use of some kind of random sampling.

They concluded that the dispute is a matter of differences in objectives. Both judgmental selection and random sampling have traditions of use, and both have strengths and weaknesses. They pointed out that:

- Small random samples do not provide precise estimates.
- Even small random samples may, however, permit unanticipated vulnerabilities to appear.
- Samples selected using judgment are by definition biased, and they cannot be used directly to make general statements about the population they are taken from, in this case the vulnerability of a whole target.
- The magnitudes of any biases in judgmental samples are difficult to assess.

LANL treated the question of efficiency from a research design perspective, and pointed out that:

- Sampling theory recommends oversampling certain classes of events, including those where there is more uncertainty. This is consistent with the judgmental approach.

- Simple random sampling is inefficient, which may be important if the tests are very costly (e.g., some shots may be nearly duplicative of others, while other areas may not be sampled).

In the LANL report, experimental design provides a framework for thinking about shot selection, but we find there are problems with the proposal:

- LANL treats the judgmental approach as a form of stratified sampling, focusing on low, medium, and high vulnerability, suggesting that the oversampling might be done in high vulnerability areas, where variability and uncertainty are greatest. In fact, variability may be greatest at intermediate vulnerabilities, where for example the probability of penetration of armor is close to .5.
- It does nothing to resolve the conflict between the objectives of the competing positions.

3) BAST shot selection proposal. The interim method for choosing shots worked out by the BAST and adopted by the Army's Bradley test officials does in fact employ random sampling from a combat distribution. It also attempts to meet the criticism that some randomly chosen shots will be wasteful, by constraining the random selection in three ways:

- If two shotlines are close together, one will be discarded. The criterion for this is not specified.
- In order to focus on the mechanisms producing crew casualties all shotlines will be constrained to pass through the crew compartment.
- Shots that will clearly be catastrophic in effect need not actually be fired but simply scored as K-kills.

Requiring shotlines to pass through the crew compartment limits the generalizability of any conclusions to—at most—shotlines that pass through the crew compartment. The BAST procedure sacrifices potential knowledge of all those vehicle vulnerabilities in other locations, including any casualty mechanisms that are likely to originate in materials or components outside the crew compartment.

The remaining features of the BAST procedures include:

- random selection of attack azimuth.
- random choice of the left or right side of the vehicle.
- selection of an aim point which is always the apparent center of mass when viewed from the chosen attack angle.

- random selection of a range from a distribution based on war games and limited combat data.

However, the actual aim points and ranges in the live fire shot are not those selected by the above process. Instead, the actual aim point (termed the "proposed impact point") is determined by randomly selecting a right-left distance and an up-down distance from the apparent center-of-mass, using dispersion data from tests. These two values define the point at which the shot is actually fired. The actual range is set sufficiently close to ensure hit accuracy, with kinetic energy rounds downloaded to match the impact velocity at the randomly selected range. That is, the randomly selected range is used only to determine the appropriate dispersion and the impact velocity for kinetic energy rounds.

Some experts claim that downloading is unrealistic. They argue that while the downloaded round's impact velocity may match that of the selected range, its yaw and spin do not. Consequently, its penetration characteristics are different.

There are two main differences between the BAST shot selection method and the one proposed earlier by the Army. The first is the random selection of azimuths and displacements from the center of mass rather than allowing the test designer to select shot locations. This prevents even inadvertent biases from entering into the selection. The second difference is that the dispersion data that form the basis for selecting actual aim points are carried out to three standard deviations, so that shots will not be restricted to a one standard deviation ellipse. However, the use of bivariate normal distributions will still tend to place shots more frequently toward the center of the target than at the periphery, and the question of whether the selected shots would resemble the distribution of combat hits remains. BAST does not claim that the resulting samples of shots will represent combat distributions.

Our Analysis

These proposals are complex compromises whose consequences for the interpretation of the test data obtained are uncertain and difficult to assess. We believe that technical solutions to shot selection problems like those proposed by the Army, LANL, and BAST can make some progress toward working out acceptable departures from realism and the avoidance of bias in shot selection, but they continue to ignore the competing agendas of participants in the controversy:

- Proponents of judgmental shot selection have confidence in engineering judgment and vulnerability models, and emphasize efficiency and reliance on expertise. They emphasize the test design features that serve vulnerability models.
- Proponents of random shot selection do not trust model predictions, and emphasize combat realism and the avoidance of bias in tests, at all costs. They do not require that vulnerabilities be the quantitative $P_{K/H}$s of vulnerability models, but want to locate unexpected effects under realistic conditions.
- Something is gained from developing formal selection methods to prevent inadvertent bias in estimates, but this does not solve the problem of mistrusting motives.

We do not believe that technical solutions alone can resolve the shot selection problem. Some sort of random selection is the only way to avoid even the appearance of bias, yet simple random selection is an inherently inefficient way to select shots. Sampling efficiency is paramount in live fire test design because of the expense and scarcity of targets. Additionally, random selection removes the legitimate expertise of test designers along with their biases, i.e., the "baby with the bathwater". An interim solution might be to designate that some proportion of shots be selected judgmentally and others randomly (although constrained by rules for excluding clearly wasteful shots). A more satisfactory solution is difficult without a decision on the objectives of live fire testing and their priority.

Characterization of Human Effects

Accurate estimation of human effects is essential to estimating casualties. Typically, plain plywood or instrumented mannequins are used to estimate the effects, and the raw damage data is interpreted through models based on combat data and/or animal experiments.

The January, 1985 JLF/Armor plan states that in general, personnel vulnerability is "well known," and the JTCG/ME chief told us that casualty estimates can be obtained by taking some shots and running the data through models. However, other sources cast doubt on the casualty estimates currently being produced.

Validity of Damage Assessment

First, the validity of using mannequins to assess personnel damage is questioned. IDA reported the following observations from the September, 1985, armor tests:

- the assessor teams had difficulty agreeing on the extent of personnel damage that could be inferred from the mannequin damage.
- spall damage can be assessed from mannequins, but not pressure, temperature, or flash damage.
- mannequins are unrealistic in their flammability and shielding effects.
- certain reactions found in mannequins but not in people (e.g., splintering) make it difficult to assess the accuracy of model predictions.

The mannequins used in those particular tests and most live fire tests to date were the plywood, non-instrumented, non-anthropomorphic type. The principal arguments for this type of mannequin are cost (around $8), weight, and comparability with past test results. Instrumented anthropomorphic mannequins clearly yield more realistic estimates of personnel effects. They are being used for selected shots in the Bradley Phase II testing, but according to the draft revised JLF/Armor plan, only the plywood mannequins are scheduled for use in JLF.

Validity of Animal Testing

Second, the validity of the animal testing is questioned. In the 1984 Bradley vaporifics tests, test personnel reported they could not get into the vehicle to release the animals for 20 to 30 minutes following the tests because of the bad post-test environment; a similar phenomenon occurred in BRL tank tests in the 1950's and M113 tests in the 1970's. Such delayed observations do not provide a good picture of the real-time behind-armor effects of flash, overpressure, and other phenomena on personnel. Additionally, the animals must be heavily drugged for humanitarian reasons, exacerbating the usual problems in generalizing to humans from animal experiments. Finally, behind armor effects affect crew members psychologically as well as physiologically. The sudden introduction of brilliant light, choking fumes, swirling gases, flying metal, jarring motion, loud noise, high temperature, and overpressure is likely to have a severe psychological effect. The animals cannot be interrogated as to their psychological condition following a test, and current vulnerability models do not include any psychological effects caused by the penetration; nor do they include the psychological effects on non-casualties from observing casualties in the crew compartment. Reportedly, the Israelis have some observational data from combat on psychological effects, but nothing quantifiable.

Validity of Analytic Methodology

Third, the analytic methodology used for casualty estimation is questioned. According to a BRL paper, there is presently no generally accepted quantitative measure of incapacitation from the prime blast threat. The currently used measure—employed in the Bradley tests and endorsed by the Army Surgeon General—tends to underestimate casualty production from blast. The JLF/Aircraft Master Plan likewise calls the method "not fully satisfactory." Specifically, the analysts' practice of equating casualty production to a one percent lethality curve omits the casualties that would certainly result from lesser levels of blast pressure-duration than defined for this curve.[6] The BRL paper offers a more conservative alternative, but admits that it is subjective, neither supported nor contradicted by the literature.

There is also some disagreement over how much ear damage incapacitates a crew member. The BRL paper maintains that eardrum rupture, and its accompanying pain and hearing loss, can render a soldier ineffective in performing certain tasks, and therefore offers the 50 percent eardrum damage curve as a threshold for incapacitation. However, there is little evidence in the literature to support or reject the claim that ear damage results in a casualty. There is also reportedly little data on the effects of spalling, overpressure, etc. in combination, i.e., as they occur in combat.

Level of Emphasis in JLF

The JLF/Aircraft Preliminary Plan and draft revised JLF/Armor plan are both sketchy on crew effects. The objectives for the aircraft plan are written more as statements of need than objectives, with no explanation of how they will be carried out. None of the FY1985 or FY1986 tests addressed crew effects. According to the JLF/Aircraft program manager, they will not be looking at what kills a pilot, but that the effect of pilot loss will be "factored in" to $P_{K\ H}$. He could not offer any more details. The armor plan lists personnel as a component (this is common in the V L community) and says only that the number of casualties will be assessed with appropriately clad plywood mannequins. The current OSD program manager has asked for more attention to crew effects in the next revision.

The current OSD program manager believes that historically, the vulnerability community has shown insufficient interest in crew survivability issues. As noted earlier, he has cited crew survivability as his principal concern, and has asked the JLF program managers to emphasize it more in their test programs. Given the current state of the art, however, we

[6] The one percent lethality curve is the curve of pressure-duration levels that is lethal one percent of the time.

do not believe that precise estimates of casualties can be a realistic expectation of JLF.

Incentive Structure

Strictly speaking, DOD's incentive structure is not a methodological issue. However, T&E of high methodological quality requires an environment that features and facilitates objective, realistic, and adequately financed testing. Consequently, it merits discussion here.

No Overall Requirement for Aircraft Testing

Though some aircraft are undergoing live fire testing as part of the qualification process (described below), and test officials claim that vulnerability issues have higher priority than in the past, but according to JLF/Aircraft officials, there were still no overall requirements for vulnerability testing of aircraft in the acquisition process prior to the passage of the live fire legislation.

The gap is potentially bridged by JLF/Aircraft, but they have not scheduled any full-up firings before FY1989. Given that the program is already behind schedule and further delays are expected, the actual date may be still later. The phasing logic is perfectly reasonable and appropriate from a tester's viewpoint. However, if there are serious vulnerabilities that can only be detected by full-up firings, they will remain undetected while the system in question continues to be procured.

Lack of Linkage Between Live Fire Testing and Procurement

DOD has not established any linkage from JLF and related live fire testing to the procurement cycle. There is nothing requiring the SPOs to use the test results to improve their systems, and no requirement that production be stopped or slowed down if serious problems are found. The only exception is the Bradley, where the input of live fire testing to procurement has been Congressionally mandated. The new live fire legislation mandates this input for new systems, but will not affect currently fielded systems such as those being tested under JLF or the Army (Bradley excepted). These systems are expected to be in the inventory at least though the end of the century.

Threatened Interests

It is clear from the Bradley situation that realistic, full-up live fire testing can represent a real and unpredictable threat to the "business as usual" of procurement. It logically follows that some interests within DOD—specifically, those rewarded for successfully managing a system through the acquisition cycle—are threatened as well. We do not question the integrity of any of the individuals involved. However, despite claims that the needs of the soldier come first, the current incentive structure seems to support other competing goals.

[material deleted]

Comparison Programs

Comparisons With Past Live Fire Testing Programs

Armor

1) CARDE trials. JLF/Armor documents refer to these tests as the last comprehensive series of live fire tests involving armored targets. They were conducted in the late 1950s at the Canadian Armament Research and Development Establishment (CARDE) as a joint Canadian, U.K. and U.S. effort. Their purpose was to provide data to be used in the selection of the size and type of warhead to be used in the Shillelagh and Swingfire missiles. These tests:

- were intended to assess the lethality of experimental 5, 6, 7, and 8-inch shaped charge warheads rather than fielded munitions;
- consisted of 68 static detonations against M-46, M-47, and M-48 tanks rather than real threat vehicles; and
- employed non-functioning target vehicles that were missing components.

Early JLF/Armor planning called for JLF to be a modern CARDE trial. But the JLF goals of testing fielded weapons and vehicles, with at least some fully combat loaded shots represents an improvement in the realism of test conditions, as compared to CARDE. Nonetheless the CARDE data base formed a substantial part of the foundation for the primary computer vulnerability model (Compartment-Kill) used over the past 25 years.

The function that was settled on to represent the CARDE data was a curve relating the assessed loss of vehicle firepower or mobility to the size of the exit hole produced by the shaped charge at the interior surface of the armor. An analysis by the System Planning Corporation (SPC) of the way the data from these trials were aggregated shows that without the very large 8" charges, which almost always produced a large hole and very large assessed kill values, there is very little trend in the data. Although there is variation in the size of holes for the remaining charges, this variation is not related to assessed loss of mobility or firepower. It appears therefore that all lethality assessments conducted using the computer model constructed around the CARDE data have contained a bias in favor of those anti-armor weapons producing large exit holes behind armor, without consideration of any other effectiveness factors (e.g., blast, overpressure).

2) A-10 GAU-8 (LAVP) tank tests. Between 1978 and 1980 a total of 410 aerial gun firing passes were made by A-10 aircraft against arrays of U.S. M-47 tanks as part of a program to test the effectiveness of the ammunition for the GAU-8 gun.[7] Seven passes were also made against [material deleted] and eleven were made against [material deleted] The tests were called the A10/GAU-8 Lot Acceptance Verification Program (LAVP). The LAVP tests were the only major live fire tests in the U.S. with armored vehicles after the CARDE trials and before JLF. The LAVP tests were a model for the initial proposal of JLF, according to the former OSD program manager. They illustrate some of the potential uses of full-scale live fire tests, as well as some of the difficulties in conducting them and interpreting the results.

The design of the LAVP tests was guided by the effort to be as realistic as possible, with a higher priority placed on realism than on scientific reproducibility. The M-47 tanks were loaded with main gun ammunition, diesel fuel, lubricating oil, and crew mannequins (plywood for the early tests and mild steel for the later tests). They were arrayed in groups simulating [material deleted] and were originally in operating condition. The pilots making the firing passes were instructed to fly at low altitudes and low dive angles to simulate movement through a hostile air defense system but were otherwise unconstrained (recall that in JLF and other current live fire testing, munitions are not being fired from actual operating weapon systems).

After each pass a combat damage assessment team documented aircraft flight parameters and a large amount of information about the location and effects of each projectile hit on the tanks The data were published by the Naval Postgraduate School and were eventually stored in an information retrieval system at Eglin AFB so that the V L community

[7] LAV P tested effectiveness rather than lethality, because actual operating weapon systems (flying aircraft) were used to fire the shots, and targets (tanks) were placed in operational formations

has access to the test results. The results have been used to check the predictions of vulnerability models.

Despite the effort to be as realistic as possible, the LAVP tests did depart from combat realism in a number of ways. In addition, they raise some other issues that are relevant to live fire testing in general:

- Although the M-47 is in the same class of tanks as the [material deleted] of interest, most of the targets were surrogates rather than [material deleted] vehicles.
- The electrical systems and engines of the tanks were not operating during the tests and were not tested before or after the firings.
- The tanks often were missing some components which, while not themselves critical, are thought to shield critical components in combat configured vehicles, so that damage levels could be higher in the tests than they would be in combat.
- For safety reasons high explosive warheads could not be installed in the main gun ammunition rounds stored in the vehicles, so conventional assumptions about the consequences of hits on these rounds were used.
- Fuel and oil were not heated to operating temperatures, and so were less sensitive to fire than they would be in combat.
- The use of multiple-hit firing passes, while realistic in allowing any synergistic effects of several nearly simultaneous impacts to occur, made it difficult to estimate a kill probability for single hits.
- The use of a non-standard damage assessment procedure has made some vulnerability analysts reluctant to take account of the results in their work. However, the former OSD JLF program manager and at least some outside experts have a different explanation: they think modelers have avoided and discounted LAVP out of fear that it would expose weaknesses in their models.

Aircraft

1) Test and Evaluation of Aircraft Survivability (TEAS). As noted earlier, TEAS was the only U.S. systematic live fire testing program of aircraft. According to the JLF/Aircraft program manager, the TEAS test providing the best model for the upcoming JLF full-up testing was the F-4 test. In this test, Soviet projectiles, were fired at a full-up operational F-4A, with the emphasis on fuel system vulnerability. However, several key environmental factors were not simulated: airflow, altitude, altitude history, maneuver load, and slosh. It is not clear why airflow was omitted, given that other TEAS tests included it (e.g., A-7D). As noted earlier, airflow is being simulated in relevant JLF tests, but there is no still no satisfactory capability to simulate the other environmental factors.

