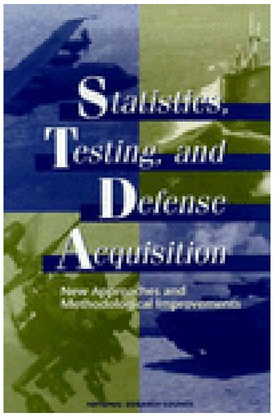

Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements

Michael L. Cohen, John E. Rolph, and Duane L. Steffey, Editors; Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics, National Research Council

ISBN: 0-309-51964-0, 240 pages, 6 x 9, (1998)

Statistics, Testing, and Defense Acquisition

New Approaches and Methodological Improvements

Michael L. Cohen, John E. Rolph, and Duane L. Steffey, *Editors*

Panel on Statistical Methods for Testing and Evaluating Defense Systems

Committee on National Statistics

Commission on Behavioral and Social Sciences and Education

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1998

NATIONAL ACADEMY PRESS • 2101 Constitution Avenue, N.W. • Washington, D.C. 20418

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

The project that is the subject of this report is supported by Contract DASW01-94-C-0119 between the National Academy of Sciences and the Director, Operational Test and Evaluation at the U.S. Department of Defense. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

Additional copies of this report are available from the National Academy Press, 2101 Constitution Avenue, NW, Lock Box 285, Washington, DC 20055. (800) 624-6242 or (202) 334-3313 (in the Washington Metropolitan Area).

Library of Congress Cataloging-in-Publication Data

Statistics, testing, and defense acquisition : new approaches and
methodological improvements / Michael L. Cohen, John E. Rolph, and
Duane L. Steffey, editors ; Panel on Statistical Methods for
Testing and Evaluating Defense Systems, Committee on National
Statistics, Commission on Behavioral and Social Sciences and
Education, National Research Council.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 0-309-06551-8 (pbk.)
1. United States—Armed Forces—Weapons
systems—Testing—Statistical methods. I. Cohen, Michael L. II.
Rolph, John E. III. Steffey, Duane L. IV. National Research Council
(U.S.). Commission on Behavioral and Social Sciences and Education.
Panel on Statistical Methods for Testing and Evaluating Defense
Systems.
UF503 .S727 1998
358.4′1807′0973—ddc21 98-9065

Printed in the United States of America

PANEL ON STATISTICAL METHODS FOR TESTING AND EVALUATING DEFENSE SYSTEMS

JOHN E. ROLPH *(Chair)*, Marshall School of Business, University of Southern California
MARION R. BRYSON, North Tree Management, Monterey, California
HERMAN CHERNOFF, Department of Statistics, Harvard University
JOHN D. CHRISTIE, Logistics Management Institute, McLean, Virginia
LOUIS GORDON, Private Consultant, Palo Alto, California
KATHRYN BLACKMOND LASKEY, Department of Systems Engineering and Center of Excellence in C^3I, George Mason University
ROBERT C. MARSHALL, Department of Economics, Pennsylvania State University
VIJAYAN N. NAIR, Department of Statistics, University of Michigan
ROBERT T. O'NEILL, Division of Biometrics, Food and Drug Administration, U.S. Department of Health and Human Services
STEPHEN M. POLLOCK, Department of Industrial and Operations Engineering, University of Michigan
JESSE H. POORE, Department of Computer Science, University of Tennessee
FRANCISCO J. SAMANIEGO, Division of Statistics, University of California, Davis
DENNIS E. SMALLWOOD, Department of Social Sciences, U.S. Military Academy

MICHAEL L. COHEN, *Study Director*
DUANE L. STEFFEY, *Study Director* (to July 1995); *Consultant*
ANURADHA P. DAS, *Research Assistant*
ERIC M. GAIER, *Consultant*
CANDICE S. EVANS, *Senior Project Assistant*

COMMITTEE ON NATIONAL STATISTICS
1997-1998

Acknowledgments

The very nature of this study required the panel and staff to attend meetings (nine plenary and ten working group meetings) all over the country to collect information on test designs and evaluations on a wide variety of systems; a list of the systems we studied, including several we used as case studies, are in Appendix A. Locations we visited include the Army Test and Experimentation Command Headquarters at Fort Hunter Liggett in California; Eglin Air Force Base at Fort Walton Beach in Florida; the Air Force Operational Test and Evaluation Center in Albuquerque, New Mexico; the Navy Operational Test and Evaluation Force in Norfolk, Virginia; the Army Operational Test and Evaluation Command in Alexandria, Virginia; RAND in Santa Monica, California; and the Institute for Defense Analyses in Alexandria, Virginia. We were extremely fortunate to meet with so many people—from the military services, the Office of the Secretary of the U.S. Department of Defense (DoD), and private organizations in the testing community—willing to share their expertise with, and extend their hospitality to, the panel: to all these individuals, we are grateful.

We particularly wish to acknowledge the support of Philip Coyle, director, and Ernest Seglie, science adviser, DoD Office of the Director, Operational Test and Evaluation (the study sponsors); Henry Dubin, technical director, U.S. Army Operational Test and Evaluation Command; Steven Whitehead, technical director, U.S. Navy Operational Test and Evaluation Force; Marion Williams, technical director, U.S. Air Force Operational Test and Evaluation Center; and Robert Bell, technical director, U.S. Marine Corps Operational Test and Evaluation Activity.

In addition, many people went beyond the call of duty to assist the panel in

its work. We thank, first: Kwai Chan, Christine Fossett, Jackie Guin, Louis Rodrigues, and Robert Stolba, General Accounting Office; Ric Sylvester, Office of the Deputy Undersecretary of Defense for Acquisition Reform; Lee Frame and Austin Huangfu, Office of the Director, Operational Test and Evaluation; Margaret Myers and Ray Paul, Office of the Secretary of Defense; Patricia Sanders, Office of the Undersecretary of Defense for Acquisition and Technology; Dean Zerbe, Senator Grassely's Office; Donald Yockey, former Under Secretary of Defense for Acquisition.

From the Air Force, we thank: Howard Leaf, Director of Test and Evaluation, U.S. Air Force; Suzanne Beers, David Blanks, Lyn Canham, Michael Carpenter, Charles Carter, Angie Crawford, David Crean, William Dyess, John Faris, Tim Gooley, Anthony "Shady" Groves, Ken Hebert, Brian Ishihara, Jeff Jacobs, Eric Keck, Roderick Leitch, Scott Long, Mike Malone, Michael McHugh, Donald Merkison, Terence Mitchell, Herbert Morgan, Ken Murphy, Sharon Nichols, Steve Ordonia, Ronald Reano, Mark Reid, James Sheedy, Brian Simes, Chuck Stansberry, Cecil Stevens, Robert Stovall, Frank Swehoskey, Scott Weisgerber, Larry Wolfe, and Dave Young, U.S. Air Force Operational Test and Evaluation Center.

From the Army, we thank: Susan Wright, Army Digitization Office; Cy Lorber, Army Materiel Command; Will Brooks, Sam Frost, Dwayne Nuzman, Jim Streilein, and Bill Yeakel, Army Materiel Systems Analysis Activity; Charles Pate, Training and Doctrine Command; Larry Leiby, Scott Lucero, John McVey, and Hank Romberg, U.S. Army Operational Test and Evaluation Command; Michael Hall, Greg Kokoskie, Ed Miller, Harold Pasini, Patrick Sul, and Tom Zeberlein, U.S. Army Operational Evaluation Command; Michael Jackson and Carl Russell, U.S. Army Test and Experimentation Center.

From the Navy, we thank: James Duff, former technical director of the Navy Operational Test and Evaluation Command; Donald Gaver, Naval Postgraduate School; Karen Ahlquist, Mike Alesi, Jeff Gerlitz, Kevin Smith, and Cynthia Womble, Navy Operational Test and Evaluation Force.

And from other institutions and agencies, we thank: Nozer Singpurwalla, George Washington University; Robert Boling, Peter Brooks, William Buchanan, James Carpenter, Thomas Christie, Gary Comfort, Robert Daly, Robert Dighton, Richard Fejfar, Arthur Fries, David Hart, Kent Haspert, Anil Joglekar, Irwin Kaufman, Richard "Hap" Miller, Michael Shaw, David Spalding, Bradley Thayer, Alfred Victor, Charles Waespy, and Steve Warner, Institute for Defense Analyses; Dale Pace, Johns Hopkins University; and Patrick Vye, RAND.

This report has been reviewed by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the NRC's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the authors and the NRC in making the published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The content of the review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their participation in the review of this report: David S.C. Chu, RAND, Washington, D.C.; Phil E. DePoy, National Opinion Research Center, University of Chicago; Gerald P. Dinneen, consultant, Edina, Minnesota; William Eddy, Department of Statistics, Carnegie Mellon University; Alexander H. Flax, consultant, Potomac, Maryland; David R. Heebner, Heebner Associates, McLean, Virginia; Robert J. Hermann, United Technologies Corporation, Hartford, Connecticut; James Hodges, Division of Biostatistics, University of Minnesota, Twin Cities; William Howard, consultant, Scottsdale, Arizona; Joseph B. Kadane, Department of Statistics, Carnegie Mellon University; Patrick D. Larkey, Heinz School of Public Policy, Carnegie Mellon University; John L. McLucas, consultant, Alexandria, Virginia; General Glenn K. Otis (ret.), Newport News, Virginia; and Warren F. Rogers, Warren Rogers Associates, Inc., Middletown, Rhode Island.

While the individuals listed above provided many constructive comments and suggestions, responsibility for the final content of this report rests solely with the authoring committee and the NRC.

The panel was fortunate to have an extremely able staff who both supported and led the panel through the past four years. I would like to particularly acknowledge the contributions of Duane Steffey, study director for the first phase of the study, and Michael Cohen, study director for the final phase. Their research and organizational skills, combined with their ability to develop contacts and foster relationships in the testing community made them an important asset to the success of this study.

The panel is grateful to Eugenia Grohman, Associate Director for Reports of the Commission on Behavioral and Social Sciences and Education (CBASSE), for her fine technical editorial work, which contributed greatly to the readability of this report.

Anu Das, research assistant for the study, provided invaluable support to the panel, particularly to the work of the software testing working group, assisting greatly in the preparation of Chapter 8. The panel's senior project assistant, Candice Evans, along with the difficult job of coordinating many offsite meetings—further complicated by the necessity for security clearances (and a hurricane!)—also handled all aspects of report production, most notably helping prepare drafts of Chapters 1 and 2 and Appendix D.

Finally, no acknowledgment would be complete without thanking the panel members themselves: they traveled extensively to military bases and test facilities, contributed their time and expert knowledge, and drafted many of the sections of the report.

John E. Rolph, *Chair*
Panel on Statistical Methods for Testing
and Evaluating Defense Systems

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. William A. Wulf is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce Alberts and Dr. William A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

Contents

PART I
APPLICATIONS OF STATISTICAL PRINCIPLES TO DEFENSE ACQUISITION

PART II
APPLICATIONS OF STATISTICAL METHODS TO OPERATIONAL TESTING

APPENDICES

Preface

The Committee on National Statistics of the National Research Council (NRC) has had a long-standing goal of encouraging the development and use of state-of-the-art statistical methods across the federal government. In this context, discussions began several years ago during meetings of the Committee on National Statistics about the possibility of providing assistance to the U.S. Department of Defense (DoD). Mutual interest between the committee and the DoD Office of Program Analysis and Evaluation in the greater application of statistics within DoD led to a meeting of key DoD personnel and several NRC staff. As a result of this meeting, system testing and evaluation emerged as an area for which improvement in the application of statistics could prove useful.

Consequently, at the request of DoD, the Committee on National Statistics, in conjunction with the NRC Committee on Applied and Theoretical Statistics, held a 2-day workshop in September 1992 on experimental design, statistical modeling, simulation, sources of variability, data storage and use, and operational testing of weapon systems. The workshop was sponsored by the Office of the Director, Operational Test and Evaluation (DOT&E) and the Office of the Assistant Secretary of Defense for Program Analysis and Evaluation. Defense analysts were invited to write and present background papers and discuss substantive areas in which they sought improvements through application of statistical methods. Statisticians and other participants responded by suggesting alternative approaches to specific problems and identifying program areas that might especially benefit from the application of improved statistical methods. The overarching theme of the workshop was that using more appropriate statistical approaches could improve the evaluation of weapon systems in the DoD acquisition process. The workshop findings were published in *Statistical Issues in Defense Analysis and Testing: Summary of a Workshop* (Rolph and Steffey, 1994).

Workshop participants expressed the need for a study to address in greater depth the issues that surfaced at the workshop. A multiyear panel study was undertaken by the Committee on National Statistics in early 1994 sponsored by DOT&E. The Panel on Statistical Methods for Testing and Evaluating Defense Systems was established to recommend statistical methods for improving the effectiveness and efficiency of testing and evaluation of defense systems, with emphasis on operational testing. The 13-member panel comprised experts in the fields of statistics (including quality management, decision theory, sequential testing, reliability theory, and experimental design), operations research, software engineering, defense acquisition, and military systems.

The panel's interim report, *Statistical Methods for Testing and Evaluating Defense Systems* (National Research Council, 1995), presented some preliminary findings, but it did not offer any recommendations. Key chapters were devoted to experimental design of operational test, operational testing of software-intensive systems, operational test and evaluation for reliability, availability, and maintainability, and use of modeling and simulation to assist in operational test design and evaluation.

This report presents the conclusions and recommendations resulting from the panel's 4-year study. The report is structured to accommodate various types of readers. Chapters 1-4 are for a nontechnical audience. Chapter 1 discusses the panel's scope of work and how this was adjusted to deal with constraints on the application of statistics in the test and acquisition of military systems. Chapters 2 and 3 assess the current use of testing in system development and identify key elements of a new paradigm for the use of testing as part of the development of defense systems. Chapter 4 summarizes the substantial benefits that defense operational test design and evaluation would obtain from the use of statistical methods that reflect current practices. The changes recommended in Chapter 4 do not assume that the new paradigm recommended in Chapter 3 will be adopted and therefore can be implemented immediately.

Chapters 5-9 explore in more detail and more technically the topics covered in Chapter 4 as well as additional issues concerning the application of state-of-the-art statistical methods to defense operational test design and evaluation: experimental design for operational testing (Chapter 5), operational test evaluation (Chapter 6), test design and evaluation for reliability, availability, and maintainability (Chapter 7), software testing (Chapter 8), and modeling and simulation for use in operational test design and evaluation (Chapter 9).

Chapter 10 considers the need for the defense test and acquisition community to develop greater access to statistical expertise and how to do so.

Though the panel was not charged with the development or execution of technical work related to operational testing and evaluation, the panel decided that further exploration of certain technical issues would be useful for its deliberations. Thus, three technical papers were prepared: "Strategic Information Generation and Transmission: The Evolution of Institutions in DoD Operational Testing," "On the Performance of Weibull Life Tests Based on Exponential Life Testing Designs," and "Application of Statistical Science to Testing and Evaluating Software Intensive Systems." The panel has drawn from the papers, which will be published separately; abstracts of them are presented in Appendix B.

John E. Rolph, *Chair*
Panel on Statistical Methods for Testing
and Evaluating Defense Systems

Executive Summary

The Panel on Statistical Methods for Testing and Evaluating Defense Systems was charged with examining the statistical techniques currently used in the design and evaluation of operational tests in the U.S. Department of Defense (DoD) and making recommendations for improvement. Operational testing and evaluation is an independent assessment of whether a system is effective and suitable for its intended use; it is a key part of military system development. The DoD's Office of the Director, Operational Test and Evaluation (DOT&E) is responsible for providing advice to the Secretary of Defense on whether military systems are ready for full-rate production and for prescribing policies and procedures to the military services regarding operational test and evaluation. Given the importance of the nation's defense, the high cost of many weapon systems, and the substantial cost of testing them, even modest improvements in operational testing by using the most appropriate statistical methods can lead to more efficient use of public funds and considerable improvements in the reliability and effectiveness of the systems deployed.

The panel's examination comes at a time of substantial change in the development and testing process and, more broadly, in military acquisition. There are five major components of that change: decreased testing budgets; more complicated systems; more software-intensive systems; more upgrades to existing systems ("evolutionary procurement"); and greater interest in system reliability, availability, and maintainability.

Early in its work the panel realized that a narrow focus only on the use of sophisticated statistical techniques in test design and evaluation would not provide the best advice to DoD. Rather, the panel decided it must also consider the aspects of the acquisition process as a whole that affected the application of statistical techniques. It became clear to us that adopting effective statistical practices that command wide and consistent support within the DoD acquisition community would yield substantial gains.

CONCLUSIONS ABOUT THE CURRENT PROCESS

The panel's main conclusions concerning the current use of operational testing as part of system development cover broad aspects of DoD operational testing and evaluation.

Currently, operational testing is a collective final event in system development. Since many design flaws in both industrial and defense systems become apparent only in operationally relevant use, some testing with operational realism needs to be carried out as early as possible. Inexpensive, small, focused preliminary tests with operational aspects could help to identify problems before a design is final. Such tests could also identify scenarios in which the current design performs poorly and which system characteristics should be the focus of subsequent tests.

In addition, it is currently uncommon to use developmental test data, or test data for related systems, to augment data from operational testing except for direct pooling of reliability test data. This omission derives in part from understandable concerns about the relevance of developmental test data or test data from related systems for the evaluation of a system's operational performance, but it also originates in part, from a lack of statistical expertise about how to use the information and a lack of access to the information.

> **Conclusion 2.1: For many defense systems, the current operational testing paradigm restricts the application of statistical techniques and thereby reduces their potential benefits by preventing the integration of all available and relevant information for use in planning and carrying out tests and in making production decisions.**

Also, the incentive structure in military procurement provides each major participant—including the program manager for a system, the test director, the contractor, various activities in the Office of the Secretary of Defense (OSD), and Congress—with strong, often differing and even competing perspectives. This set of complicated dynamics affects a variety of aspects of the test and evaluation process—budgets, schedules, test requirements, test size, which test events should be excluded because of possible test abnormalities, and even the rules for scoring a test event a "failure." It is critical that the perspectives of the participants are understood and taken into account in decision making concerning test design and test evaluation.

Further, for operational tests of most complicated systems, the required sample sizes that would support significance tests at the usual levels of statistical confidence typically are not affordable. As a result, there is a common perception in the DoD acquisition community, including middle- and high-level management, that the design and conduct of "statistically valid" tests are unaffordable. This impression stems in part from a lack of communication between the test and acquisition and statistical communities about how statistical theory and methods can be applied to complex problems. For example, effective use of statistical methods is not limited to a determination of the appropriate sample size so that a test yields interval estimates at the required level of statistical confidence. It would often be preferable, given a fixed test budget, to design a test aimed at maximizing the amount of information from the resulting test data in order to reach supportable statements about system performance. Failure to use such a test is tantamount to using more test cases than needed or to throwing away information from test cases after conducting the test.

> **Conclusion 2.2: The operational test and evaluation requirement, stated in law, that the Director, Operational Test and Evaluation certify that a system is operationally effective and suitable often cannot be supported solely by the use of standard statistical measures of confidence for complex defense systems with reasonable amounts of testing resources.**

However, operational test and evaluation is a crucial element of assessing whether a system is ready for full-rate production.

> **Conclusion 2.3: Operational testing performs a unique and valuable function by providing information on the integration of user, user support (e.g., training and doctrine), and equipment in a quasirealistic operational environment.**

Moreover, we believe that:

> **Conclusion 3.1: Major advances can be realized by applying selected industrial principles and practices in restructuring the paradigm for operational testing and the associated information gathering and evaluation process in the development of military systems.**

A NEW PARADIGM

These conclusions support a new approach to more effectively employing testing as part of military system development. We therefore propose a new paradigm in which testing is more fully integrated as a part of military system development and recommend that Congress change the law to reflect the shift from testing primarily to confirm to one of testing to learn. Specifically, we recommend:

> **Recommendation 3.1: Congress and the Department of Defense should broaden the objective of operational testing to improve its contribution**

to the defense acquisition process. The primary mandate of the Director, Operational Test and Evaluation should be to integrate operational testing into the overall system development process to provide as much information as possible as soon as possible on operational effectiveness and suitability. In this way, improvements to the system and decisions about continuing system development or passing to full-rate production can be made in a timely and cost-efficient manner.

With this redirection of focus, the new DoD paradigm for test and evaluation as part of acquiring defense systems would have the following four essential characteristics:

1. It is a continuous process of information gathering and decision making in which operational testing and evaluation play an integral role. This orientation is consistent with contemporary trends in industrial practice de-emphasizing "inspecting defects out" in place of integrated development processes and "building quality into the product."

2. There is an environment of systematic data collection, analysis, and documentation. All sources of data and information should be clearly documented, and the process made consistent across the military services so that development teams are able to identify and learn from other relevant studies and findings. To create this environment, DoD could institute multi-service developmental and operational test and evaluation standards similar to those contained in ISO 9000.

3. Efficient statistical methods are used for decision making based on all available, relevant data. An environment of continuous assessment and systematic data collection should enable the use of efficient statistical methods, including decision-theoretic methods, yet would still allow the decision maker to be held accountable for judgments on system procurement.

4. The life-cycle costs of acquiring and maintaining new military capabilities are reduced. Integrating operational test and evaluation in a process of continuous assessment (and improvement) means that problems of consequence will be discovered earlier in the development process when they are easier and cheaper to solve. A system would therefore not be likely to enter the final, confirmatory stages of operational test until it clearly is ready for such a stage, and therefore also likely ready for full-rate production. The production of higher quality and more reliable systems will reduce the amount of logistic support required to maintain the system in the field, a major contributor to total cost over the lifetime of the system.

We note that these characteristics have not been totally missing from all aspects of military system development. They are consistent with several recent initiatives in DoD. They are found in various places, for particular items, at some times. But what has *not* existed is a consistent, department-wide, and institutionalized application of these features to the acquisition process. Adopting the new paradigm will enable DoD to obtain the best possible information at the lowest possible costs from operational testing and evaluation.

In the new paradigm, the role of DOT&E should be expanded to include providing assistance to the test and evaluation service agencies in the application of statistical and quality management principles to the overall decision-making process. This direction and oversight role involves the following:

- Careful definition and documentation of the key steps in the overall test and evaluation process, and what each step involves, to ensure that there is less variability in the process (from program to program while permitting tailoring for different types of programs) and that best current practices are used widely and consistently.
- Through the use of uniform terminology, protocols, and standards for data collection, archival, and documentation, ensuring that information from different sources can be compared and combined when necessary for both test design and evaluation. Information from both developmental and operational tests on the same and related systems can best be used for improved test design and test evaluation if it is accessible and usable in a test archive. It is most useful if data are collected and archived in standard ways using standard definitions and that essential information on every test exercise and condition is included. Adding information on field use would support the operation of feedback loops that would help improve the development of individual systems as well as the process of test design and evaluation. The use of standard terminology and practices will also facilitate the work of DOT&E in approving test plans, verifying test results, and assessing the validity of the results.
- Ensuring the full use of the information that is collected, archived, and documented through the use of state-of-the-art statistical methods and models. These methods and models permit use of data from different sources to design tests more efficiently and to evaluate tests by incorporating the specific information from each test data set that is relevant for the assessment of operational performance. In addition, statistical models can be constructed to provide such specific information as which test factors have the greatest effect on system performance and which test results seem most anomalous. The costs of operational testing and the importance of correctly deciding whether to proceed to full-rate production makes it extremely important to base decisions on all of the available, relevant information.
- Requiring careful documentation of each test evaluation and inclusion of confidence statements, along with sensitivity analyses and external validation of modeling and simulation used to augment operational tests. Careful but appropriately condensed documentation of the entire testing and evaluation process permits all the participants in defense acquisition to see how each evaluation was made, where there is still uncertainty, and whether and what further testing would

be useful. It also allows current and future users to better understand causes of variation and unanticipated problems and how they were addressed, to better manage the process.

Consistent with the new paradigm and the expanded revised role for DOT&E, the panel recommends:

Recommendation 2.1: The Department of Defense and the military services should provide a role for operational test personnel in the process of establishing verifiable, quantifiable, and meaningful operational requirements. Although the military operators have the final responsibility for establishing operational requirements, the Operational Requirements Document would benefit from consultation with and input from test personnel, the Director, Operational Test and Evaluation, and the operational test agency in the originating service. This consultation will ensure that requirements are stated in ways that promote their assessment.

Recommendation 2.2: The Director, Operational Test and Evaluation, subject to the approval of the Secretary of Defense on a case-by-case basis, should have access to a portion of the military department's acquisition funding reserves (being set up as a result of the first quadrennial defense review) to augment operational tests for selected weapon systems.

Recommendation 3.2: Operational evaluations should address the overall performance of a defense system and its ability to meet its broadly stated mission goals. When a system is tested against detailed requirements, these requirements should be shown to contribute to necessary performance factors of the system as a whole.

Recommendation 3.3: The Department of Defense and the military services, using common financial resources, should develop a centralized testing and operational evaluation data archive for use in test design and test evaluation.

Recommendation 3.4: All services should explore the adoption of the use of small-scale testing similar to the Army concept of force development test and experimentation.

ADVANCES IN STATISTICAL METHODS

At a more technical level, the panel also concludes that current practices in defense testing and evaluation do not take full advantage of the benefits available from the use of state-of-the-art statistical methodology. We present here our general conclusions and recommendations. Additional, detailed recommendations are in Chapters 5-10, including a recommendation in Chapter 10 that DoD form a Statistical Analysis Improvement Group, using the best statisticians in OSD and the military services on a part-time basis to advise senior decision makers.

Test Planning Comprehensive operational test planning is not uniformly conducted. Test planning collects and uses information on the purpose of the test, the test factors and how they need to be treated in the test design, the test environment and associated constraints, and preliminary data on system performance (e.g., comparison of system variation within and between scenarios) to help determine how large a test should be and which test scenarios (in which order) should be used to achieve the test goals. It is also used to resolve such questions as how and what test data are to be recorded. Proper test planning can avoid serious mistakes in test design and can help identify designs that produce substantially more information at less cost.

Estimates of Uncertainty All estimates of system performance from operational test should be accompanied by statements of uncertainty. Such statements would identify which levels of system performance were consistent with the test results, so that the extent of the information provided by the test is clear to decision makers. Estimates of uncertainty make it possible to use all information about the performance of a system in combination for the purpose of deciding on full-rate production. If more testing is needed, statements of uncertainty will usually make that clear. Estimates of uncertainty for performance in important individual scenarios should also be provided, as should approximate estimates of uncertainty of the information from use of modeling and simulation.

Use of All Relevant Information All relevant information from tests and field use of related systems, developmental tests, early operational tests, and training and contractor testing should be examined for possible use in both the design and evaluation of operational tests. Given the importance of the decision on whether to proceed to full-rate production, state-of-the-art statistical methods for combining information should be used, when appropriate, to make tests and their associated evaluations as cost-efficient as possible.

Experimental Design Methods State-of-the-art methods for experimental design should be routinely used in designing operational tests. A comprehensive literature is available to use in producing test designs that accommodate a broad array of complexities, especially various constraints, that arise in testing associated with system development. Routine reliance on this literature will produce more effective tests and help make efficient use of operational test funds.

Appropriate Statistical Models Statistical models that are based on empirically supported assumptions should be used in test design for and evaluation of system reliability (as well as for system effectiveness). It is typical for the service test agencies to make use of a specific model (the exponential model) of the time to first failure, in reliability design and evaluation. Use of this model, when inappropriate, can often result in unnecessarily large test sizes and inappropriate results for significance tests. There are many other models that are often more appropriate for this kind of application.

Software-Intensive Systems Operational tests of software-intensive systems should employ a usage-based perspective, in order to demonstrate that the system is fit for its intended use. When testing schedules and budgets are tightly constrained, usage-based testing yields the highest practical reliability because if failures are identified, they are likely to be the high-frequency failures.

In summary, by making testing a more useful and integrated component of military system development, by adopting up-to-date statistical practices, and by changing the paradigm in which test and evaluation is used in defense system development, the acquisition of military systems will become even more effective and efficient, and benefit from considerable savings in system life-cycle costs.

PART I

APPLICATIONS OF STATISTICAL PRINCIPLES TO DEFENSE ACQUISITION

1

Introduction

The Panel on Statistical Methods for Testing and Evaluating Defense Systems was formed to assess current practice related to operational testing in the Department of Defense (DoD) and to consider how the use of statistical techniques can improve that practice. This chapter describes the scope of the panel's work. It begins with a hypothetical example of the complex task of operational test design, implementation, and evaluation. The panel then discusses changes in the DoD acquisition process that guided its expanded scope of work. It concludes by detailing how statistics can effectively contribute to acquiring high-quality defense systems.

A HYPOTHETICAL SCENARIO IN THE TESTING AND EVALUATION OF A COMPLICATED SYSTEM

Consider the following scenario. A new major has just rotated into a service test agency and has been assigned to oversee the operational test design and subsequent evaluation of the test results of QZ3, a major military system. The results of the test are to inform the assessment of the system's operational readiness. Operational readiness is determined by the system's performance in an operational setting, which itself can only be measured in the military contexts of training, doctrine, and the scale in which the system is to be used. Of necessity, operational testing will be limited in time, the level of realism that can be attained, and the quality or quantity of measurement data obtainable from what at best is an adequate approximation of the chaotic and uncontrollable environment in which the system will actually be used. In addition, the size of operational testing is limited by the level of funding set by the system's program manager. The test report will help decide (along with other considerations, especially cost) whether the system should enter full-rate production, continue to be developed, or be terminated. If it is ready for full-rate production, the costs of production are likely to exceed $5 billion.

QZ3 is a substantial modification of an existing system, QZ2. One modification incorporates a substantial amount of new software, which requires more than 1 million new lines of code. QZ3 is designed for four different threats, three different types of engagement scenarios, and, possibly, three different environments (e.g., snow background, verdant temperate, and desert). There are 15 different measures of performance or effectiveness that will be used to evaluate QZ3, linked only loosely to the higher-level mission outcomes. For each of the 15 measures, and for each combination of threat, engagement scenario, and environment, the major must evaluate and compare the relative performance of QZ3 and QZ2, with the ultimate goal of answering: "Does the test confirm that a unit equipped with QZ3 accomplishes its mission in a better way than one equipped with QZ2."

To be approved for full-rate production, QZ3 must outperform QZ2 in several of the 15 measures, and it cannot perform poorly in others. Unfortunately, many of the requirements for performance were vaguely worded in the Operational Requirements Document (ORD) and are essentially untestable in their original form. For example, the ORD states that the new system should be "easier to use," and a crucial component should have reliability of 1 failure for every 2,000 hours, yet the total time on test is limited to 600 hours. Because such requirements are not testable, the major has to negotiate with various parties to specify new requirements that are testable. Some of the 15 measures relate to how effective the system is in carrying out its function (system effectiveness), and some relate to how long or often the system is in a usable state (system suitability or reliability, maintainability, and availability). Moreover, a test designed to measure effectiveness by itself may be substantially different than one designed to measure only suitability.

To further complicate his task, the major is faced with numerous difficulties when trying to implement the QZ3 tests, including: scheduling of test facilities and soldiers; the availability and use of threats; the availability of environments; and safety, noise, and other environmental restrictions. In addition, the test runs are very complicated to coordinate, and uncontrolled events in the field often cause atypical outcomes that must be identified and analyzed separately. The entire test and evaluation process must be carried out in less than 2 years, including 2-3 months to evaluate the test data and write the summary test report. The report will focus on several key results from various statistical significance (or hypothesis) tests for differences of means or differences of percentages for individual measures of effectiveness or performance.

The program manager had originally budgeted for 15 test scenarios, with 5 test prototypes each. This permits a maximum of 75 replications to be distributed in 36 possible test scenarios (4 threats × 3 engagement scenarios × 3 environments). Thus, only an incomplete assessment of variability across scenarios and prototypes (potentially substantially incomplete) is possible. Given unanticipated system development costs, there is pressure to reduce the number of test replications. Various earlier developmental tests of this system, not conducted under operationally realistic conditions, have given results that differ substantially among themselves, and so there is a general reluctance to use this earlier test data to supplement the information from operational test. Simple pooling of the earlier data with operational test data will require unreasonable assumptions (e.g., operational conditions have no effect on performance) so combination will require sophisticated statistical models that can accommodate the effects of different test circumstances. Furthermore, it is difficult to use information from developmental and operational testing and field performance for similar systems either to help design QZ3's operational test or to supplement QZ3's operational test evaluation because all the conditions under which the developmental data were taken, and the developmental and operational test results, are not archived.

Given these difficulties in augmenting operational test information, it is hard to imagine how confirmatory testing with such small sample sizes, in the form of significance testing even at the levels of statistical confidence used for this purpose (typically 80%), could be accomplished. Statistical theory does offer experimental designs that use relatively small sample sizes efficiently. However, constraints and uncontrolled events complicate the design (and subsequent analysis) and often require nonstandard techniques that would need substantial technical training to understand or develop. At the analysis stage, test events that resulted from uncontrolled circumstances may need to be removed, and the vast amount of collected data must be organized and summarized before statistical significance tests are carried out. Even if the major were comfortable using various exploratory data analytic and multivariate techniques for the QZ3 analysis, there is little time to determine whether a method can be used and, if not, to develop an alternative approach.

The decision rule for deciding whether QZ3 is ready for full-rate production generally does not accommodate use of information with respect to a conditional satisfaction of requirements, such as conditional on threat, type of engagement, environment, or other covariates of interest. It also does not indicate how to deal with partial satisfaction of requirements. For example, the evaluator is given no guidance on what to do if only various subsets of the 15 requirements for QZ3 are met in various environments or against various threats, nor was the contractor told how such tradeoffs would be made. The production decision must, of course, be made, so the evaluator must commit.