A 1973 evaluation by DOD's Weapons Systems Evaluation Group (WSEG) found a general absence of systematic planning of tests, with few written test plans available for review, and tests underway with no written plans. In this regard, JLF/Aircraft has clearly improved upon TEAS; all individual test plans for FY85, FY86, and FY87 were written before testing began. However, the major substantive concern stressed in the TEAS evaluation remains unresolved in JLF. This was that the then current methodology for estimating vulnerable areas had two major deficiencies: it did not provide for the validation of estimates ($P_{k/h}$'s) and it failed to reveal the uncertainty in the estimates. WSEG regarded these as "grave omissions", casting doubt on any estimate produced and making it impossible to resolve disputes over estimates (estimates can vary considerably; for example, a JTCG/Aircraft report estimated the vulnerability of the bottom aspect of the A-7D to be 3.6 times higher than the aircraft manufacturer's estimate for the same threat). They concluded that the deficiencies would continue unless the TEAS program developed the necessary scientific discipline and data base to substantiate vulnerability estimates. Fourteen years later, the conclusion is equally applicable to JLF/Aircraft.

In sum, there appear to have been improvements in program planning and simulating realistic environments, but little or no improvement in producing scientifically valid vulnerability estimates.

2) Qualification testing. Between the termination of TEAS in the mid-70s and the beginning of JLF, there were no live fire testing programs with the objective of quantitatively assessing vulnerability. However, all three services have used live fire testing in the qualification process for at least some new aircraft. Typically, an aircraft will have a specific survivability requirement, e.g., the engine must survive a 12 7 mm threat. Live fire testing can be used to validate that this requirement has been met. Examples include survivability of the A-10 (fuel cells and structures), the F/A-18 (engine fuel ingestion), and the UH-60 (various components).

Many of these tests are live fire by any definition, with Soviet rounds, full-up components, running engines, and airflow where appropriate. However, the objectives are much more limited than in TEAS or JLF. For example, such a test will attempt to determine whether a particular threat kills an aircraft or component; it will not necessarily attempt to characterize the kill mechanism, extrapolate to different size or type threats, or generally enhance the vulnerability data base in a systematic way. Although targets are provided by the program office, availability of targets is still a serious constraint on test design.

Comparisons With Foreign Live Fire Testing Programs

[material deleted][8]

U.K.

The U.K. has a live fire test program for aircraft. It was described as similar to JLF, with component vulnerability trials followed by "proof" tests on a complete aircraft. However, the U.K. testers reportedly prefer using surrogate munitions (developed from captured threat munitions) to using actual threats, and redesigning or hardening current aircraft does not appear to be a program goal: rather, the emphasis is on validating methodology for estimating vulnerability. Target availability may be less of a problem than in the U.S; as soon as an aircraft becomes surplus, the live fire program has first claim on it. This contrasts with the U.S. program where numerous other programs would compete. On the other hand, U.K. testers typically have less money for testing.

JLF Aircraft has a cooperative agreement to exchange live fire test plans and reports with the U.K. Two of the aircraft being tested are common to both nations (AV-8 and UH-60). A U.S. tester described the U.K. plans as ingenious, particularly with respect to design efficiency.

There is no analogous cooperative agreement between the U.K. and JLF Armor. A BRL official in contact with the European V L communities had not seen any evidence of British live fire testing of actual armored vehicles. A British embassy official told us they are very concerned with armored vehicle vulnerability, and do some live fire testing. From their description, they may focus more than the U.S. on finding improvements than on satisfying the requirements of the system being tested. Other sources report extensive U.K. testing against instrumented armor-bound simulated vehicles, leading to important contributions on vaporifics and other behind-armor effects.

[material deleted]

[8] [material deleted]

Conclusions

In this chapter we addressed the evaluation question, "What has been the methodological quality of the test and evaluation process?" Our conclusions follow.

Program Objectives

- With the exception of the model validation objective, the four JTCG objectives for JLF are not stated in an specific evaluable way. There are no specified comparisons to be made or criteria to be met, only a statement that the state of knowledge on the vulnerability or lethality of weapon systems will somehow be improved. This vagueness means that three of the four objectives can appear to have been accomplished regardless of the methodological quality, cost-effectiveness, or usefulness of the program.
- More specific objectives, such as performing empirical comparisons between V L improvements and baseline configurations (as in Bradley Phase II), would allow more useful information to be produced.
- The model validation objective will not be accomplished in a scientifically defensible way; however, it is likely the models will at least be improved. The extent of the improvement will depend on the test results and how they are interpreted by the V L community.

Armor

Overall Planning

- The persistent failure of OSD and JLF Armor test officials to reach an agreement about the approach to be taken to live fire test design has caused delays of implementation and waste of JLF resources in repeated plan revision. Consequently, the first DTP was still not in final form after two years.
- In important respects, the October, 1986, draft revised JLF Armor master plan resembles the 1984 version, which had been rejected by OSD because of inconsistency with the objectives of JLF. The 1986 version specifies that 1) approximately two-thirds of the shots will be warhead characterizations or studies of behind armor debris, rather than shots on vehicles, and 2) target condition is mostly inert or semi-inert, with only 20 percent of shots on vehicles to be full-up.

Setting Test Objectives

- In their present draft form, JLF/Armor outline test plans generally repeat one or more of the JTCG statements of overall JLF objectives as major test objectives.˙Each outline plan also has one or more specific objectives. Some of these may be infeasible.
- The Bradley Phase II objectives are much more specific.
- The objective of training damage assessors as part of a single test was unrealistic.

Test Planning

- Early DTPs were primarily driven by target availability and the data needs of modelers. Newer plans are more realistic in their inclusion of lead time for obtaining or developing hardware, but some specify that targets may not be available.
- The Bradley controversy has led to a very rigidly specified live fire test plan which leaves little to the judgment of testers on the range. The Bradley Phase II plan is the most detailed and thoroughly specified live fire test plan produced to date.
- The Bradley Phase II plan places greater explicit emphasis on casualties and on fire and explosion than previous Bradley or JLF/Armor live fire tests. However, it misstates to some extent the position of the BAST group assigned to develop the test's shot selection methodology, and proposes the use of a questionable statistical test.
- Testers are very sensitive to test efficiency from an engineering standpoint, i.e., designing tests to conserve targets and prevent testing effects.

Implementation

- Within JLF/Armor, a training and demonstration test has been implemented. It departed from the plan in a number of ways, primarily due to changes in target availability.
- The implementation of Bradley Phase I was a source of considerable controversy. To avoid recurrences, the DTP for Bradley Phase II requires explicit OSD approval for departures from the plan.

Analysis and Results

- Surrogate munitions stored in the [material deleted] vehicle may have reacted more violently than actual munitions would have, potentially biasing the Bradley vs. [material deleted] comparison in favor of the Bradley; however, this was not reported to Congress. We believe that the use of surrogates and questions about their equivalence to actual Soviet munitions should have been reported.
- The two JLF/Armor reports were preliminary drafts that provided little indication of how the data will eventually be analyzed.

- Other than assigning damage assessment values to the shots, no attempt was made to analyze the M-48 tank training test.
- The IDA report's treatment of the M-48/[material deleted] test as a methodological comparison is questionable.

Aircraft

Overall Planning

- In general, JLF/Aircraft planning has been well organized and thorough.
- JLF/Aircraft established a formal process to designate test priorities; however, test priorities were actually driven by more pragmatic concerns (target availability and the need to ensure tri-service cooperation).
- The principal constraint on realism is the inability to simulate flight conditions on the ground. Airflow is used to simulate airspeed but the coverage area is small, and other environmental factors affecting fire are not simulated at all.

Setting Test Objectives

- In FY85 and FY86 DTPS, JLF/Aircraft specified objectives congruent with the version of the program objectives they had established. These were generally feasible, with the exception of objectives related to determining probabilities.

Test Planning

- JLF/Aircraft test designs are generally congruent with test objectives, efficient with respect to conserving targets, and realistic given their limited objectives.
- Some DTPS specified target requirements which exceeded the availability of those targets.
- Testers are highly sensitive to test efficiency from an engineering standpoint, i.e., designing tests to conserve targets and prevent testing effects.
- DTPS omit key information (e.g., data analysis plans) and are inconsistent in selection of threat velocities.

Implementation

- To the limited extent we could observe them, departures from test plans have generally been reasonable.

Analysis and Results

- Only one draft report has been completed—the F100 engine steady state fuel ingestion test. This report omitted key information, overstated the generalizability of results, and presented a highly questionable model. Recommendations were congruent with results and sensitive to the likelihood of user acceptance.

General Issues

Conflict Over Objectives

- The JLF charter did not define live fire testing well enough to give test designers a clear direction.
- There have been several conflicting versions of the objectives of JLF and live fire testing in general. This appears to have in part resulted from the decision to task the JTCG's to implement JLF.
- The conflict over objectives reflects underlying differences between the interests of proponents of full-up testing and those of modelers, resulting in largely incompatible approaches.

Availability of Targets

- The principal constraint faced by all JLF test officials is a lack of targets. This is in part a result of inadequate planning; there is no assigned responsibility to provide targets and related support to JLF. Consequently. test officials have had to spend a substantial portion of their time "selling" the program to skeptical service components.
- The systems and components that JLF does receive are frequently in poor condition, yet JLF provides no funds for restoration.
- JLF has been further hindered by competing governmental and non-governmental interests and negative attitudes toward destructive testing.

Statistical Validity

- In general, the sample sizes of JLF and related live fire testing have not been sufficient to produce statistically reliable results. This would be a problem even if the number of targets listed in the test plans could be obtained.
- The statistical input to JLF has been minimal and had little effect, and the few applications of statistical analysis to live fire test data thus far are highly questionable. Several efforts are underway to make live fire tests more statistically interpretable.
- As a substitute for statistical analysis, engineering judgment—which is heavily relied upon throughout the V L assessment process—has little scientific validity, being subject to individual and collective biases.

- The most common form of vulnerability/lethality indicator—probability of a kill given a hit $P_{k h}$—has not been demonstrated to be reliable or valid.

Shot Selection

- Controversy over shot selection is to some degree a conflict between sampling efficiency and the desire to avoid bias at all costs.
- Random sampling from combat distributions is a reasonable way to preclude intentional or inadvertent bias in shot selection. However, sampling from a uniform distribution avoids tester bias <u>and</u> biases in the combat data.
- The shot selection problem will not be resolved by technical solutions alone. An interim solution might be to designate that some proportion of shots be selected judgmentally and others randomly, but ultimately, it appears impossible to agree on how to select live fire shots without first deciding on test objectives.

Human Effects

- JLF plans do not provide an adequate treatment of human effects.
- The claims of some JLF officials that personnel vulnerability is well known are overstated
- Given the current state of the art, it is unlikely that JLF will produce precise estimates of casualties.

Incentive Structure

- DOD's incentive structure is not entirely conducive to realistic live fire testing. [material deleted]

Comparison Programs

Past Programs

- The state of the art of live fire testing has improved since prior live fire testing programs, but some potentially solvable problems raised earlier have not been solved. For example, little progress has been made in the empirical validation of V L estimates ($P_{k h}$s).

Foreign Programs

- [material deleted]

Summary Conclusion

- There is little completed testing on which to base a methodological evaluation. However, it is apparent that the technical capability to do full-up testing is not well developed. This is partly due to the historically low emphasis on live fire testing in the U.S.

What Are the Advantages and Limitations of Full-Up Live Fire Testing, and How Do Other Methods Complement Full-Up Testing?

Advantages of Full-Up Live Fire Testing

Full-up, live fire testing offers a unique advantage over all other methods of V L assessment. It is the only method providing direct visual observation of the damage process caused by a weapon/target interaction under realistic combat conditions. Consequently, it is less reliant on unverified assumptions than other methods (e.g., ammunition is on board; therefore, analytic estimates of what *would* have happened had ammunition been on board are unnecessary).

A single shot can identify an excessively vulnerable component or unexpected kill mechanism, reveal model flaws, and generally provide qualitative insights into the vulnerability or lethality of the system and how to improve it. A few shots can tentatively characterize these phenomena, e.g., show how vulnerability progresses with threat size. Matched comparison shots, such as those being fired at the Bradley M3 and M3(HS), can be particularly useful because absolute measures of vulnerability are not required. The descriptions of directly observable damage that full-up testing provides are regarded as highly beneficial to users.

Despite the meager amount of live fire testing to date, there are already several examples of live fire "surprises", i.e., results that were not predicted, and might not have been detected by other means of testing or analysis.

- The Air Force introduced a new hydraulic fluid, 83202, which laboratory tests had demonstrated to be less flammable than their standard hydraulic fluid, 5606. However, the JLF F-15/16 hydraulic fluid live fire tests with airflow suggested the opposite: 30 percent of shots on 83202 resulted in fires, compared to 15 percent of shots on 5606.
- In the A-6 dry bay foam tests (pre-JLF), the effectiveness of reticulated foam in preventing fires was tested. As expected, the foam reduced the likelihood of the dry bay catching fire, relative to the baseline (no foam) condition. However, when the foam *did* catch fire, the fire was more severe than in the baseline condition The results were later used to persuade NAVAIR and the aircraft's manufacturer against using that particular foam.
- In the Bradley Phase I tests, the automatic fire suppression system (AFSS) false alarm rate proved to be unexpectedly high. The halon bottles discharged even though there was no fire on 45 percent of the shots into the space protected by it.
- In the Bradley Phase I tests, direct hits by primary penetrators on the explosive or propellant sections of on-board ammunition were shown to pose the most significant threat to the Bradley and its crew, though some impacts appear to be survivable. Although previous tests and

other data indicated the threat posed by penetrator impacts on stored ammunition, the findings that some impacts will not produce catastrophic results were disclosed only by the live fire tests.

Limitations of Full-Up Live Fire Testing

Cost

The primary limitation of full-up live fire testing is cost, principally target costs. High testing and restoration costs also contribute substantially to the total. The Bradley testing was estimated to have cost, as of December, 1986 (midway into Phase II), $30-35 million. The M1/M1A1 tank testing is expected to cost $55 million.

Target Costs

U.S. front line armored systems currently cost as much as $3 million, aircraft as much as $35 million. Threat systems pose problems of availability as well as cost. By its nature, full-up live fire testing is destructive so reuse of targets is limited. Target preservation is the principal rationale for departures from full-up configurations in V L testing, i.e., to exclusion of fuel, ammunition, and/or hydraulic fluid. Unfortunately, these ingredients are generally considered to be the principal contributors to casualties and target kill, and therefore the main reason for doing the test. And they are subject to complex interactions (i.e., synergistic effects) that cannot be assessed by separate component tests.

Testing Costs

Setting up and conducting a full-up test requires elaborate facilities (including stringent safety precautions). It also requires time consuming post-shot data reduction and analysis, and the close attention of senior officials.

Restoration Costs

When the target is salvageable, damage repair is also costly. Considerable care must be taken to ensure that targets are fully restored between shots, so as to minimize testing effects. When tests are conducted with flammable substances on board, the time and cost of restoration are greatly increased. [material deleted]

Testing Costs as a Proportion of Program Costs

We believe that testing costs need to be viewed in the context of total program costs. By one recent estimate, the total cost of acquiring 6,882 Bradleys will be $10.74 billion. Thus, even if the total cost of live fire testing of the Bradley were to reach $50 million it would still be less than one-half of one percent of the total program cost. And given the interruptions and redirections that have plagued the Bradley testing, and the fact that it was the first armored system tested (i.e., initial costs of instrumentation, etc., had not been absorbed), its testing costs may be unrepresentatively high.

Limited Information Yield

Full-up testing has been criticized within the V/L community on the grounds that catastrophic kills (i e., shots that result in the destruction of the target) yield very limited information. In their view, a vehicle or aircraft is sacrificed for essentially one data point—whether or not it blew up. The reason is that all evidence of the kill mechanism, as well as much of the instrumentation for recording it, is destroyed with the target. Proponents of full-up testing maintain that the information that is obtained is the key information needed, both for assessing and reducing vulnerability, and that full-up testing is the only way to obtain it.

Different viewpoints aside, it is clear that a completely destroyed target leaves little record of the means of its destruction. Shotlines cannot be traced, component damage cannot be studied, and little is produced of use to V/L modelers. To the extent that this information is important to the V/L assessment process, it is better obtained by other means. However, the "one data point" argument is somewhat extreme. A full-up target is not completely destroyed or even significantly damaged each time it is shot. When it remains intact, it can provide much of the same detailed damage assessment as an inert target, without the uncertainties of analytic assumptions required by the inert target.

Limited Generalizability of Findings

Testing provides point estimates of damage for selected values of controlled variables. However, V/L estimates (qualitative as well as quantitative) are required over a range of values for these variables, so these point estimates must be generalized to make inferences about conditions not tested. Where a reasonable approximation to a true continuum can be tested, as in some component and subcomponent testing, generalizability is less problematic. It is particularly problematic with full-up live fire testing, where typically only a small proportion of relevant test conditions can actually be tested.

Generalization takes two forms: interpolation within the range of tested conditions, and extrapolation outside the range. Extrapolations are clearly the riskier of the two. They must receive additional guidance from experience or from an understanding of the physical principles involved, yet experts in the V L community admit that these principles are not well understood with respect to full-up testing. The IDA report noted that generalization by interpolation is useful and usually valid. However, they neglected to point out that even interpolation is complicated in live fire testing by the statistical unreliability of the point estimates. If one of those estimates is an atypical result, then interpolations computed from it will be atypical as well.

Limited Redesign Opportunities

JLF aircraft and armor test officials and outside experts believe that live fire testing of developed systems can have, at best, limited impact on those systems. From preliminary development on, designs tend to be "frozen", making major changes prohibitively expensive. For example, a live fire test might reveal an aircraft wing to be excessively vulnerable. Even if the aircraft is not yet in production, the wing design would have been frozen, possibly for years. Any change would change the performance of the fuselage and all related stresses. Weight tolerances of tactical aircraft are extremely narrow, so additions of even a few pounds are problematic. This is not to suggest, however, that important vulnerability modifications are never feasible after development (e.g., the addition of reactive armor tiles to the Bradley vehicle).

The new live fire legislation (Section 910 of the FY87 defense authorization act) specifies that live fire testing must be completed before proceeding beyond low rate initial production (LRIP). LRIP will produce targets that are reasonably representative of the final version. This is desirable from the standpoint of realism; however, testers claim that generally, it is already too late to incorporate significant vulnerability reductions into designs. They recommend live fire testing of components during development (also specified in the legislation) and "proof" testing at the end. For the same reason, they see the main benefit of JLF and other live fire tests of fielded systems as reducing vulnerability of future systems through lessons learned.

How Do Other Methods Complement Full-Up Live Fire Tests?

In view of the above limitations, other methods are brought into the V/L assessment process. Two of these—subscale testing and inert testing—are types of live fire testing. Two others—analysis of combat data and modeling—are not. We discuss all four here, but in light of the controversial role of modeling in live fire testing, modeling is the main focus.

Subscale Testing

Test firings against system components are a common source for vulnerability data. They can generally support larger sample sizes than full-scale tests, and are useful in determining the boundaries of effects and providing input to prediction models. Like full-scale targets, components can be inert or full-up, depending on the objective (e.g., a fuel tank can be tested inert to assess structural damage or full-up to assess probability of fire). More basic than component tests, but still related, are terminal ballistics tests (for vulnerability) and munitions performance tests (for lethality). Terminal ballistics tests, in which armor plate or other materials are the objects of test firings, are not necessarily associated with specific developmental items. but rather, contribute to the data base and provide insights on component vulnerability. They are particularly useful in the development of physical theory for munition effects on target elements (e.g., armor plate). Munition performance tests provide fragmentation distributions of space, mass, velocity, etc. Their use parallels that of terminal ballistics tests, but from a lethality perspective.

The principal limitation of subscale tests, whether full-up or inert, is their failure to provide direct evidence of interactive (i.e., synergistic) effects on realistic targets. For example, an aircraft fuel tank can be damaged, leak fuel, but not result in aircraft loss. An engine can ingest fuel, stall, recover, and not result in aircraft loss. However, when the two systems are integrated by the engine inlet, aircraft loss can occur. The fuel now leaks into the inlet, flows into the engine, and detonates; the resulting flame propagates forward in the inlet to the damaged fuel tank, and ignites the leaking fuel. Even though the engine recovered, the aircraft is lost due to the fuel system fire. Subscale testing can supplement full-up, full-scale testing, e.g., in design and interpretation, but cannot substitute for it.

Inert Testing

Inert testing of full-scale targets is superior to full-up testing in characterizing mechanical damage to individual components caused by the residual penetrator and spall. Assessors can directly compare the functioning of individual components before and after each shot. Since flammables are not onboard the target in inert testing, targets remain in testable condition longer before having to be discarded. Additionally, fewer internal components require replacement between shots. All are reasons why the V L community tends to prefer inert over full-up testing.

Advocates of full-up testing view mechanical component damage as of secondary concern. To them, making the vehicle inert removes the primary contributors to casualties and target kill: flammable substances causing catastrophic damage. Catastrophic damage cannot be directly observed from shots on inert targets. In the M-48 tank tests in September, 1985, assessors were instructed to score a K-kill if and only if the casing of any ammunition round was penetrated. This particular method—which followed the standard damage assessment guidelines for armor—can underestimate the true likelihood of a K-kill because catastrophic fires may also result from hits in other locations. However, even if a more realistic method were used, inert testing could still only be used to infer catastrophic damage through an indirect analytical process, intrinsically limited by the current state of knowledge. Such information is not a direct result of the test shot, as it is with full-up testing.

Combat Data

Combat data provides information from realistic, full-up interactions of weapons and targets. By definition, combat data provides greater realism than any other source. It can be used to obtain aggregated survivability measures, such as kill or loss rates, as well as likely direction of fire, distribution of hits, vulnerability of subsystems, and critical vulnerability interactions. Analyzing available data is considerably less expensive than testing. The Israelis report frequent use of combat data to improve the survivability of armored vehicles.

One philosophy of live fire testing argues that testing should approximate combat in any way possible, including tactics and formation: scientific control and related technical concerns are secondary. According to this view, combat data is a more useful tool for V L assessment than the controlled testing characteristic of JLF. Nonetheless, combat data are obviously limited to munitions and systems that have actually been employed and may not represent systems of interest. JLF includes some systems that have been in combat, while the live fire tests required by the FY87 authorization legislation will, by definition, be confined to nonfielded systems. Combat data can be useful for designing live fire tests of such systems, as discussed earlier. But even with systems having been employed in combat, combat data provide less scientific control than testing, and offer no view of the damage process, only the results.

Modeling

As noted previously, the assessment of the survivability and effectiveness of U.S. weapons has come to depend increasingly on computerized V-L models over the past twenty-five years. This trend has corresponded with a period of rapid cost growth of weapon systems, which has tended to limit the amount of full-up live fire testing that is feasible. V-L models are seen as a potential solution to the problem of high testing costs. Assuming a model is valid, it can greatly increase the generalizability of a few live fire test shots. Additionally, models have the unique advantage of applicability to systems not yet built. This permits a greater range of redesign possibilities than tests on completed systems. The output from V-L models is also used in a variety of other activities in DOD, including war games, simulation models, weapons design, and logistical planning for repair times and stocks of spare parts.

Both live fire tests and models are ways of assessing vulnerability or lethality. A considerable part of the controversy over the planning and direction of the JLF/Armor tests stems from differing positions over the proper role of V-L models in live fire tests and the relative value of models and live fire tests in determining vulnerability and lethality. Similar models are used in assessing aircraft vulnerability, but their role in the JLF/Aircraft tests has been less controversial. This is largely because they are viewed as less central to the design and interpretation of the aircraft tests.

The Role of Models in Live Fire Testing

The position taken by the vulnerability analysts at BRL is that live fire test data alone are not sufficient for determining target vulnerability and weapon lethality. Targets are complex in their geometries and composition, they have many different components, and there are many different types of munitions. It would therefore be prohibitively expensive and time consuming to conduct live fire tests of the effects of all types of munitions over all the surface area of all potential targets.

Because of the practical limitations, BRL argues that live fire test data should be used to "provide critical input and ultimate calibration of evaluation models" which currently exist. Theoretically, if there were well-validated models of the effects of threat on targets it would be possible to provide accurate predictions about the vulnerability and lethality of at least some of the targets and munitions which have not been directly tested, including those which do not yet exist.

1) Armor. The V L models currently used by armor modelers are intended to permit the integration of ballistics test results, geometric descriptions of the targets and the characteristics of munitions to permit predictions of the results of the impact of a particular weapon at a particular location, angle, etc. on a target. Calculation of such predictions for all possible shots then permits general statements about the overall vulnerability of a U.S. armored vehicle to a particular threat, or the lethality of a weapon against a threat vehicle. The mapping of a vehicle's vulnerability over its surface or the determination of an aircraft's "vulnerable area"—or some summary index derived from the maps—constitute the quantification of vulnerability or lethality called for in the JLF charter.

The approach advocated by the BRL for assessing the vulnerability of combat vehicles is illustrated in Figure 4.1. It shows the place of full-scale live fire tests among modeling and subscale tests, including the testing of components and armor. BRL acknowledges that they have followed this approach for the past two decades, but without benefit of full-scale testing. In this sense, the modelers have been operating in an "open loop." Unable to realistically test the accuracy of their predictions, they have instead relied on engineering judgment and subscale tests to provide input values to the models.

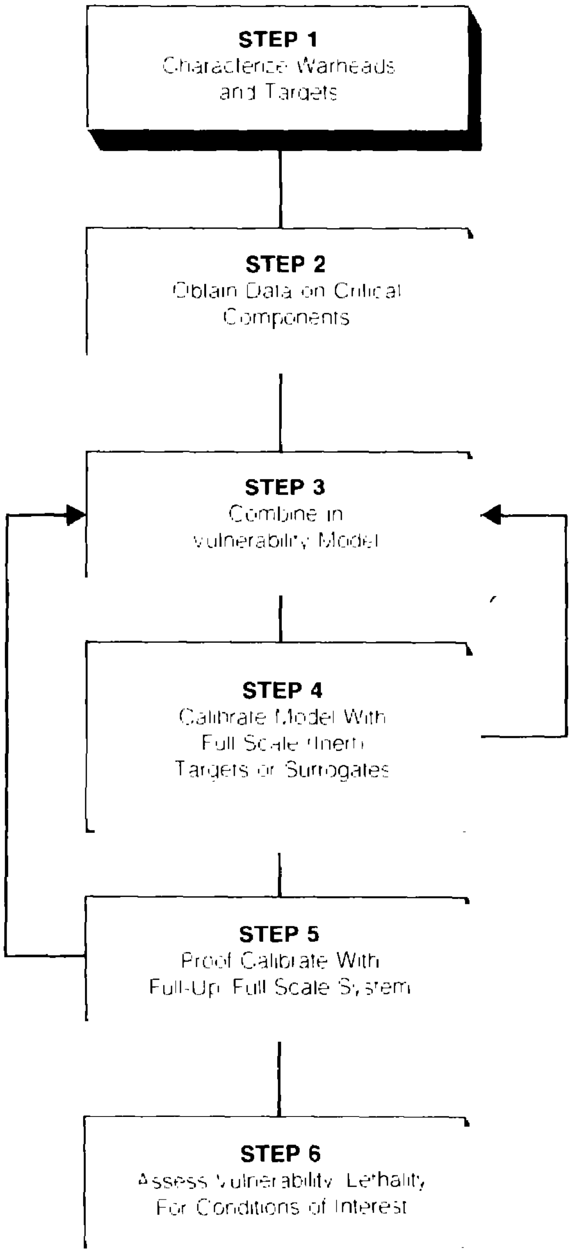
From this point of view the role of V L models in live fire testing is first to "support" the tests. The models:

- guide shot selection in test planning,
- provide pre-shot predictions of the effects of the shotlines selected, and
- assist in the graphic display of test results.

The Bradley Phase I live fire tests, for example, used the output from a vulnerability model in each of these ways.

The second aspect is the use of the tests to improve the models. The vulnerability modelers argue that if live fire test results can be used to improve the models, the value of the tests is spread to the other uses of the V L models, in more accurate V L predictions for weapons that cannot be tested, and in enabling vulnerability reduction programs to examine

Figure 4.1: Approach to Vulnerability Assessment Preferred by BRL


STEP 1
Characterize Warheads and Targets
STEP 2
Obtain Data on Critical Components
STEP 3
Combine in Vulnerability Model
STEP 4
Calibrate Model With Full Scale Inert Targets or Surrogates
STEP 5
Proof Calibrate With Full-Up Full Scale System
STEP 6
Assess Vulnerability Lethality For Conditions of Interest

the effects of proposed design changes. Any improvement in the predictive accuracy of models would thus also benefit the other uses to which V L models are put.

2) Aircraft. The JLF/Aircraft tests employ models similar to those used by armor analysts. These models generate shotlines through geometric descriptions of targets, and predict the effects of specific shots—what will be hit, penetrated, etc. Models are also used to calculate the vulnerable area of an aircraft perpendicular to a shotline, based on the vulnerability of components and their geometries. There are also more specialized models to simulate operational aspects of aircraft such as the gravitational loads on wings. There is some use of non-computerized models, such as formulas on what happens to metal under ballistic threats.

There is, however, less evidence of a unified position on the role of models in aircraft testing among the members of the aircraft survivability community we interviewed. Officials told us that TEAS had led to a skeptical attitude toward existing vulnerability models in JLF/Aircraft, because fire and explosion were unpredictable. The JLF/Aircraft Preliminary Plan does acknowledge that a number of the models currently in use are inadequate or lack validation and modelers described to us several ways in which they anticipate improvements to models as a result of JLF tests. It appears that models and the concerns of modelers play a less central role in the planning, design, and interpretation of aircraft survivability tests than in the armor tests.

3) Critics of modeling. The former OSD program manager and other critics of computerized V L modeling maintain that vulnerability modeling is not credible. They claim that:

- U.S. vehicles and aircraft procured on the basis of computerized vulnerability assessments have been proven excessively vulnerable.
- anti-armor weapons have also proved to be less effective against threat vehicles than had been predicted by the models.
- fire and explosion, which are among the most important sources of casualties, are among the phenomena handled least well by vulnerability models.

The former OSD program manager claimed that the models could only be relied on if they were to be thoroughly validated by test results. Such model validation would, he claims, require thousands of shots. He argued that it is not practical to conduct these tests, and so models should not guide the selection of live fire shots. If modelers are able to use the results of randomly selected shots to revise or validate their models, this is a secondary benefit of JLF But, he claimed, the calibration of models is not JLF's primary purpose, and identifying vulnerabilities and testing improvements does not require models.

We believe that the two sides in this dispute are to some extent arguing at cross-purposes. It is impossible to determine the "proper" role of computerized vulnerability/lethality models in JLF unless there is agreement on the purposes of the tests. As we have noted, almost all the statements of JLF/Armor objectives explicitly mention the calibration or validation of vulnerability models.

We have not found evidence that vulnerability models have played as great a role in the design of live fire tests as some statements by modelers would indicate. In fact engineering judgment and the intuitions of those familiar with the design of particular systems are reported to play a major role in the selection of live fire shots, for both the armor and the aircraft tests planned so far in JLF. The draft revision of the JLF Armor plan does clearly reflect the interests and data needs of the modeling approach. It would devote a substantial part of JLF resources to the subscale tests that are mainly of use to vulnerability modelers.

Models Currently Used in Vulnerability/Lethality Assessment

We outline the two main armor models and describe some of the aircraft models, in order to make clear their assumptions, requirements, and limitations

1) The Compartment-Kill model. The first major computerized vulnerability model for armored vehicles was based on the data from the CARDE trials. It is called the Compartment-Kill model. Although it is more than two decades old, the Compartment-Kill model is still used to produce most of BRL's input to war games. By one estimate, 95 percent of such requests are still met by output from the Compartment-Kill model.

The Compartment-Kill model traces a shotline through a geometric representation of an armored vehicle to the point at which it enters either the crew compartment or the engine compartment. At that point it treats the inside of a tank as a "black box." Empirical relationships are then used to determine the effects of the round hitting that point. The most important of these is the relationship between the expected size of the hole produced by a penetrator and the expected loss of mobility or firepower. The curves expressing these "damage correlations" were originally derived from the CARDE trials. Ammunition (and in some versions fuel) are the only internal components directly assessed in the vehicle interior by the Compartment-Kill model. If they are intersected by the shotline, a K-kill is generally declared.

2) Internal Point-Burst model. In the early 1970s, dissatisfaction with the Compartment-Kill approach's lack of detailed modeling of damage mechanisms and reliance on older data sets led modelers to develop a more sophisticated modeling approach. It is called "internal point-burst modeling." In this approach the internal interactions of the penetrator and the armored vehicle's components are simulated in great detail. This approach includes tracing the main penetrator all the way through the armor of the tank and any of the internal components it impacts. The cone of spall particles produced at the inside of the armor is also modeled in a separate submodel. The approach further requires that each component in the vehicle that may be impacted by a penetrator or spall undergo a vulnerability analysis of its own to determine the mass and velocity or energy of an impacting object required to damage it to a certain level, a component P_{kill}.

3) Aircraft models. COVART is the model used by JLF/Aircraft that is most like the vulnerability models used in the armor tests. It simulates impacts on a target by warhead fragments, armor-piercing projectiles, and armor-piercing-incendiary projectiles. The program calculates target vulnerable area, component vulnerable areas, and expected repair times. It depends on a detailed geometric target description, and probes it with shotlines like the ones in the armor V/L models. It calculates the effects of a penetrator using the standard JTCG/ME penetration equations. In addition to the slowdown of the penetrator as it passes through components, the equations compute slowdown in fluids and, for projectiles, change in yaw angle, incendiary functioning and the break up of the penetrator's core. Like the point-burst armor model, COVART requires input of P_{kill}s based on vulnerability analyses of components.

FASTGEN, another computer model used by aircraft testers, gives a grid of paths through a component, generating a large number of possible shotlines. It is not a vulnerability model per se, but is used in the selection of shotlines, as is GIFT. Selection of actual shotlines for testing is a matter of several considerations including combat data, engineering judgment and the goal of the tests.

MAGNA stands for "Materially and Geometrically Nonlinear Analysis." It is strictly a structural model of components. It has been used mainly to try to determine the residual strength (of, say, a wing) after a shot.

Loads are put into the model, just as they are in the test, then the modelers "damage the model" (simulate a shot into the wing) and check the stresses that changed between the undamaged and damaged states for evidence of failure.

Assumptions and Limitations of Vulnerability Models

One way of assessing models apart from their ability to predict test data is to examine their assumptions and limitations.