As part of QZ3's suitability study, it is also necessary to test the new software. In less than 2 years, a small staff must check a "representative" set of 1.5 percent of the lines of code (approximately 15,000) for errors. There are no precise rules about structuring the code in a way that would facilitate this operation.

Given the complexities and restrictions that are built into the current process (e.g., time, sample size, combination of relevant information, etc.), it is hard to even conceive of a series of tests with a rigorous statistical basis that could confirm that QZ3 was superior to QZ2. Yet with respect to the decision, the stakes are extremely high, immediately in terms of money and perhaps in the long run in terms of the best defense of the United States. The result is a problem that would require sophisticated statistical reasoning, coupled with insightful judgment of the real needs of the military user, as well as intimate knowledge of the system, its track record, and related systems being developed now or in the past. A tall order for our major.

The major, because of particular interest in statistical problems, took one course more than the 1-year undergraduate sequence of probability and statistics that is required. However, even 1-1/2 years of undergraduate training is not likely to provide an understanding of the distinctions among assuming an exponential distribution for the distribution of waiting times to system failures; or using the Weibull distribution to model these waiting times; or modeling failure times as a nonhomogeneous Poisson process, which might have important implications for how much time to test the system for suitability (see Chapter 7). The major also had a course on systems engineering at the undergraduate level, but it is not likely to have provided a full understanding of how to test sophisticated software for errors.

The above scenario describes many difficulties often encountered by those responsible for designing, carrying out, evaluating, and presenting results from operational tests on defense systems. Several, but not all, of the problems that are (implicitly) raised here are addressed in this report, but the key message is that the individual in charge of an operational test has an extremely difficult job, which is made even more difficult by the fact that defense system testing and evaluation has an essential statistical component.

Yet the hypothetical major should *not* be a statistician (or systems engineer) since the person responsible for test design and evaluation has to be knowledgeable about the system being tested, the military application, the management of a large enterprise, and various DoD rules and regulations governing testing and acquisition. Indeed, the first priority of operational testing should be to conduct a comprehensive series of tests under realistic and relevant conditions and environments; statisticians are generally not qualified to implement such tests. Without a well-designed, realistic test, any post-test statistical analysis, whether using standard or sophisticated techniques, can easily lead to incorrect decisions. It is important to emphasize that most of the statistical methods that the major would need to use are not immediate applications of standard formulas and procedures. Many of the problems that would occur—such as trend-free designs to accommodate player learning, analysis of non-orthogonal experimental designs, analysis of variance with missing values, censored observations, combination of information across different types of experiments, analysis of failures with non-exponential lifetime distributions with small sample sizes—are nonstandard statistical problems. It would not be reasonable to expect the major to possess this level of analytical expertise. It would be reasonable to expect the major to have this expertise available, either through use of civilian employees, consultants, or ties to academia. This level of expertise is available to the major through the national laboratories, federally financed research and development centers, and support contractors, but it is not used as frequently as would be desired: the quantity of expertise is limited, the major doesn't know how to get help, and the statisticians who are available are not knowledgeable about or are not interested in the system under test. These issues are discussed in Chapter 10.

Would more statistical expertise throughout the acquisition process result in superior decision making? If so, how should it be accomplished? What aspects of operational testing are likely to be improved through the application of state-of-the-art statistical techniques? These are some of the questions that we examine in this report. To complete the picture of why the answers to these questions are so important to defense system acquisition, we first examine the changes that have been occurring in the acquisition community that led to our study.

DYNAMICS IN THE ACQUISITION OF MILITARY SYSTEMS

The military systems under development and the development and testing process for military acquisition are continuing to change. Because of those changes, DoD will have to rethink the way that tests are designed, systems are evaluated, and, possibly, how the acquisition process is structured.[1] These changes highlight the importance of statistical expertise within DoD.

Decreased Testing Budgets The current DoD budget is much smaller than it was 10 years ago, and there have been substantial budget cuts for testing (see U.S. Department of Defense, 1997). (Decreased testing budgets have also reduced government developmental testing.) Consequently, tests are often smaller,

[1] Some of this is already happening. The 5000 series was updated in 1996 as part of DoD's acquisition reform effort (U.S. Department of Defense, 1996). These revised directives emphasized several major themes, including the importance of cross-functional teams, tailoring the acquisition process to better fit each system, empowering program managers and other acquisition professionals, and incorporating best commercial practices when appropriate. In addition, the Army has developed and implemented the Army Performance Improvement Criteria (APIC) based on the Malcolm Baldrige National Quality Award Criteria for Performance Excellence and the Presidential Quality Award Criteria (U.S. Department of Defense, 1998a). The goals for APIC include its use as: (1) a resource to guide planning, assessing, and training; (2) helping raise performance expectations and standards; and (3) implementing state-of-the-art business practices.

shorter, and must be executed with fewer prototypes. The Director, Operational Test and Evaluation (DOT&E) must by law determine the number of low-rate production items needed for test but this is done without clear, published guidelines. However, more sophisticated statistical methods can help to make the most effective use of whatever resources are available. When appropriate, methods for combining test data with information from other sources—including developmental test results, knowledge about similar systems, and modeling and simulation results—can be used to provide additional information for decision making. Also, statistical decision theory can be used to help determine cost-effective test budgets.

More Complicated Systems Today's military systems are growing more complex, as modern warfare requires greater effectiveness and flexibility in defense systems. Complicated systems mean more measures of performance and effectiveness, which increases the complexity of test design and test evaluation, which in turn requires sophisticated statistical analysis.

More Software-Intensive Systems The automation of system control, and greater use of information management, fault checking, and other software-supported features, has substantially increased the number of software-intensive systems and the size of the software code embedded in them. Nearly all ACAT I systems that have gone to full production in the last 3 years—systems that are estimated to cost more than $2.135 billion (in fiscal 1996 constant dollars)—have had a substantial software component. More software-intensive systems, therefore, require the latest techniques in software engineering, which in turn depends on up-to-date statistical techniques.

More Upgrades to Systems, "Evolutionary Procurement" As the development process—from concept formulation to full production—becomes more complicated and costly, more development programs have become upgrades of existing systems. For example, the Longbow Apache helicopter is a substantial upgrade to the original Apache helicopter, primarily through the addition of a radar system.

Evolutionary procurement is an acquisition concept that was developed in an attempt to cope with difficulties inherent in procuring very complex systems, especially systems with a significant software component. In evolutionary procurement, a system is procured in stages, with additional functionality and features provided at each stage. An evolutionary procurement cycle can be regarded as a series of rapid upgrades. More system upgrades require the use of archived information from both previous stages in the system's development and from related systems both for test design and the ability to appropriately combine that information.

Greater Interest in System Reliability, Availability, and Maintainability Many of the changes and trends just described (especially decreased defense budgets) and the potential for retaining systems in military forces for more years (extending lifetimes) have created greater interest in developing systems that have higher reliability and require less maintenance. This increased emphasis on reliability, availability, and maintainability emphasizes the importance of test design and evaluation that places greater weight on reliability considerations.

These changes create major challenges for operational testing to meet its goals to ensure the effectiveness and suitability of defense systems. This report identifies (see Chapters 5-8) a number of ways that statistical analysis and statistical methods can contribute to more cost-effective defense testing. As a result of this examination and related implications, the report also considers changes beyond the constraints of the existing acquisition process.

PANEL'S CHARGE AND SCOPE OF WORK

The panel was charged with examining the applicability of statistical methods to defense system testing and evaluation, particularly operational test design and evaluation. The statement of work reads:

> The panel would explore (1) developing measures of effectiveness, (2) designing operational tests and experiments with guidelines for determining the extent and types of testing required, (3) developing test and models that incorporate information from previous analyses in the acquisition process, and (4) representing and characterizing uncertainties in presenting results. In addition to making recommendations on how to improve operational testing under current requirements and criteria, the panel would also consider whether and to what extent technical criteria and organizational and legal requirements for testing and evaluation constrain optimal decision making.

The panel focused on statistical techniques in four primary areas: (1) assessment of reliability, availability, and maintainability; (2) use of modeling and simulation; (3) methods for software testing; and (4) use of experimental design techniques. We believe that our recommendations for improvements in these areas could be implemented immediately.

In the course of its work, however, the panel discovered that it was necessary to expand its charge. To provide the best advice to DoD, the panel examined the acquisition process as a whole and considered changes that might better support information gathering and, therefore, the development of more effective and suitable systems. The panel also examined the application of statistical science to operational testing, specifically, the process that leads to the decision whether to let a system proceed to full-rate production. The small sample sizes used in operational tests led the panel to consider whether and how developmental test activities (and operational tests for earlier systems) could be used to augment operational testing through combining information across experiments. Operational test design and evaluation is more efficient when based on information from earlier tests, including developmental tests. Developmental and earlier test results provide information on the problems a system might have, estimates of variances for determining the statistical power that a planned operational test would have, and which distributional assumptions about performance measures are likely to be valid.

Questions about the appropriate size of operational tests raise further issues: How are operational test budgets determined? What is the purpose of operational testing in defense system acquisition: that is, how is it established that a system passes or fails an operational test and what happens when a system fails? Recent trends in industry point to a procurement philosophy in which testing is viewed as an integral part of an acquisition process focused on designing and building quality into a product. In this view, end-of-the-line inspection is de-emphasized in favor of continuous information gathering to find and fix problems before a product is complete. The panel decided that it would be valuable in meeting its charge to consider whether these industry trends can be adapted to the DoD environment and if so, whether the result would be similar improvements in quality and cost-effectiveness to those observed in industry. As a result, our primary task naturally generalized to whether statistical principles could be used more effectively as an integral part of the broader defense acquisition process.

Statistics is more than a collection of individual techniques; it is a science concerned with the efficient use of information for effective decision making. Therefore, the panel examined whether the broader acquisition process was consistent with statistical principles that have been found useful in system development in other contexts. For example, the acquisition process, when attempting to take uncertainty in operational tests into consideration, makes use of significance testing as a key decision-making tool.[2] We detail the limitations of this approach and suggest additional methods to enhance decision making. The panel also examined what constraints exist in the entire acquisition process that limit the use of statistical methods or broad statistical principles in operational testing and in associated operations (including developmental testing, the development of requirements, and post-production testing). The panel acknowledges that the recommendations in this area involve organizational change. Implementation of these changes requires careful thought and planning and will take considerable time and effort. In the panel's view, the potential benefit is well worth the effort.

Many defense systems are extraordinarily complex and have idiosyncratic characteristics, at times using novel technologies in unique situations. It could be

[2] Significance testing, discussed in Chapters 5 and 6, is a method for assessing, in a statistically defensible way, whether an observed statistic related to some parameter is consistent with an assumed hypothetical value for that parameter.

premature to draw conclusions on the basis of the panel's in-depth examination of only a few systems. However, we believe that the systems we have studied are illustrative of best current practice. We are confident that our recommendations have applicability beyond the systems that we examined.

The defense testing community views statistics mainly as estimating means or percentages and of testing hypotheses. These are important and valuable tools, but they are often too simple to deal with important and complex issues involving enormous costs. Even when these tools are used, the expected costs and benefits must be evaluated for each competing strategy, and must be used in deciding what experimentation to do and how much to spend on it.

The panel has been told many times that the "statistical answer" is prohibitively expensive and so cannot be used. This often reflects severely reduced budgets determined without proper consideration of costs and benefits. The response should not be to disregard the statistical approach, but to use it either to evaluate what is the best that can be done under the given financial constraints, or to reset the constraints.

The complex problems arising in operational testing require sophisticated and well trained statisticians who understand how to develop methods appropriate for the underlying problems.

The report outlines the statistical methods that, both in the short- and long-term, will have a dramatic effect on the quality of defense systems acquired by DoD, and the costs of this acquisition. These include new approaches to software testing, new methods for test design for reliability, availability, and maintainability, methods for validation of modeling and simulation, and methods for obtaining more information from operational and other testing. This also includes a new view for the role of testing as part of military system development. In addition, the panel believes that a regular interaction between the statistical and defense testing communities must be established so that new statistical techniques and applications will be developed specifically tailored to meet the challenges unique to defense testing and evaluation.

2

Operational Testing and System Acquisition

The main objective of defense acquisition is to obtain quality weapon systems, in a cost-effective and timely manner, that meet an operational need. Two broad types of testing are used to assist in this goal. Developmental testing covers a wide range that includes component testing, modeling and simulation, and engineering systems testing. Developmental testing presents the first opportunity to measure the performance and effectiveness of the system against the criteria developed in the analysis of alternatives (formerly cost and operational effectiveness analysis). Operational testing and evaluation examine the performance of a fully integrated set of systems, including the subject system, under realistic operating environments, perhaps for the first and only time before implementation. It is the process that the U.S. Department of Defense (DoD) uses to assess whether a weapon system actually meets it planned capability before deciding whether to begin full-rate production. Operational testing and evaluation is an independent assessment of whether a system is effective and suitable for its specified use; this independent assessment is vitally important to acquisition.

The first two sections of this chapter briefly describe operational testing and evaluation and discuss the differing perspectives of the parties involved in defense acquisition. The panel then considers how well the current operational testing and evaluation structure supports the effective acquisition of defense systems and how modern statistical practices, combined with fundamental changes in the current paradigm (the topic of Chapter 3), could result in substantial improvements.

A SHORT HISTORY OF OPERATIONAL TESTING

The Office of the Director, Operational Test and Evaluation (DOT&E) in DoD was established in 1983. Before then, there were several organizations in the Office of the Secretary of Defense (OSD) involved in one or more aspects of operational testing; none had the responsibility of managing or monitoring operational testing and evaluation as a whole. In addition, each service was generally responsible for planning, conducting, evaluating, and reporting its own operational tests. With limited DoD oversight, each service developed unique procedures and regulations for operational testing and evaluation. They differed in the flexibility of operational test guidelines, in the extent to which test results were used iteratively to improve weapon systems, and in the degree to which operational test agencies were subordinate to organizations responsible for system development.

Without any unified structure or policy for operational testing and evaluation, DoD began to be strongly criticized in the late 1960s. Testing agencies often suffered from high turnover and were often subordinate to organizations responsible for new system development. There were excessive layers of bureaucracy separating the test agencies from the chiefs of staff (direct reporting of test results to decision makers was all but nonexistent), and operational commands generally set aside insufficient funding, personnel, or facilities to accomplish adequate testing and evaluation. (For detailed historical information prior to 1970, see Blue Ribbon Defense Panel, 1970.)

In 1971 Congress enacted Public Law 92-156 which, among other things, required DoD to begin reporting operational test results to Congress. The Deputy Secretary of Defense directed the military services to designate field commands, independent of the system developers and the eventual users, to be responsible for planning, conducting, and evaluating operational tests (U.S. General Accounting Office, 1986). These agencies were instructed to report directly to the appropriate chief of staff.

The next major change in DoD came in 1983, with the congressionally established DOT&E. DOT&E is headed by a director, who is the principal department adviser to the Secretary of Defense on operational testing and evaluation and is responsible for prescribing policies and procedures for its conduct. By law, a major defense acquisition program may not proceed beyond low-rate initial production until initial operational test and evaluation is completed. The law requires that the director shall analyze the results and prepare a report stating the opinion of the director as to:

(A) whether the test and evaluation performed were adequate; and

(B) whether the results of such test and evaluation confirm that the items or components actually tested are effective and suitable for combat.

This provision has led many officials to view operational testing and evaluation as a final series of test events; the system either passes and is produced or fails and is either returned to development or terminated.

A 1987 General Accounting Office evaluation of the effectiveness of DOT&E found several important deficiencies (U.S. General Accounting Office, 1987b). First, DOT&E was criticized for failing to sufficiently monitor the services: on-site observations of operational tests were inadequate to ensure compliance with DoD testing policy. Second, DOT&E's independent assessments of operational test results relied too heavily on the services' own test reports. DOT&E is required to conduct independent analyses using the actual test data and not the services' reports; however, in many cases DOT&E reports were copied verbatim from service documents. Third, DOT&E failed to maintain accurate records of its principal activities. DOT&E officials concurred with all of the General Accounting Office findings, citing understaffing as the major problem.

DOT&E had attempted to upgrade the conduct of operational testing and evaluation by concentrating its early efforts on improving test planning. The General Accounting Office reported that these efforts were largely successful (U.S. General Accounting Office, 1987b). Traditionally, the services had not given much emphasis to preparing test plans; with increased DOT&E oversight, GAO evaluations support the assertion that the quality of test design was improved. A recent GAO report (U.S. General Accounting Office, 1997) has strongly endorsed the value of the different functions of independent review, approval, and assessment of the various stages of operational test design and evaluation provided by DOT&E.

Since the establishment of DOT&E, the conduct of operational testing has undergone several minor changes. With the assistance of DOT&E, two successive Secretaries of Defense rewrote the directives that govern the acquisition of major systems (including Directive 5000.1, *Defense Acquisition*, and Directive 5000.2, *Major System Acquisition Policies and Procedures*). These documents acknowledge and emphasize the importance of testing and evaluation in the acquisition cycle. In the past 10 years, DOT&E has eliminated many of the deficiencies noted in the 1987 evaluation: for example, DOT&E assessments now rely less heavily on service test reports.

Congress has imposed some additional changes on the acquisition process. Operational testing and evaluation of major defense acquisition programs may not be conducted without formal approval from DOT&E. In addition, vulnerability and lethality testing must be completed under the oversight of DOT&E.

Currently, DOT&E has identified several service initiatives as priorities, including: (1) earlier involvement of operational testers in the acquisition process through the use of early operational assessments and integrated product teams; (2) more effective use of modeling and simulation before operational testing and evaluation and as part of early operational assessments; (3) combining developmental and operational tests and using information from all viable sources to make operational testing and evaluation more effective; (4) operational command partnerships between testers and users; and (5) combining testing and training (U.S. Department of Defense, 1997). The panel supports these initiatives and encourages the services, along with DOT&E, to vigorously pursue them.

THE DECISION CONTEXT: ROLE OF INCENTIVES

To understand the use of testing as part of system development, it is important to have a general understanding of the acquisition process. We very briefly describe the different milestones below; for a more detailed description, see *Statistical Methods for Testing and Evaluating Defense Systems* (National Research Council, 1995:Appendix A).

The Acquisition Process

The procurement of a major weapon system follows a series of event-based decisions called milestones. At each milestone, a set of criteria must be met if the next phase of acquisition is to proceed; see Figure 2-1. For all acquisition programs, test planning is supposed to begin very early in the process and involve both the developmental and operational testers. Both should prepare the Test and Evaluation Master Plan, which is a requirement for all acquisition programs. For ACAT I programs, this plan must be approved by DOT&E and the Director, Test, System Engineering, and Evaluation (the OSD office with oversight of developmental testing, which reports to the Defense Acquisition Executive).

Milestone II follows the demonstration and validation phase. At this point the milestone decision authority, in consultation with DOT&E, determines the low-rate initial production quantity to be procured before initial operational testing is completed. The number of prototypes required for operational testing are also specified by the service test agencies and by DOT&E for ACAT I programs.

After obtaining milestone II approval, the system enters the engineering and manufacturing development phase. One of the testing objectives during this phase is to demonstrate that the system satisfies the mission need and meets contract specifications and minimum operational performance requirements. Based on what has been outlined in the Test and Evaluation Master Plan, the testers and evaluators prepare detailed planning documents that are subjected to OSD review. However, resource constraints may prevent certain system characteristics from being evaluated. In such cases, the testers identify what they can accomplish given the constraints.

The operational test results are interpreted by many separate agencies, including the relevant service's operational test agency and DOT&E. DOT&E prepares independent operational test and evaluation reports for the Defense Acquisition Board, the Secretary of Defense, and Congress.

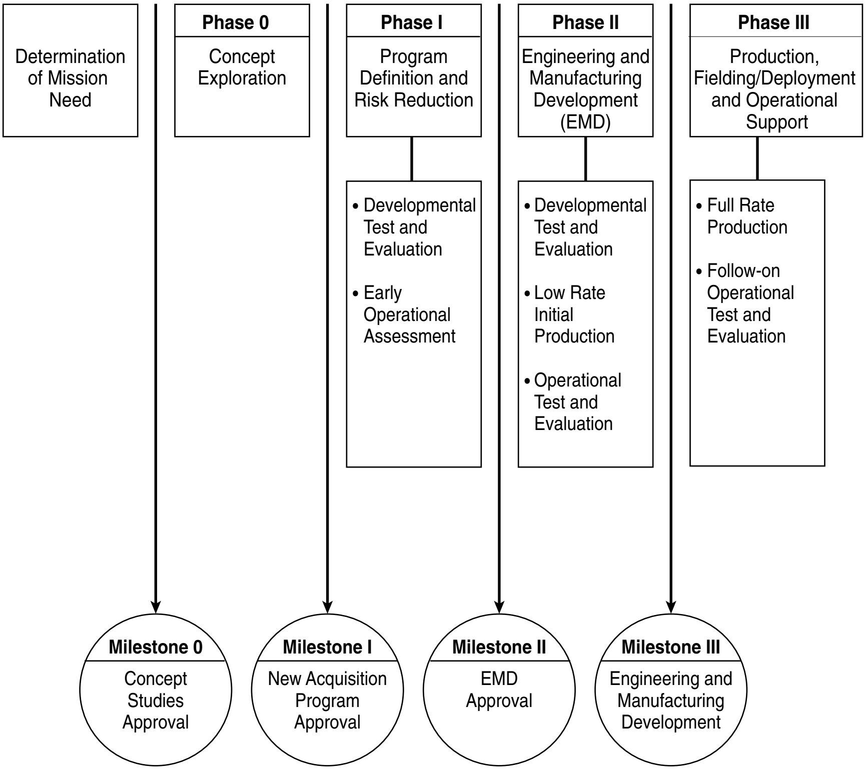
Determination of Mission Need
Phase 0
Concept Exploration
Phase I
Program Definition and Risk Reduction
• Developmental Test and Evaluation
• Early Operational Assessment
Phase II
Engineering and Manufacturing Development (EMD)
• Developmental Test and Evaluation
• Low Rate Initial Production
• Operational Test and Evaluation
Phase III
Production, Fielding/Deployment and Operational Support
• Full Rate Production
• Follow-on Operational Test and Evaluation
Milestone 0
Concept Studies Approval
Milestone I
New Acquisition Program Approval
Milestone II
EMD Approval
Milestone III
Engineering and Manufacturing Development

FIGURE 2-1 DoD Acquisition Process. SOURCE: U.S. General Accounting Office (1997:35).

In making the milestone III recommendation to initiate full-rate production for ACAT I systems, the Defense Acquisition Board considers among other things (e.g., costs or changes in a threat) the developmental test results and the reports from DOT&E and the service test organizations. If the Undersecretary of Defense for Acquisition and Technology approves full-rate production, the full-rate production contracts are then awarded, consistent with any DoD or congressional restrictions. Follow-on operational testing is performed during the early stages of the production phase to monitor system performance and quality.

Incentives in the Acquisition Process

Everyone directly or indirectly involved in defense testing and evaluation faces constraints and rules that give each a unique perspective and result in different incentives. While neither good nor bad, these different incentives have a profound impact on the test and evaluation decision-making process (e.g., budgets, schedules, access to information). Therefore, when examining how these decisions are made, it is important to understand and take into consideration each participant's role in that process.

For example, a scoring conference is a group of about six to eight people—including testers, evaluators, and representatives of the program manager—who review information about possible test failures and determine whether to count such events as failures or exclude them from the regular part of the test evaluation. Each member brings a different perspective to the table and may have an interest in the test results that are unrelated to the value of the system under consideration. These different perspectives and motivations must be examined and considered; if they are not, improving the statistical methods used in testing and evaluation are likely to have no or limited effect.

As studies of organizational behavior have long demonstrated, the goals of a large organization need not be exactly concordant with the smaller organizational units they comprise. It is entirely possible to confront a situation in which individually attractive behavior can lead to collectively undesirable consequences. In the rest of this section we describe the various actors in defense testing and acquisition in a stylized and somewhat oversimplified manner by drawing a slightly exaggerated representation of conscientious individuals' best efforts on behalf of their country and their positions. In describing the generation and flow of information and how the incentives of those involved in the process affect decisions, our goal is clarity and not offense. We realize that there are oversimplifications in this description, but we are convinced that the points made are fully relevant to the current process of DoD acquisition. At a minimum, the arguments offered demonstrate that incentives, if ignored or not balanced or otherwise taken into consideration, can have a substantial impact on the acquisition process.

The primary organizations involved in operational testing and evaluation are DoD and the military services, contracting firms, legislators, and the news media; many of them have diverse parts, such as test personnel, the program managers and their staffs and immediate supervisors, the DOT&E staff and managers, the project managers for the contractors, members of Congress and their staffs, General Accounting Office personnel, and finally, the U.S. public. For the purpose of this stylized exercise, we limit our scope to OSD management, the program staff, test personnel, DOT&E, the contractor, the public, legislators (which broadly includes Congressional members, their staffs, and General Accounting Office auditors), and the news media. This and other such simplifications allow us to provide a succinct and simplified description that gives some insight into the effects of each participant's objectives and incentives.

Ultimately, national security is being provided for the U.S. population. Using game-theoretic terminology, they are the "principals" while the legislators are an "agent." Citizens delegate DoD oversight to Congress; therefore, in the congressional-DoD relationship, the legislators are the principal and DoD the agent. It is important to note that this vernacular should not be interpreted to mean that legislators have "better" or "purer" objectives than DoD. The only group with presumably pure objectives is the U.S. population.

Contracts determine the nature of relationships between principals and agents. No agent will ever act exactly as the principal desires unless (1) the agent can be perfectly monitored or (2) the agent acquires the entire stake or interest of the principal, thereby becoming the principal. In the context of DoD acquisition, neither one of these occurs. So the contract under which a "player" functions will almost certainly not lead the player to act exactly as the principal would want.

Legislators are concerned with issues that can affect their reelection, such as employment in their districts. Program managers receive promotions when they can take their program to the next milestone or full-rate production. Consequently, they have a strong incentive to meet set schedules. System deficiencies are typically viewed as problems that will be fixed in the future.[1] The tester is, ideally, an independent and objective professional who is not affected by the pressures built into the process. Such independence may be compromised, however, if a tester is influenced in some way by the program staff. When this happens, the tester attempts to craft a test that (partially) reflects the program staff's perspective.

A rather dramatic tension exists between the program staff and their superiors and legislators. The program manager must comply with regulatory and reporting requirements, but the goal is to have the program approved for full-rate production, and so program staff have an incentive not to convey all known data about the system to legislators if negative information could delay or perhaps even terminate the program. (This incentive to conceal potentially unfavorable data was recognized by Congress in the early 1980s and was a primary motivation for the creation of DOT&E.) At the same time, the desire to avoid negative publicity creates an incentive for additional regulation and oversight of the acquisition process.

There are legitimate reasons, however, for a program staff to be concerned with complete disclosure to legislators. For example, a system may have a problem that the program staff knows can be fixed without jeopardizing the overall program. Legislators, however, may find these assurances to be unverifiable and not believable. For legislators, the risk of supporting a program with a

[1] A recent report (U.S. General Accounting Office, 1997) provides another perspective on the interaction between acquisition and test officials: "In reviews of individual weapon systems, we have consistently found that testing and evaluation is generally viewed by the acquisition community as a requirement imposed by outsiders rather than a management tool to identify, evaluate, and reduce risks, and therefore a means to more successful programs. Developers are frustrated by the delays and expense imposed on their programs by what they perceive as overzealous testers."

major flaw is greater than the program staff's risk: if a program is flawed but funding continues, it is Congress and OSD management that will be held responsible by the public. Yet if the program becomes a great success, DoD will enjoy the accolades.

Another source of tension lies in the fact that participants in the acquisition process have different amounts of, access to, and abilities to process information. A reasonable ranking from "least informed/least capable to process" to "most informed/most capable to process" would be: public, news media, legislators, DOT&E, DoD management, contractor, program staff. Testers have specific information that makes it difficult to assess where they fit in the ranking. The news media often report only on weapon system deficiencies. Although these reports may be part of a bigger story that cannot be told without compromising national security, people are naturally upset at this apparent misuse of public funds. Members of Congress know that only a part of the information has been made public; nevertheless, they still tend to be responsive to a public outcry because of reelection concerns.

Operational testing and evaluation provides the program staff with valuable information and is the means by which Congress determines (through DOT&E certification) that it is getting the product that was promised. Even though operational testing and evaluation has a demonstrated value, the program staff commits limited resources to it because of overall program budget constraints, choosing instead to reserve the majority of available funds to program development. In fact, those programs with serious problems may have the smallest testing budgets since more funds will likely be allocated for development, leaving less for operational testing and evaluation.

How does this system of conflicting incentives operate? For one example, consider the potential go/no go[2] decision that is associated with the current milestone system and operational testing and evaluation, particularly the negotiations between the program staff and the contractor. If there were no such decision point, we conjecture that the contractor would be less focused on research and development, more willing to let problems go unresolved, and generally less concerned about job performance. If so, then the go/no go decision plays a role in advancing systems to production (see next section). This incentive structure also explains why it is common for system requirements to be aggressive or optimistic; the goal is to get the system approved for development. One can also use the above considerations to show that additional oversight could result in less testing, thereby making the acquisition process worse (Gaier and Marshall, 1998). Therefore, while this game theoretic framework is oversimplified, it might, for

[2] In this report, we use the phrase "pass/fail" to mean "produce/terminate": a decision is made to either proceed to full-rate production or terminate the program. We use the phrase "go/no go" to mean "produce/delay decision/terminate": a decision is made to proceed to full-rate production, conduct further testing or development, or terminate the program.

example, be able to assist Congress in examining potential legislation to see if it might have unintended consequences. Further efforts to expand this approach to represent more of the complexity of decision making in acquisition could have real benefits.

Recommendations

The above discussion suggests how incongruent and conflicting incentives may have a negative effect on defense acquisition, and specifically on testing and evaluation. Although simple models of portions of the acquisition process are instructive to develop and examine and should be further pursued (see, e.g., Gaier and Marshall, 1998), the number of participants and the complexity of their interaction poses too difficult a challenge to permit elaborate modeling. Therefore, recommendations related to individual participants or steps in the process that seem sensible when that person or step is examined in isolation run the real risk of ignoring important interactions. This consideration led the panel to be cautious in its recommendations related to incentives and information flow. Overall, DoD should consider the complex, real situation when developing and assessing proposals for acquisition reform and should strive for structures in which the incentives of the participants are as congruent as possible.

There are two areas affected by this interplay of incentives that are easy to conceptualize and are in need of improvement: determining operational requirements and allocating resources. As noted above, there is a strong incentive to be optimistic with system requirements to ensure program initiation approval. The result can be parameters that are untestable or requirements that cannot be met. If operational testers reviewed the Operational Requirements Document, it would help build integrity and quality into the process from the start. It might also discipline the test community to focus on the operational needs as expressed in the Operational Requirements Document and not expand "test requirements" with measures of performance that are included in other documents.

Recommendation 2.1: The Department of Defense and the military services should provide a role for operational test personnel in the process of establishing verifiable, quantifiable, and meaningful operational requirements. Although the military operators have the final responsibility for establishing operational requirements, the Operational Requirements Document would benefit from consultation with and input from test personnel, the Director, Operational Test and Evaluation, and the operational test agency in the originating service. This consultation will ensure that requirements are stated in ways that promote their assessment.

The federal government is often criticized for mandating work to the states but not providing the necessary funds. As part of its oversight function, DOT&E can order that more tests be performed, but with budget constraints affecting all of DoD, the services cannot be expected to have the additional resources needed to cover the cost. A reserve fund would ensure that an appropriate number of tests are performed when necessary. Additional tests may show serious system flaws that require a return to development, something a program staff clearly wants to avoid. If program managers know that DOT&E will be able not only to mandate additional tests, but will have access to a source of funds to make sure they are actually implemented, they might be encouraged to undertake more complete operational testing at an earlier stage in the process. Finally, just the use of these funds should be seen as an indication of a potentially substantial problem either with how the system is being tested or even about the worthiness of the system itself.

> **Recommendation 2.2: The Director, Operational Test and Evaluation, subject to the approval of the Secretary of Defense on a case-by-case basis, should have access to a portion of the military department's acquisition funding reserves (being set up as a result of the first quadrennial defense review) to augment operational tests for selected weapon systems.**

PROBLEMATIC CHARACTERISTICS OF THE CURRENT SYSTEM

In addition to the problems posed by the incentives in operational testing and evaluation, there are several other characteristics of the acquisition system that must ultimately be changed if operational testing and evaluation and acquisition are to meet their greatest potential.

Operational Testing as a Go/No Go Decision

The results of operational testing are critical inputs for the decision makers who will determine the future of the system under test. Significance testing is sometimes used to give a rigorous statistical grounding to this decision process: in those instances, if a system fails a significance test for a key operational requirement or for a large collection of other requirements, the milestone decision can be portrayed as resulting in the termination of the system. (For a detailed discussion of the potential problems in using significance tests, see Chapter 6.) However, the actual decision process is much more continuous than the idealized milestone system would indicate, as it should be in light of the large investment in system development.

It is extremely unusual for a defense system to be terminated. Because of the momentum built up during the many years of developmental activity (and because perfection is a process and not a state), a system must convincingly fail one or more operational tests or have other problems to be terminated. The panel is aware of only four ACAT I systems in the last 12 years whose failure during operational testing has been believed by many (though not all) observers to have been a primary element in its subsequent termination (DIVAD]Sergeant York], ASPJ, AQUILA, and ADATS). As a result, systems typically must undergo corrective action when they fail operational test and are later retested.