1) Compartment-Kill. Although its relative ease of use means that the Compartment-Kill model is still used by BRL for the majority of requests for vulnerability assessment it has a number of severe limitations:

- Its dependence on a single parameter, hole size diameter, is highly questionable. Many other features of warhead/target interaction may be important in producing damage.
- It is based on data from one kind of warhead and obsolete vehicles that did not contain many of the kinds of components introduced into newer vehicles.
- Because it does not treat components in detail it is not suitable for vulnerability reduction tests, and in general is not thought suitable for support of full-up live fire tests like those in JLF.
- It has been shown to fail to predict combat and test results in a number of cases.

2) Point-Burst. The point-burst model approach is in principle capable of much more detailed representation of the events that occur when a munition impacts an armored vehicle and penetrates the armor, but it has additional limitations

- It requires much more input data from armor and component testing to function as it was intended.
- The component data and warhead characterization data are often lacking for newer items, and the input to the models is based on engineering judgment instead.
- It requires much more detailed geometric description of the vehicle and its components than the Compartment-Kill model.
- It requires much more computer time and is therefore more expensive to run, although this is less of a problem with newer computers.

3) General. The armor and aircraft vulnerability models also share assumptions and limitations.

- Target geometries are assumed to be correct at the level of detail required by the model. The Bradley tests indicate that this is not always the case.
- The basic physics of warhead-target interaction is not well understood. The characteristics of munition and armor types cannot currently be inferred from physical laws. The modeling approach requires that interactions of warheads and armor be "characterized" in extensive subscale tests. Even if such data have been obtained, whenever new armor or warhead designs are developed it is necessary to conduct new tests to characterize their interaction. Many of the existing armors and munitions have not yet been tested, so the models are currently dependent on engineering estimates rather than test data.
- There are different versions of the models in existence in several locations, and frequent modification of their code has caused the versions to diverge in ways that sometimes produce very different results.
- Many effects of importance in producing damage and casualties are not yet well modeled or are not included in the models at all. Among these are fires and their propagation; explosions; the effects of multiple hits on a component; the synergistic effects of different damage mechanisms such as shock and fragment hits; ricochets in the interior of a vehicle; and effects on humans from blast, shock, flash, overpressure, acceleration, etc.

$P_{k/h}$s as Measures of Vulnerability

Although V/L models can generate different forms of output, the form most commonly used is $P_{k/h}$. The original formulation of armored vehicle assessment methodology defined three kinds of kill: mobility (M), firepower (F) and catastrophic (K). Mobility and firepower $P_{k/h}$s are often not true probabilities or even subjective estimates of probability:

- They are generated by comparing the damage caused by a hit to a Standard Damage Assessment List (SDAL) and reading the percent loss of function associated with the loss of a particular component.
- A 50 percent M-kill does not mean that the model predicts a 50 percent chance of the vehicle losing all of its mobility, but that the assessed damage to components results in a 50 percent loss of mobility, according to the SDAL. The percent loss-of-functions were the products of consensus judgments by a panel of three armor officers produced more than 25 years ago. They are therefore subjective judgments with an unknown reliability and validity.
- The rationale given for the development of a standard damage assessment list was that assessors are generally unable to decide on the percent loss-of-function implied by the loss of a component and in any case

find it impossible to maintain consistency in such judgments from assessor to assessor, test to test, vehicle to vehicle, and year to year. In other words, the list had to be standardized because the assessment process was so unreliable.

The SDAL is contained in the vulnerability models used in live fire testing, so the armor models typically produce output in the form of the three types of $P_{k\ H}$ for each shot. Because it has become a NATO standard the SDAL continues to be used. Damage assessments that use other criteria for assessing kills cannot be directly compared with those performed using the SDAL, and at least one notable live fire test, the LAVP described earlier, has been ignored by many in the armor vulnerability community in part because it used different criteria.

Vulnerability analysts have begun to substitute the term "expected loss of function" for the misleading term $P_{k\ H}$ in, for example, the detailed test plan for the Phase II Bradley live fire tests. The users of output from vulnerability analysis have sometimes been unaware of the nature of $P_{k\ H}$s; their strong subjective, judgmental component; and their unknown validity. Some analysts have proposed that live fire test results be compared to model prediction at the level of physical damage, foregoing the use of the SDAL to produce $P_{k\ H}$'s.

K-kills are catastrophic events such as explosions and sustained fires that are judged to be likely to result in the complete loss of the vehicle and its crew. $P_{k\ H}$s for K-Kills do therefore have an interpretation as a probability, and the predictions of K-kill by vulnerability models do represent estimates of the probability that such an event will occur as the result of a particular shot.

Component $P_{k\ H}$s are required by the point-burst model and a number of the aircraft models. Although these are sometimes based on data, the required tests have often not been conducted, especially for new components, and engineering judgment is substituted.

Validation of Vulnerability Models

In spite of the claims that V L models have been shown to be poor predictors of test and combat data or that they have shown good or acceptable agreement with data, we found few instances of serious attempts to compare V L model predictions with tests or combat data sets. We have reviewed the main studies cited by critics as examples of the models' inability to predict live fire or combat results, and the studies claimed by modelers to show better prediction or the reasons for misprediction.

1) MEXPO study. There was at least one effort to validate the Compartment-Kill model, as part of a program called MEXPO (Materiel Exploitation Program). [material deleted] The first used computations that had been done previously, in 1970, selecting shots that were closest to those in the MEXPO data. The results were then compared to new calculations done in 1973 using the precise shotlines in the MEXPO data. Finally, the damage correlation curve used for U. S. tanks in the Compartment-Kill model was replaced by one felt by the analysts to be more representative of [material deleted]. A summary of the results and the 1973 predictions appears in Table 4.1.

Table 4.1: Comparison of Average P_K From MEXPO Data and Predictions From Compartment-Kill Model

		Predicted	
Kill criterion	**Assessed**	**1973**[a]	**1973**[b]
K (Catastrophic)	22	33	65
M (Mobility)	92	.67	88
F (Firepower)	77	74	74

[a] Based on damage correlation curve for U S tanks

[b] Based on damage correlation curve for [material deleted]

Source Hafer, T Lethality Model Evaluation (Briefing) Alexandria, VA Systems Planning Corporation 1985

The 1973 simulation results did show fairly good agreement for firepower kills, but the observed mobility and K-kills were in poor agreement with the model predictions. The use of the newer damage assessment curve designed for [material deleted] only made the predictions of K-kills worse. These were not, however, assessed statistically or using any normative criteria for validating models.

MEXPO validation data were also analyzed by IDA in support of JLF. Focusing on just those K-kill predictions that were unambiguous ($P_{K,H}$ = 0 or 1) the IDA analysis examined shot-by-shot comparisons of the predictions with the MEXPO shots. The data as presented by IDA appear in Table 4.2. The report concludes that the model and combat assessments "were in general agreement when averaged over all shots." The model predicted that 26 percent of the shots would result in K-kills, and 30 percent of the combat shots were in fact K-kills. IDA called the overall percentages "global estimates," but it can be seen that on a shot-by-shot basis, the model correctly predicted each outcome (i.e., K-kill or no K-kill) only 59 percent (16/27) of the time. The report concludes that the computer model "does not take advantage of the unique features of each shot" in its prediction even though its "global estimates" are similar to the combat results. IDA noted that the model correctly predicted only 59 percent of the shots, but focused on comparing this figure to the 58 percent one would correctly predict by guessing 30 percent of the time that a K-kill would occur, having seen the combat data. We believe that the meaning of these data can be made more evident by noting that the 59 percent correct prediction rate with 27 cases is not significantly different from chance (50 percent).

Table 4.2: IDA Presentation of MEXPO Data Comparing Combat Results and Model Predictions

		Model predictions	
Outcome	**Combat results**	**Number predicted**	**Number correctly predicted**
K-Kill	8	7	2
No K-Kill	19	20	14
Total	**27**	**27**	**16**[a]

[a] 16 Correct predictions out of 27 cases = 59% 7/27 = 26% K Kills predicted, 8/27 = 30% K Kills observed On the average' the model predicts fairly well

Source Smith G et al The Joint Live Fire (JLF) Test Background and Exploratory Testing (Draft) ' Alexandria Va Institute for Defense Analyses March 1986

The IDA presentation of these data obscures their true implications. We have rearranged the data in standard (2 X 2) format in Table 4.3. It is clear that the "global assessment's" agreement with the combat data merely reflects the similarity of the marginal distributions of the table. Both the observed and the predicted data contain roughly the same percentage of K-kills. But the important numbers in assessing the prediction accuracy are the ones that fall on the diagonal of correct predictions (predicted K-kill and observed K-kill, and predicted No K-kill and observed No K-kill). The fact that the marginal frequencies are similar is irrelevant to the accuracy of prediction. This point is illustrated by the hypothetical data in Table 4.4, in which the marginal frequencies are identical (i.e., both the predicted and observed K-kills were 50 percent) but every single shot is predicted incorrectly.

Table 4.3: IDA Table Rearranged in Conventional 2 X 2 Format

	Model predictions		
Combat Results	K-Kill	No K-Kill	Total
K-Kill	2	6	8 (30%)[a]
No K-Kill	5	14	19 (70%)
Total	**7 (26%)[a]**	**20 (74%)**	**27 (100%)**

[a] "On the average" refers to the similarity of marginal percentages (30% vs. 26% K Kills). the number of correct predictions = 2 + 14 = 16. percent correct = 16/27 = 59%. 59% correct is not significantly different from chance (50%)

Source: Adapted and rearranged from G. Smith et al., "The Joint Live Fire (JLF) Test: Background and Exploratory Testing (DRAFT)", Alexandria, Va.: Institute for Defense Analyses, March 1986

Table 4.4: Hypothetical Counterpart to Table 4.3

	Model predictions		
Combat Results	K-Kill	No K-Kill	Total
K-Kill	0	5	5 (50%)[a]
No K-Kill	5	0	5 (50%)
Total	**5 (50%)[a]**	**5 (50%)**	**10 (100%)**

[a] "On the average" the model predictions look good. 50% K-Kills are predicted and 50% are observed but every single prediction is wrong, so the use of the phrase "on the average" is misleading

Claims that "on the average" the models predict well can be misleading and must therefore be examined carefully. Such claims have been made, for example, in reporting of the Bradley Phase I tests to Congress. After pointing out correctly that the outcome of an individual test firing is influenced by a number of variables such as round-to-round variation in warhead penetration and yaw and random variation in spalling, the Army report then notes that on the average the predictions agree fairly well with the test results. But what is being averaged here is a set of shots from various impact points on the vehicle. This is not the same thing as averaging over repetitions of a single shotline. There are in general two sources of variation in live fire shots. One is the random variation which would occur if one shot were repeated many times. A single model prediction is the expectation, or average, result of these shots. The other source of variation is not random, but results from the real difference in vulnerability in different locations on the vehicle. It is these variations that are important in locating a vehicle's vulnerabilities.

2) SPC point-burst study. We were able to obtain data from one systematic effort by System Planning Corporation (SPC) to validate one version of the point burst model called VAST. VAST was used to predict the

results of the Bradley Phase I tests. In the validation test, eighteen scatterable mines were fired against M-48 tanks and the resulting assessments of damage were compared to predictions from the internal point burst model. These results are presented in Table 4.5.

Table 4.5: [material deleted]

In general these predictions look fairly well-matched to the observed results, yet there are some striking discrepancies between predictions and observations for individual shots (4 and 14, for example). And unlike other bodies of live fire test data there are some shots that are very nearly replications, in that there is variation in the results of these shots, but because the location is nearly the same, the model makes similar predictions in each case (e.g., shots 6 and 15). So some of these predictions will look rather good and others will diverge from the observed results. One of the authors of the model pointed out that the average of the predictions for shots 6 and 15 is close to the average observed result.

We believe that there are too few shots in these data to assess the variability between nearly identical shots, but these kinds of near repeats are important. If live fire tests generally do not include repeated firings at the same location on a target and if vulnerability models include stochastic components, the size of the variability in shot outcome incorporated in the model will not be based on test data. The aggregation of data from "nearly identical shots" is one way to approximate the stochastic variation to be expected from repetition of the same shot. There is some effort to base the stochastic component of the current point burst model on test data but the size of the variances in the model is largely dependent on engineering judgment.

There are other limitations of the SPC study:

- The vehicles were not fully combat loaded, so the results, especially for K-kills are dependent on the inferences made in damage assessment about combustibles that were not present.
- The results are limited to one munition (a mine); so their generalizability to other types of anti-armor weapons, impacting on surfaces other than the bottom of the tank, is in question

Although the VAST validation study is the most complete validation study of the current armor vulnerability methodology that we have seen, it is not a systematic validation of the model's predictive accuracy under realistic conditions and over a variety of weapons. It also does not appear that the results were the occasion for considering any revision in the model.

3) LAVP studies. In the LAVP tests of the A-10/GAU-8 gun conducted between 1978 and 1980, pilots fired from the A-10 close-support aircraft against simulated Soviet tank companies in order to evaluate the effects of the GAU-8's 30mm antitank ammunition. [material deleted] most of the targets were combat-loaded U. S. M-47 tanks. Though the LAVP tests were not intended primarily as model validation studies, the data have nevertheless been compared to vulnerability model predictions in more than one analysis conducted since the original tests were done.

Initial analyses at the Air Force Armament Laboratory (AFATL), showed notable discrepancies between predictions from a version of the Compartment-Kill model and test data, especially for K-kills per pass by the A-10. Table 4.6 contains one summary of LAVP data and model predictions for U. S. M-47 tanks and a small number of [material deleted] [1] These average K-Kills per pass are an exception to the small number of cases generally found in vulnerability model validation studies. This initial comparison has been cited by critics of vulnerability models as a notable example of model misprediction.

Table 4.6: [material deleted]

In a later study using the LAVP data, analysts at AFATL made a systematic effort to account for the original discrepancies between model predictions and test data. Some of the factors they considered involve differences between the assessment procedures used in LAVP and those assumed by the model(s) and traditionally used by armor vulnerability analysts. Others had to do with the data input to the models on the characteristics of the target tanks, such as the thickness and hardness of the armor. Estimates were generated using both a Compartment Kill model (called CONIC) and a version of the point-burst model (called PDAM). The ability of the point-burst model to account for component damage in detail was expected to result in better predictions than the compartment-kill model. AFATL concluded that:

- Nearly every hit on a major critical component was apparently assessed as a kill of that component. These component damage assessments differ

[1] And note that, while good predictions "on the average" may mask poor shot-by-shot prediction, there is no way for poor predictions "on the average" to mask better shot-by-shot predictions.

from conventional assessments and may represent somewhat optimistic estimates of the GAU-8's lethality.[2]

- All the plywood crew mannequins were assessed by LAVP as killed whenever a propellant fire started, and a 100% casualty was assessed whenever any spall fragment impacted a mannequin. Standard damage assessment procedures do not score a complete casualty for every impact by a spall fragment.
- LAVP assessors scored a 100 percent firepower kill whenever the turret ring was penetrated. Standard assessment procedures do not score a 100 percent firepower kill unless the turret cannot be rotated (loss of turret rotation could not be determined from the LAVP tests because hydraulic power was not operational).
- Each named component such as "the turret" or "hull side" was assigned a single thickness in the model, while actual armor thickness can range from two to three inches on the hull side, for example. Additionally, small "soft spots" on the tanks such, as the lips of hatches and suspension attachment points were not modeled separately.
- The CONIC model treated all the armor as having the same hardness [material deleted] but actual measurements showed that the armor of M-47s like those used in LAVP was substantially softer. This difference accounted for greater observed penetration than had been originally predicted.

Adjustments were made to model inputs and assumptions to reflect some of these discrepancies. For example, adjustments were made to the modeled armor thickness, as a way of accounting for the measured differences in hardness. Hardness adjustments were at first based on previously published armor plate test data, but the adjustment factors were then modified to maximize the fit to the LAVP results on penetrations. The PDAM point-burst results were also adjusted so that every predicted hit on a crew mannequin was assessed as a casualty, to match the apparent LAVP assessment procedure.

Table 4.7 presents one set of comparisons between the revised model predictions and LAVP test data. Note that there is now very close agreement between model estimates and test data for K-kills. The M-kill estimates did not appear to improve, and the F-kill estimates may be slightly worse. (Further analysis indicated that much of the discrepancy

[2] The ignition of fires in the propellant of main gun ammunition was an apparent exception Model predictions were closest to LAVP assessments if it was assumed that only one penetration per pass caused the ignition of propellant, even if several penetrations had occurred.

could possibly be attributed to the model's assumption of the independence of shots in a burst). These data are not otherwise directly comparable with those in Table 4.6, because data from the last six missions, consisting of 83 passes, had not yet been entered into the data base when the earlier analysis was performed, and passes at dive angles greater than ten degrees and rear attacks are excluded from the reanalysis. Moreover, our earlier caution about aggregated or "global" estimates is relevant here.

Table 4.7: Comparison of A-10 GAU-8 (LAVP) Test Data and Revised Model Predictions[a]

Kill criterion	Test (LAVP)	Model predictions
K (Catastrophic)	20	21
M (Mobility)	44	57
F (Firepower)	56	50

[a] Kill probabilities for a single pass by the A-10 aircraft (based on 183 passes in which the side of a U.S. M-47 tank was attacked)

Source: Derived from Flint, James E., GAU-8 30mm API damage to U.S. M-47 tanks: Tests versus analytical estimates, Eglin AFB, Florida: Air Force Armament Laboratory, March 1984.