Clearly, a system redesign should be done when it is relatively straightforward and cost-effective (i.e., modest in scope and expense), or when there are no alternatives to providing the operational requirements. However, when this is not the case and termination should be considered, there are strong disincentives to do so in the current system. There are very few people with decision authority in the acquisition community who would advocate terminating a marginal system in their own areas of responsibility. Therefore, if the choice is between termination and redevelopment, a deficient system most likely will be returned to development.

There is the serious related problem that results from the costs of retrofitting a system because its deficiencies were found after full-rate production had begun. Postproduction redesign is particularly troublesome with software-intensive systems. As more systems incorporate extensive software components, it is becoming more common for software problems to arise during operational tests. These problems are often erroneously considered to be easier to fix than hardware problems; as a result, decision makers are inclined to pass a system with the expectation that any software errors will be subsequently solved in parallel with full-rate production. This assumption has caused problems because software failures are often very serious. (Software test issues are discussed in Chapter 8.)

Constraints on Statistics from the Current Application of Test to Military System Development

There are many specific problems presented by the need to effectively carry out operational testing and evaluation that can be substantially improved by state-of-the-art statistical practices. However, any improvements gained through the application of modern statistical practices (detailed in Chapters 5-9) are limited by the current structure of operational testing within military system development.

Because of both limited statistical modeling expertise and the fact that developmental test data, absent statistical modeling, are typically not relevant to evaluation of a system's operational readiness, it is not typical for data other than operational test data (except for pooling data for assessing reliability, availability, or maintainability) to be used to assess whether a system is ready for full-rate production. Furthermore, information from tests on related systems and from developmental test, including information on test circumstances, are not available to operational test evaluators since there is no well-maintained repository of data on test and field performance. This lack of information restricts the use of modern statistical techniques for effective test design when designing operational tests and for the combination of information at the evaluation stage. As a result, test efficiency is diminished because statistical practices cannot be effectively used and test costs are increased.

A lack of continuous test and evaluation of systems throughout development with respect to their operational performance—which would result from developmental testing having more operational realism or through the use of small, early operational tests—also hampers test design and limits additional possibilities for combining information. Finally, systems are at times permitted to change during operational testing, especially software systems. All of these practices restrict opportunities for full use of statistical methods. (Chapter 3 presents more detail on these information and statistical practice constraints.)

How Much Testing Is Enough?

The 1983 legislation effectively mandates that a decision to proceed beyond low-rate initial production must be based on operational testing that is adequate and that confirms the system's effectiveness and suitability for its intended uses. Assessing whether a test is adequate and has confirmatory power leads naturally to one question: How much testing is enough? We examine this question from a statistical perspective. (See Chapter 5 for some of the decision-theoretic considerations.)

As noted by Seglie (1992), the question has several dimensions. In it simplest form, "how much testing is enough" might be regarded as a relatively straightforward exercise in determining sample size requirements. However, if operational testing is viewed as a pass/fail step in the milestone process, then the costs and benefits of this decision must be considered. As Vardeman (1992) observes: "even in apparently simple situations, producing an honest answer to the question requires hard work and typically involves genuinely subtle considerations."

Significance-test-based statistical justifications result in sample sizes for operational tests that often exceed credible resource limits. For example, consider a prospective missile system for which the total projected procurement would be 1,000 missiles. Assume that the missile should come within lethal range of a primary target at least 80 percent of the time. In order to estimate the missile's ability to deliver a warhead within lethal range to within 5 percentage points with 90 percent confidence, approximately 148 missiles would have to be fired in destructive testing (174 would result from ignoring the finite population correction) (see, e.g., Cochran, 1977:75-76). A proposed operational test design in which 15 percent of the projected arsenal would be consumed in live-fire testing would be justifiably challenged as an inappropriate allocation of military acquisition resources. This problem is even more complicated for defense systems that need to be evaluated in many different operating scenarios, which effectively reduces the sample size in any given scenario.

The resource constraints on operational tests come in direct conflict with the information demands and test sample sizes that are required for rigorous statistically supported confirmation of system effectiveness and suitability. The panel recognizes this places DOT&E in a difficult position when it attempts to carry out its work as mandated by Congress.

Although supportable certification of system effectiveness and suitability can be and is made on the basis of the information (and judgment) generally available at the conclusion of operational testing, operational testing can also perform a different but extremely valuable function. Operational testing produces information about system performance in settings that come closer to real use than any other testing activity, and therefore operational testing uniquely informs about problems in system design and limitations in system performance from an operational perspective. It has long been recognized that operational testing involves a "system" that includes not only the prototype piece of hardware, but also the human operators, the training of those operators, and the military tactics and doctrine used in the weapon's deployment. Thus, a system's failure to perform in an effective or suitable manner may be attributable to poorly trained operators or inappropriate tactics. It is important to understand the causes of ineffective or unsuitable performance in real use, which can best be accomplished with operational test and evaluation. Elements of the planned operational tests that the panel reviewed at Fort Hunter Liggett were extremely useful for understanding how the Longbow Apache helicopter would function with realistic challenges, typical users, in day and night, and under different scenarios of attack. The panel is not arguing against use of operational testing. Rather, the panel is arguing that in a new view of operational testing and evaluation as part of defense system development, operational tests might play an expanded and more informative role in evaluating new defense systems.

Conclusions

The panel believes that neither the U.S. public, Congress, nor DoD is well served by perpetuating the fiction that, if done properly, operational testing will always provide sufficient information to make definitive statistical assessments of an individual system's effectiveness and suitability.

> **Conclusion 2.1: For many defense systems, the current operational testing paradigm restricts the application of statistical techniques and thereby reduces their potential benefits by preventing the integration of all available and relevant information for use in planning and carrying out tests and in making production decisions.**

Conclusion 2.2: The operational test and evaluation requirement, stated in law, that the Director, Operational Test and Evaluation certify that a system is operationally effective and suitable often cannot be supported solely by the use of standard statistical measures of confidence for complex defense systems with reasonable amounts of testing resources.

Conclusion 2.3: Operational testing performs a unique and valuable function by providing information on the integration of user, user support (e.g., training and doctrine), and equipment in a quasirealistic operational environment.

In remarks to a symposium of the test and evaluation community, Paul G. Kaminski, former Under Secretary of Defense for Acquisition and Technology, stated his belief that "a cultural change is necessary . . . one that can only begin by reexamining the fundamental role of test and evaluation in the acquisition of new military capabilities" (Kaminski, 1995). Similar remarks suggest that others share the panel's view that operational testing and its role as part of system development should be reconfigured to increase its effectiveness and efficiency in producing information about prospective military systems.

The panel believes that substantial advances can be realized by modifying the current defense acquisition and operational testing paradigm by approaching operational testing as an information-gathering activity. Although we do not offer a complete blueprint for reorganizing defense testing as part of system development, we believe we can contribute to ongoing discussions about the role of test and evaluation in defense acquisition. Chapter 3 describes improvements that the panel believes could result from moving toward a new paradigm in which modern statistical methods support a quality-focused acquisition process. The approach we advocate is consonant with recent trends both in industry and within DoD.

3

A New Paradigm for Testing and Evaluation in Defense Acquisition

The message of the preceding chapter, simply stated, is that the current paradigm for how operational testing is used as part of defense system development is not coherent. In an increasingly complex environment, the current paradigm does not address the goals originally specified for operational testing and evaluation. The incoherence of the current paradigm stems from three fundamental difficulties. First, the use of operational testing and evaluation as a final test event prevents the identification of operational problems early in system development. Second, while this testing event is often portrayed in the milestone system as supporting a decision to pass or fail a defense system in development, systems are almost never terminated as a result of tests (nor should they be). When doing destructive testing or testing complex systems with high unit costs, it is often infeasible to have a big enough operational test to make clear pass/fail decisions with any reasonable level of statistical confidence. Third, current limitations on using all sources of relevant information also reduce the efficiency of testing and evaluation; policies restricting such use or limitations in expertise increase test costs.

These conclusions have led us to propose a new paradigm to operational testing and its role in system development. We are not the first to come to this conclusion. As discussed below, we have seen encouraging signs at the highest levels of the testing and acquisition community of an emerging new view of the role of operational testing and evaluation in the acquisition process. The panel has focused its deliberations around the question: How should operational testing and evaluation work to fully realize potential gains in efficiency and effectiveness through the improved use of statistical methods? This chapter presents a new view of the process of system development in which the role of operational testing and evaluation as a late series of tests for go/no go decisions is broadened to a role of continuous information gathering to support the development of effective and suitable systems. We do not presume to offer a blueprint for such a process. Instead, we discuss key ideas and principles for such a paradigm, and draw conclusions with respect to improvements that can be made within that paradigm, trusting details of their adoption and implementation to the defense acquisition community.

ESSENTIAL FEATURES OF A NEW PARADIGM

A new and more effective paradigm for the use of testing and evaluation as part of defense system development should possess four essential characteristics:

1. *A continuous process of information gathering and decision making in which operational testing and evaluation plays an integral role.* This orientation is consistent with contemporary trends in industrial practice, deemphasizing "inspecting defects out" in place of integrated development processes and "building quality into" a product.
2. *An environment of more systematic data collection, archival, and documentation.* All sources of data and information should be archived and clearly documented and the process made consistent across the military services so that development teams are able to identify and learn from all relevant studies and findings. To create this environment, DoD could institute multiservice operational test and evaluation standards similar to those contained in ISO 9000 (see Appendix D).
3. *The use of efficient statistical methods for decision making.* An environment of continuous assessment and systematic data collection would facilitate the use of efficient statistical methods, including decision-theoretic methods, yet would still allow the decision maker to be held accountable for judgments on system procurement.
4. *Reductions in the life-cycle costs of acquiring and maintaining new military capabilities.* Integrating operational testing and evaluation in a process involving more systematic data collection, analysis, and documentation plus continuous assessment (and improvement) means that problems of consequence will be discovered earlier in the development process, when they are easier and cheaper to solve. A system would therefore not be likely to enter the final, confirmatory stages of operational testing until it is clearly ready for that stage, and, therefore, also more likely to be ready for full-rate production. The production of higher quality and more reliable systems will reduce the amount of logistic support required to maintain the system in the field, a major contributor to total cost over the lifetime of the system.

We acknowledge that the ideas in our "new paradigm" are not completely new with respect to DoD test and acquisition. These ideas have often individually been suggested, and they are applied in various specific, narrow instances throughout DoD test and acquisition. What is new and important is the consistent, widespread, institutionalized use of the processes discussed here, with full recognition of their importance and value.

Our proposed new paradigm of testing as part of system development should not change the fundamental independent advisory and assessment role of DOT&E. There is still need for an independent assessment of the test and evaluation process as part of defense procurement. In fact, the many opportunities in this new paradigm for collective decision making—including establishing requirements, setting test budgets, evaluation of test results, and setting up the data archive—create increased responsibilities for DOT&E. DOT&E officials would continue to provide an independent evaluation of all operational test results. A paradigm shift to focus the test community on providing as much information as possible (as early as possible) on operational effectiveness and suitability does not preclude the independence and accountability of DOT&E. The importance of this independent advisory and assessment role was strongly supported in a recent report (U.S. General Accounting Office, 1997) that considered this independence key to the effectiveness of DOT&E.

Such a shift is also consistent with changes that are evolving in the role of government in defense system development. Changes in technology, including DoD's greater use of technology developed by industry, congressional pressure to reduce the size of the DoD acquisition workforce, DoD's use of integrated product teams that include industry members, and increasingly popular views that industry should have a greater role in such areas as DoD systems logistics support imply a larger role for the industrial sector in defense acquisition in general. In many systems developmental and support areas, DoD's prime contractors, subcontractors, and support contractors will become the sources for much of the data needed to design test and evaluation programs. In addition, contractors could be specifically asked for information, explanations, or analysis during the course of an operational test. To preserve the desired and desirable independence, contractors should not be involved in the testing itself, but might be able to provide worthwhile assistance both in planning tests and in the interpretation of test results. Congress should consider lifting the constraint on contractors' participation in operational test and evaluation activities and permit or encourage DoD to propose new guidelines for the limited use of contractor data and personnel in the operational test process. The independent evaluation role can and should remain in DoD and DOT&E.

In the sections that follow we describe a new paradigm for operational testing and evaluation as an integral part of defense acquisition. We begin by noting that discrepancies between operational testing on paper and in practice suggest how the role of operational testing might be constructively redefined. Next, we discuss the consonance of such a redefinition with successful ideas from private industrial practice. Finally, we lay out the implications for specific components of the new paradigm for DoD acquisition and testing, as well as some recent and promising developments in test and evaluation.

INSIGHTS FROM CURRENT PRACTICE

The goal of operational testing as stated in the authorizing legislation is "to confirm the effectiveness and suitability of the system." This has been interpreted by some members of the test and acquisition community, as well as its critics, to mean that operational testing is intended to provide the basis for deciding between procurement and cancellation of a prospective defense system. In actual use, however, the third milestone decision regarding operational test and readiness for production is almost never a simple dichotomy between either canceling a program or moving into deployment with no further development or evolution of the system design.

The manner in which operational test results are used in practice suggests that the most valuable functions of operational testing are to identify deficiencies in the performance of a system, to characterize the conditions under which deficiencies are likely to occur, and to isolate and remove the causes of such deficiencies. Thus, a testing program can most usefully focus on such questions as: "What needs to be fixed?" "How likely is it that the needed fixes will be technologically feasible and affordable?" "What sequence of tests can be performed that will show the weaknesses, as well as the strengths, of a system?" These activities can and should be undertaken throughout system development. If operational testing is performed to determine whether a system needs further development, then it is extremely wasteful to conduct a very expensive test after the system design is final. It would be much more efficient to create test policies and procedures that identify operational deficiencies earlier in the development process.

Since it is well recognized in DoD that the implied objective of operational test in DoD acquisition to provide statistical certification of whether a system is operationally effective and suitable is generally unrealistic, some more specific alternative objective is implicit. We believe that the existing implicit objective and the new paradigm we propose have much in common. Operational testing should be viewed less as a final exam used before deciding whether to approve full-rate procurement. Instead, its overriding objective, using testing and evaluation with operational criteria, should be to provide decision makers with timely necessary information in the form most useful to make effective decisions regarding the need for further development.

LEARNING FROM INDUSTRIAL PRACTICES

It is now generally acknowledged that quality cannot be "inspected" into a product. Product testing is often performed too late in the process to change deficiencies in a timely and cost-efficient manner. Rather, quality must be designed into any complex product at virtually every stage of its development. This is accomplished by creating acquisition processes that make use of continuous monitoring and testing.

Applicability to Defense Acquisition

We acknowledge that there is a danger in drawing analogies between developing new products in industrial and other private- and public-sector areas and the development of new defense systems. There is no question that certain aspects of defense acquisition are unique. For example, a new military technology might be used in an emergency before it is completely ready for routine operation. A defense system may also serve deterrent purposes even if it is only partly successful. On the other side, commercial concerns often must commit to a product's development to stay in business, so a pass/fail decision may be less relevant in commercial applications. The development of defense systems involves security and classification issues (though industry also values secrecy), and requires evaluation that is independent of the contractor. Finally, many military systems, especially ACAT I systems, and their associated tests are much more complicated and costly than for nonmilitary systems. So in examining industrial practice for developing new products, one must be cautious about transferring lessons learned to the unique situation of defense acquisition.

However, there are also substantial similarities. A recent report (U.S. General Accounting Office, 1996) pointed out that, in an industrial manufacturing process for a product that had both military and nonmilitary customers, the production line for the nonmilitary customer used automation and process control throughout production. This reduced the need—as far as the non-military customers were concerned—for end product testing. However, the military customer required a potentially wasteful 100 percent end product testing.

In spite of conventional wisdom, all of the decision makers for an industrial corporation do not have a common profit objective. Just as in the military defense system development process, the incentives and performance measures of design engineers, testers, manufacturing and production personnel, and marketers are often disparate. Some of the procedures developed by industry to account for and overcome these disparate interests are relevant to DoD.

Quality Improvement and the Role of Statistics

U.S. industry, especially the manufacturing sector, has been forced to undergo major shifts in the process of system development and reengineering over the last two decades in order to become more competitive in the international marketplace. Managers have recognized that achieving long-term improvements in quality and productivity is one way to regain their competitive edge. Statistical thinking, methods, and practices have played a critical role in these developments. Much of this renewed emphasis on quality and productivity was in response to the competitive success of Japanese industry, sparking an associated interest in studying world-class quality practices and in adapting them to fit the U.S. environment.

The result has been a reexamination of old, accepted notions about quality. End-of-production inspection activities and the use of "military standards," developed in response to the needs of war-time activities, were popular in the 1960s and 1970s, and inspection activities were based on the philosophy that defined quality as "conformance to specifications." The first of these techniques was aimed at preventing bad products from being shipped to customers, but it did not contribute to the more important task of improving (where needed) the processes that led to the production of bad products.

The other technique, the "specification limit" definition of quality, ignores several critical realities. First, products closer to a design target generally perform better. Moreover, they are less likely to drift out of the specification limits over time than those that are further from the target but still within design specifications. Second, the cost of quality is not necessarily a "0-1" function of the amount by which a product deviates from its design specifications. Third, quality is defined more by the needs of the customer than a designer with some predetermined view of manufacturing capabilities.

The concepts of inspection and 0-1 cost structures have been replaced by concentration on "quality by design" and reorganizing in order to make it right the first time, philosophies that emphasize that quality should be built into the product at the design stage. The focus of quality improvement has shifted to the design and development phases of both products and processes. There has been a concordant emphasis on developmental testing, often with operational aspects, continuous product improvement, and increased emphasis on the use of sophisticated statistical methods. Proactive techniques such as extensive use of statistical design of experiments and accelerated reliability testing are used in the design and development phase to optimize the product and process design. Feedback control, process monitoring, and failure and process diagnostics are used to control and reduce variation. In particular, specified statistical methods are now used extensively during the production realization process, in order to understand and manage the different sources of variation: they range from simple techniques, such as the "basic seven tools" (Ishikawa, 1985), the "new seven tools" (Mizuno, 1988), and quality function deployment to computer-implemented advanced methods of real-time and sequential decision making, Bayesian methods, and interactive and adaptive experimental designs mentioned earlier. There has also been an emphasis on reducing the length of the product development cycle in order to be more competitive. Reducing product development cycle time in DoD is important as well both to reduce or control costs and to provide soldiers with new technology quickly.

Quality Management and the Role of Statistics

Although the use of technical development is important in its own right, what has made these tools more effective are corresponding improvements in quality management. The success of Japanese industry has clearly demonstrated that technological superiority alone is not enough to be competitive in the long run. Although advantages of technological innovation can be compensated for by the ability to perform product reengineering, competitive success is based on an ability to understand and manage processes and the commitment to long-term quality improvement.

Statistical thinking has been at the heart of much of this development of quality and process management methods over the last two decades. In particular, total quality management is one of several modern management tools. It consists of an overall philosophy as well as methods for leading and managing an organization efficiently and effectively in order to meet the customers' needs. Some fundamental principles of total quality management relevant to our discussion here include:

1. Each decision in the systems development and acquisition process is a series of interconnected subprocesses.
2. All these processes have random variations that underlay outcomes;
3. Understanding (and reducing) these variations is a key to success; and
4. Effective decisions must be based on accurate and agreed-upon data.

Since statistical concepts and statistical thinking play a major role in studying, managing, and reducing variation, they are at the heart of process and quality management. This is, in fact, one of the basic messages in Deming's management theory (Deming, 1986).

Reduction and management of process variation is also the underlying theme in internationally accepted efforts at standardization, such as ISO 9000 (see Appendix D). These standards are based on several fundamental operating principles:

1. Processes affecting quality must be documented: documenting often identifies important ad hoc methods that are inconsistently applied. Adhering to consistent applications of the process reduces the variability in the product. When processes are well documented, customer requirements are more fully understood, the staff is more capable of satisfying the customer because of an improved

understanding of how products and services are designed and produced, and consistent means are achieved for investigating root causes and resolving customer complaints.

2. Records of important decisions and valuable data must be retained. Decisions must be periodically reviewed during audits and scrutinized for authorization. Data describing information about the quality of the product or service must be retained for a defined time period. These data form a valuable source of information used for improvement and development of new products or services.

3. Processes must be in control: effective documentation and record retention result in bringing processes under control. This is a necessary step before a process of continuing improvements can be implemented.

4. Examining a process and documenting it leads to better understanding of the causes of variation, variability reduction, and better process management. The resulting continuous improvement leads to identification and root-cause analysis of systemic problems and suggests corrective action.

The panel believes that some of the lessons learned in industry are indeed applicable to defense acquisition.

IMPLICATIONS FOR DEFENSE ACQUISITION AND OPERATIONAL TESTING AND EVALUATION

There are several implications of adopting the proposed new paradigm as the foundation for DoD testing and evaluation as part of the system development process. To avoid unintended consequences, DoD should attempt to understand how each change will affect aspects of the acquisition process before adoption.

Evaluating a System Against Broadly Stated Mission Needs

All parties from the operational requirements community should avoid using unnecessarily simplistic metrics and detailed specifications as key measures of merit or success. Specifically, all parties—including program managers and the test community—should accept the goal of measuring progress against broadly stated mission needs. This might involve evaluating and making tradeoffs (using modeling and simulations, for example) as well as judgment.

For example, consider the early operational testing that showed flaws in the Intervehicular Information System (IVIS) of the M1A2 battle tank. This system was an integral part of the upgrade that defined the new equipment as the M1A2 tank. The unsatisfactory performance could have been attributable to one or more of three possible causes: inadequate training of the tank crews, hardware immaturity, or software errors. Although that performance could have been attributable to one or more of these causes, it was appropriate to use overall system performance as the measure of the problem, rather than, say, adequacy of training. (In this case, the key issue was whether to cancel the entire upgrade, cancel only the IVIS subsystem, or attempt to fix the problems with IVIS while accepting the upgrade.)

The recent operational test and evaluation of the Longbow Apache helicopter seems to be representative of a healthy change in this direction. According to presentations to the panel by the test manager, this helicopter was evaluated against broad statements of operational needs rather than against detailed measures of performance that did not translate into higher level mission outcomes. He noted that there were a few individual performance characteristics that failed the operational test but they were considered unimportant for operational needs.

If DOT&E is to function more as an information-gathering agency, then absolute requirements are less meaningful than measures, evaluated continuously, linking performance to some explicit notion of military value or utility. Decisions would not be made on individual acquisition programs without simultaneous consideration of other related programs and their expected joint contribution to military capabilities.

When a new system or an improvement to an existing system is under development, responsibility is assigned to a single service, even though the system may be useful to other services. Many observers believe that DoD should give more attention to joint testing, not only for systems with joint use, but also for systems that will be used by a single service in joint operations. There are notable examples of successful joint tests—successful in the sense of cooperation between or among the services. An obvious one is the Air Force C-17: although the Air Force had the lead, the Army had a full-time presence at Edwards Air Force Base during the test. In contrast, the High Mobility Multipurpose Wheeled Vehicle, used by all the services, was essentially an Army only test. DoD should consider establishing a point of contact at the Joint Chiefs of Staff level for operational testing, with representation on all ACAT I operational tests.

There are many views that need to be considered in the design and testing of defense systems. Each can offer valuable information on appropriate and objective measures of performance and effectiveness and should have input into the definition of these measures. In particular, the setting of requirements should involve representatives from the program manager, the service test agency, and DOT&E (see Recommendation 2.1).

Archiving and Using Performance Data

DoD could benefit in a variety of ways from standardized test data archival practices. These include:

1. providing the information needed to validate models and simulations, which in turn could be used to plan for (or reduce the amount of) experimentation needed to reach specified operational test and evaluation goals;

2. facilitating the "borrowing" of information from past studies (if they are clearly documented and there is consistent usage of terminology and data) to inform the assessment of a system's performance, by means of statistical methods;
3. making data from developmental testing widely available for efficient operational test design;
4. facilitating learning from best current practices across the services; and
5. lead to an organized accumulation of knowledge within the Department of Defense.

To accomplish all of this, the test data archive should include both developmental and operational test data, and, possibly, training data; use of uniform terminology in data collection across services; and careful documentation of development and test plans, development and test budgets, test evaluation processes, and the justification for all test-related decisions, including decisions concerning resource allocation. In addition, the critical circumstances that produced all the data must be clearly documented. While this is extremely important for data from developmental tests, it is also important for training data, which could be used to alert DoD about the necessity of a post-production review. Finally, it is important that the in-use performance of systems, when available, should also be included in the archive.

We point out that the trade-off between the utility of the information in operational testing and that from real use is not completely clear. Data from field use can at times be less useful than those from operational testing since the circumstances surrounding field use are less controlled than in operational testing. Furthermore, careful data collection is often extremely difficult in real use.

A key benefit of documentation and archival of test planning, test evaluation, and in-use system performance is the creation of feedback loops to identify system flaws for system redesign and to identify when tests or models have missed important deficiencies in system performance. The performance of systems in the field can then be compared with developmental and operational tests, and modeling results to help improve future system development, test design, and improve modeling and simulation techniques, and to better understand the limitations of various approaches to testing systems. More specifically, the archive could provide information about: (1) operational test successes and limitations—by comparing results from operational tests with observations on actual use; (2) sources of system problems—by comparing observed system (especially suitability) problems with problems observed in test history; and (3) the validity of modeling and simulation—by comparing the results from modeling and simulation to those from actual use. Such comparisons would be especially relevant for reliability, availability, and maintainability issues (see Chapter 7). Of course, the relevance of feedback loops requires that differences between the systems being compared (e.g., related systems, systems in development, systems once fielded) and the test circumstances are either not important or are taken into consideration in the analysis. To help ensure comparability in comparisons, the archive should include a description of any substantial system modifications and when they were instituted.

Access to such an archive would necessarily have to be restricted because of the sensitive nature of a system's performance while it is in development. The question of access can be addressed by arranging different levels of information access, including access to abstracts, study summaries, and data. Operational performance information might be available only to DOT&E and the service operational test agencies while the system is still in development. However, once a system has entered full-rate production, access to its data should be broadened to those with a legitimate "need to know."

Standardization, documentation, and data archiving are important because they facilitate use of all available information for efficient evaluation and decision making. The service test agencies should investigate the use of industrial models as examples of ways of collecting and archiving test data. This includes adherence to modifications of ISO 9000. Although data archives may have been of limited value in the past, they can be made much more useful with modern technology. Since individual programs would not be able to support such an undertaking, DoD should support investments in test infrastructure by centrally funding this common warehouse for test and performance data.

In developing such an archive, it will be important to explicitly acknowledge the costs and benefits of data collection and to develop an incentive structure that ensures the effective participation of all involved parties. A recent RAND study (Galway and Hanks, 1996) discusses the difficulties that can arise when data are collected by one unit of an organization for use by another unit. The production and flow of useful information across organizational units can be hampered by poor communication and understanding of multiple purposes of data collection and archiving, even when such activities serve the collective good of the organization. If Recommendations 3.3 and 10.2 are adopted, the Statistical Analysis Improvement Group (see Chapter 10) should be tasked to recommend the party or parties to be responsible for managing the test data archive and to address the complex issue of who should and who should not have access to it.

Establishing a Continuum of Experimentation, Testing, Evaluation, and Reporting

Testing should not serve solely as a "final examination" of a system. Testing is more effective when it identifies system problems as they occur so that design changes can be instituted before substantial resources are committed to a flawed design. To identify system problems earlier that are unique to operational experience, the continuous testing must mimic aspects of operational use. This means using smaller scale testing with typical users, interactions with enemy systems, and use of less scripted activities. This approach would be similar to the Army's force development test and experimentation, which provide insights into possible new operational concepts and doctrine for equipment before it is fully developed, and recognize system limitations and their causes related to operational performance as soon as it is feasible. These operationally-oriented early tests could provide important timely information about operational deficiencies that could help in operational test design and evaluation, and system redesign. These early tests would also be used to identify the key factors that would limit system performance beyond the defined operating conditions. Along these lines, if the above were not feasible, some real benefit would often result from taking existing developmental testing and modifying it to have whatever operational aspects were practicable. Operational test personnel might assist developmental testers in incorporating operational-type scenarios in their tests, possibly in part through providing them with operational test strategies earlier in program planning.

In this approach, operational evaluation reports would be prepared in recognition of the need for multiple assessments of the operational performance of a system under development, and the reports would be issued continuously throughout all operational test activities. Such reports will be helpful in providing information for the feedback loop to inform system development and will minimize system faults that are discovered in late-stage operational testing. These reports could also, in extreme circumstances, be used to support early termination of an unsuccessful program.

POSITIVE DEVELOPMENTS IN RECENT PRACTICE

The fiscal 1996 annual report of the Director, Operational Test and Evaluation (U.S. Department of Defense, 1997) discusses in some detail efforts to implement former Secretary of Defense Perry's five themes of operational testing and evaluation in the "new world" of DoD acquisition:

Earlier involvement of operational testing in the acquisition process: "to take advantage of operational test insights early in systems acquisition programs, to identify problems and fixes early, and to avoid the program disruption and costs which can come when problems are found later."

More effective use of modeling and simulation: "modeling and simulation can help us determine . . . when the probability of test success is high enough to warrant actually beginning the test or whether some additional work needs to be done before valuable test resources are employed."

Combining tests, including developmental tests and operational tests: "This theme includes combining developmental tests (DT) with operational tests (OT) and sometimes combining tests in different programs. It also includes making all feasible use of data gathered in DT, in EOAs [early operational assessments], and in any other way that makes sense during OT."

Combining testing and training: "Training exercises are often quite realistic and can help provide the kind of test environment needed for operational tests. Similarly, operational tests can add a richness and complexity that can be valuable in a training environment. Together, testing and training employ many of the same resources, often at the same range."

Advanced concept technology demonstrations (ACTD) (testing for "insight and understanding, not a 'final exam' grade"): "ACTD programs are wide-ranging in size and scope and require tailored technical and managerial approaches and strong conceptual and operational links to the warfighter."

These themes are consistent with the ideas discussed above, and we encourage efforts within DoD to attain the objectives. Statistics and statisticians can and should play a key role in the identification and development of appropriate methodologies for each of these themes, but we also argue that organizational changes are needed to fully realize potential gains. We applaud the expression of commitment from the highest levels of policy within DoD.

Establishing the general goals of a new approach to operational testing and evaluation is important, but it must be followed by implementation and tangible success in practice. The panel has had little direct exposure to recently innovated approaches, but some are cited in the DOT&E report (U.S. Department of Defense, 1997), and we have seen some evidence that the lessons learned in private industry are at times now being put into practice by the military services in defense acquisition. One example is the development of electronic warfare systems. Historically, most electronic warfare systems have not met all their requirements, so now new electronic warfare programs are being exposed to operational test and evaluation planning, and operational contexts are being addressed early as part of developmental testing. We encourage wider application of these initiatives.

A NEW SYSTEM: TWO STORIES

To help develop a picture of the advantages that stem from adopting a new paradigm for defense acquisition, we compare what might happen to a fictitious system, QZ5, under the current paradigm to what might happen under a new paradigm containing the features we recommend. While QZ5 is fictitious, each of the problems below have occurred with real systems.

The Current Paradigm

QZ5 is a high priority ship-based missile system representing a technological advance that gives it a substantial advantage over the system it is designed to replace. The program manager's staff and the relevant service test agency know that the technology in QZ5 has been used successfully in an existing land-based system, QZ4, and so are enthusiastic and confident about the possibility of the system's success both in operational testing and when fielded.

The design of QZ5 is fundamentally sound, but there are several reliability problems with the new technology that could be solved with relatively minor design alterations. These reliability problems—involving certain components when they are exposed to cold, wet conditions—are much more likely to be demonstrated in an operational setting rather than in a laboratory. The primary measure of effectiveness for QZ5 is "p," the probability of hit against a specific enemy threat. The value of p for QZ5 must be greater than 0.8 to justify its acquisition. Developmental testing has shown that QZ5 has a value of p much greater than 0.8 for the commonly used, unstressful levels of a particular kind of obscurant, and therefore should ultimately be acquired once the design changes to enhance reliability are implemented. However, developmental testing also has shown that the value of p is only 0.3 if there are extreme levels of the obscurant.

QZ5 had undergone developmental testing by expert users, but it had not been subjected to any operational exercise before beginning operational testing. The program manager is aware of both the possible reliability problems and the problem with extreme levels of obscurants, but he does not believe that the former is serious or the latter is relevant, and he hopes that the system will be certified for full-rate production. At worst, a few minor design modifications may ultimately be needed, and since the system is a distinct improvement over the currently available system, the program manager argues strongly that it quickly be brought into service, with the recognition that there may be some potential small costs for retrofitting later on.

System QZ5 then enters into operational testing. Because there is no test archive, the commander responsible for the design and evaluation of QZ5's operational test is unaware of the test results for QZ4, and so designs a general purpose operational test that does not include certain test scenarios—including testing under simultaneously cold and wet conditions—that were troublesome for this technology in the previous system. In addition, even though the results of the developmental test were known, the levels of obscurant that were used in the operational test scenarios were modest compared with the extreme levels used in the developmental test since those levels would only rarely be observed in practice.