In some ways, this use of live-fire data to investigate and modify a vulnerability model was exemplary. It systematically and carefully compared various sets of assumptions about the interpretation of test results and adjustments to the models, seeking those that accounted for features of the data or improved prediction. Penetration and component damage predictions were examined as well as overall kill predictions. A sensitivity analysis of the debris model was also conducted. It varied average number and mass of the spall particles, and showed that this submodel did not contribute to poor prediction, because variations of as much as 150% of the original values had little effect on predictions of damage. The test data were disaggregated by attack angle (left and right side, or rear), where appropriate, and by the estimated impact velocity of the GAU-8 projectiles (one of the main variables affecting the probability of penetration and damage, apart from impact location). This aided efforts to locate and identify the causes of poor prediction.

This is also the only model validation study we have seen in which there was some effort at statistical assessment of the fit between model and test data, made possible by the large number of test shots. For each of the impact velocities and for each attack direction, most of the predictions from the adjusted point-burst model were within the 90% confidence limits of the mean LAVP $P_{K/L}$s.

There are, however, limitations to the analysis as a model for the use of live fire tests for vulnerability model improvement.

- Most of the investigations of reasons for discrepancies were only possible because of the large number of LAVP test shots. Analysts can be confident that observed average $P_{k\,H}$s or kills per pass are not just statistical outliers (oddities that might appear in small samples but be expected to disappear in larger samples).
- Ad hoc "data fitting" can make a revised model's predictions look good, but not generalize to new test data sets. Some of the hardness adjustments appear to have this character. A proper test of whether the vulnerability models have been improved in a general way rather than just being adjusted to the LAVP data set would involve using the modified model to predict new tests on comparable targets. Given the large number of shots available, this sort of cross-validation could also have been performed by holding out a portion of the LAVP data from use in modifying the model, and then testing the modified model on that portion.

4) Aircraft $P_{k\,H}$ study. We did not learn of any large scale studies in which aircraft model predictions were formally compared with data. We were told by modelers that modifications are made in light of test results on a smaller scale. In one small-scale validation study described as fairly typical, the aircraft $P_{k\,H}$ methodology for assessing damage to components was compared with seven shots into push-pull tubes used in aircraft controls. The data appear in Table 4.8. Although the measures of residual capability (R) and the assessment of $P_{k\,H}$ show reasonable agreement, the assessment is a fairly uncomplicated one.[3] The meaning of the results of these seven shots for aircraft $P_{k\,H}$ methodology in general is questionable.

Table 4.8: [material deleted]

In another test, a model was used to predict the likelihood that a projectile would penetrate the rear wall of a jet engine with enough force to damage components beyond it. For lack of test data on nonhomogeneous components such as wiring bundles and avionics, those components were modeled as equivalent densities of steel or aluminum, based on the component's density. Test results from 31 shots indicated that a model based on such simplifications did not predict actual ballistic resistance.

[3] R is a measure of the post-damage stress in the material relative to the pre-damage stress, and is dependent on the original diameter of the tube and the post-damage remaining area

Adjustments of the ballistic penetration constant were not satisfactory. The testers were forced to conclude that the model was too simple; a more complex representation of the engine would be needed to match the test data.

5) Bradley Phase I tests. The vulnerability modelers maintain that the point burst model used in the Phase I Bradley tests showed acceptable agreement with the full-up test data, while the former OSD program manager in a separate report characterized the results as showing dramatic model misprediction. He found reasonable agreement with the predictions in 40 percent of the Bradley shots and 62 percent of the shots on the [material deleted] and stated that the predictions were "grossly incorrect" in the remaining cases.

These results are presented in Table 4.9. The former OSD program manager's designation of shots on which model predictions were in reasonable agreement with test results is shown in the table. The criterion was that no prediction should differ from the assessed kill by more than 30 percentage points. While perhaps reasonable, the 30 percentage point criterion is arbitrary. The percentages resulting from it are likely to be unstable with such small numbers.

Table 4.9: Point-Burst Model Predictions Compared to Full-Up Live Fire Test Results on Bradley Vehicle and [Material Deleted][a]

	$P_{K/H}s$[b]						
	M (Mobility)		F (Firepower)		K (Catastrophic)		"Reasonable
Shot[a]	Test	Model	Test	Model	Test	Model	Agreement"[c]
1	100	100	20	100	0	100	No
2	28	0	98	90	0	0	Yes
3	0	100	0	100	0	100	No
4	0	0	45	0	0	0	No
5	25	100	52	100	0	100	No
6	5	69	15	6	0	6	No
7	100	100	100	100	100	100	Yes
8	0	0	0	0	0	0	Yes
9	100	100	0	0	0	0	Yes
10	0	0	0	4	0	0	Yes
11	26	10	100	100	0	0	Yes
12	0	0	100	20	0	0	No
13	100	100	100	100	100	100	Yes
14	100	2	100	5	100	0	No
15	0	100	40	100	0	100	No
16	100	100	100	100	100	100	Yes
17	23	25	79	95	0	0	Yes
18	44	0	49	49	0	0	No

[a] Results from two test series in random order

[b] M and F are "percent loss of function" K is probability of catastrophic fire or explosion (x 100)

[c] OSD program manager's determination

Source U S Army Ballistics Research Laboratory Bradley Survivability Enhancement Program—Phase I Results Aberdeen Proving Ground Maryland December, 1985

6) Our analysis. The few model validation studies we have reviewed do not represent a firm basis for concluding either that vulnerability models are uniformly poor predictors of combat or test data or that they have shown acceptable agreement with such data.

- There are some instances of notable failure to predict the results of particular shots or series of K-kills, but this is more often true for older models than those to be used in live fire tests.
- The numbers of validation shots are almost without exception too small to provide a sound basis for statistical assessments of the goodness of fit of models to data.

- What constitutes good prediction depends on the use to be made of the predictions and the level of aggregation of the data appropriate for that use. Model validation therefore has to be viewed in that context. The IDA report suggested that knowing that predictions of vulnerability were accurate to within 30 percentage points would be adequate for typical users of vulnerability estimates: those interested in vulnerability aggregated over a vehicle. But these users are primarily war gamers, who use the data as input to their simulations. But for users who are concerned with vulnerability reduction, greater precision may be needed. For example, decisions to introduce specific design changes would require greater accuracy about the expected effects of a shot in a particular location.

Use of Live Fire Data in Calibrating or Revising Models

The sample sizes of live fire tests are likely to remain restricted. If live fire tests as currently planned will not generate the kind or quantity of experimental data needed for formal model validation, there is a question about how live fire tests will actually be used by the vulnerability modelers. Apart from formal attempts to validate vulnerability models by comparing their predictions with test or combat data, we sought out instances of the use of data as a basis for revising vulnerability models, as some indication of how live fire data may actually be used by the modeling community.

1) Bradley Phase I tests. We were told that after the Bradley Phase I tests the modelers realized that the critics were comparing the results of single test shots to model predictions, but the predictions are really of the average outcome to be expected if a large number of repeat shots were taken. Accordingly, the modelers do not believe such comparisons are appropriate. They have begun to emphasize the complex, highly variable results that can be expected from live fire shots under realistic conditions

Even if attempts were made to repeat all the conditions of a shot as exactly as possible, there would be round-to-round variations in manufacture, impact velocity, yaw, etc. for munitions, and variations in target configuration, armor thickness, the pattern of spall, and whether components are broken or fires started. Such variation is accounted for in a stochastic, or probabilistic, model. The modelers have incorporated a number of stochastic or randomly varying elements into the point burst model in use at BRL. Predictions from this stochastic model (now called SQUASH), rather than being a single number are distributions of predicted outcomes. There is also some effort by aircraft analysts to introduce stochastic components into their models.

We note several features about the move to stochastic vulnerability modeling:

- It does in fact appear to reflect the highly variable nature of full-up live fire tests.
- It was prompted in part by live fire test results, specifically the results of Phase I Bradley tests and the claim that the models did not predict them well.
- One consequence of making the model stochastic is that it is explicitly protected from invalidation by the results of a single live fire shot. Most damage states are likely to be consistent with the distribution of expected outcomes.

We were told that with small numbers of live fire shots it is difficult to claim that any one result is incompatible with the model and the stochastic revision of the point burst model just makes this fact explicit. But, according to the author of the model, it is possible to detect biases in submodels over a series of shots. For example, if the number of spall fragments striking components is consistently smaller or larger than the number predicted by the model, it might be necessary to revise the spall model.

There were other model revisions as a result of the Bradley Phase I testing:

- Errors in the geometric description of the Bradley vehicle were discovered during the Bradley Phase I tests: A critical wire that was cut in one test shot, immobilizing the turret, had been left out of the geometry, and the length of the TOW missile had at first been entered incorrectly. Failure to predict the results of a test shot led to revisions in the target description that is input to the model.
- Unspecified improvements in the way the model handles electrical systems resulted from the Bradley Phase I tests, according to the Army report to Congress.

2) Other instances. Other instances in which models or input data have been revised to accommodate test results include:

- The MEXPO results led to abandoning the proposed [material deleted] curve for the Compartment-Kill Model.

- Discrepancies between LAVP results and predictions from a version of the point burst model led to revisions in assumptions about the hardness of the target's armor and recalculation of the predictions.
- Aircraft modelers said that they learned during a series of live fire tests that one wing model was "always on the stiff side" because it left out rotational forces. They anticipate getting data from JLF which will enable them to revise it.
- The JLF/Aircraft AV-8B flight control tests are reportedly being used to revise a model.
- [material deleted]

3) Model revision as an informal process. The connections between live fire data and model revision in all of these examples are more tenuous than is suggested by terms like "model validation" and "model calibration" and the process is more informal and more dependent on the modeler's judgment. Our interviews with vulnerability analysts along with statements in the Bradley Phase II plan and the draft revised JLF/Armor plan suggest that future model revisions will generally not be based on the results of any single live fire shot. Determination of the need for model revision will continue to be based on trends observed in a series of shots. The decision that a model is in need of revision will remain in the hands of the modelers themselves.

The potential problems with this approach are that:

- The process by which live fire test results will be used to update or revise models is underspecified.
- Stochastic models provide an unknown level of protection from invalidation by test data. It is not clear exactly what degree of discrepancy between model predictions and test results is required to show that the model is incorrect.
- A large part of the modeling and model revision process is closed to outside analysts, including weapons designers. This has led to claims that modelers ignore or misspecify important V L mechanisms, or that they are accountable only to their own community.

Conclusions

In this chapter we addressed the evaluation question, "What are the advantages and limitations of full-up live fire testing, and how do other methods complement full-up testing?" Our conclusions follow.

Advantages of Live Fire Testing

- As the only method providing direct visual observation of the damage process caused by a weapon/target interaction under realistic combat conditions, full-up live fire testing offers a unique advantage over all other methods of V/L assessment.
- The descriptions of directly observable damage that full-up testing provides are regarded as highly beneficial by users.
- Full-up testing has already demonstrated some value by producing several "surprises", i.e., results that were not predicted, and might not have been detected by other methods of testing or analysis.

Limitations of Live Fire Testing

High Cost

- The primary limitation of full-up, full-scale live fire testing is cost. On a per shot basis, it is considerably more expensive than inert or subscale testing, primarily due to the high cost and limited availability of targets. Testing and restoration costs are also higher, as are their associated time requirements. Nonetheless, live fire testing costs are a very small percentage of total program costs.

Limited Information

- Full-up testing potentially yields less information about damage mechanisms per shot than inert or subscale testing, primarily because catastrophic kills destroy the target and its components, along with much of the instrumentation used to record the damage. However, not all full-up shots result in catastrophic kills; such shots potentially yield more interpretable information than equivalent inert shots.

Limited Generalizability

- Full-up live fire test results typically are less easily generalized beyond the specific test conditions than inert or subscale testing. Full-up testing brings a larger number of variables into play that potentially affect outcomes, yet because full-up testing destroys targets, a smaller proportion of relevant test conditions can be examined.

Limited Redesign Opportunities

- The impact of live fire testing of developed systems is limited by "frozen" designs which are prohibitively expensive to change. For this reason, test officials see the main benefit of JLF and related programs as reducing vulnerability of future systems through lessons learned. This is not to suggest, however, that important V L modifications are never feasible.

Other Methods That Complement Full-Up Live Fire Tests

Subscale Testing

- Subscale tests can support larger sample sizes than full-scale tests (whether full-up or inert), and are useful in bounding effects and providing input to models. Certain types of subscale testing are also useful for developing generic characterization of munitions effects.
- Subscale tests can provide only indirect evidence of synergistic effects on realistic targets, which must be inferred through an unproven analytical process (modeling). Therefore, subscale testing can supplement full-up, full- scale testing but not substitute for it.

Inert Testing

- Inert testing of full-scale targets is superior to full-up testing in characterizing mechanical damage to individual components and in conserving both components and targets.
- Catastrophic damage cannot be observed directly from shots on inert targets, and the standard method for inferring a K-kill underestimates its true likelihood. Like subscale tests, inert tests can provide only indirect evidence of effects on realistic (i.e., full-up) targets, inferred through models acknowledged to be weak on combustibles. Therefore, inert testing can supplement full-up, full-scale testing but not substitute for it.

Combat Data

- Analysis of combat data, if available, has several advantages over V L testing: it provides greater realism, includes information above the level of vulnerability and lethality (e.g., aggregated survivability measures), and is considerably less expensive.
- Combat data provide less scientific control than testing, are limited to munitions and systems that have been employed in combat, and offer no direct view of the damage process or the conditions of firing. Like subscale and inert testing, combat data can supplement full-up, full-scale testing but not substitute for it.

Modeling

- V L models support the design and interpretation of live fire tests, and are potentially useful in extrapolating beyond test results. A unique advantage of models over testing is their applicability to systems not yet built.
- Models are widely used in V L assessment generally, but play a more central role in the design and interpretation of armor tests than in aircraft tests.
- It does not appear that models have as yet played as great a role in the design of live fire tests as some statements by the modelers would indicate.
- Input data on warhead, armor interaction and behind-armor debris required by the logic of the models is often lacking. The planned use of JLF resources by JLF Armor to obtain this kind of data primarily serves the input needs of models.
- Current vulnerability models share numerous limitations; specifically, fire, explosion, multiple hits, ricochets, synergistic effects, and human effects are not yet well modeled.
- Many of the most important mechanisms for producing casualties are poorly modeled, if at all. Without specific efforts to bring these casualty mechanisms into the modeling process, V L models can be expected to be of limited utility in predicting casualties or providing insights into the casualty reduction.
- Currently used V L models are inadequately validated.
- A large part of the modeling and model revision process is closed to outside analysts, including weapons designers. This has led to claims that modelers ignore or misspecify important V L mechanisms, or that they are accountable only to their own community.
- Claims that "on the average" models predict well can be misleading, and in general such claims must be examined carefully.
- Because there are no clear criteria for success and failure in model prediction, proponents and opponents of modeling can both claim support from the same data. This happened in the reporting of Bradley Phase I.
- Claims that vulnerability models predict poorly are somewhat overstated, often referring to predictions from older models not expected to be used in live fire tests, and insufficient test or combat data to permit unqualified conclusions. Additionally, little attention has been paid to the different levels of accuracy required for different users' purposes.
- The stochastic components introduced into vulnerability models after the Bradley Phase I tests provide an unknown level of protection from invalidation by test data.
- There are no clearly specified mechanisms for using live fire test data to calibrate or revise models. Models frequently are revised on the basis of

test data, but the process is more informal and judgmental than the terms "validation" and "calibration" would suggest.

- It is doubtful that JLF or any future live fire testing will produce the kind or quantity of live fire data that would be required to validate sophisticated V L models. However, the body of live fire data that accumulates should provide a basis for checking on whether model revisions do in fact improve predictions.

Chapter 5

How Can Live Fire Testing Be Improved?

Some important concerns arising from the uncertainties of JLF will be resolved for future systems by the new live fire testing legislation. Specifically, the act establishes:

- service responsibility for supplying targets,
- linkage to the procurement process, and
- a requirement for full-up and full-scale testing.

Future live fire tests should improve as a result. However, other areas for improvement remain.

We believe that the following improvements should be considered. We divide these into technical improvements—improvements in the design, conduct, and interpretation of live fire tests—and general improvements—improvements to facilitate realistic live fire testing and the usefulness of its results. These are suggestions, not recommendations. Our recommendations appear in the next chapter.

Technical Improvements

We suggest that DOD

1. [material deleted]

2. improve the estimation of human effects. Begin by replacing non-instrumented plywood mannequins with the instrumented anthropomorphic type.

3. improve the reliability and validity of quantitative V L estimates. For example, interrater agreement studies could determine the magnitude of the reliability problem, and provide insights into reducing it.

4. expand efforts to improve statistical validity, and establish guidelines for the statistical interpretation of small-sample live fire test results.

5. concentrate model improvements on currently weak areas vital to casualty estimation—fire and explosion and human effects.

6. establish guidelines for how models can better support the design and interpretation of live fire tests.

7. establish guidelines for how live fire test results can be used in the revision of models.

8. allow outside analysts into the modeling and modeling revision process, and provide better documentation of the process for use by those analysts.

9. accumulate comparisons of model predictions with live fire test results over multiple tests in order to assess improvements in models, and make results available to outside analysts; also, redo predictions of earlier live fire shots after models have been revised in order to validate improvements.

10. require that detailed test plans include shotlines, munitions, sample sizes, predictions, analysis plans, rationales for decisions, and other critical information to enable proper oversight. Keeping plans unclassified should not be a justification for omitting key information.