When the operational testing of QZ5 was concluded, the statistical significance tests for system suitability were all acceptable: that is, nothing indicated the suitability problems. This "false negative" occurred because (1) the test design failed to use the information about scenarios related to reliability problems from QZ4, (2) the results of QZ4's operational test were not used in conjunction with the operational test results for QZ5, and (3) the small sample size (due to the expensive nature of the expendables) of the operational test resulted in a test that only had a 60 percent chance of identifying the existing problems. This 60 percent chance of identifying the problem was not communicated to the decision makers, since only the results of the statistical significance test were presented in the evaluation report. Furthermore, the test performed had 80 percent power for testing p, the primary measure of effectiveness at the required level, so the low power with respect to suitability was not challenged in test design. Finally, the levels of obscurant tested were modest, providing no information about the system's ability to handle more intense levels of obscurants.

System QZ5 was approved for full-rate production. Unfortunately, the reliability problems necessitated that the system undergo an expensive retrofitting, which took 12 months to implement, during which time the missile system was out of service. After retrofitting, a potential enemy had developed a new obscurant, which represented a higher level than that tested. There was interest in using the QZ5 system in this situation, but it was not clear whether it would be effective given the low levels of obscurant used in the operational test. The best guess (but erroneous conclusion) was that it might be effective, given the fact that it had easily passed the operational test.

The New Paradigm

Under the new paradigm, the problems with QZ4 with respect to wet and cold conditions were well known since they were documented in a test and performance archive. Throughout QZ5's development small-scale tests with operational aspects were used, revealing the causes of the reliability problems, and the system design was modified early to overcome these faults. These tests were designed to use scenarios with low temperatures and high precipitation that were similar to those for which the QZ4 system had previously experienced reliability problems.

When the system was subjected to its final operational test, it was anticipated that the system would experience no remaining problems. The operational test was therefore an opportunity to learn as much as possible about the characteristics of the system. The system passed the reliability-related statistical significance tests easily due to the design changes. The operational tests were much shorter and less expensive than previously since data from QZ4's operational test were used to supplement the information from the tests of QZ5, and it was known which scenarios to focus on. Rather than being concerned with passing statistical significance tests on scenarios that fell well within the definition of ordinary use, QZ5 was also placed in test scenarios that explored how the performance of the system varied with typical but also atypical levels of stress. It was discovered that the system was sensitive to high levels of obscurant and, in particular, would not be appropriate for use in an operational setting in which this countermeasure was likely to be used.

System QZ5 passed its operational test, as fully expected, and needed no postproduction retrofitting, saving considerably more funds than were used both to continuously test it throughout development and to contribute to the test data archive. The results of the testing of the QZ5 system were archived to assist in the test design and evaluation of any related systems in the future.

CONCLUSION AND RECOMMENDATIONS

Conclusion 3.1: Major advances can be realized by applying selected industrial principles and practices in restructuring the paradigm for operational testing and the associated information gathering and evaluation process in the development of military systems.

Recommendation 3.1: Congress and the Department of Defense should broaden the objective of operational testing to improve its contribution to the defense acquisition process. The primary mandate of the Director, Operational Test and Evaluation should be to integrate operational testing into the overall system development process to provide as much information as possible as soon as possible on operational effectiveness and suitability. In this way, improvements to the system and decisions about continuing system development or passing to full-rate production can be made in a timely and cost-efficient manner.

Recommendation 3.2: Operational evaluations should address the overall performance of a defense system and its ability to meet its broadly stated mission goals. When a system is tested against detailed requirements, these requirements should be shown to contribute to necessary performance factors of the system as a whole.

Recommendation 3.3: The Department of Defense and the military services, using common financial resources, should develop a centralized testing and operational evaluation data archive for use in test design and test evaluation.

Recommendation 3.4: All services should explore the adoption of the use of small-scale testing similar to the Army concept of force development test and experimentation.

4

Upgrading Statistical Methods for Testing and Evaluation

Chapter 3 outlined a new paradigm for integrating testing into defense system development. This new paradigm reflects state-of-the-art industrial models and is based on applying statistical principles throughout the system development process. Other improvements to the testing and evaluation process itself could be realized by applying current views of statistical methodology in a more widespread and appropriate way.

> **Conclusion 4.1: The current practice of statistics in defense testing design and evaluation does not take full advantage of the benefits available from the use of state-of-the-art statistical methodology.**

This chapter presents an overview of that methodology so that test planning, design, and evaluation are as effective and efficient as possible; chapters 5-9 discuss these issues in greater detail. The adoption of many current techniques can be accomplished at minimal expense, and some discussion of how this can be accomplished is presented in Chapter 10. These changes are implementable in the short term and do not, generally speaking, require the institution of the new paradigm recommended here, although some of them would be more effective if implemented concurrently.

Detailed recommendations related to test design, test evaluation, design and evaluation for reliability, availability, and maintainability, software test methodology, and use of modeling and simulation are in the chapters that follow; this chapter presents a less technical review of the inadequacy of current statistical practice in defense testing and the benefits to be gained from use of the best current methods and practices. In this chapter we highlight issues in need of *immediate* attention, identifying areas in which current Department of Defense

(DoD) practice differs substantially from best practice, to the detriment of effective operational test and evaluation.

We focus on operational (rather than developmental) testing, especially for ACAT I systems. However, many of the issues raised and recommendations made apply to developmental (or other forms) of testing, and to systems in other acquisition categories.

KEY ISSUES ILLUSTRATING THE USES OF STATISTICAL METHODS IN OPERATIONAL TESTING AND EVALUATION

Test Planning and Design

Test planning consists of collecting specific information about various characteristics of a system and the anticipated test scenarios and environments and recognizing the implications of this information for test design. Test planning is crucial to a test's success. Test planning comprises several elements (see, e.g., Hahn, 1977; Coleman and Montgomery, 1993).

Defining the Purpose of the Test Operational tests often have multiple objectives, for example: to measure "average" or "typical" performance across relevant scenarios, to identify sources of the most important system flaws and limitations, or to measure system performance in the most demanding scenario. Each of these objectives could be applied to several performance measures. Different objectives and measures can require different tests. Some test designs that are extremely effective for one purpose can be quite ineffective for others; therefore, agreeing on the purpose of the test is necessary for test design. One must also identify those performance measures that are most important (however defined) so that the operational test can be designed to effectively measure them.

Handling Test Factors Test factors include the defining variables for: environments of interest (temperature, terrain, humidity, day/night, etc.), tactics and the use of countermeasures, the training and ability of the users, and the particular prototype used. Clearly, how a system's performance varies across different values of some factors (e.g., in day or night, or against various kinds of enemy tactics) is crucial to an informed decision about procuring the system. Some test factors are under the control of the test planner and some are not, and some test factors are (or are not) influential in that varying them can cause substantial changes in system performance. Considering each test factor with respect to whether or not it is controllable and/or influential, may require different approaches to its use in testing. A serious problem arises from the failure to consider some influential factors in the test design: such a failure can make the test ineffective since those factors may vary during the test, causing performance differences.

Specifying Test Constraints Budgets, environmental constraints, and various limitations concerning the scheduling of test participants and test facilities are just a few of the constraints on an operational test. The ability to change the level (or category) of test factors in time for successive test runs may also be constrained. To effectively design a test, one must fully understand the various test constraints.

Using Previous Information to Assess Variation Understanding the degree of variation in performance measures from repeated tests *within* a single scenario, compared with the variation in measures *between* scenarios (i.e., the sensitivity of the system to changes in environment, tactics, users, prototypes), is needed to decide how to allocate prototypes to scenarios and the number of replications needed for each scenario. This information is also needed to decide on the maximum number of scenarios to use if estimates of performance for individual scenarios are needed. Collecting information about the sources and degrees of variability is crucial for effective test design.

Establishing Standardized, Consistent Data Recording Procedures Operational tests are, in part, unscripted activities for which data collection is clearly complicated. To use test results from a given system for test design or evaluation of another (related) system and to properly evaluate whether a test result was due to unusual circumstances, test data need to be recorded in a form that is accessible and useful across the military services. Such data are also extremely valuable in improving operational test practice, and in the validation of modeling and simulation when used in conjunction with data on real use, by comparing observed performance with test performance. The standardization of data recording should be done in a manner analogous to that of industry, using ideas from industrial standards (ISO 9000).

Using Preliminary Tests for Test Planning Information, and Running Operational Tests in Stages Effective decisions about operational test design and test size require information that is often system specific, especially with respect to operational characteristics. For example, data needed to determine an operational test's sample size is sometimes not available from other systems or developmental test results. Moreover, issues often arise in the test of a complicated system that are difficult to anticipate.

For these reasons, some form of inexpensive and operationally realistic preliminary testing would be extremely valuable. Such tests would help ensure that operational tests will be as informative, effective, and efficient as possible. Preliminary testing is broader than the Army concept of a force development test and evaluation; it includes the collection of other information concerning test design and test conduct. Additional benefits of preliminary testing are discussed in Chapter 3 (as well as in Part II), in terms of a more continuous assessment of operational system performance. Although there are aspects of operational testing that do not easily scale down to small tests (e.g., the number of users and systems needed for the test of a radar system), we are convinced that the use of preliminary testing has not been adequately explored.

The recommended continuous process of information gathering with respect to the operational performance of a weapon system under development could be accomplished in many ways. Operationally realistic, small, focused tests could be conducted earlier in system development. Developmental testing could incorporate aspects of operational realism. Training exercises could be organized to make objective performance assessments. Although some or all of these may not be feasible for individual systems, some of these methods will be feasible and should be used for most systems.

Additional Considerations Five other issues need to be addressed before designing a test:

1. the proper experimental range for the controllable variables;
2. the statistical relationships between performance (effectiveness) measures and test factors;
3. the desired degree of precision of the estimates;
4. the existence of previous benchmarks of performance; and
5. the statistical techniques that will be required to analyze the data.

Comprehensive test planning, including the elements described above, should be a routine, early step in the design of operational tests. (For more details on test planning, see Chapter 5.) Those in charge of an operational test must work with appropriate statistical experts to understand how the above information can be used in the operational test's experimental design. One useful technique, suggested by J. Stuart Hunter (personal communication) as a means to address many of the issues of test planning, is to try to guess the test results a priori. This is a quick way to communicate objectives of the test, test factors, and the expected between- and within-scenario variability. It will also be useful in estimating the statistical power of the test. Modeling and simulation could be very useful in organizing this information (see Chapter 9). The above components of test planning would be extremely helpful for the hypothetical major of Chapter 1 as a checklist when designing a complicated operational test and would assist the major in communicating with an experimental design expert. Templates for this purpose exist in the statistical literature and could be modified to be more specific to the operational testing of defense systems.

Recommendation 4.1: Comprehensive operational test planning should be required of all ACAT I operational tests, and the results should be appropriately summarized in the Test and Evaluation Master Plan. The

following information should be included: (1) the purpose and scope of the test, (2) explicit identification of the test factors and methods for handling them, (3) definition of the test environment and specification of constraints, (4) comparison of variation within and across test scenarios, and (5) specified, consistent data recording procedures. All of these steps should be documented in a set of standardized formats that are consistent across the military services. The elements of each set of formats would be designed for a specific type of system. The feasibility of preliminary testing should be fully explored by service test agencies as part of operational test planning.

Test Analysis and Reporting

Reporting Estimates of Uncertainty

In order to fully use the information collected in an operational test, it is important that all reported test results—typically averages and percentages—be accompanied by an assessment of their uncertainty. This information would alert decision makers about the variability of performance estimates, which in turn can help to determine the risks and benefits of proceeding to full-rate production on the basis of the results of the operational test. Such reporting should be in the form of confidence intervals, at a typical level of 90 or 95 percent, for each measure of performance or effectiveness. For example, instead of reporting that a missile system obtained an estimated average hit rate of 0.85, the report would state that there is a 95 percent confidence interval for the hit rate in the range between 0.65 and 0.92. Presenting the test results in this way helps to raise important questions, such as: If the true hit rate is as low as 0.65, would one decide to go ahead with procurement or to perform more testing to rule 0.65 out (one hopes)?

If significance testing for major measures of performance or effectiveness is used to decide whether to proceed to full-rate production (a method criticized in Chapter 6), the probability of "passing" and "failing" an operational test given that the true system performance is at various levels (the "operating characteristics" of the test) should also be provided to the decision makers. This should be done for several hypothesized performance levels. Four levels that would be particularly informative would be (1) at a level higher than the requirement, (2) at a level equal to the requirement, (3) at a minimally acceptable level, and (4) at a clearly unacceptable level. Along with information on the costs of further testing and the consequences of incorrect decisions about further system development, these probabilities would provide valuable information to decision makers about the risks involved in deciding to pass a system to full-rate production.

The evaluation report should also include the uncertainty of results for estimates for each important individual scenario. For example, if a test includes two replications at a scenario of great interest, the confidence interval of the test results that are specific to that scenario should be reported. If an evaluation uses modeling and simulation, results from an analysis of the variability due to model misspecification, and its effect on the simulation results, should also be reported, providing important information about the reliability of input from the simulation model. (For more details on test evaluation, see Chapter 6.)

> **Recommendation 4.2: All estimates of the performance of a system from operational test should be accompanied by statements of uncertainty through use of confidence intervals. If significance testing is used, the operating characteristics of the test (along with the costs of additional testing and the consequences of incorrect decisions) should also be reported. Estimates of uncertainty for performance in important individual scenarios should also be provided, as should information about variability due to model misspecification and its effect on simulation results.**

Combining Information from All Appropriate Sources for Test Design and Evaluation

Sources of information available on the performance of a defense system under development, before operational testing, include the test results and field use of the system that is intended to be replaced, the performance of similar components on other systems currently in use, the results of developmental tests, data from possibly less controlled situations such as training exercises or contractor test results, and early operational assessments or the preliminary testing suggested above.

Test designers need to make use of all of the available information about system performance in order to make an operational test as effective and efficient as possible. The above sources can be used to provide information on: the variability of system performance across replications within a single scenario; the sensitivity of the performance of the system to particular changes in the environment, tactics, countermeasures, etc.; the variability of system performance across prototypes; and components of the system or system design issues that might need focused attention due to perceived problems. This information can also help identify whether modeling or simulation is likely to be effective in various aspects of operational testing, by understanding whether it was effective in testing the baseline or related systems. Information about these characteristics of system performance are not now uniformly collected and used for test design. For example, in the Longbow Apache, the test design described in Appendix B calls for only a modest increase in the number of replications in nighttime scenarios over the number in daytime scenarios (320 versus 256). However, it is likely that the variability of the difference in performance between the Longbow Apache and the baseline Apache would be much greater in nighttime than in daytime scenarios, suggesting that a more efficient test design would have been to use relatively more replications of the Longbow during nighttime. Given the importance of test design, especially for ACAT I systems, this information needs to be collected and used.

Test evaluators can also make important, productive use of this auxiliary information, particularly when measuring system suitability. There is an understandable concern about combining operational test information with information from tests either on different systems or from non-operational settings (such as developmental tests). Combining such information for test evaluation will sometimes be inappropriate. However, there are statistical methods to identify when various forms of pooling or weighting data are appropriate, and there are statistical methods that can account for the difference in test circumstances that might be directly relevant. The applicability of these methods should be investigated.

Most methods for using auxiliary information, either for test design or for test evaluation, are beyond the technical expertise of our hypothetical major, who will need access to statistical expertise in order to use this auxiliary information effectively. The assumptions that underlie these methods require extensive understanding of the systems and tests in question; therefore, collaboration between those with statistical and system expertise is vital.

> **Recommendation 4.3: Information from tests and field use of related systems, developmental tests, early operational tests, and training and contractor testing should be examined for possible use in appropriate combination, when defensible, and with operational test results to achieve a more comprehensive assessment of system effectiveness and suitability.**

It is important to stress the use of the term *appropriate*. For example, simple pooling and other use of developmental test data may be inappropriate since the test circumstances can be so different from operational use. (A good example concerning the Javelin is given in U.S. General Accounting Office, 1997:12; see also Table 9.1.)

Clearly, the ability to combine data on system performance from these various sources, including developmental test, is strongly dependent on the establishment of standardized methods of documenting test circumstances. Therefore, the archive put forward in Recommendation 3.3 (and the ideas in the supporting text of Chapter 3) is crucial to support combination of information. Otherwise, it is extremely difficult to understand the operational relevance of the information.

FOCUSED USES OF STATISTICAL METHODS

Experimental Design

The operational test of a defense system is an experiment with a key objective of determining whether its performance satisfies stated requirements or exceeds those of a control or benchmark system. Since operational testing, particularly of ACAT I systems, is important and costly, operational tests must be designed so that they are as informative and efficient as possible. They must produce results that permit the best decisions to be made with respect to proceeding to full-rate production.

The statistical field of experimental design has made enormous advances in addressing precisely this broad issue, and specific techniques have been developed to design a test to either maximize the information gained given fixed costs or (equivalently) to minimize costs while providing information that permits a decision with acceptably small risk. Methods relevant to this problem include the use of randomization and controls, various specific designs (including fractional factorial and Plackett-Burman designs), and response surface methods. In addition, several principles have been discovered that broaden the applicability of these specific techniques (when not directly applicable) to test situations. Two general principles that can be applied to a wide variety of testing problems are: (1) test relatively more where variation of what is being measured is greatest, and (2) choose (some) values for test factors that are close to the limits of typical use. (For more details on test design, see Chapter 5.)

Although some of these techniques and principles are finding their way into DoD's standard operational test design, current practice is still substantially distant from the state of the art. This has resulted in inefficient test designs, wasted resources, and less effective acquisition decision making.

> **Recommendation 4.4: The service test agencies should examine the applicability of state-of-the-art experimental design techniques and principles and, as appropriate, make greater use of them in the design of operational tests.**

Appropriate Models for the Distribution of Failure Times

The distribution of times to first failure or of times between failures of defense systems can depend on the cause of the failure, the age of the system, whether the system had experienced previous failures and then been repaired, the amount of time the system has been in continuous operation, the users, the stress of the environment, and the specific prototype used. It is common (though some exceptions exist) for testers in the defense community to assume that the time to first failure or between failures follows a common exponential distribution, probably because of the simplicity of the methods that result. But the use of this distribution amounts to an implicit assumption that reliability does not depend on the age of the system or the amount of time on test. The assumption of exponentiality is used for test design (to decide how much time on test is sufficient), and it is used for test evaluation (e.g., to provide estimates of the variance of estimates of mean time to failure). When this assumption is completely accepted, it does not make any difference how many different prototypes are used or whether the systems being tested were new systems or systems that had been repaired. (For more details on test design and evaluation of system reliability, see Chapter 7.)

There are systems and components for which the assumption of exponential failure times is appropriate, such as some types of electronic systems. However, when it is not, there are many alternative failure-time distribution models that will be more appropriate for many systems. Use of these alternatives (when valid) can have the following benefits:

1. smaller test size when assessing reliability for systems that have the characteristic that their reliability decreases with use (which is frequently the case), since the data are more informative than if reliability is not a function of time in use;
2. test size that is relevant to the process being tested;
3. test design that explicitly addresses the treatment of repaired and unrepaired systems; and
4. significance tests that are more valid, since they are based on a more accurate model of the failure-time distribution.

> **Recommendation 4.5: Service test agencies should examine the use of alternative models to that of the exponential distribution for their applicability to model failure-time distributions in the operational tests of defense systems.**

Software Testing

Since ACAT I defense systems essentially now all have a software component and since software reliability is a common troublespot in recently developed defense systems, the proper testing of software in defense systems has a high priority. At least one service uses the resources available for operational testing of a system with a software component to check individual lines of code. Although this is a useful activity in developmental testing, it is very resource intensive and does not serve the purposes of operational testing.

The proper method for the operational testing of software is usage-based testing, which tests whether the system is fit for its intended use. This testing involves using specific user profiles to develop a statistical distribution of how the software is going to be used, that is, how commands or inputs are chosen. These distributions are then drawn from and applied to the software to determine if it performed correctly. When testing schedules and budgets are tightly constrained, usage-based testing yields the highest practical reliability because if failures occur, they will be high frequency failures. (The existence of rarer, but particularly crucial failures can be separately tested.) Companies such as AT&T, IBM, and Texas Instruments are increasingly using such strategies, and there has also been some use in the Army, Navy, and Air Force. The panel is convinced that this is an effective strategy with many benefits and urges DoD to expand its use of this strategy. (For more details on test design for software-intensive systems, see Chapter 8.)

> **Recommendation 4.6: The Department of Defense should expand its use of a usage-based perspective in the operational testing of software-intensive systems.**

Use of Modeling

Modeling and simulation are now being widely considered and occasionally used by DoD to augment operational testing. Given the benefits of decreased cost, enhanced safety, and avoidance of environmental and other constraints, it is natural to explore the extent to which such use can contribute to operational testing. (Note that the use of modeling and simulation cannot, by itself, replace operational tests.) One method for using modeling and simulation to augment operational tests is referred to as model-test-model. A model is developed, an operational exercise is carried out for modeled scenarios, and differences between the observed and the modeled performance are used to modify the model so that it is more in conformance with the observations.

This procedure will likely (but not necessarily) result in a revised model that is valid for situations very close to or the same as the tested scenarios. However, if the model is used to extrapolate performance to untested scenarios that are relatively distinct from those tested, there is a great risk that the predictions will be of very poor quality, possibly even worse than those from the unmodified model because a small number of operational experiences are used to refit what is typically a very large, complicated model. Such "overfitting" can skew the model to represent only the particular scenarios tested.

The size of the differences between the predictions and the observations are an indication of the quality of the original *and* the modified model.[1] A recommended method for measuring the degree to which the model has been overfit in

[1] There are experimental design implications for model-test-model. Clearly, the more representative the test scenarios are of all situations of interest, the less danger there is of overfitting.

the model-test-model approach is to test the system again, but on new, possibly andomly selected scenarios and then, without refitting the model, measure the average distance between the model's predictions and the observed performance of the system. This approach could be given the name model-test-model-test, and it may be appropriate to use in augmenting some operational tests. (For more details on use of modeling and simulation for test design and evaluation, see Chapter 9.)

Part II

Applications of Statistical Methods to Operational Testing

5

Improving Operational Test Planning and Design

The operational testing of systems under development is a common industrial activity, in which the assessment of system performance is needed for a wide variety of intended environments.[1] The statistical field of experimental design has provided the basis, through theoretical and applied advances over the last 80 years, for important improvements over direct but naive methods in test planning and test design. Methods now exist to guide the planning and design of efficient, informative tests for a wide variety of circumstances, constraints, and complications that can arise in actual use. As stated in Chapter 3, we recognize that there is a danger in drawing analogies between the problem of developing new products in industrial and other private- and public-sector areas and the development of new defense systems. Although defense systems have unique aspects, there are also substantial similarities in the operational testing of a defense system and the testing of a new industrial product. Given the very high stakes involved in the testing of defense systems that are candidates for full-rate production, it is extremely important that the officials in charge of designing and carrying out operational tests make use of the full range of techniques available so that operational tests are as informative as possible for the least cost possible.

The panel examined plans for the operational tests of several defense systems, including the Longbow Apache helicopter, the Common Missile Warning System, and the ATACMS/BAT system, as well as components of plans for

[1] The term environment is used in a very broad sense to include, in addition to weather and climate, classes of users and types of application.

several other systems. The panel identified two broad problems from this examination of test designs. First, there is no evidence of a methodical approach to test planning, which is an important requisite to successful test design in industrial applications. Second, although we found many examples of the proper use of specific techniques of experimental design—including simple ideas, such as the benefits of randomization and control, and some more sophisticated designs, such as fractional factorial designs—there were also test designs that were clearly not representative of the state of the art. Our assessment is that the current level of test planning and experimental design for operational testing in the Department of Defense (DoD) is neither representative of best industrial practice, nor takes full advantage of the relevant experimental design literature. Yet with fairly modest effort and minor modifications in the process of operational test design, the quality of operational test design could be substantially improved, resulting in improved decisions about moving to full-rate production, as well as reduced costs for testing.[2]

This chapter identifies the primary ways in which operational test planning and test design can be improved through use of state-of-the-art methods. We first outline some general test principles, including the important role of test planning as a requisite to efficient and effective test design. We then present a discussion of some of the methods of statistical experimental design that should be examined for applicability to the design of operational testing. Fries (1994a) provides supporting material for much of that discussion. Finally, we address the key question of how to determine the appropriate size of an operational test.

GENERAL PRINCIPLES OF TEST DESIGN

Systematic Test Planning

As stated in Chapter 4, test planning is the collection of specific information about various characteristics of a system and the test environment, and the recognition of the implications of these characteristics for test design. Because a single, unanticipated restriction can render worthless an otherwise well considered design, test planning is extremely important as an input to test design. Without repeating the discussion in Chapter 4, we outline here the components of operational test planning and add some technical details. Again, much of this discussion is taken from Hahn (1977) and Coleman and Montgomery (1993).

The components of test planning include:

1. defining the purpose of the test;
2. identifying methods for handling test factors;

[2] The following material may also be applicable to developmental test design.

3. defining the test environment and specification of constraints;
4. using previous information to compare variation within and across environments, and to understand system performance as a function of test factors;
5. establishing standardized, consistent data recording processes; and
6. use of small-scale screening or guiding tests for collecting information on test planning.

Defining the Purpose of the Test

It is important to keep the decision context in mind when designing an operational test. The decision that a military service faces is whether or not to pass a system on to full-rate production. Therefore, those in charge of test design must remember that the test needs to provide as much useful information as possible, at minimal cost or at a given cost, for that decision. More specifically, an operational test is currently used to satisfy two broad objectives: (1) to help certify, through significance testing, that a system's performance satisfies its requirements as specified in the Operational Requirements Document (ORD) and related documents and (2) to identify any serious deficiencies in the system design that need correction before full-rate production. The panel recommends (see Chapter 3) that objective (2) receive greater emphasis in operational testing.

The choice of test objective has important implications for test design. One issue is how close should a system be tested to the limit of normal stress with respect to various factors (at times referred to as how close to the "edge of the envelope" to test). For example, if the requirement is that the system be able to deliver a payload of 20 tons, should it be tested at 25 tons? Testing at higher payloads has the advantage of providing information on system performance for higher payloads, with the possible disadvantage that somewhat less information (but very possibly acceptably less for the given purpose) is obtained on the performance for 20 tons. This might be a desirable tradeoff, especially if the test objective is (2) above, since, although present demands might make 20 tons seem like a reasonable upper bound for typical use, it is difficult to anticipate future needs and demands.

A related question is the following: Should one allocate a substantial fraction of test runs to a high probability scenario(s) (as specified in the Mission Profiles and Operational Mode Summary) thereby obtaining high-quality understanding of system performance in that scenario(s), or should one test using a wide variety of scenarios that represent potentially very different stresses to the system? The tradeoff is understanding system performance for a specific requirement across a more limited number of stresses, with little understanding of system performance for some lower probability environments, versus understanding system performance when used in a wide variety of environments that represent a broader array of stresses to the system, with possibly reduced understanding of system performance for the most likely environments.[3] Again, the answer lies in recognizing the goal of the test.[4]

Identifying Methods for Handling Test Factors

Test factors can be influential or not influential, and they can be under the control of the experimenter or not controllable. Influential factors are factors for which system performance varies substantially with changes in the level of the factor. Clearly, misdiagnosis of which factors are and are not influential is a serious problem—often leading to wasted test resources. Therefore, the use of results from developmental tests and tests of related systems, as well as early operational assessments, if they are available, are helpful for avoiding this misdiagnosis.

Factors that are under the tester's control include testing in day or night, choice of test terrain, and choice of test tactics. Examples of factors that are not under the tester's control are unknown differences in system prototypes and weather changes during the course of the test. Obviously, factors that are not under the tester's control are not considered in test design, except for the important recognition that a baseline system may be extremely useful as a control. (In this case one is either interested in estimating the differences in performance between the two systems to find out the amount of improvement represented by the new system but not specifically the level of performance of the new system, or one is mainly interested in how good the new system is and one has excellent information on the performance of the old system.) If factors that are not controllable are influential, it is crucial during test planning that the presence of these factors be recognized so that information about their levels during testing can be collected and their contribution to test outcomes can later be accounted for through the use of statistical modeling. Failure to take the variability of these uncontrolled factors into account at the evaluation stage would likely result in substantially biased estimates.

[3] In stating the choice in this way, we are ignoring the fact that one can sometimes gain considerable information about performance in one scenario from its observed performance in others through statistical modeling; accelerated testing is one example.

[4] Operational test requirements, while typically unspecific about how to treat performance across scenarios, are often analyzed from the perspective that the requirement is to estimate the average performance across test scenarios. So, if a missile is to have a hit rate of 0.80, that hit rate could be measured as a weighted average of hit rates in individual scenarios, weighted by their frequency and military importance. Even though the objective of operational testing is often to determine whether this average is consistent with the system requirement, testers are also understandably very interested in examining performance in individual scenarios because system deficiencies might only be sensitive to some test scenarios. However, DoD typically cannot use such separate measurement as a test objective because of test size limitations. (See related discussion in Chapter 2.)

Furthermore, when using a baseline system to account for uncontrollable factors, and when it is not possible to test the system of interest and the baseline system under identical circumstances, any remaining choices in test design could be decided either through the use of randomization, or preferably, through a "balanced" experimental design. For example, when it is not possible to test the systems with the same users, it is good to design the selection of users to systems, since if one group of users is either more adept or learns more quickly than the other group, the resulting test result can be misleading. Randomization or balanced designs can be used to select users to systems, where the balancing made use of the performance of users on a pretest to select them to systems. These methods protect against (even an unconscious) selection of more expert users to either the control system or the system of interest.

Another example occurs if there are several comparisons of the new and the baseline systems in several environments or scenarios, where one system's test has to precede the other test. The decision of which system to test first in each of these several scenarios could either be made through use of randomization, or again through use of a balanced design which would mean, in this case, that each system is tested first an equal number of times. Randomization or use of a balanced design in this case helps to account for factors such as changes in the weather between tests, that might affect either system's performance. The panel's understanding is that such uncontrolled effects are not always taken into consideration in operational test planning and test design.

In addition to understanding *which* factors influence system performance, it is also important to understand *how* test factors affect system performance, i.e., it is important to understand the variability of system performance resulting from changes in the levels of various test factors. (Of course this understanding is likely to be fairly incomplete at the test planning stage.) For example, the shape of the function relating the probability of hit of a missile system to the distance to target is likely to be a non-linear curve, where for small distances the hit probability is quite close to 1, the hit probability decreases quickly as distances increase, and then for great distances the hit probability slowly approaches 0. Many test factors that can be considered to be stresses on a system have a non-linear impact on the level of performance, e.g., temperature or degree of obscurance. It is useful to understand where the active part of the non-linear functions are since it is important to test more where system performance is more variable. Therefore, collecting information that is relevant to understanding how system performance relates to the levels of test factors is an important input to test design, in that it is used to determine what levels of each factor to test at, and how many replications should be taken at each level for each test factor.

Related to the above, it is also important to compare the variability of system performance within a scenario to the variability of system performance between scenarios. Within-scenario variability can be due to such effects as idiosyncratic differences in user performance across tests, changes in the weather, or random differences in system prototypes. Information on the relative size of between- and within-scenario variabilities helps to answer important questions concerning how many different scenarios one can include in the test and the number of replications needed for each scenario. We do not provide details as to how to do this, but it is obvious that with less within-scenario variability in each replication one obtains more information on system performance for a given scenario. Therefore, small within-scenario variability would make it feasible to have either fewer trials or more test scenarios. Obviously, it is important to decide such key issues on the basis of as much information as possible.

Use of Small-Scale Screening or Guiding Tests

Proper decisions about (optimal) operational test design and test size require information that is often system specific, especially with respect to operational characteristics. This information often cannot be fully gained through examination of related systems or the developmental test results for the system in question. The information needed includes many of the factors listed above concerning test constraints, test factors, and the relative size of within- and between-scenario variability, but it also includes additional data needed to ascertain the proper operational test sample size (discussed below), which is often not available through examination of other systems or developmental test results.

Instead of relying on this possibly limited, perhaps somewhat irrelevant information to decide on test design, small-scale screening or guiding tests—preliminary tests with operational realism—can be used to collect information concerning system performance. The information includes various parameter estimates, including estimates of within- and between-scenario variation, and logistics information. Other information that could be collected includes: information as to any potential flaws in the system under development; the expected difference in performance between a system under development and any control or baseline system; information on the plausibility of various assumptions; and information that might be helpful in determining the usefulness of modeling and simulation to supplement results from operational testing.

The information from these screening tests can be used to estimate costs on the basis of the number of test runs, the standard error of prediction for estimates of system performance for individual scenarios and for average system performance across scenarios, and the statistical power if significance testing is used. In addition, other advantages and disadvantages of various plans (such as the reduction in opportunity for player learning) could also be considered, and data collection can be tested. Issues can also arise in these screening tests that are often difficult to anticipate, such as difficulties in measuring various aspects of system performance. As a result, information from the screening test can be used to facilitate measurement.

Screening or guiding tests performed in a relatively inexpensive way, can help to ensure that operational tests are as informative, effective, and efficient as possible. Testers can use these tests to understand how to anticipate unusual circumstances that might arise when first conducting operational tests on a defense system. The notion of screening tests is related to the Army concept of a force development test and experimentation, but it is broader in that the collection of other information concerning test design and test conduct (besides force development) is the objective. Additional benefits of this idea are briefly discussed in Chapter 3, where the notion of a more continuous assessment of operational system performance is advanced.[5] Certainly, there are aspects of operational testing that do not scale down to small tests easily, for example, the number of users and systems needed for a test of a theatre radar system. As was mentioned in Chapter 3, if these screening or guiding tests are not feasible, there will often be some real benefit from taking existing developmental testing and modifying it to have operational aspects where practicable. However, we are convinced that the use of screening tests has not been adequately explored, especially given their potential value.