11. develop, modify, or procure instrumentation to yield more information from catastrophic shots. For example:

- ways of hardening instrumentation to survive in a full-up live fire environment should be developed (e.g., employ localized fire suppression systems that do not affect conditions of adjacent components).
- ways to use cheaper instrumentation for shots likely to go catastrophic should be explored (e.g., low quality pressure transducers, lower grade ammunition in ammunition stores, expendable remote video cameras).
- in general, the state of the art of live fire testing instrumentation should be improved (e. g., there is currently no unobtrusive method to measure fuel/air ratio, or dynamically measure fuel ingestion rate).

12. improve methods for simulating in-flight conditions; specifically altitude, altitude history, maneuver load, and slosh.

General Improvements

We suggest that DOD

1. avoid requiring unrealistic or incompatible objectives in future live fire tests (e g., combat realism <u>and</u> model validation).

2 consider total program costs in considerations of target costs, including the for example concept of a percentage set-aside for live fire testing.

3. determine whether the live fire testing infrastructure is adequate to implement the legislation, or has to be expanded. For example, only two facilities in the U.S. currently have high speed airflow capability.

4. determine to the extent possible the cost of live fire testing of new systems, and the relative costs and benefits of different approaches to live fire testing. Currently, there are claims and counter-claims about the costs of full-up vs. subscale tests, but little data.

5. promote awareness of the benefits to be obtained from destructive testing to top level military and civilian officials.

6 with the legislation as a foundation, continue to strengthen incentives that support realistic live fire testing.

Recommendations and Agency Comments

Recommendations to the Secretary of Defense

In addition to the improvements noted in Chapter 5, there is a need to resolve current conflicts about the purpose of live fire tests and to make clear that the objective of reducing vulnerability and increasing lethality of U.S. systems is the primary emphasis of testing. Accordingly, we recommend that the Secretary of Defense

1. conduct full-up tests of developing systems, first at the subscale level as subscale systems are developed, and later at the full-scale level mandated in the legislation. This will minimize vulnerability "surprises" at the full-scale level, at which time design changes are more difficult and costly.

2. establish guidelines on the role live fire testing will play in procurement.

3. establish guidelines on the objectives and conduct of live fire testing of new systems, with particular attention to clarifying what is to be expected from the services.

4. ensure that the primary users' priorities drive the objectives of live fire tests. Modelers are secondary users.

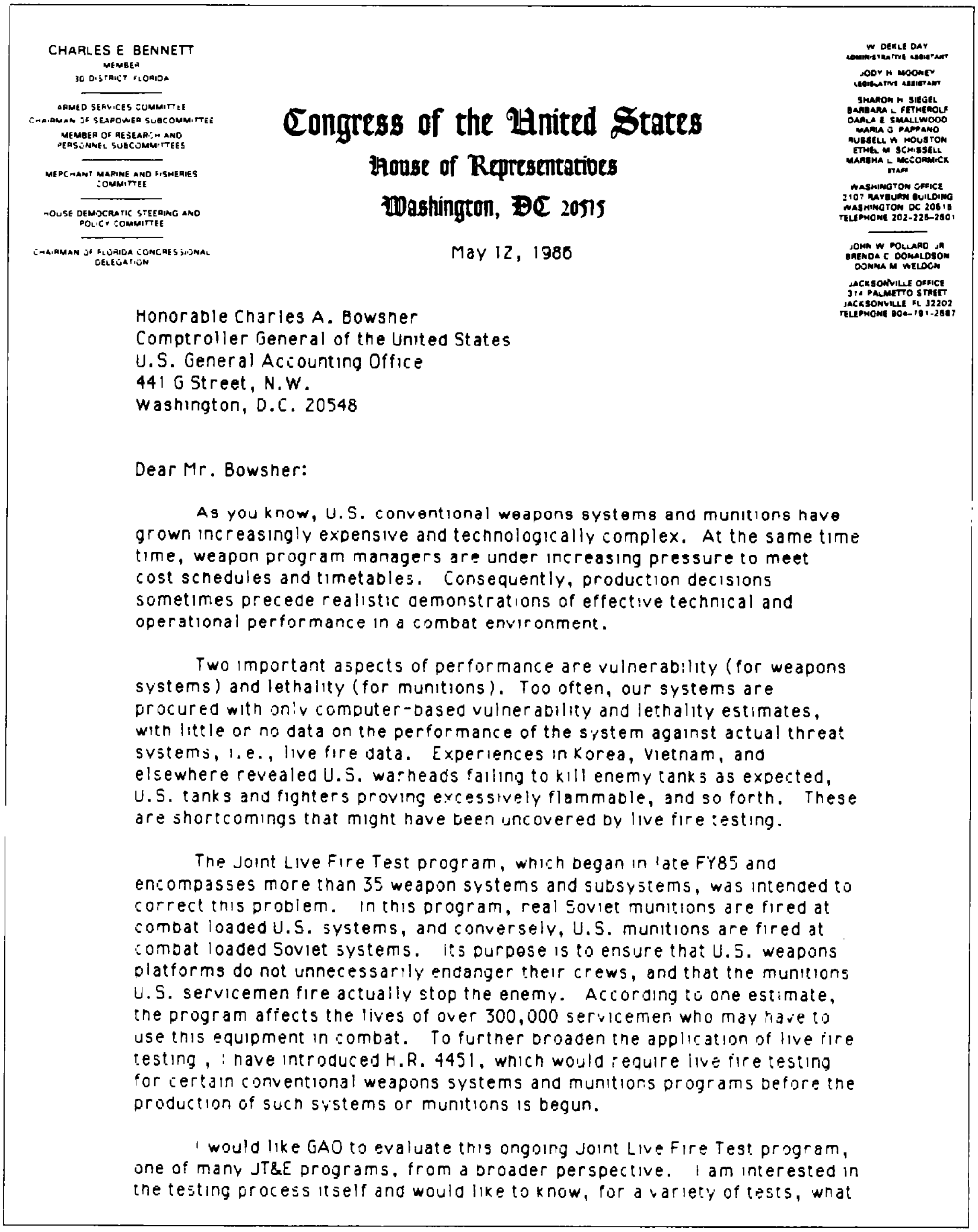
Recent live fire legislation requires the services to provide targets for testing new systems, but there is no similar requirement for the fielded systems in JLF, where lack of targets has impeded testing. Accordingly, we recommend that the Secretary of Defense

5. provide more support to JLF for obtaining targets.

Agency Comments

DOD provided oral comments on the report. DOD concurred with all recommendations and most findings, and made several suggestions to improve technical accuracy. GAO made changes based on these suggestions where appropriate.

Request Letter



CHARLES E BENNETT
MEMBER
3D DISTRICT FLORIDA

ARMED SERVICES COMMITTEE
CHAIRMAN OF SEAPOWER SUBCOMMITTEE
MEMBER OF RESEARCH AND
PERSONNEL SUBCOMMITTEES

MERCHANT MARINE AND FISHERIES
COMMITTEE

HOUSE DEMOCRATIC STEERING AND
POLICY COMMITTEE

CHAIRMAN OF FLORIDA CONGRESSIONAL
DELEGATION

Congress of the United States
House of Representatives
Washington, DC 20515

W DEKLE DAY
ADMINISTRATIVE ASSISTANT

JODY H MOONEY
LEGISLATIVE ASSISTANT

SHARON H SIEGEL
BARBARA L FETHEROLF
DARLA E SMALLWOOD
MARIA G PAPPANO
RUSSELL W HOUSTON
ETHEL M SCHISSELL
MARSHA L McCORMICK
STAFF

WASHINGTON OFFICE
2107 RAYBURN BUILDING
WASHINGTON DC 20515
TELEPHONE 202-225-2501

JOHN W POLLARD JR
BRENDA C DONALDSON
DONNA M WELDON

JACKSONVILLE OFFICE
314 PALMETTO STREET
JACKSONVILLE FL 32202
TELEPHONE 904-791-2587

May 12, 1986

Honorable Charles A. Bowsher
Comptroller General of the United States
U.S. General Accounting Office
441 G Street, N.W.
Washington, D.C. 20548

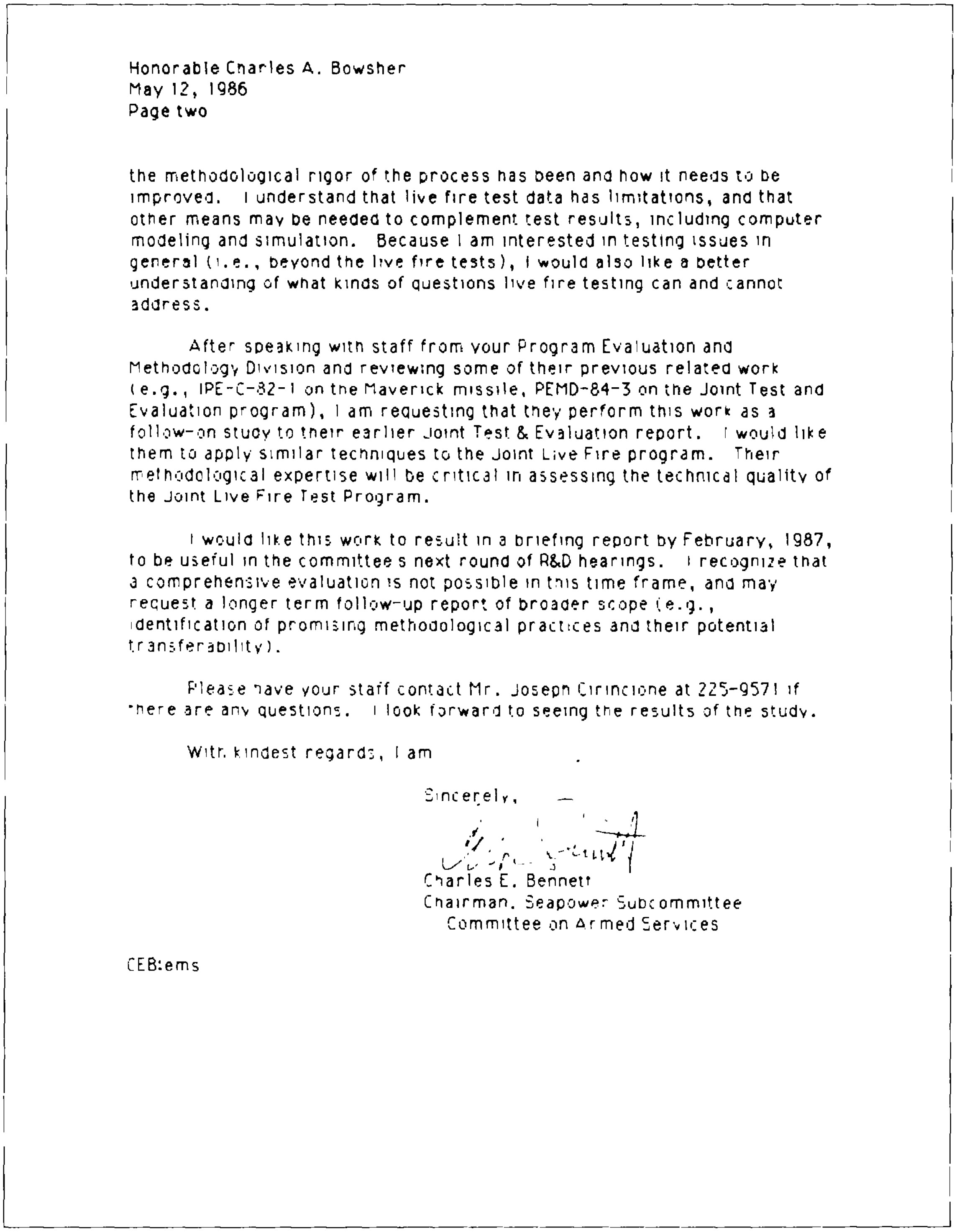
Dear Mr. Bowsher:

As you know, U.S. conventional weapons systems and munitions have grown increasingly expensive and technologically complex. At the same time time, weapon program managers are under increasing pressure to meet cost schedules and timetables. Consequently, production decisions sometimes precede realistic demonstrations of effective technical and operational performance in a combat environment.

Two important aspects of performance are vulnerability (for weapons systems) and lethality (for munitions). Too often, our systems are procured with only computer-based vulnerability and lethality estimates, with little or no data on the performance of the system against actual threat systems, i.e., live fire data. Experiences in Korea, Vietnam, and elsewhere revealed U.S. warheads failing to kill enemy tanks as expected, U.S. tanks and fighters proving excessively flammable, and so forth. These are shortcomings that might have been uncovered by live fire testing.

The Joint Live Fire Test program, which began in late FY85 and encompasses more than 35 weapon systems and subsystems, was intended to correct this problem. In this program, real Soviet munitions are fired at combat loaded U.S. systems, and conversely, U.S. munitions are fired at combat loaded Soviet systems. Its purpose is to ensure that U.S. weapons platforms do not unnecessarily endanger their crews, and that the munitions U.S. servicemen fire actually stop the enemy. According to one estimate, the program affects the lives of over 300,000 servicemen who may have to use this equipment in combat. To further broaden the application of live fire testing , I have introduced H.R. 4451, which would require live fire testing for certain conventional weapons systems and munitions programs before the production of such systems or munitions is begun.

I would like GAO to evaluate this ongoing Joint Live Fire Test program, one of many JT&E programs, from a broader perspective. I am interested in the testing process itself and would like to know, for a variety of tests, what



Honorable Charles A. Bowsher
May 12, 1986
Page two

the methodological rigor of the process has been and how it needs to be improved. I understand that live fire test data has limitations, and that other means may be needed to complement test results, including computer modeling and simulation. Because I am interested in testing issues in general (i.e., beyond the live fire tests), I would also like a better understanding of what kinds of questions live fire testing can and cannot address.

After speaking with staff from your Program Evaluation and Methodology Division and reviewing some of their previous related work (e.g., IPE-C-82-1 on the Maverick missile, PEMD-84-3 on the Joint Test and Evaluation program), I am requesting that they perform this work as a follow-on study to their earlier Joint Test & Evaluation report. I would like them to apply similar techniques to the Joint Live Fire program. Their methodological expertise will be critical in assessing the technical quality of the Joint Live Fire Test Program.

I would like this work to result in a briefing report by February, 1987, to be useful in the committee s next round of R&D hearings. I recognize that a comprehensive evaluation is not possible in this time frame, and may request a longer term follow-up report of broader scope (e.g., identification of promising methodological practices and their potential transferability).

Please have your staff contact Mr. Joseph Cirincione at 225-9571 if there are any questions. I look forward to seeing the results of the study.

With kindest regards, I am

Sincerely,

Charles E. Bennett
Chairman, Seapower Subcommittee
Committee on Armed Services

CEB:ems

Review of Individual Tests

[material deleted]

Armor

[material deleted]

[material deleted]

Setting Test Objectives

[material deleted] and 3) use the models to generalize the results to the overall vulnerability of each newer vehicle).

- Test the difference between "static" detonation of shaped-charge warheads fixed to the surface of vehicles and "dynamic" firings from realistic ranges. If the results did not differ, most subsequent shots would have been fired statically, to take advantage of the lower cost and increased precision of static detonations.

The October 1986 draft revised JLF/Armor plan dropped the proposed comparison of static and dynamic firings, as guidance from OSD had emphasized the dynamic firing of munitions. The issue of when the less expensive and more controllable static firings are equivalent to dynamic launches at combat ranges has been a persistent source of controversy in live fire testing methodology. We believe that cost considerations in live fire testing do warrant an empirical comparison of the two methods, as originally proposed. The Army has recently conducted some such tests as part of the Bradley Phase II tests, and we were told that further static/dynamic comparison tests will be included in the JLF/Armor tests.

Test Planning

[material deleted]

Again, although negotiations preceded this version of the plan OSD and the JTCG/ME test planners were still not in agreement about the role of inert and full-up tests in JLF. The DTP was rejected by the former OSD program manager in part because of the proposal to shoot at inert targets, with guidance that a revision should include a majority of full-up shots. The test outline in the October 1986 draft revised plan submitted to the current OSD program manager returns to the use of a majority of inert vehicle tests.

[material deleted]

The sequencing of shots was sensitive to target conservation, and "testing effects." (Testing effects are biases resulting from shooting repeatedly at the same target. A target which has been degraded by earlier shots reacts differently.) Firings were to progress from the smallest and least damaging to the larger munitions. There were two reasons for this: to preserve the targets' condition as long as possible and to permit data acquisition and test procedures to be "perfected" on the less expensive tests so that all the desired data would be obtained on the more expensive tests with the larger munitions.

Implementation, Analysis and Reporting

The JLF [material deleted] tests had not been implemented as of December 1986, and hence there is neither analysis nor a report to assess.

U.S. M-48 Tank

This test was not in the JLF/Armor 1985 Plan and does not appear in any JLF schedule. It was set up in order to have some testing occur within JLF in FY 85.

Setting Test Objectives

The objectives of this test were to train damage assessors and add to the data on the effectiveness of long-rod kinetic energy penetrators on tanks of the M-48/[material deleted] class. The U.S. M-48 is an older tank (the one tested was built in 1953) unlike those first-line vehicles identified as the main interest of JLF.