Test planning needs to become a regular, early step in the design of operational tests. Recommendation 4.1 is an important requisite to the improvement of operational test design. We repeat that recommendation here:

> **Recommendation 4.1: Comprehensive operational test planning should be required of all ACAT I operational tests, and the results should be appropriately summarized in the Test and Evaluation Master Plan. The following information should be included: (1) the purpose and scope of the test, (2) explicit identification of the test factors and methods for handling them, (3) definition of the test environment and specification of constraints, (4) comparison of variation within and across test scenarios, and (5) specified, consistent data recording procedures. All of these steps should be documented in a set of standardized formats that are consistent across the military services. The elements of each set of formats would be designed for a specific type of system. The feasibility of preliminary testing should be fully explored by service test agencies as part of operational test planning.**

To assist in the implementation of this recommendation, we strongly suggest the development and use of templates for test planning. Fairly straightforward modifications of the templates that have appeared in the literature (see, e.g., Coleman and Montgomery, 1993) should be applicable for operational testing of defense systems. The development and use of these templates should help raise the quality of operational test design, especially over time, since improvements in the template will have associated benefits in test design.

[5] An interesting example of this approach is the pilot studies performed by the Bureau of the Census to better understand how to take the decennial census.

Linking Design and Analysis

In current practice, the focus of the quantitative evaluation of operational tests involves calculation of means and percentages computed across environments, followed by a direct comparison of these summary statistics with requirements, which is sometimes accompanied by simple significance tests comparing these summary statistics either with required levels or with baseline systems. However, various characteristics of the individual test environments undoubtedly affect system performance, and these effects are hidden when reporting (and subsequent decision making) focuses only on the summary statistics. Given the small number of test runs typical of operational tests, the benefit from more disaggregate analysis or statistical modeling may be limited. However, some information on questions such as whether the system's performance in one scenario is significantly lower than in another, and which factor used to define test scenarios is most related to a decrease in system performance, are important to answer, even in a very approximate manner. (These types of questions are closely related to the panel's recommended modified objective of operational test; see Chapter 3.) These questions require various kinds of statistical analysis to answer, possibly including analysis of variance or logistic or multiple regression. To make these analyses as effective as possible (given the test sample size and other competing test goals) the test design must take these intended analyses into consideration.

The selection of scenarios to test in, the number of test replications in a given test scenario, and the selection of levels of test factors in a test scenario, can be very inefficient and costly if determined without recognition of the intended analysis. As an example, if the difference in performance across two specific test scenarios is key to the evaluation of a system, it is crucial that there be enough information in the test to evaluate that difference. Also, if multiple regression is used to identify which test factors have the most important effects on system performance, it is useful to select factor levels so that each test factor is relatively uncorrelated with the remaining test factors. It is important to recognize that test design and the subsequent evaluation are closely linked: the Test and Evaluation Master Plan and subsequent test planning documents should include not only the test plan, but also an outline of the anticipated analysis and how the test design accommodates the evaluation. If averages and percentages remain the primary analysis associated with an operational test, this approach is much less important. But, if more thorough analysis becomes more typical, as we recommend, this approach is more important.

Objective, Informative Measures

Test design and successful testing depends on objective, informative measures. In the operational testing of defense systems, objective measurement is even more important since, as described in Chapter 2, the participants have different incentives. Thus, any ambiguity in the definition of measures could be the source of disputes. Objective definitions—e.g., in scoring rules, to determine what constitutes an engagement opportunity, a successful engagement, a valid trial, a "no-test," or various types of failures—are important to agree on in advance. Other examples of measures that require objective definitions include: what constitutes time on test for assessments of system suitability; a rule for deciding how and whether the data generated before a problem occurs in an operational test should be used (for example, when a force-on-force trial is aborted while in progress due to an accident, weather, or instrumentation failure); and, a rule for determining how to handle data contaminated by failure of test instrumentation or test personnel. (It should be pointed out that complete objectivity may be difficult to achieve in some of these cases.) Objective scoring rules reduce the possibility of the effect of individual incentives in this important decision-making process. Deciding how to enhance objectivity should be done in consultation with representatives from the Office of the Director, Operational Test and Evaluation (DOT&E).

Clearly defined measures also facilitate test design. Precise definitions allow one to link information from developmental tests and testing on related systems to questions concerning operational test design.

Some measures are more informative in a statistical sense than others. For example, measures that are based on 0-1 outcomes, such as a target miss or hit, generally speaking provide less information than continuous measures, such as distance from target at closest approach (which will not always be feasible to measure). (In addition, continuous measures also can be more useful than 0-1 measures by supporting more effective models of missed distance which can be useful in diagnosing why a target was missed; see Box 5-1 for additional information on 0-1 data.) The identification of alternative measures for various performance characteristics and full understanding of their advantages and disadvantages is an important step in test design.

STATISTICAL EXPERIMENTAL DESIGN

This section considers both some broad issues related to the methods used for experimental design of operational tests of defense systems and some specific issues raised in association with four "special problems" that the panel was introduced to while examining the application of experimental design to operational testing.

The operational test of a defense system is an experiment, one of whose key objectives is to determine whether the system satisfies performance requirements, or that its performance exceeds that of a control system. As noted above, the stakes involved are extremely high, especially with ACAT I systems, and the

Box 5-1 Comparisons of 0-1 and Continuous Data

To investigate the extent to which 0-1 data are less informative than the associated continuous data in real applications, consider the following. Assume that we have a data set that indicates, for n shot attempts, both whether or not a missile came within at most one unit of a target at closest approach, designated as a hit (the 0-1 measure), and also the horizontal and vertical missed distances from the target (at closest approach).

We can directly estimate the probability p of hitting the target by dividing the number of target hits by the total number of attempts, n. This estimate has variance p(1–p)/n.

A more complicated estimate of p is as follows. We assume that the (signed) horizontal and vertical missed distances are normally distributed with mean zero and variances σ_v^2 and σ_h^2. (This is often likely to be a reasonable assumption.) We use the data set of horizontal and vertical missed distances to estimate these two associated variances. These, in turn, can be used to estimate the probability that the sum of two independent squared normals each with mean zero and the associated variances is less than or equal to one, which is the probability of a hit. This is done through numerical integration. An estimate of the asymptotic variance of this estimate can be derived using the so-called delta method. Finally, an estimate of the relative efficiency of this complicated estimate of the probability of a hit in comparison to the above direct measurement can be computed by taking the ratio of estimated variance of the procedure described here to p(1–p)/n. An efficiency of .6 would mean that using a 0-1 measure is losing 40% of the information by not using the distances, and that 40% of the sample size could be saved by moving to the new estimate. The following relative efficiencies were computed for various values of σ_v^2 and σ_h^2:

σ_v^2	σ_h^2	p	Relative Efficiency
0.2	0.55	0.92	0.61
0.3	0.70	0.82	0.64
0.5	0.60	0.80	0.64
0.6	0.70	0.69	0.61
0.7	0.80	0.59	0.55
0.8	0.90	0.50	0.48
0.9	1.10	0.39	0.38

These efficiencies indicate that there may be real cost savings in operational tests by making use of continuous measures. Unfortunately the gains realized in practice may be less than indicated above since these efficiencies are based on the assumption that missed distances are normally distributed, which may not obtain. Some preliminary work suggests that the assumption of normality is important in that longer-tailed distributions will produce results with much less efficiency for the new procedure. Finally, researchers in the DoD community have conducted research that is somewhat related to this problem. One reference to this literature is Taub and Thomas (1983).

decision whether to pass a system to full-rate production is of great consequence. Operational testing itself of ACAT I systems involves a considerable amount of money. Therefore, it is vital that operational tests be designed so that they are as informative and efficient as possible, and that they produce results that permit the best decisions to be made with respect to proceeding to full-rate production given the level of test funding.

Experimental Design for Operational Testing

Experimental design is concerned with how to make tests more effective. Design includes how to choose the number of test scenarios, the factors used to define each test scenario, at what level the factors are tested, the allocation of prototypes to scenarios, and the time sequence of testing. The field of experimental design has developed substantially since the work of Fisher, Box, Finney, Bose, Kiefer, Wolfowitz, their collaborators, and others (see, e.g., Steinberg and Hunter, 1984). In the twentieth century, industrial, agricultural, and military experimentation have raised a wide variety of problems, such as: how to accommodate various test constraints; how to accommodate trend effects; how to understand how a relatively large number of test factors affect system performance when one is limited to a relatively small number of test scenarios (e.g., fractional factorial and Latin Hypercube designs); how to learn about the effects of individual test factors and estimate performance well in a key scenario; and how to efficiently test to simultaneously inform about several measures. Methods that have been developed to address these and other problems have resulted in a rich technical literature. *Technometrics* is one of several journals that devote a substantial percentage of their pages to methods of optimal experimental design and the practice of experimental design in applied settings. Some of this literature is devoted to precisely the issue of how to design a test in a variety of scenarios to evaluate a system under development.

Many problems in experimental design are solved by identifying designs that are optimal for a stated criterion. Much of the classical theory of optimal design for a single measure yields asymptotically optimal designs for estimating a single unknown parameter. This theory for a single measure has limited application for several reasons: it is a large-sample theory, with unknown applicability for very small samples; the optimal designs often depend on the value of parameters (relating performance to test factors) that are costly or otherwise difficult to estimate; the optimality depends on the assumed model for the response being (approximately) correct; and most systems are judged using multiple measures, and an optimal design for one measure might not be the same as an optimal design for another.[6] However, even with limited direct applicability, the theory of optimal design yields insights into good designs for nonstandard problems that have features that approximate those for standard problems. So, although standard techniques will often be inadequate for very complicated systems, the principles developed from answers to related, simple problems can be helpful for suggesting approaches to the more complicated problems.[7] For example, one such principle is to test more where variation is greatest.[8]

As in industrial applications of experimental design for product testing, circumstances can conspire either to cause some test runs to abort or to require the operational test designer to choose alternative test events that were not planned. These unanticipated complications can affect the utility of a test design (e.g., by compromising features such as the orthogonality of a design, and thereby reducing the efficiency of a test; see discussion of these issues in Hahn, 1984). Again, techniques have been developed both to minimize the problems that result from aborting or altering of a test run and to adjust test designs when this occurs.

Unfortunately, the current practice of experimental design in the operational testing of defense systems does not demonstrate great fluency with this variety of powerful techniques that comprise the state of the art of applied experimental design, though there are exceptions. The result is inefficient test designs, wasted resources, and less effective decision making concerning the test and acquisition of defense systems. The knowledge of the existence of these techniques and this literature needs to be more widespread in the test and acquisition community. When it is appropriate, expert help will then be sought so that these techniques can be used to greater advantage in operational test design. We repeat Recommendation 4.4 from Chapter 4:

> **Recommendation 4.4: The service test agencies should examine the applicability of state-of-the-art experimental design techniques and principles and, as appropriate, make greater use of them in the design of operational tests.**

[6] A more recent Bayesian approach to the question of optimal design is, in contrast, a small-sample theory.

[7] An example is for a test design in which the response is assumed to be the slope of a linear function of a single input. The optimal design is to test only at both extreme values for this input (though one would generally include tests in between the extremes to investigate model misspecification). Now, if one is faced with a problem in which the response is a linear function of multiple inputs, and one is interested in estimating all regression coefficients, in a multiple regression test design problem one could use the principle from the case with a single factor to infer that it would be reasonable to oversample points with extreme values for one or more of the inputs. (Optimal theory can be applied if a criterion to be optimized is selected.)

[8] Understanding how to apply optimal design theory for nonstandard problems requires relatively advanced understanding of the field of experimental design; see recommendations in Chapter 10 regarding the level of statistical expertise needed in the DoD test community.

Experimental Design for Operational Testing: Special Problems

The remainder of this section describes some special problems raised in the application of experimental design to the design of operational testing of defense systems and the statistical techniques that have application to them.

Problem 1: Dubin's Challenge

Dubin's challenge[9] is how to select a very small number of test scenarios and then allocate a possibly moderately larger number of test events to these test scenarios in order to understand system performance in a larger number of scenarios.[10] That is, how should one allocate a given number of tests, t, to m different scenarios of interest, with t substantially smaller than m, both to best understand average system performance across the m scenarios and performance in the individual scenarios?

As a start, we examine a related problem. Considering the different scenarios as strata in a population, one knows from the literature on sample surveys that the best allocation scheme depends on the variability of the measure of interest, Y, within each scenario (σ_s^2), and the cost of testing in each scenario (see, e.g., Cochran, 1977:Ch. 5). Assuming first that the goal is to estimate the average performance across all scenarios of interest, possibly weighted by some probabilities of occurrence or importance, one can work out an optimal allocation scheme as a function of the costs, variances (which would have to be estimated, possibly through the use of small-scale testing), and weights. If one assumes that the strata are not related and that no borrowing of information across strata would be possible, this approach requires that t is at least as large as m. For a simple example, if the importance weights, variances (σ^2), and costs happen to be equal for all scenarios of interest, then equal allocation is optimal, each scenario receives t/m test units (assuming this is an integer), and the variance of the sample mean is equal to σ^2/t. However, one cannot estimate Y well in each individual scenario as the variance of those estimates is $m(\sigma^2/t)$.

If t is smaller than m, the only way to proceed is to assume that the test scenarios have features in common and use statistical modeling.[11] Modeling has

[9] Dubin's challenge is so named because the problem was suggested to the panel for study by Henry Dubin, technical director of the Army Operational Test and Evaluation Command.

[10] Note that focusing on a single measure is a major simplification, since most military systems have dozens of critical measures of importance. Clearly, if different measures have different relationships to the test factors, designs that are effective for some measures may be ineffective for others. We briefly discuss this problem below.

[11] One can show the benefits from modeling with respect to estimation for individual scenarios by a simple example: Suppose $m = t = 8$, and the eight different scenarios can be viewed as the eight possible combinations of three factors each at two levels. Suppose it is also reasonable to assume

TABLE 5-1 Missed Distance at Eight Test Scenarios for Simple Example

	Distance (in miles)			
Obscurants	1	2	3	4
0	6	10	20	32
1	15	23	32	40

the advantage of relating system performance across different scenarios, but there is the need to validate the assumptions used by a model.

We digress to demonstrate the advantages and disadvantages to be gained by combining information across scenarios for a related problem. Suppose one is interested in the performance of a missile system for four different distances (1, 2, 3, and 4 miles), and for two different levels of obscurants (denoted 0 and 1 for off and on). One performs one test for each of the eight scenarios represented by the crossing of these two factors. The (fictitious) missed distance for each of these eight tests is provided in Table 5-1.

We will discuss two models. The first model is less stringent than the second. The experimental design, assuming the models are correct, can be shown to reduce the variance of estimated performance by a substantial amount. However, this advantage comes at a potential cost, since there is a possible specification bias if the models are wrong. First, a two-way analysis of variance model that can be used to describe the data is

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} ,$$

where Y_{ij} is the missed distance (the response), μ is the overall mean, α_{i} (i = 1, . . . , 4) is the effect for being distance i from the target (where it is assumed that the sum of these equals 0), β_j (j = 0,1) is the effect from obscurants being at level 0 or 1 (again summing to 0), and e_{ij} represents an error term with unknown variance σ^2. (A model similar to this might have been useful for the design and evaluation of the Common Missile Warning System; see Appendix B.) This model assumes that each distance and the level of obscurance has an additive impact on the response. However, nothing else is assumed about how Y_{ij} changes

that the interactions among these three factors are all 0, the variance of the measure is σ^2 in each scenario, and the costs of testing are equal in each scenario. Then the optimal allocation is one test unit to each scenario. Here, even though the means for the measure of interest *Y* in each of the eight scenarios are all different, one can estimate each with a variance of $\sigma^2/2$ rather than σ^2, which would be the case if the interactions are not 0. Thus, the related structure provides additional efficiency that can be used through modeling.

as a function of the continuous variable distance (e.g., linear or quadratic). This model has six parameters (μ, α_1, α_2, α_3, β_1, and σ^2). The estimates from this model for the mean performance in an individual scenario have variance of $0.625\sigma^2$.

Second, one can also assume, in addition, that mean performance is a linear function of distance, which results in the regression model:

$$Y_{ij} = \alpha + \beta_{\mathrm{d}} \text{ (distance)} + \beta_{\mathrm{o}} \text{ (indicator for whether obscurant used) } + e_{ij} ,$$

where β_{d} is the impact on expected missed distance of a 1-mile increase in the distance from the target, and β_{o} represents the impact of the use of obscurants. This model makes a more restrictive assumption about the impact of distance from target on mean performance, but in return there are only four parameters to estimate (α, β_{d}, β_{o}, and σ^2), producing estimates of the mean of Y_{ij} in an individual scenario that have variance of either $0.475\sigma^2$ or $0.275\sigma^2$, depending on the tested scenario. (There are similar gains in predicting mean performance for some untested scenarios.) Recall that if one assumes nothing, including additivity, about the relationship between Y_{ij} and the scenario characteristics, the result is model predictions for Y_{ij} in individual scenarios that have variance σ^2. Therefore, these two models result in important decreases in the variance of estimates of mean performance. It is important to point out that instead of reducing the variance of the estimates of performance, one could instead have maintained the variance of the estimates of performance and used correspondingly smaller test sizes, with the associated costs savings. For instance, for analogous situations in which there are variance reductions proportionate to the drop from σ^2 to $0.475\sigma^2$, the sample size could have been approximately halved, while maintaining the same quality estimates. (Of course, this assumes that a halved sample size was still a feasible test design.)

Although it is not always easy to determine whether the additional assumptions are supported, especially given the small sample sizes of operational tests and the limited information about a system's operational performance prior to operational testing, having a test data archive (as recommended in Chapter 3) would often provide much of the needed information to support these types of models. Returning to the example, Table 5-1 shows that scenarios with a distance of 2 miles have roughly an average missed distance of 6 more than scenarios with a distance of 1, 16 for distance 3, and 25 for distance 4. These values of 6, 16, and 25 seem to suggest that distance from target has a linear impact on missed distance. Of course, there are risks in making this assumption if it turns out not to be true, and therefore all additional information that can address the validity of this assumption needs to be used to reduce this risk.[12] There is an extensive

[12] There are different risks depending on the degree to which the assumption is false, and certainly, the linear assumption is ultimately false in that it cannot hold for very short or very large distances.

literature on examining the validity of assumptions used in models that can be used to help reduce the possibility of choosing an inappropriate model.

We have shown that there are large potential gains in estimating performance by relating performance in different scenarios using statistical models, assuming the models are approximately true. Returning to Dubin's challenge, two specific approaches to solving it have been suggested by the panel which use statistical models to blend information across scenarios. One approach, making use of multidimensional scaling, is described in the panel's interim report (National Research Council, 1995). A second approach involves use of Bayesian optimal experimental design theory (Pollock, 1997). Both approaches need to be tried out on real data possibly with other alternative approaches and compared for situations where the answers are known.

Problem 2: Multiple Measures

Since there are often several dozen measures of performance or effectiveness of interest, especially for ACAT I systems, an operational test needs to be effective for estimating each of them. It is clear that interest in different measures has design implications since, for example, different environments stress different system components in different ways. However, producing a test design that can accommodate the needs of several measures is difficult.

As an initial step, one can reduce the number of measures by either setting priorities for the measures or combining them. (There is already a weak prioritization of the measures implicit or explicit in the analysis of alternatives—formerly cost and operational effectiveness analysis—and in the Test Evaluation Plan.) We proceed under the assumption that prioritization and combination still results in several measures of interest.

The theory of optimal experimental design for multiple measures is well established in some specific situations, but it is difficult to apply in many applied settings for the same reasons that it is difficult to apply optimal design for single measures. In addition, the breadth of application for multiple measures in terms of the problems that are solved is much more limited. Absent examination of individual applications and recommendation of specific techniques, there is a possible omnibus technique that should have utility. For each performance measure of interest, x_i, identify two quantities: the minimum operationally meaningful difference in a measure, Δx_i, and a priority weight, w_i. For example, suppose that one is estimating the hit rate of a missile, which is expected to be around 0.8. Assume that knowing the hit rate to within 0.05 is sufficient for deciding to proceed to full-rate production. Also suppose that the hit rate is the primary measure of interest in the system (system reliability might be another measure of interest). As a result, one wants the measure of the hit rate to represent 80 percent of the weight of the operational test. Then letting $var(x_i)$ denote the estimated variance of the mean of x_i (which is a function of the experimental design), we propose use of the following criterion: choose a design that minimizes $\Sigma w_i [var(x_i)/\Delta x_i^2]$ over each possible design.[13]

This approach is sensible since one is keeping each variance small relative to the square of its associated meaningful difference and these are then combined using the priority weights. Yet it might be difficult to apply this criterion in practice since linking the $var(x_i)$ to a specific experimental design may not be straightforward. Furthermore, $var(x_i)$ would often need to be based on information from developmental testing, test results, or modeling and simulation for similar systems, which argues again for both a test archive and for the use of screening or guiding tests.

Problem 3: Sequential Experimentation and Other Sequence and Time Issues

The accommodation of time and sequence considerations is another set of technical issues raised by the operational testing of defense systems. Time and sequence issues arise in several different ways. First, players involved in an operational test often "learn" how to be successful in the test as certain aspects of the test environment and the system become more familiar to them. Second, as mentioned above, as a test proceeds, various characteristics of the test environment will change, such as the weather, the degree of tiredness of the players, or the wear on the system. Test designs need to ensure that these time trends do not compromise the comparison of systems under test. Finally, sequential testing methods, not commonly used in the operational test of defense systems, might have wider applicability than current use would imply. We discuss each of these issues separately.

Player Learning and Other Time Trends The increased familiarity over time of users with the systems under test and the test environment is often referred to as player learning.[14] Player learning can create design problems, since,

[13] This criterion is a simple version of a slightly more general criterion. This version implicitly assumes that the measures are uncorrelated. The more general version would include additional terms in the summation involving the correlations between the estimated measures.

[14] A system that undergoes an operational test consists of three major components: the device, tactics and doctrine, and personnel. Too frequently, one thinks only of the device, and if the system fails it is attributed to that. However, in a combat environment, if the wrong tactics are used or if the personnel are ineffective, the system is very likely to fail, regardless of how good the device is. Using poorly trained personnel in an operational test can be as serious a problem as using a known defective piece of hardware. Therefore, adequate personnel training is essential to a successful test. At the very outset of test planning, provision should be made to minimize the effect of human variability. This can best be done by planning for appropriate training of the people operating the systems. The higher the skill level required, the greater the importance of adequate training. For

e.g., it can complicate comparisons of system performance between scenarios if the scenarios are tested at such different times that users have developed increased comfort with the system or the test environment between test events. Also, if a baseline system is included in the test as a control, the relative comfort and training of users on the baseline system compared with the new system raises some difficult design questions, since it is not clear what is meant by equal training or experience on the two systems, given some initial familiarity with the baseline system. Failure to address this question could provide an advantage to either system. The statistical principle is to make the comparison between the two systems, or two test scenarios, as unaffected by additional test factors as possible, but this can be difficult.

There are other trends besides player learning, especially environmental factors, that can change with time, and operational tests must also take these changes into account. Otherwise the comparison of the two systems may be affected (confounded) by a change in environmental factors, such as temperature or light, rather than real differences in the systems themselves. While there are often substantial logistic benefits to successive testing, the possibility of confounding is real. To address this type of problem, there are test designs that are balanced with respect to time trends; some examples are Daniel (1976) and Hill (1960).

Sequential Tests Sequential tests are tests that rely on results from previous trials to decide whether to continue testing. Sequential significance tests have the property that at a given confidence level and power, the expected test size will be smaller than the comparable fixed sample size test. In the application of operational testing, wide use of sequential testing could result in substantial savings of test dollars and a decrease in test time.

Unfortunately, such practical considerations as the need to obtain expedited analysis of test results and the scheduling of soldiers and test facilities make sequential designs difficult to apply in some circumstances. Acknowledging that there are practical constraints, the panel believes that these methods should be examined for more widespread use. The panel is unfamiliar with the extent to which these designs are used in developmental testing, but the above practical limitations should be less common in developmental tests. The panel is concerned that the technical demands of these tests contributes to their infrequent use. This would be unfortunate given their advantages in reducing test costs.

systems requiring highly skilled operators, testing should be conducted to assess whether the training is adequate. Force development tests and experiments are excellent tools for the assessment of the adequacy of training. There are times when even a “golden crew,” an extremely competent crew, is useful to examine the upper limit of the performance of a system. The data generated by such crews should be used carefully and not advertised as the “expected” performance of a system.

Problem 4: Different Systems, Different Tests

There are an extraordinarily wide variety of systems that can begin operational testing in any year. They include payroll management systems, which are entirely software; attack helicopters; airport surface repavers, and electronic warfare systems. To say that these systems need different tests is to state the obvious. How, then, can successful experience in test design and test evaluation be used on subsequent tests? The panel believes that a taxonomy or some taxonomic structure of defense systems with respect to their operational test properties would be useful. An attempt at creating a taxonomy of defense systems for operational test design is presented in Appendix C. When this initial attempt is refined, successful case studies for each cell in the finished taxonomy could be collected to help guide the major in charge of test design and evaluation by indicating techniques and tests that have worked previously. Of course, given the dynamic nature of defense systems and the understandable focus on the use of new technologies, any taxonomy will have to be flexible and change over time; it would need to be formally reconsidered on a regular basis. The taxonomy provided in Appendix C attempts to take this into account.

DETERMINATION OF SAMPLE SIZE USING DECISION THEORY

This section concerns the determination of the sample size of operational tests. Sample size is generally considered a component of test design, but sample size calculations are often ignored in current operational test design practice.

We start with the description of the justification for the test sample size contained in the Test and Evaluation Plan for the Longbow Apache (Longbow) Initial Operational Test and Evaluation (IOT&E) for the force-on-force phase of the operational test for this system. Here a replication is an opportunity to engage, i.e., a replication is an individual opportunity to fire at an “enemy” within a test event (in which there will be multiple opportunities). (We are discussing this as an example that is consistent with best current practice in the service test agencies. We are not commenting on its appropriateness for the specific application.) The sample size for the Longbow force-on-force test was justified by the following.

Consider the two hypotheses:

H_o: $p_l \le p_b$
H_1: $p_l > p_b + \delta$,

where:

p_l = the probability of hit by the Longbow team against the Red forces;

p_b = the probability of hit by the Apache (the baseline system) team against the Red forces; and

δ = the absolute magnitude of the difference between p_l and p_b that it is considered important to detect.

The argument assumes that a t-test for the equality of two proportions—i.e., the standardized difference (to have mean zero and variance one under H_o) between the observed probability of hit by the Longbow and the Apache, which is a function of the test sample size *n*—is used to measure the statistical significance of the improvement represented by the Longbow over the Apache. Let Z_c be the *c* percentile point on the standard normal distribution (that is, the area under the normal curve to the left of Z_c is *c*). We choose the critical value $Z_{1-\alpha}$ so that when $p_l = p_b$ this t-test has significance level α; i.e., the probability of rejecting H_o when it is true is α. We would like to choose the sample size, *n*, so that the probability of rejecting H_o at the alternative $p_l = p_b + \delta$ is $1 - \beta$ for some error rate β.

The required sample size *n* is then determined through use of the following formula (which makes use of the normal approximation to the binomial distribution):

$$n = 2*[\ (Z_{1-\alpha} + Z_{1-\beta})/d\ ]^2\ ,$$

where

$$d = (2 \arcsin \sqrt{(0.5 + \delta)} - 2 \arcsin \sqrt{0.5}\ ) \le (2 \arcsin \sqrt{p_l} - 2 \arcsin \sqrt{p_b}).$$

For $\alpha = .10$, $\beta = .10$, and $\delta = 0.10$, n = 325, which was roughly the sample size used in the test for both the control and the new system. Finally, in the computations above, 0.5 was substituted as a conservative estimate of p_b, and therefore $p_l = p_b + \delta = 0.6$, which gave an upperbound to the sample size to satisfy the above test criteria.

We note that in the above:

- the conclusions associated with the hypotheses are not important compared with the action to be taken based on the conclusion; and
- more generally, the power function is the probability of concluding in favor of H_1 for various values of p_l and p_b, which is primarily dependent on the sample size *n* and the difference $p_l - p_b$. (Making the difference $p_l - p_b$ unrealistically small will require a very large *n* to achieve high probability of rejecting H_o in favor of H_1.)

As a way of determining the sample size of operational tests, analytical procedures of this type have some advantages. The use of significance testing provides an objective means of determining sample size, which, given the various incentives of participants in the test process, can be very useful. The test explicitly presents the improvement needed to justify acquisition, which is when the difference δ between the baseline and the new system is 0.10. Also, the test is designed so that a system with a 10 percentage point improvement over the baseline system will be identified correctly (H_o will be rejected) with a probability of 90 percent. Finally, the binomial percentages are transformed using the arcsine square root transformation, which is useful since it helps to symmetrize or normalize the binomial distribution and stabilize the variance.

The argument presented above is typical of current best practice in the service test agencies. If the practice results in a test that would cost less than the funds earmarked for operational testing by the program manager, the test will go forward. However, as is more likely for ACAT I systems, if the test budget of the program manager cannot support this sample size, negotiations between the program manager, the service test agency, and DOT&E take place, with the program manager's test budget limit possibly modestly augmented, but very likely resulting in an operational test size that is substantially reduced from that based on this approach.

Some appreciation of the difficulty of achieving even modest goals in inferring, for example, reliability with limited sample sizes using binomial data, may be derived from the following example. Suppose that it is desired to be able to state with confidence at least 0.8 (which we call approving the system) that the reliability of a system is at least $r_0 = 0.81$, based on n trials. To achieve this result

TABLE 5-2 True Reliability r* Needed to Have Probability 0.79 of Deciding that the System Has Reliability of at Least 81% with Confidence of at Least 80%

n	S	s = S/n	r*	p
8	8	1.00	0.971	0.533
10	10	1.00	0.977	0.407
25	23	0.92	0.905	0.273
50	44	0.88	0.885	0.241
100	85	0.85	0.862	0.267
400	332	0.83	0.841	0.205

S = Number of successes required to certify at least 81% reliability with confidence at least 80%
r* = True reliability required to assure probability of at least 79% of certification
p = Probability of certification when true reliability is 81%

we need to have greater than or equal to S successes out of n trials, where S depends on n and is given in Table 5-2.

Now let us ask how reliable (denoted r*) the system must be so that we have probability at least $p_0 = 0.79$, that the data will lead to approving the system. Finally, what is the probability p of certifying the system when the reliability is $r_0 = 0.81$? Table 5-2 presents S, s = S/n, r* and p for various values of n.

From Table 5-2 it is clear that for small sample sizes, the true r* has to be very high for us to be moderately likely to approve the system. In particular, for n = 50, the true reliability has to be at least 88.5 percent for us to be only 79 percent sure of approving the system as having reliability of at least 81 percent, when approval requires only 80 percent confidence. In short, we need either large samples or modest goals in inference or both.

The above approach to determining test size ignores the costs associated with the decision at hand (as further discussed in Chapter 6). The program manager and test manager should address the question about setting test sample size by explicitly considering the costs associated with the decision as to whether to pass a system to full-rate production and the consequences of that decision. Two kinds of errors can be made: (1) passing a system to full-rate production when it was not (yet) satisfactorily effective and suitable (which would involve a broad spectrum of performances classified as unsatisfactory), and (2) returning a system for further tests or development when it was satisfactorily effective and suitable (which would again involve a broad spectrum of performances classified as satisfactory). The costs associated with these two kinds of errors can be dramatically different. Since the probabilities of these errors occurring are a function of the test sample size, one can justify test sample size by balancing the costs of additional testing against the reduction in the risk (probability of error multiplied by cost of error) achieved through additional testing. The principle is that the marginal cost of additional testing should be less than or equal to the marginal benefit gained through use of the information collected in the additional testing.

DoD needs to examine the feasibility of taking such a decision-theoretic approach to the question of "how much testing is enough." Although a cost-benefit comparison is at times difficult to make, the acquisition process will be improved through the attempt at explicitly making it. Other approaches to determining the size of operational tests, such as allocating a fixed percentage of program dollars to operational testing, do not use test dollars effectively.

Unfortunately, the decision-theoretic approach is difficult to apply since the costs associated with the wrong decisions are difficult to quantify and must be based, at least in part, on subjective judgments, which ought to be made explicit. For example, if one includes the decreased safety of military personnel as a cost either through an increased probability of a failure of the system itself or through a decreased probability of surviving a conflict given an ineffective system, the life of a soldier (or other service person) would have to be compared with test dollars. Also, the entire defense procurement budget is essentially fixed at the level of this argument. Thus, an increase in the test budget for one system would cause a decrease in another system's test budget. So a full decision-theoretic approach would require measuring the value of additional runs of any system currently in operational testing. However, it is typical in problems such as this that precise cost estimates are not very important since good procedures are relatively insensitive to variations in the costs. Another complication is that the accounting procedures used to track costs in test and acquisition make it difficult to determine how much was spent in developmental and operational tests (see Rolph and Steffey, 1994). This difficulty is partially understandable since costs of soldiers, etc., are difficult to attribute. However, in order to undertake any cost-benefit comparison, the costs of testing must be available for analysis. Therefore accounting and documentation of costs must be adjusted to enable these types of analyses to be carried out.