This test's objective of training damage assessors who could be used in subsequent JLF tests does not follow simply from the program objectives and is not mentioned in the master plan. The assessment of damage from combat or live fire shots is difficult, and apparently there are insufficient numbers of people with experience in it at armor test facilities. One experienced tester thought it would not be possible to train damage assessors during the course of a test. The test proved the objective to be unrealistic. The brief course conducted at the test site did not succeed in training the damage assessment teams to fill out forms with acceptable accuracy or consistency across assessors. The training consisted of only a few class sessions supplemented by discussions with the instructor between the shots. One tester suggested that a full year's experience might be necessary to produce an acceptable level of competence in damage assessment.

To take advantage of a previously tested, burned-out hulk of a [material deleted] tank that had become available, it was decided to include several shots against this target during the time the M-48 was being tested. These shots were not included in the M-48 plan.

An M-48c tank was used in lieu of obtaining an actual M-48. A training version of the M-48, the M-48c has a "mild steel" hull rather than actual armor, rendering it unsuitable for combat. Realistic shots were therefore restricted to its turret or external components other than the hull. The [material deleted] hulk had been tested and burned before, and was missing many internal components. Because a target that has been degraded by earlier tests reacts differently from one which has not, and because internal components affect the results, little information was produced that could be generalized to combat-ready [material deleted]. The use of these vehicles as targets represents a departure from the original JLF goal of shooting at operational first-line vehicles. It was in part dictated by the failure to obtain actual [material deleted] in good condition, to conduct the first scheduled JLF tests.

The [material deleted] hulk was loaded with combustibles before shots were fired at it, while the M-48 was kept inert, in part to explore the consequences of the different target conditions for the conduct of tests and damage assessment.

The test was also to add to the data base on the effectiveness of long-rod penetrators on tanks of the M-48 [material deleted] class. However, the munition employed against the M-48c was not an [material deleted] munition but a simulant produced by a foreign country. It is not known whether or how the effects of the simulant differ from the [material deleted] munition.

The draft plan does not give a rationale for the shotlines selected. Three shots were to be fired from the front, two at the turret and one at the hull in the location of the driver. The other was to be fired from the right side at the turret.

Implementation

The test implementation departed from the draft plan for the M-48 test in a number of ways:

- As noted above, the plan was written for the M-48 alone. We were told that the [material deleted] hulk was made available only later.

- The personnel to be trained as damage assessors were used to assess only the first three shots. BRL personnel assessed the final shot. Test officials told us that they realized it was not possible to train damage assessors in the time available (limited to a few classroom sessions), concluding that such training could take as much as a year.
- Because the M-48 available for testing, an M-48c, did not have a hull of actual armor, a shot against a track had to be substituted for the one planned against the hull at the driver's position.

Analysis and Reporting

The preliminary draft of the M-48 JLF report contains raw data in the form of photographs and damage assessment forms. The text describes the test and efforts to train damage assessors. Values for M, F, and K-kills were assigned to each shot by BRL personnel, but the results were not analyzed further. It is not possible from the draft to assess the way the data will eventually be analyzed.

The results of the shots against the [material deleted] hulk were not published in a JLF report by JTCG ME, but only in the March 1986 IDA report on the FY85 JLF activities. This report treats the [material deleted] M-48 test as a comparison of two methodological approaches to live fire testing: inert vs. full-up. The report states that the [material deleted] hulk was fired at three times, fully loaded with ammunition and fuel (though hydraulic lines were not present), and each shot resulted in a catastrophic kill in the form of an explosion or uncontrolled fire. The M-48, by contrast was inert, with water in its fuel tanks and dummy ammunition stowed aboard. IDA concluded that there are tradeoffs in the methodologies: inert testing provides detailed information on behind armor effects, while full-up testing provides unambiguous information on catastrophic kills.

We believe that there were too many differences between these two targets to regard the test as a fair comparison of the inert versus full-up test approaches. The two vehicles, although of the same general class of tank, were different models, and their state of repair and completeness were very different even before the [material deleted] was loaded with combustibles and the M-48 was left inert. The [material deleted] was little more than an empty hulk: that is, it was in poor condition and most of its internal components were missing. Sheet metal was used to simulate some of the internal components only after some question was raised about the possible masking effects of components, after the first shot. The fact that it was impossible to trace component damage in the [material deleted] after the shots does not imply that this would generally be the case in full-up tests of tanks, as the IDA report suggests.

[Material Deleted] Vehicle

The Army conducted two series of full-up live fire shots, against the [material deleted] vehicle and the M901 Improved TOW Vehicle (ITV) version of the U.S. M113 armored personnel carrier, in late FY85.

Setting Test Objectives

These tests were intended to provide baseline data for assessing the relative vulnerability of the Bradley and these vehicles. The rationale explicitly mentions the biased or "skewed" perceptions created in observers by the results of tests against the Bradley with overmatching munitions. The comparison tests were designed to "anchor" perceptions of the vulnerability of U.S. vehicles such as the Bradley by demonstrating the lethality of U.S. munitions against comparable [material deleted] vehicles.

Test Planning

These ten Bradley comparison shots on the [material deleted] are not the entire series of live fire tests originally scheduled. The January 1985 plan indicates that there were to have been between 225 and 291 shots at the [material deleted] with 22 different munitions. Rather than systematically exploring the lethality of a range of U.S. munitions against the [material deleted], as had been originally proposed, the ten comparison shots (funded and conducted by the Army rather than JLF) merely provided some context for interpretation of the Army's Bradley shots. The DTP for the [material deleted] (written as part of JLF) was a bare outline, five pages of double-spaced text plus shot diagrams.

There are two versions of the [material deleted] vehicle, the [material deleted]. The main differences are in the turret armor and gun Only [material deleted] were available for JLF testing. The [material deleted] was simulated by modifications to the turret of a [material deleted], including the main gun. This appears to have been a reasonable simulant of the [material deleted].

The plan specified that the tests were to be full-up. The fuel cells were to be 3/4 full of diesel fuel and live ammunition of the type and quantity used for the particular model of [material deleted] was to be stowed in the vehicle

The munitions selected for firing at the [material deleted] were the same type as those selected for the Bradley tests, which had focused on "overmatching weapons," more powerful than those the system was designed to withstand. They are therefore considerable overmatches for the [material deleted] as well, almost guaranteed to penetrate and make the [material deleted] look vulnerable. The smaller munitions, perhaps more representative of the munitions that infantry would be firing at [material deleted] were not included. While the selection is appropriate for the Bradley comparison, it limits the generality of the results and leaves open the possibility that perception of the [material deleted] vulnerability will be "skewed" by the lack of context. Other munitions are scheduled for testing against the [material deleted] later in JLF.

The ten shots of shaped-charge munitions were selected to match the full-up shots taken on the Bradley. The matchups were not restricted to precise duplications of impact locations, because of differences in vehicle design (e.g. a shot into the right rear door would impact stored fuel in the [material deleted] but not in the ITV.) The effect of these departures from strict matching is not known. That is, it is uncertain whether differences in the assessed vulnerability are the result of the failure to find truly comparable impact locations, or differences in the vehicles' "true" vulnerability at equivalent impact locations. This issue will arise whenever two vehicles are compared on shots that are not representative (or randomly chosen) samples of the hits to be expected on the vehicle type in combat.

Implementation

The DTP stated that [material deleted] munitions would be stowed on the [material deleted] if they were available. But the short deadline led to the use of US munitions in place of [material deleted]. The list of munitions chosen and some of the reasons for their choice are reported in the draft. Overall, the selected munitions were thought to approximate the placement, size and general explosive layout of the threat vehicles. But

> "The explosive weights are generally greater than the threat systems and in some cases, the explosive type is of higher brisance [shattering or crushing effect]. The overall effect of these differences could make the [material deleted] vehicles respond more violently to overmatching projectiles."

The draft report then argues that because the [material deleted] munitions on the [material deleted] are packed very densely "the effects on the vehicle should be quite similar to the known battlefield effects of overmatching penetrations on the [material deleted]." The surrogates were for this reason "considered adequate for these tests."

But the stated purpose of this test series was to provide a comparison for the Bradley tests, because "the very nature of those tests—overmatching munitions which are guaranteed to perforate the armor shell—tends to skew the perception of some observers as to the vulnerability of US equipment to enemy munitions." So the series was intended to show that the [material deleted] was also vulnerable to overmatching weapons, and to compare its vulnerability to the Bradley's. The results were to be used to put the Bradley's vulnerability in context. Any bias in the tests that tends to make the [material deleted] look more vulnerable has the effect of reducing the Bradley's apparent relative vulnerability. Although the test report argues that the effects of the surrogates on the vehicle should be similar to that of the [material deleted] ammunition, in fact the magnitude of any difference in reactivity of the surrogates is unknown. Because the direction of any departure from the actual vulnerability of the original [material deleted] munitions is thought to be in the Bradley's favor, this possible bias could constitute a threat to the validity of the [material deleted]-Bradley comparison.

The [material deleted] test report was in fact candid about the potential problem with the surrogate munitions, though it relied on engineering judgment to conclude that the problem was probably not serious. Test officials also argued that only one of the ten comparison shots could have been strongly affected by the use of more vulnerable surrogates, so any bias was insignificant. However, one year after the completion of the test series the JLF report of the [material deleted] tests existed only as a rough preliminary draft. It has not therefore been generally available to decision makers who must assess the vulnerability of the Bradley or other experts who might have alternative interpretations. The only form of these results made available to Congress, for example, is their treatment in the Bradley Phase I report of December 1985. In that report there is no mention of the reservations about the possibly excessive explosiveness of the surrogate munitions stowed on the [material deleted] that were expressed in the unreleased draft of the full [material deleted] report. In fact the report to Congress states that both the [material deleted] and the ITV "were loaded with the full complement of ammunition and supplies they carry into combat." Furthermore the Bradley report specifically concluded that the Bradley was less vulnerable than the [material deleted], citing the comparison tests as evidence.

Testers acknowledge that the choice of surrogates for [material deleted] munitions is complicated by different design philosophies. For example, [material deleted] shaped charges have generally used more vulnerable explosives, but less vulnerable metal cases around them. In the absence of tests demonstrating the equivalence of surrogates and threats, the validity of test results must be open to argument. Therefore we believe that departures from combat realism, such as the use of surrogate loadings of questionable equivalence to actual threat munitions, should be included in reports of live fire tests (including reports to Congress), along with arguments and data concerning the choice of surrogates and the practical effects of their use.

U.S. Bradley Vehicle Phase I and II

There was much controversy over the objectives, planning, design, and reporting of the Bradley Phase I live fire test funded and conducted by the Army.[1] These are treated in some detail in a previous GAO report and a report by the staff of the HASC. Briefly, concerns were expressed over:

- The proper priority of test objectives. The JLF OSD program manager stated that the predominant purpose of JLF was to locate sources of casualties and provide insight into modifications that would reduce casualties. In contrast, the army testers at BRL focused on obtaining information useful for calibrating or checking computerized vulnerability models.
- Vehicle selection. There are two versions of the Bradley, the M2 infantry version, which carries nine troops, and the M3 cavalry version, which carries five. The infantry version, which is more susceptible to larger numbers of casualties, was not tested.
- Shot selection. Reflecting the conflict over the test objectives' priority, the Phase I Bradley test shots conducted during 1985 were selected to resolve vulnerability uncertainties. All ten full-up shots were aimed so as to avoid ammunition, because it was felt by Army testers that there was little uncertainty about the effects of shots into stored ammunition. This shot selection represents a departure from the standard of combat realism emphasized by the OSD program manager, but not necessarily from the JLF objectives in the charter and the approved January, 1985 Master Plan. The shots were not selected randomly or to be representative of combat hits. The procedure was justified by the test planners in this case as efficient for obtaining information they judged to be most useful without excessive risk of destroying test assets.

[1] This is not to be confused with the so-called Bradley vaporifics test of 1984

The previous investigators did not conclude that the selection of shots had been intentionally biased to make the Bradley appear less vulnerable. However, the general issue of bias in shot selection is that any judgmental or intuitive basis for selecting shots is open to the influence of inadvertent bias or at least its appearance. Some form of random selection is the only way to avoid charges of bias. The Board on Army Science and Technology (BAST) of the National Research Council was requested by OSD and the Army to develop a method for selecting unbiased and combat-realistic shots in live fire testing. Their recommendations were followed in producing the final form of the Bradley Phase II test plan.

Test Planning

The Army's Phase II Bradley plan is the most detailed and thoroughly specified of the live fire test plans we have reviewed. Its six volumes include detailed descriptions of the procedures to be followed in all the subtests, predictions generated by vulnerability models for all proposed impacts, detailed diagrams of the vehicle configuration and stowage plans and a detailed evaluation or analysis plan. It removes all test implementation decisions from the informal judgments of testers, specifying elaborate contingency plans for departures from planned procedures during the tests.

We note five of the plan's features:

1) Emphasis on casualties. This is the only DTP we saw that explicitly emphasized casualty estimation in the objectives. The objectives for the full-up firings include:

- to generate baseline data on the number of casualties expected for comparable firings at the M2 and M3 Bradleys (infantry and cavalry versions).
- to generate baseline casualty and vehicle vulnerability data for the M2.

2) Emphasis on fire and explosion. All but one test, including component tests, involve fire and/or explosion in some way, despite the risk to test assets. This is in contrast to the earlier Bradley testing which was criticized for avoiding shots resulting in fire and explosion.

3) Shot selection. The plan adopts a shot selection methodology based on random selection of impact locations from combat distributions, developed by the BAST group specifically for use in the Bradley testing. The plan claims that BAST selected the shotlines for the Army, but strictly

speaking, this misstates BAST's role in the process. The BAST report and its cover letter make it clear that BAST was only suggesting an interim method for selecting shotlines. BAST specifically stated that the shotlines appended to their report were the results of a "trial use" of the method, and did not constitute recommendations as to which shotlines should be used for the Bradley. They stated that the responsibility for choosing shotlines was the Army's. The plan states that the BAST selections met the Army's goals and so the complete set of BAST shots was adopted by the Army without modification.

4) Sample size The plan states that BAST recommended the minimum number of shots required per munition type to establish with reasonable confidence that observed vulnerability differences between specified test targets are true differences. But BAST explicitly stated in their report that the number of live fire shots required for reliable vulnerability assessment is an open question, and their selection of 20 shots was dictated by the OSD request, not by statistical considerations. OSD originally requested that BAST provide a method for selecting 13 shots for comparative tests. BAST felt that 20 shots distributed among four munition types would be "more representative" but cautioned that the use of 20 shots should not be interpreted as an indication that this sample size is adequate for reliable vulnerability assessments or statistically meaningful conclusions.

5) Statistical analysis. The plan proposes a questionable statistical analysis. A procedure known as the sign test will be used to compare twelve matched shots on the standard version and the high survivability version of the Bradley M3. However, a sign test with only twelve pairs of shots will fail to detect differences between the two vehicles unless the differences are very large. The procedure has additional problems, detailed in the general discussion of statistical validity in Chapter 3.

Aircraft

As noted earlier, only one JLF Aircraft test has been completed and written up. Consequently, only one report was available for our review, and it was in draft form. This was the FY85 test on steady state fuel ingestion in the F100 engine, which powers both the F-15 and F-16 aircraft.

F100 Engine Steady State Fuel Ingestion Test

Fuel ingestion is a potential kill mechanism experienced by jet aircraft when a projectile (small arms, warhead fragment) penetrates a fuel cell in such a manner that fuel is injected into the engine inlet

Setting Test Objectives

In the original DTP, this test had two objectives:

- Determine the fuel ingestion tolerance of the F100 turbofan engine when exposed to controlled steady state fuel leakage.
- Compare the results with previous vulnerability analyses and identify the degree of enhancement which might be required in applicable models of the F100.

The draft report retained these two objectives and added a third:

- Identify possible pilot responses, different from those published in flight manuals, which might improve the probability of survival during a fuel ingestion incident.

All were consistent with the program objectives as specified by JTCG AS.

Test Planning

The F100 DTP was congruent with the test objectives. A test matrix was specified, but sample size (number of runs) was not, on grounds that the specific number of runs cannot be known until the tests are conducted (the test ends when the engine is destroyed).

Three positions for fuel injection were selected along the inlet duct—20, 80, and 120 inches from the engine face. However, no rationale was provided for these positions, and no mention was made of hole size. In the draft test report, the test matrix showed the hole size and its position on the inlet duct for each test run, and also stated that clean round holes were selected over other hole geometries, in part to meet the primary objective of "controlled" fuel ingestion. However, there was no mention of the size or type of ballistic threat the holes are supposed to be simulating, even though the report is classified; nor was there any rationale for the choice of hole sizes and positions.

The DTP contained no pretest predictions, but stated that predictions would be produced later.

Implementation

The report revealed that the basic test conditions had changed since the DTP. Inlet ram pressure originally was to be supplied to simulate flight at Mach .8, 3,000 ft. above sea level. In implementation, conditions were set at Mach .7, 2,230 ft. above sea level. No explanation for the change was provided.

Test officials appeared to be making all reasonable efforts to maintain the realism of test conditions. For example:

- The engine was "trimmed" to establish nominal relationships among engine temperatures, pressures, and rotor speeds. This was done to ensure that the engine's reaction to fuel ingestion would be representative of engines currently in use.
- Some equipment problems were observed in the early runs. (i.e., flood lamps and cameras failing due to engine vibration, difficulty establishing and maintaining the desired fuel injector pressure drop), but were eventually solved.