Some issues that could ultimately be addressed by the comparison of the marginal benefits and marginal costs of test events are:

1. When is it preferable to use a sequential method of testing, rather than a fixed sample size test? For example, should one use 80 test events in one set of tests, or instead use 40 test events, evaluate and possibly decide to stop testing, or decide to conduct 40 more test events?
2. When is it preferable to allocate test units to a variety of environments of interest, and when is it better to concentrate the test units in a single environment—possibly the most stressful environment? For example, should one use two test replications in the desert, two in the tropics, two in the Arctic, and two in mountainous terrain, or should one use five events in the desert and one in each of the remaining environments?
3. Similarly, when is it preferable to test two factors considered related to system performance, leaving the third factor constant over tests, and when is it preferable to also vary the third factor, thereby gaining understanding of the third factor's effect on system performance but reducing the understanding of the effect of the first two factors?

These are hard problems that will require experience to address. However, there will be a number of important ACAT I systems where a cost-benefit comparison will either clearly support, or argue strongly against, the budget provided by the program manager.

Finally, if the objective of operational testing is viewed as not only a statistical certification that requirements have been met, but also to gain the most information about operational system performance and identification of deficiencies in system design, the assessment of the benefit of additional testing will involve broader considerations, such as:

- identification of problems in operational use;
- improvement of tactics and doctrine;
- understanding of heterogeneity of units and performance of units in heterogeneous situations, including environment, tactical situations, etc.;
- relating of deficiencies in performance to operational conditions;
- understanding of training issues; and
- generally, improving the understanding of whether the basic design is flawed, even for focused applications, thereby saving money required for retrofitting after full-rate production has begun.

When the decision rule encompasses all of these considerations, the decision-theory argument suggested here is even further complicated, since the decision rule is more difficult to analyze. To better understand how to implement the comparison of marginal cost with the value of the information gained through additional testing, a feedback system is needed, in which the service test agencies maintain records of the costs of testing and the costs of any retrofitting, along with the precise test events that were conducted and the reason why retrofitting was required. Discovered defects or system limitations might then be traced back to inadequate testing. In this way, the process of comparing costs and benefits can be improved so that similar problems do not recur. This approach should be systematically investigated.

6

Analyzing and Reporting Test Results

Chapter 5 argued that substantial improvements in the cost-effectiveness of operational testing can be achieved by test planning and state-of-the-art statistical methods for test design. It was also noted that achieving the full benefit of improved test design requires a design that takes account of how test data are to be analyzed. This chapter focuses on how data from operational tests are analyzed and how results are reported to decision makers. The panel found potential for substantially improving the quality of information that can be made available to decision makers by applying improved statistical methods for analysis and reporting of results.

In its review of reports documenting the analysis of operational test data, the panel found individual examples of sophisticated analyses. However, the vast majority focused on calculating means and percentages of measures of effectiveness and performance, and developing significance tests to compare aggregate means and percentages with target values derived from operational requirements or the performance of baseline systems.

The panel found the following problems with the current approach to analysis and reporting of test results:

- While significance tests can be useful as part of a comprehensive analysis, exclusive focus on these tests ignores information of value to the decision process. There is often no explicit estimate of the variability of estimates of system performance. This makes it difficult to assess whether additional testing to narrow the uncertainty ranges is cost-effective. It also makes it difficult to evaluate the relative risk associated with moving ahead to full-rate production on a system that may still have problems, versus holding back a system that is performing acceptably.

- Decision makers typically are concerned not just with aggregate measures, but also with the performance of systems in particular environments. While the panel found this concern reflected in the analyses we examined, the reporting of results for individual scenarios and prototypes was almost always informal. The panel was often told that statistics could not be applied to the problem of performance of individual scenarios because sample sizes were too small. Analysts were frequently unaware of formal statistical methods and modeling approaches for making effective use of limited sample sizes.
- Information from operational tests is infrequently combined with information from developmental tests and test and field performance of related systems. The ability to combine information is hampered by institutional problems, by the lack of a process for archiving test data, and by the lack of standardized reporting procedures.

This chapter discusses problems with current procedures for analyzing and reporting of operational test results, and recommends alternative approaches that can improve the efficiency with which decision relevant information can be extracted from operational tests. The evaluation of the operational readiness of systems can be substantially improved by implementing these changes. In addition, money can be saved through more efficient use of limited test funds.

LIMITATIONS OF SIGNIFICANCE TESTING

Significance testing has a number of advantages for presenting the results of operational tests and for deciding whether to pass (defense) systems to full-rate production.[1] Significance testing is a long-standing method for assessing whether an estimated quantity is significantly different from an assumed quantity. Therefore, it has utility in evaluating whether the results of an operational test demonstrate the satisfaction of a system requirement. The objectivity of this approach is useful, given the various incentives of participants in the acquisition process. Certainly, if a system performs significantly below its required level, it is a major concern.

[1] In several places in this report, a result from a significance (hypothesis) test, specifically a t-test, is put forward as an operational testing output of primary interest to decision makers for a given measure of performance or effectiveness. When these are produced, that is the role they can play (besides often being used to determine a statistically-based sample size in sample design). However, we point out that in many or most cases, summary statistics (such as means or percentages, especially when they exceed a required level) are viewed as sufficient for input to the decision process; use of significance testing is not customary. Even though significance testing is not uniformly applied, the panel views it as representing current best practice, and that is what we choose to comment on.

However, significance testing has several real disadvantages as the primary quantitative assessment of an operational test: knowing only the acceptance or rejection of a test's outcome simply does not provide enough information to decision makers. The panel has found several problems with the reliance on significance testing, which are detailed in the remainder of this section.

Significance Tests Focus Inappropriately on Pass/Fail The decision maker is typically interested not just in whether a system has passed or failed, but in what is actually known about how well the system performed. Did it almost pass? What range of performance levels are consistent with the test results? How much uncertainty still remains about its performance? What are the relationships between characteristics of the environment and test results? If there is doubt about whether the system meets a requirement, what is the operational significance of that doubt? What is the impact on the system's mission from any doubt the test casts on whether it meets the stated requirement?

Significance Tests Answer the Wrong Question A significance test answers the question, "How unlikely is it that the results, as extreme or more extreme than I have observed, would occur if the given hypothesis is true?" Consider the problem of testing whether a missile system meets a specified requirement for accuracy. The hypothesis selected as the "null hypothesis" might be that the system fails to meet a minimum acceptable level of performance that would justify acquisition.[2] What is reported to the decision maker is whether the system passed or failed the significance test (passing the test would mean the null hypothesis was rejected).

In the defense testing context, the term "null hypothesis" has sometimes been used to indicate the compound hypothesis of performance at or above the required level—which already represents an improvement from the baseline or control system's performance—with rejection indicating a substandard performance. This is a nonstandard use of the term. A preferable way of considering the decision structure of the testing problem is to consider the various hypotheses in a symmetric way by setting values $a1$ and $a2$ such that if the system's performance is less than $a1$, the system should be considered deficient; if the system's performance is greater than $a2$, it should be considered satisfactory; and if the system's performance is between $a1$ and $a2$, one should be indifferent to the assessment. This is a more natural framework to this decision problem. Related

[2] We refrain from using the term null hypothesis as much as possible, though we are obligated to do so at times since it is often used by the test community. The expression "null hypothesis" is usually reserved in the statistical literature to indicate "no effect" or "no improvement," when the expectation is that a substantial improvement has been made, but this needs to be objectively demonstrated.

to this approach, one could set the null hypothesis to be that the system is less than a minimum acceptable level of performance that was below the required level but above the level of the baseline. Then, the power of the test at the required level of performance should be relatively close to one for an effective test. The null hypothesis should not generally be set equal to the performance of the baseline system since only a modest improvement over the baseline system might not be worth acquiring. Also, the null hypothesis should not be that the performance is less than the requirement, since there is then a substantial probability of rejecting systems worth acquiring.

We believe that decision makers often interpret a passing result as confirming that the system probably meets its requirement and a failing result as an indication that the system probably does not meet its requirement. This interpretation is sometimes valid, but there are times when it is not. The correct interpretation of a passing result, when the null hypothesis is that the performance is less than minimally acceptable, is that the results that were observed are inconsistent with the assumption that the system does not meet the minimal acceptable level. While it is clear that the question answered by a significance test is *related* to a problem decision makers care about (whether the system meets its requirement), the significance test does not directly address this question. The panel found that test analysts and decision makers alike often failed to understand this point.

Significance Tests as Applied Often Fail to Balance the Risk of Rejecting a "Good" System Against the Risk of Accepting a "Bad" System This issue is not a criticism of significance testing per se, but of how it is often applied. In deciding whether to pass a system, one can make two different types of error: one can "fail a good system" or "pass a bad system."[3] The two types of error need to be compared with each other and both related to the cost of testing. Assuming that one hypothesis is that the system meets a minimum acceptable level of performance, making the test criteria more stringent increases the first kind of error and reduces the second; making the criteria more lenient does the reverse. By setting the cutoff between passing and failing, the tester trades off one type of error against the other. There are certainly situations where one error is more problematic than the other and so should be set lower. For example, with a missile system, requiring more hits on test shots to pass increases the probability of rejecting a good system, but reduces the probability of accepting a bad system, and vice versa. To reduce both types of error simultaneously requires more test shots and more test funds.

[3] The quote marks indicate both that a number of factors go into decisions to pass a system to full-rate production and that a system that satisfies or fails to satisfy a requirement is not necessarily worthy of being passed or held up. However, the language is useful for this discussion.

The panel stresses the importance of considering both types of error when determining pass-fail thresholds and of balancing cost against the ability to discriminate between levels of performance when selecting test sample sizes. What we found, however, is that test thresholds are usually set to achieve error levels set by arbitrary convention. We often heard comments to the effect that although everyone knows that statistical tests *should* have significance levels of .05 or .10, operational testers were forced by resource limitations to use significance levels of .20. Although viewed as regrettably high, .20 is often commonly used. Similarly, tests with power .80 are also commonly used. We found it rare that test designers made explicit tradeoffs between the two types of error.

Sufficiently Powerful Significance Tests May Require Unattainable Sample Sizes The most frequent criticism of the use of statistics in operational test and evaluation was that tests with a high ability to statistically discriminate good from bad systems cost too much. We often heard comments such as, "We can't afford the 'statistical answer,' " or "We can't use statistics because we can't afford large samples." There was a concern that test budgets do not allow the tests to make use of more standard error levels for many situations. Yet, the use of tests with significance levels of 20 percent and power of only 80 percent for important alternatives, which seems typical, is troubling to system developers and the general public: using a test that is designed to incorrectly identify as deficient one-fifth of a group of systems that actually deserve to pass will result in unnecessary, costly retesting or further development; using a test that is designed to incorrectly identify as satisfactory one-fifth of a group of systems that actually deserve to fail will result in the acquisition of deficient systems and potentially expensive retrofitting. Unfortunately, when faced with the complex task of assessing the tradeoffs, the acquisition community has reduced the emphasis on the use of statistics for these problems, precisely for those situations where statistical thinking is most critical to making efficient use of the limited information that is available on system performance, system variability, and how performance depends on test conditions. It is extremely important to understand the tradeoffs between decreases in both error probabilities and increases in test costs, and when this trade-off supports further testing, and when it does not.

INSUFFICIENT FORMAL ANALYSIS OF INDIVIDUAL SCENARIOS AND VARIABILITY OF ESTIMATES

The focus on the reporting of significance tests as summary statistics for operational test evaluation and their prominence in the decision process deemphasizes important information about the variability of system performance across scenarios and prototypes. For example, understanding which scenarios are the most challenging helps indicate how system performance depends on characteristics of the operating environment and which types of stresses are the most problematic for the system. Because of smaller sample sizes for individual scenarios than for the overall test, there will be more uncertainty about performance estimates in particular conditions than for an overall aggregate estimate of performance. Understanding this variability is important to the question of whether a system's poor performance in a given environment should be attributed to a serious problem that must be addressed or simply to the expected variability in test outcomes. Unfortunately, given the resources that can realistically be devoted to testing, answering this question conclusively may be difficult for many systems.

Information on which prototypes performed worst may indicate a faulty manufacturing process. To address this, one may want to analyze the reduced test data set through excluding data for some prototypes, which would demonstrate the performance that one might expect if the manufacturing process were improved. It is our impression that scenario-to-scenario variability will dominate prototype-to-prototype variability for the vast majority of systems, but this impression would be useful to investigate to identify the kind of systems for which this is not true.[4]

FAILURE TO USE ALL RELEVANT INFORMATION

Chapter 4 discussed the need, especially given the cost and therefore limited size of much operational testing, for making use of data from alternative sources (tests and field performance of related systems and developmental tests of the given system). Except for the pooling of hours under operational testing with those from developmental testing for reliability assessment (see Chapter 7), the panel has seen little evidence, in the evaluation reports it has examined, of a consistent strategy of examining the utility of these other data sources, an analysis of how they might best be combined with information from operational test, or the use of these alternate sources of information.

This failure has at least four explanations: (1) there is a justifiable concern as to the validity of this combination of information from different experiments; (2) there are perceived legal restrictions, stemming from this concern, to the use of developmental test and other information in evaluation of a system's operational performance; (3) there is no readily accessible test and field use data archive; and (4) the testing community lacks the expertise required to carry out more sophisti-

[4] It is common for reliability requirements to specify two reliability thresholds, the first for the full-rate decision, and the second for some time later when the production process has stabilized. So, while operational test articles may be representative of the current production process, they may not reflect future methods of production. An interesting question is whether the operational test community should be more involved in the monitoring of the degree to which these operational requirements have been demonstrated.

cated statistical analyses. Although each of these arguments involves real hurdles, each can be overcome.

As we note in other chapters, there is no archive for test data and, therefore, it is often very hard to combine information from operational testing with that from developmental testing or to combine developmental and operational testing with that from test and field experience of related systems (possibly systems with identical components). The lack of use of such information is often inefficient since, when the test circumstances are known, each of those sources can provide relevant information about operational performance. An archive would also make it possible to share information about problems in measurement techniques and threat simulation: Are there good ways to present data or to analyze missile hit distances? Are there useful things to know about design effects (e.g., learning) in trials run over multiple days? We note that there are (at least perceived) legal obstacles to using some of these sources of information for operational evaluation.[5] Finally, even if these other data sources were available and accessible, there is a scarcity of statistical modeling skills available in the test community that would be needed to make full use of this data; this issue is discussed in more detail in Chapter 10.

LIMITED ANALYSES AND STATISTICAL SKILLS

There are also some limitations concerning the use of sophisticated statistical techniques to fully analyze the information provided by operational tests and alternative data sources. These limitations stem from the acquisition process itself and the lack of requisite skills in those charged with the evaluation of operational tests. The analysis of operational testing data is typically accomplished under strong time pressure. Systems have often been in development for as much as a decade or more and when operational testing is concluded, especially if it is generally believed that the system performed adequately, there is understandable interest in having the new system produced and available. There is therefore pressure to carry out the evaluation quickly. Given current schedules, large amounts of data need to be organized, summarized, and analyzed in as little as two months. This schedule greatly reduces the time to investigate and understand anomalous results, to try out more sophisticated statistical models, to validate any assumptions used in models, to explore the data set graphically, and to generally understand what information is present in the operational test data set. This work goes well beyond the computation of summary statistics and the asso-

[5] The panel was unable to determine the precise extent to which developmental test data and data from related systems are actually or merely perceived to be restricted for use in operational evaluation for the various services. It is known that some operationally realistic testing must be conducted for ACAT I systems.

ciated significance tests which are relatively straightforward in their application. The section above on significance testing describes a number of analyses that would be worth undertaking. Failure to carry out these analyses due to time pressures may result in a failure to completely represent the information provided by operational testing and therefore lead to an incorrect decision.

The individuals charged with operational test evaluation often have limited expertise with the analysis of large, complicated data sets. The ability to examine a large data set, understand what is in it, identify its major features and more subtle details, and represent what has been discovered in an accessible form to decision makers, requires considerable training and guided experience. Knowledge of some of the more sophisticated modeling tools—for instance, nonhomogeneous Poisson processes, classification and regression trees, robust analysis of variance, and others that are likely relevant to the analysis of operational test data—are unlikely to be familiar to those currently involved with operational test evaluation. This limited statistical expertise is very understandable, given what test managers need to know about the acquisition process and the system under test. In addition, given the typical career path, that is, how long test managers are likely to spend with their operational test agency, the test manager's lack of statistical expertise is probably impossible to overcome.

RECOMMENDATIONS FOR IMPROVING ANALYSIS AND REPORTING

Perform More Comprehensive Analysis

Test evaluation should provide several types of decision relevant information in addition to point estimates for major measures of performance and effectiveness and their associated significance tests. First, a table of potential test sizes and the associated error levels for various hypothesized levels of performance for major measures should be provided to help guide decisions about the advantages of various test sizes. Since this could be constructed before testing has begun, this should be a required element of more detailed versions of the Test and Evaluation Master Plan, with additional knowledge acquired during the test about distributions and other characteristics used to update the table in the evaluation report. This tabulation of the impact of test size on error levels of a significance test is referred to as the test's operating characteristics. We advocate its calculation in Recommendation 4.2 (in Chapter 4); for an example of a table of operating characteristics, see Box 6-1.

In addition, there should be an understanding of the costs, as a function of system performance, one faces by making the decision on whether to proceed to full-rate production. That is, the benefits and costs of passing systems with different levels of performance should be compared with the benefits and costs of failing to pass such systems. Explicit treatment of benefits and costs is far better

Box 6-1 An Example of a Table of Operating Characteristics

Let us assume that we are designing an operational test for a missile system in which the minimum acceptable level of performance, the probability of successfully hitting a target for a specific scenario, is .8. So, any performance less than .8 is not worth acquiring. Let us further assume that the actual requirement is that the system achieve a success rate of .9. We decide to use a test with a significance level of .10, i.e., the consumer's risk is .10, or the probability of rejecting a system with a true success rate of .80 (or lower) is .90 (or more). The only thing still undecided is the sample size for the test. Given this, for various sample sizes, we can determine how many successes are needed to have a test at significance level .10. The test is as follows:

We accept that the true success rate of the system is greater than .8 if it produces the number of successes, "s", in "n" trials provided below. For example, a test with 15 replications requires 15 successes to consider the system as having at least a success rate of .8.

Given this test, assuming various performance levels above and below the required level, it is useful to determine what the probability is that a system with these levels would pass the test defined by the meeting or exceeding of these given number of successes. The table of operating characteristics is the result of this calculation, in our case produced for the true success rates of .75, .80, .85, .90, and .95.

Probability of Accepting a System Using a Test With "n" Replications*

		Actual Success Rates				
n	s	.75	.80	.85	.90	.95
15	15	.013	.035	.087	.206	.463
30	28	.011	.044	.151	.411	.812
50	45	.007	.048	.219	.616	.962
75	65	.010	.093	.418	.874	.999
100	86	.005	.080	.457	.927	1.000
200	168	.001	.099	.696	.997	1.000

This table provides important information about the planned test, which might affect a decision as to the amount of resources to allocate. For example, if the test size is 30, the probability of accepting a system with true success rate of .9—equal to the requirement—is only .411, not very good. Further, the probability of accepting a system with true success rate of .85 is only .151, which is a system that exceeds the minimal acceptable level of performance. If we instead move the test size up to 100, the probability of accepting a system with true success rate of .9 is now .927, and the probability of accepting a system with true success rate of .85 is now .457, possibly still lower than one would want, but one might decide that this was a good allocation of test resources. The important point is that one can look at the trade-off between test performance and its cost based on test size; and while tests with better performance cost more, one can examine this trade-off against the costs of making the wrong decision about proceeding to full-rate production.

*The granularity of the test has a substantial impact on the tabulated values. For example, the column associated with a true success rate of .8 would ideally be equal to .1 for all n.

than applying standard significance levels to balance the probabilities of the two types of errors (failing to pass an effective and suitable system versus passing a deficient system). Methods from decision analysis provide a natural framework for addressing such benefit-cost comparisons (see, e.g., von Winterfeldt and Edwards, 1986; or Clemen, 1991). The panel acknowledges that explicit quantitative assessment of the benefits and costs of decision alternatives can be controversial in a situation in which different parties have differing objectives, though this framework is natural for identifying the source of these differences which can lead to their resolution.

Each point estimate, including those for individual scenarios and for more targeted analysis, should be accompanied by a confidence interval (or related construct) that provides the values that are consistent with the test results. There should be an accompanying discussion of what the implications would be with respect to acquisition if the endpoints of the confidence intervals for the major measures of performance and effectiveness were the actual performance levels of the system.

Furthermore, test analyses should consider variability across test scenarios and system prototypes. The panel found that while results for individual scenarios or test conditions may be reported, their consideration is usually informal. The panel acknowledges that limits on sample sizes often constrain the ability to analyze individual scenarios. However, there exist sophisticated statistical methods that often can be used to extract useful information from limited sample sizes. For example, hierarchical Bayesian analysis can be used to develop estimates for individual scenarios, and to assess the variability of these estimates. Limited data for an individual scenario is augmented both by prior information (e.g., from developmental tests, or test or use data for related systems) and by "borrowing strength" from data obtained in other scenarios, to produce as accurate an estimate as possible. Statistical modeling approaches such as regression or analysis of variance can be used to increase the efficiency with which information about individual scenarios can be extracted from small sample sizes. Such analyses should be accompanied by an evaluation of the adequacy of the assumptions underlying the models and the impact of violations of the assumptions. As noted in Chapter 5, test design must consider the analyses to be performed in order to allocate resources as effectively as possible.

As mentioned above, decision makers need to be informed as to what range of between-scenario, between-prototype, and other types of heterogeneities are consistent with the operational test results so that they can understand the evidence for variability in system performance. For example, if an estimated hit rate of 0.80 that met the requirement was achieved by having an observed hit rate of 1.00 in three scenarios but a hit rate of only 0.20 in the fourth scenario, that might affect a decision about full-rate production. If this kind of analysis requires the use of statistical modeling, it should be performed, and the results of the modeling should be communicated in a way that is accessible to decision makers. All important assumptions underlying such modeling should be described, along with the support for those assumptions and the robustness of the methods if the assumptions do not hold.

In addition, there should be investigation of the existence and analysis of any discovered, anomalous results: features of the test data that show patterns that are unexpected. Such results may also suggest further testing to help understand their cause. Anomalous patterns could include time or order effects of unknown source, time of day effects, lack of consistency of results across user groups, and other counterintuitive results.

Finally, related to the need for use of confidence intervals, if modeling and simulation is used, there should be a full discussion of the extent to which the simulation model has been validated for the current purpose and an assessment of uncertainty due to the use of the simulation model to assist in evaluation, especially uncertainty due to model misspecification. One common technique for this purpose is sensitivity analyses to determine how changes in assumptions would affect the resulting estimates. Presentation of additional uncertainty due to the use of simulation models can be done using uncertainty intervals (discussed in Chapter 9).

The test community should adopt the view that the purpose of statistics in this setting is to provide tools for helping to extract as much information as possible from a test of limited size (and from related information sources). Statistics is useful for planning and conducting tests and interpreting test results to provide the best information to support decision making. Adopting this view means moving away from rote application of standard significance tests and toward the use of statistics to estimate and report both what is known about a system's performance and the amount of variability or uncertainty that remains. Rather than thinking of a significance test as a comprehensive evaluation of a system's performance with respect to a measure of interest, significance testing instead should be thought of as a method for test design that is very effective in producing operational tests that provide a great deal of relevant information and for which the costs and benefits of decision making can be compared. Significance testing is one of several methods of test evaluation that need to be used together to provide a comprehensive picture of the test results to inform the decision about passing to full-rate production.

Combine Information from Multiple Sources

Given the resource constraints inherent in operational testing of complex systems, it is important to make effective use of all relevant information. As noted above, the panel found that information from developmental tests, as well as operational and field performance of related systems, is rarely used in any formal way to augment data from operational tests, other than pooling of developmental and operational test data for reliability assessment.

Certainly, use of data from alternative sources can entail considerable risk, given the substantive differences in test conditions and sometimes in the systems (say, with a shared component) whose information is being combined. If the argument could be supported that data from alternative sources (for example, developmental test) was directly relevant in the sense that the data could be pooled with the operational test data, that would be a reasonable method for combining information. However, this naive sort of combination (for example, reliability estimation) will often result in strongly biased estimates of operational performance because of the unique properties of each system and the different failure modes that occur in developmental, in contrast to operational, test. As a result, appropriate combination of this information will usually require sophisticated statistical methods, in addition to a substantial understanding of the failure modes for the system. For example, knowledge of the reliability of a component from a related system can be used to help understand the reliability of the same component in a different system, but aspects of how it is used—such as the impact on the weight of the overall system through addition of the component—may cause different problems in the new system than were experienced in the previous application.

Especially with respect to suitability issues, but also with respect to effectiveness, knowledge of how a system performed in developmental testing could provide some qualitative information, such as which components were most error prone and which scenarios were most difficult for the system. More quantitatively, the measurement of the reliability or effectiveness of the system could possibly be broken into stages, with operational and developmental testing used to estimate the probability of success for each stage. This approach was taken in the assessment of the reliability of the O-rings in the space shuttle (Dalal et al., 1989).

While we are sympathetic with the need to be extremely careful in combining data from different types of tests, methods now exist for combination of information that reduces the risk of inappropriate combination, and these methods could be successfully applied in a variety of situations. (See National Research Council, 1996, for a description of many of these techniques.) When properly supported by statistical and subject-matter expertise, application of these techniques could both improve the decision process concerning passing to full-rate production and also help to save test funds.

To pursue this greater use of relevant information, expert assessments of the utility of the system's developmental tests and information from related systems should be included in the operational evaluation report to justify a decision either to use or not use such information to augment operational evaluation. If the decision is made to use the additional information, the validation of assumptions of any statistical models used to combine information from operational tests with that from alternative sources should be carried out and communicated. In addition, examination of the validity of any system-based assumptions that are used, such as linking system reliability to component reliability, should also be carried out and reported.

State-of-the-art methods currently exist and are being constantly refined and expanded to permit combination of information from disparate sources; see, for example, Carlin and Louis (1996) and Gelman et al. (1995). The applicability of these methods to the evaluation of defense systems under development remains unexplored, with substantial potential benefits for improving the decision process by making it more efficient and effective. Such methods are beyond the statistical expertise of the typical test manager, and in fact carrying out these methods in new applications often pose difficult research problems. (Development of increased access to statistical expertise is discussed in Chapter 10.)

The panel believes that the combination of information from developmental tests and from tests of other systems is particularly promising for estimating system reliability (suitability), since our experience suggests that system effectiveness is more specific to an individual system than suitability. However, the use of these methods should be examined for measuring both suitability and effectiveness.

An important final point is that if the new paradigm outlined in Chapter 3 is adopted, the use of screening or guiding tests will facilitate combination of information, since these tests will then have operational aspects that will reduce the risk from combination.

> **Recommendation 6.1: The defense testing community should take full advantage of statistical methods, which would account for the differences in test circumstances and the objects under test, for using information from multiple sources in assessing system suitability and effectiveness. One activity with potentially substantial gains in efficiency and effectiveness is the effort to combine reliability data from different stages of system development or from similar systems.**

Archive Data from Developmental and Operational Tests

In order to facilitate combination of information, the results of developmental and operational test and evaluation need to be archived in a useful and accessible form. This archive, described in Chapter 3, should contain a complete description of the test scenarios and methods used, which prototypes were used in each scenario, the training of the users, etc. To do this in a form that will be useful across services, standardization of the reporting will be needed. In addition, feedback from the performance of a system in the field can be used to inform as to whether the combination of information produced improved estimates of operational performance; see Recommendation 3.3 (in Chapter 3).

SUMMARY

The objectives of an expedited but thorough analysis are not necessarily in conflict. Analysis can begin much earlier using preliminary test data, or, better yet, data from the screening or guiding tests recommended in Chapter 3. As a result, data base structures can be set up, analysis programs can be debugged, statistical models can be (at least partly) tested and assumptions validated, and useful graphical tools and other methods for presentation of results identified. This would greatly expedite the evaluation of the final operational test data. However, when time is not sufficient to permit a thorough analysis, ways should be found to extend an evaluation if it can be justified that there is nonstandard analysis that is likely to be relevant to the decision on the system.

The members of the testing community are committed to do the best job they can with the resources available. However, they are handicapped by a culture that identifies statistics with significance testing, by lack of training in statistical methods beyond the basics and inadequate access to statistical expertise, and by lack of access to relevant information. By addressing each of these deficiencies, the test evaluation reports will be much more useful to acquisition decision makers.

The panel advocates that the analysis of operational test data move beyond simple summary statistics and significance tests, to provide decision makers with estimates of variability, formal analyses of results by individual scenarios, and explicit consideration of the costs, benefits, and risks of various decision alternatives.

7

Assessing Operational Suitability

Fielding operationally suitable systems is a prime objective of defense acquisition. A suitable weapon system is one that is available for combat when needed, is reliable enough to accomplish its mission, operates satisfactorily with service personnel and other systems, and does not impose an undue logistics burden in peacetime or wartime.[1] As noted above, operational test and evaluation is statutorily required to assess the effectiveness and suitability of defense systems under consideration for procurement.

Scarce resources, increasing technological complexity, and increasing attention to the life-cycle costs of defense systems underscore the need for assurance of suitability and its elements. Experience in the Department of Defense (DoD), similar to that of private industry, shows that the life-cycle maintenance cost of a major system may substantially exceed its original acquisition cost. For example, the total procurement cost of the Longbow Apache helicopter is estimated at $5.3 billion, which is slightly more than one-third of the total estimated 20-year life-cycle cost of $14.3 billion.[2]

[1] This informal definition is borrowed from Bridgman and Glass (1992:1). For purposes of test and evaluation, operational suitability is defined officially in DoD Instruction 5000.2 as “the degree to which a system can be placed satisfactorily in field use with consideration given to availability, compatibility, transportability, interoperability, reliability, wartime usage rates, maintainability, safety, human factors, manpower supportability, logistics supportability, natural environmental effects and impacts, documentation, and training requirements.”

[2] These estimates, in constant fiscal 1994 dollars, are provided in Annex D of the Longbow Apache Test and Evaluation Master Plan, which cites December 1993 estimates from the Longbow Program Office and the President’s fiscal 1995 budget as the original sources.

Suitability deficiencies have been responsible for many of the field problems with newly acquired systems and have generated concerns about the operational readiness of certain military capabilities. Concern about the department's success in fielding suitable systems was expressed in an October 1990 memorandum from the Deputy Undersecretary of Defense for Acquisition. A 1992 Logistics Management Institute review of seven recently fielded systems found that "several systems have not achieved their availability goals, and they consume significantly more logistics resources than anticipated" (Bridgman and Glass, 1992:ii). That study also found that crucial suitability issues are not adequately identified early, or addressed in operational test plans. Such concerns and findings have led to calls for improved assessment of operational suitability. In this chapter, we discuss statistical issues related to the conduct of operational suitability tests and their evaluations and related information-gathering activities.

SUITABILITY, TESTING AND EVALUATION, AND STATISTICS

In considering the role of statistical methods and principles in the assessment of operational suitability, it is important to establish first the context of suitability assessment and its relationship to operational test and evaluation. The judgment that a system is "suitable" implies confirmation that the system's use can be adequately supported in its anticipated environments. A comprehensive study of a system's suitability should include, for example, considerations such as its compatibility with other systems in current use, transportability, supportability (in terms of logistics and manpower), reliability, availability, and maintainability. A definitive suitability assessment addresses all of these matters, reporting both on the outcomes of formal tests and on informal observations and judgments made by those charged with the system's evaluation.

The suitability of a system should be assessed from early development through fielding, even though all elements of suitability may not be demonstrable at all times. The reliability of key equipment, especially critical components, can be demonstrated relatively early in a system's development, but some questions about logistics supportability can only be definitively answered after a system is fielded. Because the assessment of suitability after an operational test is performed is limited by the scope of such testing, it may be desirable to augment the test results with modeling and simulation, test and field data from other sources, and related analyses.

The challenges faced in designing and interpreting results of operational tests underscore the need for, and the potential value of, applying sound statistical methodology. For virtually every aspect of operational suitability, fundamental statistical questions arise concerning the duration and conditions of testing, the measurement and processing of suitability data, and the use of information from other sources, such as developmental tests and simulation models, in the design and subsequent analysis of operational tests. Possible test augmentation procedures involve such advanced statistical tools as hierarchical models, reliability growth models, and accelerated life testing.

Difficulties frequently encountered in suitability assessment can be overcome through a combination of: tests dedicated to providing information on operational suitability; other sources of information about the suitability of the system (e.g., developmental tests, data from similar systems); and statistical analysis and modeling to integrate all relevant sources of reliability, availability, and maintainability (RAM) information about the system in question. Such activities are currently undertaken in the defense testing community with varying degrees of frequency and effectiveness.

Although the panel has gathered information on, and has had substantial exposure to, many different aspects of the military services' practices in assessing the suitability of a potential defense procurement, we have restricted our formal review to the elements of operational suitability in which the statistical issues appear most prominently: reliability, availability, and maintainability.

RELIABILITY, AVAILABILITY, AND MAINTAINABILITY

Basic Definitions

Reliability is the probability that a system (machine, device, etc.) will perform its intended function under appropriate operating conditions for a specified period of time.[3] Clearly, component reliability is not the same as system reliability. The various services have their own definitions of reliability that are similar to the one presented here (see, e.g., U.S. Department of Defense, 1994b:7-1).

Availability has been defined many different ways, for example: "a measure of the degree to which an item is in an operable state at the start of a mission when the mission is called for at a random point in time," and "the proportion of time a system is either operating or is capable of operating when used in a specific manner in a typical maintenance and supply environment" (U.S. Department of Defense, 1994b:7-3).