Analysis and Reporting

[material deleted]

The report states that the pressure of the inlet air successfully simulated the specified flight conditions of Mach .7, 2,230 ft. above sea level, but the temperature corresponded to a very hot day—about 113 degrees F. at sea level.[2] To simulate Mach .7 on a cooler day would have required a different inlet pressure and temperature values. The report states that total pressure and density describe not just a single Mach altitude flight condition, but a locus of points in the Mach altitude map; therefore, the test data are applicable to flight conditions other than those tested, including some with cooler temperatures (i. e., Mach 1.51, 25,000 ft., 20 degrees F.). It also states that the flight conditions simulated in the test are within the flight envelopes of the F-15 and F-16.

We believe the draft report overstates the generalizability of the findings. No statement is possible on the effect of changing a single parameter—all 3 must change. So, for example, the effect of Mach .7, 2,230 ft. at a cooler temperature (one representative of a European scenario) cannot be inferred. The fact that the test conditions were within the flight envelopes is irrelevant; it does not make them generalizable. Only additional testing could do that.

The report presented a computer model for predicting damage from a steady flow of fuel into a turbofan inlet. It was derived principally from the F100 test results, although in part from "the author's intuition alone."

[material deleted]

[2] Jet engines are used to generate the airflow, consequently the air temperature is artificially high.

The user inputs inlet duct diameter, ingestant flow rate, altitude, Mach number, distance of wound from engine face, and various other parameters. The output indicates presence or absence of engine failure and various other information.

We question the value of this model for the following reasons:

- A small model (less than 50 lines of code), it only addresses turbine section thermal failure; no other failure modes are included.
- It was developed post hoc; the deferred predictions alluded to in the DTP never materialized in the report.
- The user varies parameters independently, even though several of these were held constant in the test; consequently, the quantitative effect computed by the model is highly speculative.
- [material deleted]

The report's recommendations were congruent with the results, and sensitive to the likelihood of user acceptance. It concluded that given the engine designers' focus on thrust-to-weight ratios, performance, fuel consumption, and signature, little opportunity existed for engine design changes Recommendations were therefore focused elsewhere, on airframe and fuel system design. These included:

- Design fuel systems to reduce effects, such as fuel cells with higher self-sealing capability and improved ballistic protection.
- Design more survivable fuel cell configurations, such as inlet isolation liners, fuel ingestion sensors, and divided concentric fuel tanks "managed" so that the tank adjacent to the inlet is emptied before arrival at the threat zone.

Bibliography

Bennett, G. B. Ballistic Resistance of Nonhomogeneous Components: J79 Engine Tests. (CONFIDENTIAL) Washington, D. C.: Joint Technical Coordinating Group for Aircraft Survivability, March, 1975.

Bennett, G. B., Cramer, R., and Rice, B. Ballistic Resistance of Nonhomogeneous Components: Avionics Components Tests. Washington, D. C.: Joint Technical Coordinating Group for Aircraft Survivability, September, 1986.

DeBold, L., Franco, J. and Bastien, P. F-4 Aircraft and Fuel System Baseline Vulnerability Test Program, Part I, Airframe and Fuel System Response. Washington, D. C.: Joint Technical Coordinating Group on Aircraft Survivability, December, 1978.

Deitz, P. H. "Solid Geometric Modeling: the Key to Improved Materiel Acquisition from Concept to Deployment." Paper presented to the Army Operations Research Symposium XXII. Ft. Lee, VA. 3-4 Oct., 1983.

———. "Vulnerability/Lethality Modeling of Armored Fighting Vehicles." Briefing, Aberdeen Proving Ground, MD.: U.S. Ballistics Research Laboratory, June, 1986.

———. "Modeling and Validation of Live-Fire AFV Tests." Briefing, Aberdeen Proving Ground, Md.: U.S. Army Ballistics Research Laboratory, November, 1986.

DOD (U.S. Department of Defense), Air Force Flight Dynamics Laboratory. "AX Fuel Tank Vulnerability Evaluation: Summary Evaluation Report." Wright-Patterson Air Force Base, Ohio: December, 1972.

———., Joint Technical Coordinating Group for Munitions Effectiveness. "Joint Live Fire (JLF) Test: Armor, Anti-Armor Systems: Part I—Detailed test plan for the Training of Armored Vehicle Damage Assessment Teams (DRAFT)." Aberdeen Proving Ground, MD.: n. d.

———., Joint Technical Coordinating Group for Munitions Effectiveness. Joint Live Fire Test Plan for Armor/Anti-Armor Systems. January, 1985.

———., Joint Technical Coordinating Group for Munitions Effectiveness. Joint Live Fire (JLF) Test: Armor/Anti-Armor Systems (DRAFT). October, 1986.

———., Joint Technical Coordinating Group for Munitions Effectiveness. "Test Plan for Joint Live Fire Test, [material deleted] (DRAFT)." Aberdeen Proving Ground, Maryland: March, 1986.

———., Joint Technical Coordinating Group on Aircraft Survivability. "F100 Engine Steady-State Fuel Ingestion Test (DRAFT)." Washington, D. C.: January, 1986.

———., Joint Technical Coordinating Group on Aircraft Survivability. Joint Live Fire (JLF) Test: Air Combat Weapons Systems Versus Ballistic Threats: Preliminary Plan. Washington, D. C.: February, 1984.

———., Joint Technical Coordinating Group on Aircraft Survivability. Joint Live Fire (JLF) Test: Air Combat Weapons Systems Versus Ballistic Threats: Master Plan. Washington, D. C.: October, 1984.

———., Joint Technical Coordinating Group on Aircraft Survivability. Joint Live Fire (JLF) Test: Aircraft Systems: Detailed Test Plans for FY86. Washington, D. C.: October, 1985.

———., Office of the Under Secretary of Defense, Research and Engineering (DDTE). "Joint Live Fire Test Charter." Washington, D. C.: March, 1986.

———., Office of the Under Secretary of Defense, Research and Engineering (DDTE). "Report to Congress on Phase I Results of the Live-Fire Survivability Testing for the Bradley Fighting Vehicle System." Washington, D. C.: 17 December, 1985.

———., U.S. Army Ballistics Research Laboratory. Bradley Survivability Enhancement Program—Phase I Results. (SECRET) Aberdeen Proving Ground, Maryland: December, 1985.

———., U.S. Army Ballistics Research Laboratory. Detailed Test Plan for Full-UP, Combat loaded Vehicle tests of the [material deleted] Infantry Fighting Vehicle. Aberdeen Proving Ground, MD: October 5, 1985.

———., U.S. Army Ballistics Research Laboratory. "M-48 Test Results (PRELIMINARY DRAFT)." Aberdeen Proving Ground, MD.: n. d.

———., U.S. Army Ballistics Research Laboratory. "[material deleted] (PRELIMINARY DRAFT)". (SECRET) Aberdeen Proving Ground, MD.: July 1986.

———., U.S. Army Ballistics Research Laboratories, U.S. Army Materiel Systems Analysis Activity, U.S. Army Armor Center and School, and U.S. Army Infantry Center and School. "COORDINATION DRAFT: The Basis for Assessing the Vulnerability of Armored Vehicles." Aberdeen Proving Ground, Maryland: 10 December, 1975.

Donnelly, T. "Pentagon, Hill Clash on Fate of Bradley Live-Fire Test Advocate". Defense News, April 14, 1986.

Flint, James B. "A-10. GAU-8 Tank Lethality Tests and Effectiveness Analyses." (CONFIDENTIAL) Eglin Air Force Base, Florida.: U.S. Air Force Armament Laboratory, September 1982.

Flint, James B. "GAU-8 30mm API Damage to U.S. M-47 Tanks: Tests Versus Analytical Estimates." Eglin Air Force Base, Florida.: U.S. Air Force Armament Laboratory, March 1984.

GAO (U.S General Accounting Office). Bradley Vehicle: Concerns About Army's Vulnerability Testing, GAO NSIAD-86-87. Washington, D. C : February, 1986.

———. Bradley Vehicle: Army's Efforts to Make It More Survivable, GAO NSIAD-87-40. Washington, D. C.: November 4, 1986.

———. How Well Do the Military Services Perform Jointly in Combat?: DOD's Joint Test-and-Evaluation Program Provides Few Credible Answers, GAO PEMD-84-3. Washington, D. C.: February 22, 1984.

———. Models, Data, and War: A Critique of the Foundation for Defense Analyses, GAO PAD-80-21. Washington, D. C.: March 12, 1980.

Greene, T. E. et al. A Review of Selected Elements of the FY73 Programs on Test and Evaluation of Aircraft Survivability (TEAS). Volume I, Technical Report. Arlington, Va.: Weapons Systems Evaluation Group, July, 1973.

Hafer, T. Lethality Model Evaluation. (CONFIDENTIAL) Briefing. Arlington, Va.: Systems Planning Corporation, 1985.

Hall, D. and Adams, B. "Survivability: An Introduction to Assessment Methodologies " China Lake, CA.: U.S Naval Weapons Center, July, 1985.

Hamilton, J. R. "Resources: The Key to Meaningful Testing". Army Research, Development and Acquisition Magazine, May-June 1983, 7.

House Armed Services Committee Staff. Inquiry into the Bradley Live Fire Test Program. Washington, D. C.: May 21, 1986.

Kennedy, D. R. "The Infantryman vs. the MBT: Can an Infantryman's LAW Defeat the Frontal Armor of Future MBTs?". National Defense, March, 1985, 27-49.

——-. "Improving Combat Crew Survivability". Armor, July-August, 1983, 16-21.

Kokinakis, W., and Rudolph, R. "An Assessment of the Current State-Of-The-Art of Incapacitation by Air Blast." Aberdeen Proving Ground, Md.: U.S. Army Ballistics Research Laboratory, n. d.

Los Alamos National Laboratory, Statistics and Operations Research Group. "Statistical Aspects of Live Fire Testing (Draft)." Los Alamos, New Mexico. April, 1986.

McClung, R. M. "First Aid for Pet Projects Injured in the Lab or On the Range: Or, What to Do Until the Statistician Comes." China Lake, Calif.: U.S. Naval Ordnance Test Station, March, 1952.

Menne, D. F. et al. "Plans for Updating the Armored Vehicle Lethality/Vulnerability Methodology and Data Base." Aberdeen Proving Ground, Md.: U.S. Army Ballistics Research Laboratory, August 22, 1977.

National Research Council, Board on Army Science and Technology, Committee on Vulnerability Analysis. Letter Report to Deputy Under Secretary of the Army for Operations Research and Deputy Under Secretary of Defense, Research and Engineering—Test and Evaluation. Washington, D. C.: October 20, 1986.

——- "Report of the Board on Army Science and Technology's Committee on Vulnerability Analysis." Washington, D. C.: June 25, 1986.

Prifti, J "Ballistic Liners Improve M113 Survivability Rate". Army Research, Development and Acquisition Magazine, July-August, 1980, 7.

Ringers, D. A., Brown, F. T., and Braierman, W. F. "Vulnerability Predictions Based Upon Direct Inferences of Lethality." Aberdeen Proving Ground, Md.: U.S. Army Ballistics Research Laboratory, March, 1982.

Smith, G., Black, K., Tonneson, L., and Stein, A. "The Joint Live Fire (JLF) Test: Background and Exploratory Testing. (DRAFT)" (SECRET) Alexandria, Va.: Institute for Defense Analysis, March, 1986.

Starling, J. F. and Freeman, A. M. "30mm GAU-8/A Anti-Armor Data Base: Summary and Related Analyses." Eglin Air Force Base, Fla.: Air Force Armament Laboratory Analysis Division (AFATL/DYLD), n. d.

Stolfi, R. H. S. and McEachin, R. R. A-10/GAU-8 Low Angle Firings Against Simulated Soviet Tank Company (Array 3) (LAVP Lot # OL78D043-016, Honeywell). Monterey, Calif.: U.S. Naval Postgraduate School, August, 1980.

———. LAVP Data Summary: A-10/GAU-8 Low Angle Firings Against Simulated Soviet Tank Formations (Arrays 3 Through 48): (2 Feb 79 Through 5 Dec 80). (CONFIDENTIAL) Monterey, Calif.: U.S. Naval Postgraduate School, March 1982.

System Development Corporation. "A-7D Survivability/Vulnerability Test Program. Volume II: Test and Analysis." Dayton, Ohio: May, 1974.

Ten Broeck, D. W. "Combat Damage Repair Analysis of Helicopters Using the COVART/HEVART Codes." Aberdeen Proving Ground, Md. U.S. Army Ballistics Research Laboratory, September, 1986.

Vitali, R. "Opinion Paper: The Need and Process for Full Scale Vulnerability/Lethality Testing." (CONFIDENTIAL) Aberdeen Proving Ground, MD.: U.S. Ballistic Research Laboratory, September, 1984.

Zeller, G. A. Attack Azimuths (CONFIDENTIAL) Aberdeen, MD.: Armament Systems, Inc., July 1983.

Glossary

Altitude History

Recent prior altitude of an aircraft. If an aircraft has been at high altitude. some of the more volatile fuel constituents would have been lost, changing the vapor composition, and possibly affecting the probability of sustaining a fire.

Attack Azimuth

The angle in the horizontal plane from which a shot is fired at an armored target, proceeding counterclockwise from 0 degrees at the front (so that shots perpendicular to the left side of a tank, for example, are at 90 degrees azimuth.)

Dry Bay

Area of aircraft fuselage around a fuel tank. containing fuel lines, vent lines, wiring, etc.

Dynamic Firing

Launch of a threat munition at a distance from the target. (Contrasted with static firing.)

Effectiveness

Ability of a weapon system to cause specified damage to a specific target, taking into account ability to acquire, track, and hit the target.

Full-Scale Tests

Those conducted on complete weapons systems rather than components or mock-ups.

Full-Up Tests

Those conducted with the full complement of fuel, ammunition, and hydraulic fluid carried by the system into combat.

Glacis Plates

Sloping armor plates that form the front or rear of a vehicle and in effect increase the presented armor thickness.

Halon

Gas used in automatic fire suppression systems such as the one in the Bradley vehicle.

Hydrazine

A corrosive, volatile liquid used in rocket fuels and in emergency power systems of some aircraft.

Kevlar

Synthetic material used to stop small metal fragments in bulletproof vests and spall liners for armored vehicles.

Kinetic Energy Weapon

Munition whose penetrator is a dense metal rod, relying for its effect on the momentum of its flight.

Lethality

The ability of a munition to produce a specified level of damage to a specific target, given that the target has been hit.

Maneuver Load

Changes in load on an aircraft wing due to maneuvering by the aircraft (e.g. in evading air defenses). Maneuver loads change the effective weight of fuel, and hence the leak rate; they also produce stress on fuel cell walls that potentially increase damage.

Overmatching Weapon

An antiarmor weapon that is thought to be almost certain to defeat the armor of a given target; a more powerful threat than those a system was designed to withstand.

Overpressure

The potentially casualty-producing increase in atmospheric pressure within a vehicle caused by the detonation of onboard explosives, by materials associated with the penetration process that are rapidly oxidized, or by the passage of a shaped-charge jet or kinetic energy penetrator into the vehicle's interior.

Penetrator

The part of an antiarmor weapon that is intended to pass through the armor, either a high-speed jet of metal from a shaped-charge warhead, or the solid metal rod of a kinetic energy weapon.

Replicas

Fabricated substitutes for unavailable threat weapons or targets.

Shaped Charge

Focused-energy warhead, in which the thin metal liner of a conical cavity is explosively formed at the moment of detonation into an extremely high velocity, continuously stretching, thin metal jet that has great armor-penetrating ability.

Shotline

Path travelled by threat weapon, determined by azimuth, elevation, and impact point.

Slosh

Movement of fuel within a fuel tank during flight. Slosh wets the tank's internal walls, changing the distributions of vapors and potentially affecting the probability of sustaining a fire.

Simulants

Fabricated substitutes for unavailable threat weapons or targets.

Spall

Fragments of armor and penetrator material thrown off at the inner surface of armor breached by a penetrator.

Spall Liner

Thick panels of composite material such as Kevlar installed at the interior surface of armor, to reduce the effects of spall.

Static Firing

Detonation of a shaped-charge warhead that has been taped or fixed in some other way to the surface of a target in order to hit a precise impact point. Contrasted with dynamic firing.

Stochastic

Incorporating randomness or chance. A stochastic model attempts to mimic the random variation of a process in the world by having some of the model outcomes determined in a random draw from a distribution of possible outcomes.

Subscale Tests

Any tests conducted on less-than-full-scale target weapon systems, such as component vulnerability tests or behind-armor-debris studies.

Surrogate

An existing munition or target substituted for one that is unavailable for testing on the basis of similarity.

Survivability

The ability of a weapon system to avoid being killed in battle, including its vulnerability if hit, but also taking other factors such as maneuverability and the ability to avoid detection into account.

Susceptibility

Comprises all the capabilities and characteristics of a target and threat that influence or determine the probability that the target is hit, including the threat capability to detect, lock on, track, and fire, and the target capability to evade the threat.

Vaporifics

A postulated casualty-causing mechanism involving the rapid combustion of 1) metals from armor, penetrator, or vehicle components after penetration by a very large shaped charge weapon or 2) the metal liner of shaped charges specifically designed to produce vaporific effects.

Vulnerability

The inability of a weapon system to withstand damage from a specific attack, given that it has been hit.

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.