Maintainability has been defined to be "a measure of the ability of an item to be retained in, or restored to, a specified condition when maintenance is performed by personnel having specified skill levels, using prescribed procedures and resources" (U.S. Department of Defense, 1994b:7-2).

Each definition requires the use of a set of methods in order to make appropriate and statistically supportable conclusions about its numerical value.

[3] The term time is used in a generic sense in this chapter and can refer to other measurements, such as mileage or cycles of operation.

Statistical and Broader Methodological Issues

RAM assessment is among the most statistically intensive subjects of investigation in the entire defense acquisition process. Six methodological issues underlie our concerns about both the design and the evaluation of tests that serve as sources of operational RAM data and the basis for reaching conclusions about RAM:

1. the relative prominence of effectiveness, rather than suitability (particularly, RAM) considerations in designing operational tests;
2. the applicability of commercial industrial practices and standards concerning system reliability, availability, and maintainability;
3. the questionable validity of some statistical assumptions commonly made in determining test resource requirements;
4. the subjective scoring of “mission critical” failures in processing data for RAM evaluation;
5. current and potential uses of multiple sources of RAM information in the design of operational tests and in the analysis of test results, including the collection and use of field data for continuing RAM assessment after a system is acquired; and
6. current and potential uses of special statistical methods for (a) assessing reliability changes over time and (b) using accelerated stress testing methods for assessing reliability.

The first two issues relate to key concepts, processes, and organizational structures that affect how RAM assessment can better realize potential gains from improvements in statistical practice. The last four issues relate to technical aspects of statistical practice in the design, execution, and analysis of related operational testing. In the remainder of this chapter we assess the current state of related operational testing in the services with respect to these issues and put forward recommendations aimed at strengthening its overall quality, efficiency, and utility.

Considerations in Operational Test Planning

Despite the major impact of suitability problems on a system’s life-cycle cost, shortcomings with respect to measures of effectiveness are generally regarded to be more “serious” problems for prospective defense systems than are deficiencies with respect to RAM-related measures of suitability. System effectiveness is a sine qua non: without its demonstration, further development and procurement of a prospective system is insupportable. Yet a lack of equally rigorous RAM testing and evaluation can quickly lead to unacceptable field performance of the system.

Plans to assess operational suitability in support of a procurement decision must consider two prominent questions: (1) How can one be sure the key RAM issues have been identified? (2) How does one make specific decisions with a reasonable degree of confidence about these issues, especially those concerning system availability and maintainability? Answering these questions is particularly important in view of deficiencies in recent field experiences with newly acquired systems. Such deficiencies raise concerns not only about military readiness, but also about opportunity costs. Unexpectedly high costs in maintaining a system after procurement reduces the resources available for additional military capability or further system development. The assessment of suitability—and RAM in particular—deserves increased emphasis at all stages of a system's life cycle, including operational test and evaluation.

> **Recommendation 7.1: The Department of Defense and the military services should give increased attention to their reliability, availability, and maintainability data collection and analysis procedures because deficiencies continue to be responsible for many of the current field problems and concerns about military readiness.**

A primary objective of such increased attention should be to improve the identification of key RAM parameters and the specification of associated requirements. These key parameters could cause unacceptable system performance or cost because they consume significant logistics resources or pose risks associated with major changes in system technology, operational concept, or support concept (Bridgman and Glass, 1992).

In accordance with federal statutes and DoD policy directives, test planners must decide how each RAM measure will be addressed, and the Director, Operational Test and Evaluation (DOT&E) must render an opinion as to whether that treatment is adequate. However, our review of case studies suggests that operational test planning does not always account explicitly for the type and quantity of information needed to address key RAM issues. For example, the collection of RAM data often depends heavily on how an operational test unfolds—in terms of "casualties," actions of the soldiers involved, the length of battles, and similar characteristics. Such a dynamic mode of data collection creates enormous statistical difficulties when analyzed as a sequential experiment. In such cases, particularly in force-on-force tests, there is little hope of predicting accurately the amount of RAM data that will be obtained during the operational test event and, consequently, little opportunity to use statistical methods in planning the test to meet RAM information requirements. A common approach in operational testing seems to be to do the testing necessary to assess effectiveness and accumulate the requisite amount of RAM data in a resourceful manner.

The Longbow Apache test appears to be typical in that the size of the force-on-force test was primarily determined by considerations of effectiveness, rather than those of suitability. The primary data sources for measuring the reliability, availability, and maintainability performance of the Longbow system were the gunnery and force-on-force tests conducted as part of the initial operational test and evaluation. Secondary data sources included developmental testing, logistical demonstrations, and force development test and experimentation. According to the test and evaluation plan, "Secondary data sources will be used to supplement [operational test] results by demonstrating historical performance, helping identify trends in anomalies found in the primary MOPs [measures of performance], and characterizing performance in conditions not encountered in the [initial operational test]" (Hall et al., 1994:2-92). The test and evaluation plan also stated that, because of the different conditions obtained under gunnery and force-on-force testing, results from the two phases would not be merged but, instead, would be presented separately in the evaluation report. For the RAM-related measures of performance, however, the operating hours appear to reflect the accumulated hours under both the gunnery and force-on-force phases.

While we acknowledge the inherently uncontrollable aspects of some forms of operational testing, such as force-on-force trials, and the incompleteness of the logistics support structure prior to fielding, we believe that more statistically designed RAM testing of prospective systems can and should be performed. Formal statistical approaches in designing these events—which can be called "operational suitability tests"—will ensure more rigorous assessment of key RAM issues. Furthermore, statistical test design concepts provide a conceptual structure within which costs and benefits of data collection can be weighed and test resources can be allocated optimally.

As illustrated in the C-130H program (see Appendix B), RAM considerations can drive the scope of operational test events, particularly in determining the number of units to be tested and the total time on test. In such cases, improvements in the application of statistical methods in test design could lead to more efficient use of resources.

> **Recommendation 7.2: The Test and Evaluation Master Plan and associated documents should include more explicit discussion of how reliability, availability, and maintainability issues will be addressed, particularly the extent to which operational testing, modeling, simulation, and expert judgment will be used. The military services should make greater use of statistically designed tests to assess reliability, availability, and maintainability and related measures of operational suitability.**

The effective statistical design of operational suitability tests requires, as a preliminary step, that the rationale behind each RAM-related requirement be adequately documented. In this way test designers can specify reasonable requirements for determining the test sample sizes needed to attain a required precision. The need for underlying documentation, rationale and support is especially crucial for RAM measures that may significantly affect field cost and performance. If the 20-year life cycle costs of the Longbow Apache are more than twice the cost of its procurement, even a modest amount of imprecision in estimating system availability might have significant consequences when estimating the system's life-cycle costs.

Historically, numerical operational requirements for RAM measures are usually specified and sometimes extensively documented in RAM rationale reports, but justification is rarely given for particular numerical goals. Reference is sometimes made to the RAM requirements or performance of a baseline system against which a new system is being evaluated; rarely, though, is there an explanation of the relationship between a RAM goal and the system's performance or cost of maintenance. Also missing is any discussion of the implications for performance—and the cost of failing—if the system's demonstrated RAM characteristics fall below the numerical goals.

Experience suggests that the RAM requirements specified in operational test and evaluation are not so rigid that failure to attain a numerical goal necessarily results (or should result) in cancellation of the system. Indeed, such criteria should permit a balancing of a variety of considerations. If operational requirement documents incorporate performance and cost information, decision makers can better determine how this balancing should be done in setting RAM criteria.

> **Recommendation 7.3: As part of an increased emphasis on assessing the reliability, availability, and maintainability of prospective defense systems, reasonable criteria should be developed for each system. Such criteria should permit a balancing of a variety of considerations and be explicitly linked to estimates of system cost and performance. A discussion of the implications for performance, and cost of failing, if the system's demonstrated reliability, availability, and maintainability characteristics fall below numerical goals should be included.**

One effective way to express how attainment of a reliability goal should interact with test size is to specify in the Test and Evaluation Master Plan, *for a given sample size*, the minimum observable suitability value that could be accepted.

STATISTICAL ASSUMPTIONS IN DETERMINING RAM TEST DESIGN AND RESOURCES

Suppose that a test is designed to assess whether a proposed new component of a system possesses a mean time to failure (MTTF)[4] of at least μ_0 hours. Moreover, μ_0 is significantly larger than the corresponding MTTF of μ_1 hours for an existing component that would be replaced by the prospective component if it

[4] We are ignoring for the purposes of this discussion, the difference between mean time between failure and mean time between operational mission failure. This difference relates to the issue of combining developmental and operational test data in that they are distinct concepts.

is acquired. (This example is, of course, a simplification; in practice, a prospective component would be expected to meet several RAM-related requirements. For ease of exposition, we ignore such multivariate aspects.)

The fundamental question is determining whether the prospective component system meets its MTTF requirement. This decision problem can be framed as a statistical *significance test*, where one hypothesis is "the true MTTF of the component under study is equal to μ_0" and the second hypothesis is "the true MTTF is μ_1, where $\mu_1 < \mu_0$." Both μ_1 and μ_0 have the property that rejecting a component with reliability greater than μ_0 is costly, as is accepting a component with reliability less than $\mu_{1.}$ For reliabilities between μ_1 and μ_0, a decision maker is relatively indifferent to the question of accepting the component. One key question that arises in this context is how much testing is enough: How many production components must be tested and for how long?

To answer this question precisely requires that an analyst make specific assumptions. Foremost of these is the specification of a statistical model for the distribution of observed time to failure of the component. The choice of this statistical model has important consequences for test design and interpretation. We comment below on the difficulties, risks, and missed opportunities attendant to the uncritical use of models and procedures based on questionable statistical assumptions. We also consider the potential for making more reliable inferences when information from various sources is fully exploited to evaluate the aptness of proposed statistical models, and we suggest reasonable alternatives.

Exponential Life Testing: Current Uses and Limitations

One particular statistical model has been quite widely used for RAM testing with applications far outnumbering those of competing approaches. The model is that of exponential life testing, in which the distribution of observed times to failure is assumed to be well described by an exponential probability distribution. The popularity of the exponential model is due to its analytical tractability, the clarity with which one can proceed, and empirical and theoretical evidence that some systems—for example, some purely electronic component systems—often have failure times that satisfy this model to a reasonable degree of approximation.[5]

The first two reasons lead, for many different experimental designs, to an exact analysis that can determine, in advance, the resources required to meet specified bounds or requirements for confidence levels or error probabilities. And given that many nonelectronic component systems have nonexponential failure times, tests based on the exponential assumption can be "conservative." This means that the actual producer's risk (the probability of incorrectly conclud-

[5] However, some electronic components suffer from "infant mortality," that is, a short initial phase in which the failure rate is decreasing, typically due to manufacturing defects.

ing that the mean time to failure is less than the requirement) and the actual consumer's risk (the probability of incorrectly concluding that the mean time to failure is more than the requirement) will be smaller than the nominal levels for which the test is planned.

Department of Defense Handbook H108 describes in detail how the μ_0 versus μ_1 hypothesis test should be carried out under the assumption of exponentially distributed times to failure. As explained in that handbook and in other references (e.g., U.S. Department of Defense, 1982; Lawless, 1982; Bain and Engelhardt, 1991), one proceeds by specifying desired levels of α (the producer's risk) and β (the consumer's risk), and then identifying the number of observed failures, r, and maximum test time, T, required to resolve the test at the prescribed levels of α and β. The test is then executed by rejecting the hypothesis that "the true MTTF is μ_0" if the total time on test T at the time of the r-th failure is less than some prescribed number. An appealing feature of exponential life testing is that it easily shows that the duration of the test can be made smaller by placing a larger number of component systems on test than the critical number r.

The simplicity that stems from the assumption of exponentially distributed times to failure derives from the fact that the performance of these tests depends on the hypothesized means μ_0 and μ_1 only through the discrimination ratio, $D = \mu_1/\mu_0$. The required number of observed failures (r) and the required total test time (T) can be computed explicitly given the two error probabilities (α and β) and this discrimination ratio (see U.S. Department of Defense, 1960:Table 2B-5, which provides the required number of observed failures r and the ratio T/μ_0 for selected values of D, α, and β). Using such tables, one can design and carry out reliability tests with relative ease.

However, a number of difficulties arise when the underlying assumption is not met, when failure times do not closely follow an exponential distribution. In cases for which the exponential test design is conservative, the opportunity to carry out a more efficient test or to save test resources is foregone. And conclusions based on exponential assumptions may differ substantially from the conclusions that would be drawn using more plausible statistical assumptions (Zelen and Dannemiller, 1961). It can thus be very important for an analyst to determine when the exponential model is of dubious validity and to use an alternative analysis in such cases.[6]

Alternatives to Exponential Life Testing

A key implication of exponentially distributed times to failure is that the conditional probability of experiencing a failure in any small time interval of

[6] When assessing whether an exponential model might be appropriate, it is often important to distinguish between different failure modes. For example, "start-up" failures might observe a different model than "operating failures."

fixed length, given survival to the starting point of the interval, does not depend on how long the equipment has been operating. In circumstances in which such an assumption of a "constant failure" or "*hazard rate*" is plausible (e.g., some electronic components), the exponential distribution model for time to failure provides a reasonable representation of equipment reliability. However, in some applications the hazard rate is relatively high in the early stages of operation due to manufacturing defects, i.e. infant mortality problems. Also, some equipment (for example, mechanical systems) experience a "wear-out" phase when the hazard rate becomes relatively high because of aging. The hazard rate of such equipment will, when plotted versus time, exhibit a U-shape, the so-called "bathtub" curve.

When the hazard rate is expected to change over time, alternatives to the exponential failure distribution model must be considered. Many alternative models have been discussed in the statistical literature, including the Weibull, lognormal, and gamma distributions. In particular, the Weibull model has an "aging" parameter that can be used to represent increasing or decreasing hazard rates, and it includes the exponential as a special case.

The exponential model is also often inappropriate when the (component) system is repairable. For repairable systems, the observed data can be in the form of either times between failures or numbers of observed failures in different time intervals. The assumption that times between failures are independent and identically distributed according to an exponential distribution, or equivalently a homogeneous Poisson process model for the number of failures, is also often inappropriate. This assumption implies that when a system is repaired, it is restored to "as good as new" condition and the rate with which failures occur over time (failure intensity) is constant. There are many alternatives to this model that have been developed in the statistical literature (see Cox and Lewis, 1972; Ross, 1996; and Ascher and Feingold, 1984). One approach is to use a general renewal process model where the times between failures are still independent and identically distributed but are allowed to be non-exponential (e.g., Weibull, lognormal, gamma, etc.). Other alternatives include modeling the number of observed failures as a mixture of Poisson processes or as a non-homogeneous Poisson process. The former leads to a compound Poisson distribution for the number of failures and is useful in situations where the variance is larger than the mean (over-dispersed Poisson). The non-homogeneous Poisson process allows the failure intensity to vary over time and hence can be used to model system degradation or reliability growth. A careful treatment of data from an operational test presumes that the models employed for the number and timing of observed failures are selected with attention to the special features of the application and to the quality of the fit of the model to the available data.

Our general thesis is that it is important to move beyond exponential life testing. We develop one particular alternative in some detail in the following section as an example of the potential benefits (and possible risks) of nonexponential modeling, but as we have already noted, there is an extensive literature on this topic.

An analyst can easily find instructions for carrying out exponential life tests—including a host of military documents like DoD's RAM Primer and a virtual plethora of military handbooks and military standards, but it is harder to find instructions for alternative analyses within the DoD reliability literature. However, these alternative models have been discussed extensively in the statistical literature in recent years. Unlike the exponential model where exact, analytical solutions can be easily obtained, these models require approximate, numerical methods for reliability test planning and evaluation. However, this is not a major concern with modern day computing resources. There are also tables and software routines that are available in the literature (e.g., see Escobar and Meeker (1994) for the LINF algorithm and Meeker and Nelson (1976) for tables and figures that can be used to plan Weibull life-tests). With some of the more recent statistical software packages, it is an easy matter to fit other distributions such as Weibull, lognormal, and gamma to failure time data (even with complicated censoring patterns, e.g., not knowing when a system was first put on test).

Concern about the overuse (and misuse) of simple models is not new. It has been amply documented in the research literature dealing with defense analysis, as well as in other fields. But certain consequences of the use of inappropriate models are less well understood than others. A question that needs to be explored in some detail, for which we have made a start, is the potential that exists for resource savings when one recognizes in advance that an alternative model is appropriate.

An Example of Alternative Statistical Analysis: Weibull Life Testing

The Weibull lifetime model is a common alternative to the exponential time-to-failure model in applications involving nonrepairable (or perfectly repairable) systems or components. Since the family of Weibull probability distributions contains the family of exponential distributions as a special case, it represents a generalization of the exponential model rather than a rejection of it. Despite the extensive statistical literature on modeling and inference based on the Weibull distribution (see, e.g., Nelson and Meeker, 1978; Meeker and Nelson, 1976, 1977; Meeker, Escobar, and Hill, 1992;), there has been relatively little work that is directly applicable to hypothesis test design and execution; Lawless (1982) mentions: "life test plans under the Weibull model have not been thoroughly investigated . . . it is almost always impossible to determine exact small-sample properties or to make effective comparisons of plans . . . further development of test plans under a Weibull model would be useful." As pointed out above, our examination of the Weibull model should not be taken as an endorsement of Weibull life testing. It should, instead, be taken simply as an example of an alternative approach having the potential for utilizing resources more effectively and/or increasing the precision of the statistical inferences made from operational tests.

Suppose that (1) a random sample of n system units is placed on test until failure; (2) the failure times are assumed to be well described by an exponential probability distribution; (3) the sample size n was determined (assuming an exponential distribution) with fixed values of α and β, in order to test the null hypothesis, $\mu = \mu_0$, against the alternative, $\mu = \mu_1$; and (4) after the data were collected, the Weibull model was shown to more accurately describe the distribution of the observed failure times.

Samaniego and Chong (1998) show that, under certain technical conditions, the test that a reliability is equal to a hypothesized value, appropriate for the Weibull model, corresponds to a modification of the test designed for the original exponential model. For systems or components with increasing hazard rates, the Weibull test problem can, in general, be resolved with greater statistical power than was expected under the original exponential model. More importantly, if the appropriateness of the Weibull model (with increasing hazard rates) can be identified at the time the operational test is being designed, and before the test is executed, then specified levels of producer's risk and consumer's risk can be achieved with a substantially smaller test sample than would be required under the exponential model, and testing might be concluded more quickly with a savings in time to deployment.

The opportunity for resource savings is illustrated in the following example. Suppose one wishes to test the null hypothesis that the MTTF of a prospective system is 1,000 hours against the alternative that the MTTF is 500 hours, and the levels of producer's risk and consumer's risk are both specified to be 10 percent. The resources required to resolve the test under the exponential model are 15 observed failures and a maximum total test time of 10,305 hours. However, if it could be determined in advance of the experiment that a Weibull model with an increasing system hazard rate—specifically, a value of 2 for the aging parameter—more likely described the distribution of system failure times, then from tables provided by Samaniego and Chong (1998), it can be shown that the corresponding Weibull test would require only 4 observed failures and a maximum total test time of 2,980 hours—a reduction of more than 70 percent in required test resources. In general, the results in this table confirm that potential resource savings are available when one recognizes an increasing hazard rate Weibull environment and carries out a Weibull life test instead of an exponential one.

It is important to note that under conditions involving a *decreasing* hazard rate—i.e., when the collection of components tends to improve with age—the Weibull life test will have diminished power, or will require additional test resources, to achieve specified levels of producer's risk and consumer's risk. Thus, adoption of the Weibull model, assuming that it is more plausible than the exponential model, will not automatically bring a reduction in test resource requirements. However, many engineering applications of the Weibull distribution involve conditions in which the equipment hazard rate increases over time, so that when this is recognized in advance there is the potential for increased power or savings of testing resources.

Summary and Recommendations

The task of designing an operational suitability test for a prospective defense system is not a simple one. Relatively sophisticated statistical methods are often called for and, in some cases, further research would be helpful in developing appropriate procedures for test design and analysis. Before a statistical analysis is executed in a particular life-testing application, it is important for the test designer or analyst to confirm that the assumption of an underlying statistical model is justified. (If no model is justified, [nonparametric] methods that are robust to model choice must be used.) Diagnostics for the exponential model are discussed in many basic texts on reliability analysis (see Barlow and Proschan, 1975) and for the Weibull analysis in Samaniego and Chong (1998).

Recommendation 7.4: Operational test agencies should promote more critical attention to the specification of statistical models of equipment reliability, availability, and maintainability and to supporting the underlying assumptions. Evidence from plots, diagnostics, and formal statistical tests—developed from the best currently available methods and software—should be used to justify the choice of statistical models used in both the design and the analysis of operational suitability tests.

The most compelling evidence to support the choice of a statistical model comes from a critical examination of the available RAM data. *Before* such summary measures as observed MTTF are computed, the history of individual failure events must be examined in order to assess the validity of proposed statistical models for equipment reliability. (If such data are scarce, information from similar systems should be examined.) Recommendation 7.4 underscores the importance of archiving RAM data because data from late developmental tests or operational tests of similar systems may be particularly helpful in identifying an appropriate statistical model on which to base the operational test design.

Because of the variety and complexity of military systems, RAM analysts cannot rely exclusively on exponential life testing methods to adequately address all problems requiring statistical modeling. An insupportable use of the exponential model could have several negative consequences: the operational test design may be inefficient and may forgo potential savings in test resources or the analysis of test results may yield misleading conclusions about system RAM performance. Although alternative methods have occasionally been used in some operational testing units, nonexponential life testing is underused.

Recommendation 7.5: In the design of every planned operational suitability test, test teams should explore the potential that exists for the design of equipment life tests that are both economically and statistically efficient. Operational testing agencies in each military service should routinely collect and examine evidence concerning the validity of the assumption that times to failure are independently distributed from the identical exponential distribution; when this model is found to be deficient or inadequate, they should develop procedures to construct test designs that use alternative models.

Testers should seek to determine whether conditions are such that an early resolution of a life test is possible or even likely. Doing so will entail the careful scrutiny of the assumptions of independence, constant mean, exponentiality, and the consideration of nonexponential alternatives. Discussions have recently been initiated within the military services regarding the need to move beyond exponential life testing as the standard or default method of analysis and to revise or supplement military handbooks and other reference materials that are based exclusively on exponential assumptions. We applaud these efforts, and encourage the development of effective networks that would enable an operational test team to seek outside statistical advice when faced with a complex modeling problem or limitations of staff resources.

SCORING RAM DATA FOR EVALUATION PURPOSES

The operational testing of a prospective military system generates complex information about the suitability of the system. Considerable advance effort goes into specifying precisely how RAM-related performance will be measured. Furthermore, after the test has been executed, extensive processing and interpretation of RAM data are usually required in order to conduct an assessment of operational suitability.

DoD and the military services have established formal policies and procedures for this activity. For prospective Air Force systems, a Joint Reliability and Maintainability Evaluation Team (JRMET) is formed to assist in the collection, processing, and analysis of RAM data during both developmental and operational testing. After JRMET review, RAM data are compiled and evaluated by a Test Data Scoring Board (TDSB) that includes representatives from the program office, test organization, supporting command, user command, and operational test agency. Contractor representatives are specifically excluded from the TDSB.

For major acquisition programs, the Army produces a RAM rationale report containing precise failure definitions and scoring criteria that will be applied in the operational test.[7] After an operational test has been executed, the major functions of the RAM scoring team are to judge the criticality of RAM-related test incidents (classification) and to assign a root cause to the incident (chargeability).

Scoring rules are clearly open to conflicting interpretations about which test events are considered "usable," how observations are rejected as "outliers," and exactly what activity constitutes a "trial." Tests involving software-intensive systems, or relatively little test instrumentation, seem especially prone to such subjective interpretation.

For example, a key reliability measure in operational testing of the Navy Tactical Command System-Afloat (NTCS-A) (a multifunction distributed system) was the mean time between mission critical software "faults." According to the test plan (U.S. Department of Defense, 1992:3-4):

> A software fault is any random interruption of system operation (other than those directly attributable to hardware) which can be restored to the preinterrupted state. A mission critical fault is one that prevents the system from performing its mission or results in the loss of some significant mission capability . . .

[7] For example, the report for the Longbow Apache helicopter includes the following paragraphs defining mission failure and mission essential functions (U.S. Department of Defense, 1994b):

> Mission failure is defined as any inherent hardware or software malfunction which causes the loss of a mission essential function specifically required for the mission in progress. This includes the inability to commence a mission or the termination of a mission prior to scheduled completion due to system malfunction or unacceptable equipment condition. While multiple malfunctions of mission essential equipment may occur on a single mission, only one mission failure will be charged. The mission failure will be charged to the initial malfunction causing loss of a mission essential function or to the most serious malfunction if not possible to determine which malfunction occurred first. The other malfunctions will be scored as potential mission failures. Malfunctions discovered during aircrew pre-flight and corrected within 15 minutes active corrective maintenance clock time will not be chargeable as mission failures.

An MEF [mission essential function] is the minimum operational tasks [sic] that the system must be capable of performing to accomplish its mission profiles (MPs). For an attack helicopter the basic MEFs are:

(1) Perform Nap-of-the-earth (NOE) flight within required performance and environmental constraints, maintaining design structural integrity and crashworthiness.

(2) Navigate using on-board navigation devices to designated locations within required accuracies.

(3) Detect, classify, prioritize, and engage moving and stationary targets in the presence of battlefield obscurant and adverse weather.

(4) Track, designate and shoot moving and stationary targets with appropriate weapons.

(5) Communicate with crew and designated supporting elements using the appropriate, preassigned radio frequencies for those elements.

(6) Survive against threat systems, using installed equipment and built-in aircraft characteristics to avoid mission termination and/or life threatening battle damage.

A disagreement arose between testers and system designers concerning how to score an event in which the critical workstation fails and operators subsequently switch to one of two noncritical workstations within 3 minutes. One procedure would charge the system with a "critical fault" in order to estimate its *reliability*, but its *availability* would be relatively unaffected since the system is recovered within 3 minutes.

In another example, initial Marine Corps plans for operational testing of the Advanced Amphibious Assault Vehicle (AAAV) involved force-on-force engagements. In the absence of extensive instrumentation to record events during the test, human observers would be strategically placed to produce a qualitative assessment of the system's performance. Because of concerns about the consistency, validity, and unbiased uses of such scoring, these plans are currently undergoing revision and may ultimately lead to testing at a well-instrumented site.

Such instances underscore the need for objective failure definitions and scoring criteria, as well as rigorous documentation of the actual scoring of RAM data and subsequent evaluation. In many cases, precise measurements of failures and repair times observed during testing are "processed" into "rough" estimates of such characteristics as mean time to failure and mean time to repair, which are then combined with assumptions about system operating tempo and logistics support to produce estimates of operational availability. The uncertainty and arbitrariness inherent in this process, and the sensitivity of the final estimates to changes in assumptions, are rarely discussed prominently; consequently, the appearance of mathematical sophistication and precision in reporting the resulting suitability measures may be very misleading.

> **Recommendation 7.6: Service test agencies should carefully document, in advance of operational testing, the failure definitions and criteria to be used in scoring reliability, availability, and maintainability data. The objectivity of the scoring procedures that were actually implemented should be assessed and included in the reporting of results. The sensitivity of final reliability, availability, and maintainability estimates to plausible alternative interpretations of test data, as well as subsequent assumptions concerning operating tempo and logistics support, should be discussed in the reporting.**

USE OF AUXILIARY INFORMATION IN RAM TEST DESIGN AND EVALUATION

A general and recurring theme in this report concerns the potential value of making better use of all information that is available and relevant to the evaluation of a prospective military system. In current RAM-specific practices, the test and evaluation process does not fully exploit opportunities for using information from disparate sources to obtain more reliable inferences (Bridgman and Glass, 1992).

Cultural, technical, and organizational barriers may prevent operational testing analysts from using data from developmental testing or other sources in assisting in making inferences about system suitability. Such barriers may result in part from the congressional language and concerns reflected in Title 10, Sec. 2399, U.S. Code, which discusses the operational test and evaluation of defense acquisition programs. Raising this issue, Bridgman and Glass (1992:4) note:

> This language discourages the use of modeling in lieu of test and speaks of the suitability of the item tested, not the suitability of the complete system to be fielded. Its intent is to prevent suitability judgments from being made on promised rather than actual performance. However, it should not be extended, in our judgment, to preclude the use of all other sources of data or of modeling to augment test data and help evaluate the results. OT&E [operational testing and evaluation] is specifically authorized by law to have access to such information.

We share the view that such uses of auxiliary information should be permissible (see Recommendation 4.3). Current practices, while understandably intended to maintain the independence of operational test and evaluation, have had the negative consequence of forcing RAM analysts either to live with higher than desirable variability or to incur substantial additional expense in executing the high numbers of replications needed to obtain the desired precision to support production decisions.

Technical barriers exist because operational test personnel may not have been exposed to available methods as part of their statistical training. A number of established statistical approaches allow combining information from different sources (e.g., meta-analysis, hierarchical Bayes, and empirical Bayes methods), but their successful application typically requires the involvement of personnel with advanced training in statistics (see Chapter 10). Such personnel may not be routinely available to participate as team members in suitability assessments.

We present here an illustration of the potential value of using statistical approaches to integrate information from multiple sources in assessing RAM performance. Table 7-1 shows failure data for air conditioning equipment on 13 Boeing 720 aircraft (Proschan, 1963; Gaver and O'Muircheartaigh, 1987). The observed hazard rates vary considerably—from about 3 to almost 17 failures per 1,000 hours. The number of observed failures in the test varies from aircraft to aircraft for two reasons: intrinsic reliability differences between the aircraft and random error due to the fact that the number of failures would differ from test to test even if the intrinsic reliability remained constant. Assuming that the 13 aircraft were homogeneous, with a common hazard rate of 10 failures per 1,000 hours, ordinary random variability would result in a range of observed hazard rates from 5 to 14. Since the observed range is only slightly greater than this, it makes sense to "shrink" these hazard rates in towards a common mean rate.

TABLE 7-1 Failures of Air Conditioning Equipment on 13 Boeing 720 Aircraft

Aircraft ID	Failures	Hours (in thousands)	Raw Failure Rate (failures/1,000 hr)
11	2	0.623	3.21
9	9	1.800	5.00
5	14	1.832	7.64
4	15	1.819	8.25
12	12	1.297	9.25
10	6	0.639	9.39
2	23	2.201	10.45
3	29	2.422	11.97
1	6	0.493	12.17
13	16	1.312	12.20
7	27	2.074	13.02
8	24	1.539	15.59
6	30	1.788	16.78

Incorporating the information available from all the aircraft and the methods discussed in Gaver and O'Muircheartaigh (1987), the estimated hazard rates for Aircraft #11 and #6—those with minimum and maximum observed hazard rates, respectively—can be reestimated to be approximately 8.5 and 13.5 failures per 1,000 hours, respectively.

The crucial assumption used in shrinking the observed hazard rates toward the mean hazard rate is that these 13 aircraft are essentially indistinguishable from one another with regard to the process that produced their air conditioning systems. If individual aircraft had air conditioning systems of slightly different designs or if data had been gathered from aircraft under different operating conditions, then alternative assumptions and resulting estimators might be appropriate. It has been demonstrated both theoretically and empirically that using a shrinkage estimator in situations like this can yield marked improvements in predictive accuracy. Such established statistical methods for combining RAM data from multiple sources, when correctly applied, lead to more effective policies and decision making. For example, Hoadley (1981) showed that Bell Telephone quality assurance decisions can be made more accurately if empirical Bayes procedures are used. Morris (1983) discusses several other successful applications to important problems:

> The most persuasive arguments for the effectiveness of the procedures actually were made via cross-validatory methods. The shrinkage techniques predicted new data in each application more accurately than did the traditional methods, and the investigators showed that better decisions would have been made had empirical Bayes procedures been used in the past.

In the later stages of developmental testing, when the process of developing a reliable system prototype has become relatively stable, there will be experimental data that can be both relevant and quite useful to the operational tester. The use of that data in combination with the data collected in designed operational tests may offer some important benefits, including the possibility of an early resolution regarding a system's suitability. There should be greater openness to the selective (and supported) use of statistical methods for combining data from developmental and operational testing, as well as other relevant information, including subjective inputs from scientists with appropriate expertise and commercial or industrial data on related components or systems.

Recommendation 7.7: Methods of combining reliability, availability, and maintainability data from disparate sources should be carefully studied and selectively adopted in the testing processes associated with the Department of Defense acquisition programs. In particular, authorization should be given to operational testers to combine reliability, availability, and maintainability data from developmental and operational testing as appropriate, with the proviso that analyses in which this is done be carefully justified and defended in detail.

Organizational barriers exist to effective sharing of information about RAM performance. Two institutional problems contributing to poor data utilization within DoD have been identified (Jim Hodges, in Rolph and Steffey, 1994:44): "First, no one has responsibility for accumulating test data in one place and ensuring proper utilization. Second, there are competing objectives among the chief players—defense contractors, the services, and the Office of the Secretary of Defense—that complicates the collection and interchange of data."

A recent RAND study of data quality problems in Army logistics (Galway and Hanks, 1996) pointed out that data quality problems are particularly acute when data are collected in one organization for use by another:

> The costs to collect data and to ensure quality (e.g., detailed edit checks at the point of entry) are often very visible to the collecting entity in terms of time and energy expended. The benefits may be very diffuse, however, particularly in a large organization . . . where data collected in one place may be analyzed and used in very distant parts of the organization with different responsibilities and perspectives. In these cases, one part of an organization may be asked or required to collect data that have little immediate effect on its own operations but that can be used by other parts of the organization to make decisions with long-term impacts. Intraorganizational incentives and feedback to insure data quality in these cases have been difficult to devise.

These authors propose a three-level framework for understanding and classifying the nature of data problems:

1. *Operational* data problems are present when data values are missing, invalid, or inaccurate.

2. *Conceptual* data problems are present when the data, because of imprecision or ambiguities in their definition, are not suitable for an intended use or, because of definitional problems, have been subjected to varying collection practices, again resulting in missing, invalid, inaccurate, or unreliable values.

3. *Organizational* data problems occur when there are disconnects between the various organizations that generate and use data, resulting in a lack of agreement on how to define and maintain data quality. One symptom of organizational problems is the persistence of operational and conceptual problems over time, even after repeated attempts at solution.

Changes during the process of RAM data collection (in such conditions as system configuration or test environment) are not always easily retrievable from existing databases. Such a conceptual data problem would critically interfere with the development of defensible methods for combining information from multiple test events.

A related complication concerns the utility of test documentation. As discussed in Chapter 6, operational test reports often put forth raw estimates of parameters of interest, with no indication of the amount of uncertainty involved. When such reports are combined with other evidence of a system's performance (regardless of the quality of the additional information), the decision maker cannot describe in a definitive or statistically supportable way the risks associated with the acquisition of the system.

In the military services, according to Galway and Hanks (1996): "Data seem insignificant compared to the physical assets of equipment, personnel, and materiel. However, data are also assets: they have real value when they are used to support critical decisions, and they cost real money to collect, store, and transmit. Improving the capabilities for archiving and retrieval of data and documentation (see Chapter 3) has potential benefits that are particularly great for RAM evaluation. RAM data on a prospective system taken from earlier stages of development, and from similar systems, can significantly improve the accuracy of conclusions drawn from operational testing and can reduce the amount of resources required for such testing. Therefore, efforts should be made to archive early RAM performance data for use in assessing operational suitability. Furthermore, the maintenance of a database that permits continuing RAM performance evaluation after the system is fielded is consistent with best industrial practices for quality management and product improvement (see below).

> **Recommendation 7.8: All service-approved reliability, availability, and maintainability data, including vendor-generated data, from technical, developmental, and operational tests, should be properly archived and used in the final preproduction assessment of a prospective system. Af-**

ter procurement, field performance data and associated records should be retained for the system's life, and used to provide continuing assessment of its reliability, availability, and maintainability characteristics.

Reliability, availability, and maintainability "certification" can be better accomplished through combined use of data collected during training exercises, developmental testing, component testing, bench testing, and operational testing, along with historical data on systems with similar suitability characteristics. Achieving this goal, however, will require commitment at the policy-making level, technical skill in the application of advanced statistical methods, and accessible data and documentation of acceptable quality.

SPECIAL STATISTICAL METHODS FOR RELIABILITY TEST AND EVALUATION

Two statistical procedures that are used in conjunction with reliability estimation of defense systems that merit individual discussion are reliability growth modeling and accelerated life testing.

Reliability Growth Modeling

In many defense acquisition programs, as in private industry, the reliability of system hardware grows during design, development, testing, and field use. Reliability growth results from continued engineering efforts to improve the design, manufacture, and operation of repairable system hardware. Formal statistical models and associated methods of analysis have been developed to represent such growth and to estimate future reliability and related quantities of interest. A reliability growth analysis typically involves fitting, to observed failure data, an equation expressing the underlying hazard rate as a (usually decreasing) function of time. For example, one popular model, developed at the Army Materiel Systems Analysis Activity (AMSAA) in the 1970s, assumes that failures occur according to a nonhomogeneous Poisson process (U.S. Department of Defense, 1981).

Model-based predictions of future reliability are potentially useful in at least three stages of the acquisition process:

1. in early development, to estimate the level of engineering effort (i.e., hours of testing and corrective action) that will be required to achieve prespecified standards of reliability;
2. in operational test planning, to estimate reliability at the time of a future operational test event, for purposes of determining the system's readiness for testing or the duration and conditions of testing; and

3. in the assessment of operational suitability, to incorporate projected post-test engineering efforts in estimating the field reliability of a prospective system.

Some organizations involved in test and evaluation, particularly AMSAA, are regularly engaged in reliability growth modeling for purposes of planning and analysis. One recent example involved the family of medium tactical vehicles, which underwent an initial phase of operational testing from September through December 1993. This phase of testing was terminated because the system failed to meet reliability requirements. During the subsequent engineering effort, reliability growth models were fit to observed failure data. These models were used for each variant (e.g., cargo truck, van) in the vehicle family to predict reliability at the time of the third operational test phase and to estimate the probability of success in the test. The family of medium tactical vehicles was ultimately successful in the third phase, conducted from April to July 1995 (U.S. Department of Defense, 1995a). At this time the observed reliability of each variant during testing significantly exceeded its model-based prediction. Results like this raise concerns about the validity of reliability growth modeling.

Reliability growth modeling has also been used to assess suitability in support of production decisions. Such cases typically involve complex, single-shot equipment with high unit costs (e.g., missiles) and highly reliable electronic subsystems that would require operational tests with a fixed configuration to be carried out over a period of several thousand hours (Rolph and Steffey, 1994). Some amount of reliability testing is done to identify deficiencies and verify corrective actions. Programmatic considerations, particularly unit cost and schedule, can then lead to circumstances in which the decision to commit production funds involves a criterion such as achieving an instantaneous reliability value at a prescribed point on the system's projected growth curve—a decision point that may occur before the completion of operational testing.

Statistical modeling of reliability growth can be a valuable tool in the system development and testing process. However, the risk of significant discrepancies between predicted and observed reliability, as in the example of the family of medium tactical vehicles, underscores the need for thorough validation of reliability growth models. Issues related to the scoring of RAM data, discussed in a previous section, take on added importance in this context.

Recommendation 7.9: Any use of model-based reliability predictions in the assessment of operational suitability should be validated a posteriori with test and field experience. Persistent failure to achieve validation should contraindicate the use of reliability growth models for such purposes.

Accelerated Life Testing

A potentially serious deficiency in operational testing is the frequent inability to identify suitability problems relating to cumulative effects (e.g., aging and corrosion). One recent study cited several examples of such problems in new military systems, noting that in each case (Bridgman and Glass, 1992:6): "The OT [operational testing] did not last long enough for these problems to show up during the test, and no attempt was made in the evaluation of test results to use DT [developmental testing] results or other information to predict what suitability problems might appear after extended use in the field."

Assessing the distribution of time to failure of systems (or components) with very high reliability or with susceptibility to cumulative effects is a difficult statistical problem because of the implied need for very large exposure times. For this reason, a variety of accelerated testing methods have been developed that allow data collected over relatively short experimentation times to be used to make inferences about system reliability over time ranges considerably longer than the experimentation times. The term "acceleration" has many different interpretations in the testing community, but it usually refers to one of three ways of making "time" (or any other scale used to measure system life) go "faster" in a test by increasing certain factors that directly affect reliability:

1. the use rate of the system;
2. the rate at which age affects the system; and
3. the degree of "stress" to which the system is subjected.

In addition, there are generally two different types of information a tester can gather:

4. system (or component) failure times; and
5. amount of degradation (physical, structural, etc.) in a system (or component).

Accelerated testing methods are potentially applicable in developmental, as well as operational, testing of prospective military systems. Lall, Pecht, and Cushing (1994) discuss the need for accelerated testing in the design and manufacture of electronic products, and they identify three difficult issues that have limited the application and acceptance of such methods: determination of the dominant failure mechanisms, sites, and modes under intended life-cycle loads; determination of appropriate stress levels for accelerated tests; and assessment of product reliability under intended life-cycle loads from data obtained under accelerated tests.

A critical underlying requirement for the acceptability of using any accelerated testing method is the need for a model that relates the acceleration factors (usage rate, temperature, physical stresses, vibration, voltage, pressure, etc.) to failure events. Such models tend to fall into two general types: physical, (based upon "first principles" of chemistry, physics, etc., and the various causative theories pertinent to the system; empirical, those based upon "fits" to experimental data obtained independent of the particular tests being performed.

Given a model, statistical approaches to extrapolating reliability statements from accelerated tests are straightforward (if somewhat complicated). However, the influence the model has on the final reliability statements is generally more important than the test data themselves. The wide variety of acceleration models available (see, e.g., Meeker and Escobar, 1993; Nelson, 1990), as well as the need for understanding their acceptability in any particular situation, creates an environment that is subject to misuse and exploitation.

> **Recommendation 7.10: Given the potential benefits of accelerated reliability testing methods, we support their further examination and use. To avoid misapplication, any model that is used as an explicit or implicit component of an accelerated reliability test must be subject to the same standards of validation, verification, and certification as models that are used for evaluation of system effectiveness.**

BEST RAM PRACTICES, TRAINING, AND HANDBOOKS

An accepted set of "best RAM practices" is a goal much closer to being realized in industrial than in military settings. Models for components of these developments include the Organization for International Standardization (ISO) 9000 series and existing documents on practices in the automobile and telephone industries.

The RAM-related documents in the ISO 9000 series address many subjects, ranging from the use of statistical principles and methods to data archiving and information feedback. ISO 9000-4 concerns dependability[8] program management and provides, for example, the following guidance for product suppliers (numbers in parentheses refer to document sections):

- prepare specifications which contain qualitative and quantitative requirements for RAM performance and clearly state maintenance support assumptions (§ 6.3);
- establish and maintain procedures for effective and adequate verification and validation of dependability requirements (§ 6.7);

[8] Dependability is a collective term used to describe availability performance and the performance of its influencing factors—reliability, maintainability, and maintenance support.

- establish and maintain access to effective statistical and other relevant qualitative and quantitative methods and models appropriate for prediction, analysis, and estimation of dependability characteristics (§ 5.2);
- establish and maintain procedures for assessing life-cycle cost elements (§ 6.8);
- establish and maintain data banks to provide feedback on the dependability of products from testing and field operation in order to assist in product design, current product improvement, and maintenance support planning (§ 5.3);
- establish and maintain procedures for handling, storage, and analysis of failure and fault data from testing, manufacturing, and field operation (§ 6.11); and
- retain for an appropriate period, defined in relation to the expected product lifetime, all documents containing dependability requirements, analyses and predictions, test instructions and results, and field data analysis records (§ 5.4).

Efforts to achieve more efficient (i.e., less expensive) decision making by pooling suitability data from various sources require documentation of the data sources and of the conditions under which the data were collected, as well as clear and consistent definitions of all terms used. Such efforts underscore the potential value of standardizing RAM testing and evaluation across the services and encouraging the use of best current practices.

DoD might draw constructively on industrial practices, particularly in such areas as documentation, uniform standards, and the pooling of information on operational suitability. As evidenced in the ISO 9000 series, documentation of processes and retention of RAM-related records (for important decisions and valuable data) are practices now greatly emphasized in industry. The same should be true for DoD, especially for the purposes of assessing operational suitability in support of major production decisions. As we stress throughout this report, effective retention of information allows one to learn from historical data and past practices in a more systematic manner than is currently the case in operational testing.

> **Recommendation 7.11: The Department of Defense should move aggressively to adapt for all test agencies the Organization for International Standardization (ISO) standards relating to reliability, availability, and maintainability. Great attention should be given to having all test agencies ISO-certified in their respective areas of responsibility for assuring the suitability of prospective military systems.**

Considerable differences with respect to RAM policy, practice, organization and methodology exist among the services, as well as within the testing community in each service. These differences may be partly attributable to variability in the training and expertise of developmental and operational testing personnel, which in turn contributes to an uncritical reliance on certain modeling assumptions (e.g., exponentiality) in circumstances in which they may not be tenable.

The manuals, handbooks and reference materials presently serving as the basis for military life-testing applications should be upgraded, the statistical level of the personnel who carry out the military's operational RAM testing analysis should be comparably enhanced, and the consideration of alternative models and methods for RAM testing should become routine in operational testing across the services.

> **Recommendation 7.12: Military reliability, availability, and maintainability testing should be informed and guided by a new battery of military handbooks containing a modern treatment of all pertinent topics in the fields of reliability and life testing, including, but not limited to, the design and analysis of standard and accelerated tests, the handling of censored data, stress testing, and the modeling of and testing for reliability growth. The modeling perspective of these handbooks should be broad and include practical advice on model selection and model validation. The treatment should include discussion of a broad array of parametric models and should also describe nonparametric approaches.**

8

Testing Software-Intensive Systems

The defense testing and acquisition community is faced with systems in development that are increasingly software intensive, making use of a wide variety of methods of software development. Software is becoming a more ubiquitous element of defense systems, and it is also playing an increasingly critical role in a system's ability to meet its performance objectives. Recently, a number of reported system failures have been attributable to the software. At the same time, the Department of Defense is faced with decreasing budgets generally, restricting the funds available for testing and evaluation. The problem is how to make use of limited defense funds to prevent software problems more effectively.

CURRENT PRACTICES, OPERATIONAL PROFILES, AND MODELS OF INTENDED USE

In reviewing the testing of software, we examined defense systems that are either software products or systems with significant software content. The focus of our efforts has been on how the services conduct operational testing and evaluation on software-intensive systems, what the special procedures are for such systems (noting wide variation in the techniques used for operational testing and evaluation across the services), and what special problems arise. We offer recommended methods to deal with testing these systems.

One of the first systems the panel examined was the Naval Tactical Command System-Afloat (NTCS-A), the Navy's premier command and control system.[1] NTCS-A is a commercial off-the-shelf evolutionary procurement system. (In evolutionary procurement, the system is developed in stages, with each stage undergoing separate testing.) This system experienced a number of problems that we believe are relatively widespread. First, because the system experienced frequent failures during testing, the goal of having the system run for a reasonable number of hours without failure was changed. Also, the large number of components for the system (approximately 40) created the potential for interaction problems each time one of the components was upgraded, since it would result in 40 different product enhancement and release cycles that would affect the whole system. Moreover, with little configuration control, the systems being tested in the operational testing and evaluation were materially different from systems being fielded; thus, the panel viewed the ability to control configuration as a key issue to be addressed.

The panel also had numerous interactions with the Air Force Operational Test and Evaluation Center (AFOTEC). AFOTEC places a great deal of emphasis on the examination of software code, approximately 2 percent of the code in large systems. We believe this is not an efficient use of analysts' time: the test sample of lines of code that are examined should never be based on a fixed percentage of total lines of code; furthermore, it is the software architecture that should be examined, not the code. We do applaud AFOTEC's efforts to communicate early in the process with software developers, and we concur that the use of software metrics, a measure based on code characteristics that evaluate code complexity, is useful for the purpose of producing estimates for support budgets.

The Army Operational Test and Evaluation Command has conducted impressive experiments in developing effective test processes, but these methods have not yet been institutionalized. Also, although the Army has developed an extensive software metrics program, it is of limited value because it is not connected to software failures in the field or to the development process.

On the basis of its review of these and other systems, the panel concludes that use of statistical science can significantly improve the test and evaluation of software and software-intensive systems. Papers that discuss some of the relevant research concerning statistical aspects of software testing and evaluation are Nair et al. (1998), Oshana (1997), and Poore and Trammell (1996). Before describing this use, it should be noted that there are several important software engineering issues involved in the defense system life cycle that are not in our

[1] The Naval Tactical Command System-Afloat is described as an all-source, all-knowing system that is installed in most ships and many more shore sites. It is designed to provide timely, accurate, and complete all-source information management, display, and dissemination activities—including distribution of surveillance and intelligence data and imagery to support warfare mission assessment, planning, and execution—and is a current segment of a large strategy system known as the Joint Maritime Command Information System.

purview but that have direct bearing on our task. An important example is configuration control during testing and after deployment. In the extreme case, the software fielded might be significantly and critically different from that scrutinized in operational testing and evaluation. It is beyond the panel's charge to address directly the fundamentals of software engineering and current best practice for creating and maintaining software across the complete system life cycle. However, for any defense system, the prototype tested must be essentially identical to that fielded for the test to be informative. Clearly, if the software engineering process is flawed, then the statistical designs, measurements, and analyses used in operational testing and evaluation may be irrelevant, and the decisions based on them misinformed.

Evolutionary procurement is designed in part to exploit the opportunities offered by the rapid pace of technology improvement. Evolutionary procurements will be increasingly associated with more complex systems and those with the largest software content. By their nature, evolutionary procurements will result in repetition of the operational testing and evaluation cycle, creating the opportunity for use of test infrastructure that might have been developed in earlier cycles, as well as use of existing operational test and field data. These facts should be taken as a mandate for investing in the creation of test infrastructure to be used and enhanced in later cycles of an evolutionary procurement.

A defining characteristic of software is that it affords unprecedented opportunities for systems in the field to be changed or reconfigured quickly and for customizing each system for its impending use and environment. In order to evaluate systems intended to be reconfigured and customized, it is also necessary to focus on the software architecture of the system. In order to test such systems adequately, it is necessary to use architecture information in the test design.

In Chapter 6 we note that for complex systems the variation from one prototype (in this case configuration) of the system to another is often of far less interest than variation between the environments of use. This is precisely the case with software: consequently, the issues discussed in Chapter 6 generally apply to software systems. Moreover, the advice given elsewhere in the report on experimental design, modeling and simulation, taxonomy, RAM, and combining information is in large part applicable to software intensive systems.

Testing based on operational profiles and models of intended use is called usage-based testing (also known as black box testing); it is appropriate for operational testing and evaluation because it is performed to demonstrate that a system is fit for its intended use. Other types of testing—based directly on examination of code to ensure that every line of code is executed, following each path of decision statements in the code and similar criteria (also known as white box testing)—are typically a part of developmental testing and performed in order to find faults in the code. These code-based forms of developmental testing are in no way preemptive of the operational, usage-based testing recommended below;

in fact, we believe usage based testing should be moved upstream, to development.

Operational profiles are developed by consideration of environments of use and types of users. Operational profiles are estimates of the relative frequency of use of various inputs to a system. For purposes of testing, the set of all possible inputs is progressively partitioned. Different profiles might be developed for each block of the partition. The partitioning continues with subprofiles and categories of test cases within them. In the fine-grain partitioning, more than one test is often run for a block of the partition, and the test cases are selected randomly. To create a test case, the input domain is sampled on the basis of the distribution described by the operational profile.

Operational profiles represent field use, so that the most frequently used features are tested most often. When testing schedules and budgets are tightly constrained, this design yields the highest practical reliability because if failures are seen they would be the high frequency failures. (Critical but infrequently used features might be in a separate block of the partition and would receive special attention.)

An alternative to the concept of developing operational profiles directly is to build detailed models of all possible scenarios of use. These models are then represented in the form of one or more highly structured Markov chains (a type of probabilistic model), and the result is called a usage model. Decisions are made to identify the states of use of the system and the allowable transitions among those states and to determine the probability of making each allowable transition. Markov chain-based operational profiles have great analytical potential for planning and managing usage-based testing. For example, the operational profile can be calculated as the long-run probability distribution of the states of the chain, which corresponds to the proportion of time the system will be in each state of use in expected operational field use; the expected sequence length corresponds to the average number of events in a test case or a scenario of use; and mean first passage times correspond to expected amount of random testing required to experience a given state-of-use or transition. These and other statistics of the chain are used to validate the model and support test planning.

In order to generate usage models, their structure can be described using a system of constraints on the transition probabilities as decision variables. Some constraints are related to field usage conditions and some to test management (e.g., to restrict transitions to previously tested states), which together with an appropriate objective function allows automatic generation of transition probabilities through use of standard linear optimization techniques. The objective functions can be related to cost, value, or other goals. This approach simplifies the construction and management of usage models.

We believe the usage-based testing strategy is appropriate in operational testing and evaluation for several reasons. The focus is on testing the defense system as it will be used and not on testing just the software. Representations of hardware components and human operators can be included in the usage models and operational profiles. It is practical to build on (previously used) models for future testing during incremental or evolutionary procurements, and the models will permit continuity of methodology and comparison of outcomes across increments. This strategy allows model building to begin very early in the defense system development. Also, a usage-based testing strategy provides a common base for communicating with the developers about the intended use of the system and how it will be evaluated. And information from developmental tests can be evaluated relative to the usage models to inform operational test planning.

> **Recommendation 8.1: The strategies for operational testing of software and software-intensive defense systems used by the service operational test agencies should be based on operational profiles and models of intended use and in doing so should consider environments of use, types of users, and types of failures.**

The statistical issues associated with this recommendation are numerous. First, usage-based testing is the basic strategy for statistical testing of software. Sampling efficiency can be gained by partitioning the set of all potential test cases on the basis of the operational profiles or usage models. Test cases can be randomly generated directly from usage models.

Second, test results are amenable to statistical analysis. At the operational testing and evaluation phase, there should be no failures in software, in which case there are statistical models by which to estimate the expected field reliability based on the extent and variety of testing done relative to the expected field use. If the operational testing and evaluation permits software failure-repair cycles, there are statistical models by which the rate of growth in reliability can be assessed.

Third, usage models can be analyzed and validated relative to properties of the defense system. Examples are the steady state probability, the long-run occupancy rate of each state, or the usage profile as a percentage of time spent in each state. These are additive, and sums over certain states might be easier to check for reasonableness than the individual values. Therefore, the model is helpful in identifying and creating test cases to inform specific situations.

Finally, usage model analysis supports test planning and estimation of the amount of testing required to achieve specific objectives, such as experiencing every possible state of use and every possible transition and experiencing various scenarios of use, reliability targets, and other quantitative criteria for stopping testing.

EXPERIMENTAL DESIGN AND TEST AUTOMATION

The complexity of the efficient selection of test cases is beyond human intuition, because the combinatorial choices are astronomical and the relationship of one test case to another compounds the complexity. The systems are so complex that it is absurd to expect satisfactory operational testing to occur on the basis of manual testing. Software testing is usually expensive, resources of time and budget are always limited, and every test case needs to be chosen with some rationale. Experimental design techniques and usage models are therefore important to guide test selection. System architecture, design, and development must anticipate testing and provide features to facilitate test automation. Automated testing of software systems requires an investment in test infrastructure.

The usage model has two aspects, the structural and the probabilistic. Many testing strategies may be based on the structure of the model, which is the graph of states of use (nodes) and possible transitions (arcs). Such strategies include, for example, state coverage, transition coverage, testing critical paths of use, and creating testing partitions based on the graph. With the addition of the transition probabilities, one can identify high-probability paths of use and partitions of the model to further guide testing.

Random sample test cases are generated by random walks through the Markov chain usage model. Test cases take the form of scripts that have been associated with the nodes or arcs, which are detailed instructions for conducting and checking test events. In the case of manual testing, the scripts are instructions to humans. In the case of automated testing, the scripts are commands to the testing system. Reliability and other quality measures are defined directly in terms of the source chain and testing experience without additional assumptions: for example, there is no assumption that failures are exponentially distributed, which permits monitoring quality measures and stopping criteria sequentially as each test case is run and evaluated.

A class of statistical experimental design methods known as combinatorial design algorithms can be used to generate test sets that cover the n-way combinations of inputs. For certain types of applications, including system testing and testing for conformance to protocols (e.g., the SNMP—Simple Network Monitoring Protocol), this approach has been used to minimize the amount of testing required to satisfy use-coverage goals. Scripts can be generated that interact with automated test facilities.

We believe test automation is appropriate for operational testing and evaluation of software-intensive systems for several reasons. First, it obviously permits testing a greater number of inputs. Second, it requires that operational testing be a primary consideration from the outset of system development. In order to avoid taking a system apart to connect instrumentation, sockets must be provided for entering inputs and retrieving outputs. Third, test automation requires a test oracle, the ultimate authority on correct behavior, which forces precise specifications, development according to specifications, and testing based on specifications.

We identify three statistical issues associated with test automation. First, either partitions must be sufficiently fine-grained that only one test case is required per block or an adequate random sample must be taken within each block of the partition. The former is very difficult to achieve at the usage level, but test automation makes it economical and practical to acquire adequate random samples within the blocks. Second, stopping criteria can be monitored to support decisions to stop testing because of adequate data to support a decision to accept a system. Finally, at any point during testing, what-if assumptions can be made regarding success or failure of prospective testing to evaluate the range of expected outcomes and the value of further testing.

> **Recommendation 8.2: Service operational test agencies should use experimental design methods to select or generate test cases for operational testing of software-intensive systems. Service test agencies should make the institution of test automation a priority for operational testing and evaluation.**

SOFTWARE ARCHITECTURE

It is necessary to focus on architecture and design principles to determine that a software system is good for the long term. The architecture of a software system defines the components of the system and the constraints on how the components interact. The design principles address standards, conventions, and practices for making the components.

Operational testing and evaluation of a software system involves more than just the evaluation of a specific version. There are issues of operational test and evaluation credibility, long-term impact on the entire life-cycle of a software system or family of systems, and of long-term impact on the operational test and evaluation process. If operational test and evaluation praises the software architecture of a system as supportive of field reconfiguration and future enhancement, the positive feedback will reinforce good software design. If operational test and evaluation points out that a design of a software system is inadequate, this is not second guessing developers after the fact; rather, this is demonstrating that operational test and evaluation can recognize the difference between competent and less than competent design, that an operational burden should be expected in the field (perhaps an issue for doctrine and training), that disproportionate cost burdens should be expected during future enhancements, and that the problems should be addressed prior to future enhancements.

Architecture most determines the future adaptability, maintainability, and reusability of software. Architecture is also of critical importance in incremental and evolutionary procurement because only a good design will facilitate successive increments. Another important point is that software architectures exist at multiple levels: use, look, and feel at the product level; design model for identification and interaction of modules, subsystem, components; and execution architecture for specific hardware implementations. A focus on architecture allows operational testing and evaluation to approach important aspects of the software on all levels, without being overwhelmed by code volume.

A focus on architecture will facilitate assessing uniformity (or lack thereof) across training systems and multiple fielded versions of the same basic defense system. And systems that are to be customized in the field will have to be developed to a design that specifically supports the customizing and associated testing.

The necessity to focus on architecture has a corollary regarding current use of software metrics. While metrics can be reasonable indicators of bad design, bad programming, and costly maintenance, the ability to identify good design, good programming, and a robust life cycle from code metrics is problematic. Use of software metrics that are not tied to the development practices that produced them and that are not calibrated by the field performance of the system are also problematic. Such metrics apply only to the software as a document (an array of characters) and neither to a software development process nor to an operational defense system.

Personnel time is scarce in operational testing and evaluation and could be better spent assessing the architecture and design principles used. Human examination of code shifts the focus and resources too far away from use of the system. To the extent that automated code analyzers can collect data and assess compliance with architecture and design principles, provide interpretations that are helpful in maintenance budgeting, or otherwise identify existing or potential problems, they should be used.

The understanding of architecture is complementary to defining operational profiles and models of use; thus, it is supportive of usage-based test planning and design. A focus on architecture will improve the statistical value of data harvested by automated code analyzers and make the metrics more meaningful. And, at any point during testing, what-if assumptions can be made regarding success or failure of prospective testing to evaluate the range of expected outcomes.

> **Recommendation 8.3: In operational testing of software-intensive systems, service test agencies should be required to evaluate software architecture and design principles used in developing the system.**

TRACKING FAILURES IN THE FIELD

Tracking failures in field use for root cause analysis is a fundamental software engineering technique for closing the loop from requirements to use. Failure analysis is the basis for process improvement and product improvement. The operational test, developmental test, and development organizations need such information in a database in order to improve their respective processes. The services need such information in order to improve defense systems. This is commonplace in industry and government, and information is readily available on defining, operating, and using such databases.

With such databases, operational testing and evaluation agencies could conduct long-term assessments of the effectiveness of their testing and evaluation practices. Root cause analysis might in some cases identify ways to improve usage modeling efforts, thus improving the effectiveness of future testing. Effectiveness of developmental testing could be improved by analysis of failures occurring in the field. If agencies' practices result in improved developmental testing, less time and money will be required in operational testing and evaluation.

Ultimately, the goal of operational testing is that defense systems experience no (software) failures in the field. This goal will be realized only through higher quality development work, not by better testing. Detailed information about failures and the development and use of feedback systems from field experience needs to be used to improve development practices.

Failures that occur in one system of a family of systems that is widely distributed and customized in many different configurations are difficult to assess with respect to the implications on the family. The ability to assess field experience and propagate field changes in a family of systems is a major issue in the overall effectiveness of the system. Operational testing and evaluation should ensure that effective mechanisms are in place for widely distributed and customized systems.

A database of field-reported failures for products and product lines is commonplace in industry and forms the basis for ongoing statistical analysis. Most industrial organizations seek to reduce product failures in the field while simultaneously shortening development cycles. The database becomes the vehicle to close the loop between field performance and development practices. A great deal of data is generated in all stages of the software life cycle. Field failure data are key to the most meaningful experimental controls and to evaluation of software engineering methods used throughout the life cycle of defense systems, including the operational testing and evaluation phase. Information on the number of systems deployed and hours (or other units) of use, together with field failure data, can be used in trend analysis to track reliability growth or decay.

> **Recommendation 8.4: Service test agencies should be required to collect data on system failures in the field that are attributable to software. These should be recorded and maintained in a central database that is accessible, easy to use, and makes use of common terminology across systems and services. This database should be used to improve testing practices and to improve fielded systems.**

This recommendation goes hand-in-hand with Recommendation 3.3 on developing a centralized test and evaluation data archive. It should be straightforward to include with each system some measure of the frequency of field failures, circumstances surrounding such failures, and their (possible) cause. This recommendation will probably be easiest to apply and most useful for software failures, but it should be viewed as appropriate and valuable for all defense systems when feasible.

9

Using Modeling and Simulation in Test Design and Evaluation

As part of system development, many industries, including the automobile industry, make substantial use of modeling and simulation to help understand system performance. Modeling and simulation (sometimes referred to here as simulation) are currently used for a number of applications in the Department of Defense, notably for training users of new systems and to help support arguments presented in the analysis of alternatives (formerly cost and operational effectiveness analyses) to justify going ahead with system development. Its success in similar applications in industry, and its cost, safety, and environmental advantages over operational testing, have raised interest in the use of modeling and simulation in operational testing and evaluation (where it enjoys the same advantages). The use of simulation to augment operational test design and evaluation (and other related purposes such as to provide insight into data collection) has been strongly advocated by DoD officials. Although a number of workshops have been devoted to the use of simulation for assisting in the operational test and evaluation of defense systems, the precise extent to which simulation can be of assistance for various purposes, such as aiding in operational test design, or supplementing information for operational test evaluation, remains unclear. Great care is needed to ensure that the information provided by modeling and simulation is useful since the *uncritical* use of modeling and simulation could result in the advancement of ineffective or unreliable systems to full-rate production, or, conversely, the delay or return to development of good systems.

USES OF SIMULATION

As the uses of simulation advance from training to test design to test evaluation, the demands on the validation of the simulation increase. Unfortunately, it is difficult to comprehensively determine whether a validation is "sufficient," since (as discussed in Chapter 5) there are a large variety of defense systems, types of simulations, and purposes and levels of system aggregation for which simulations might be used.

Types of Simulation

The defense community recognizes three types of simulations: live, virtual, and constructive. A live simulation is simply an operational test, with sensors used to identify which systems have been damaged by simulated firings, using real forces and real equipment. It is the closest exercise to real use. A virtual simulation ("hardware-in-the-loop") might test a complete system prototype with stimuli either produced by computer or otherwise artificially generated. This sort of exercise is typical of a developmental test. A constructive simulation is a computer-only representation of a system or systems.

Thus, a simulation can range from operational testing itself to an entirely computer-generated representation (i.e., no system components involved) of how a system will react to various inputs. It can be used for various purposes—including test design and developmental and operational test evaluation—and at various levels of system aggregation, ranging from modeling a system's individual components (e.g., system software or a radar component, often ignoring the interactions of these components with the remainder of the system), to modeling an entire prototype, to modeling multiple system interactions.

The panel examined a very small number of simulations proposed for use in developmental or operational testing, and the associated presentations and documentation about the simulation and related validation activities were necessarily brief. They included: RAPTOR, a constructive model that estimates the reliability of systems based on their reliability block diagram representations; a constructive simulation used to estimate the effectiveness of the sensor-fuzed weapon; and a hardware-in-the-loop simulation used to assess the effectiveness of Javelin, an anti-tank missile system.

The panel did not perform an in-depth analysis of the validation of any of these models or their success in augmenting operational experience. However, preliminary impressions were that RAPTOR would be useful for assessing the reliability of a system only if the reliability of each component had been previously well estimated on the basis of operational experience and only if the assumed independence of the system's components was reasonable; the simulation for the Javelin was able to successfully measure system effectiveness for some specific scenarios; and the simulation used to determine which subsystems in a tank would be damaged by the sensor-fuzed weapon was not likely to be informative. This is not to say that the simulation for the sensor-fuzed weapon was not reasonably predictive, given current technology; however, the physics needed to predict the path of (and damage caused by) a fast-moving piece of metal impacting various points on a tank is much more complicated than for the other two simulation applications.

Since there are both effective and ineffective applications of modeling and simulation to (developmental and) operational testing, the key issues are how to identify the models and simulations that can be safely used to augment operational test experience; how to conduct model validation; and how the augmentation should be carried out. Given the breadth of possible applications, only a few of these issues can be addressed here.

A key theme of this chapter is that simulation is not a substitute for operational testing. The chapter defines a comprehensive model validation of a constructive simulation for use in operational testing and evaluation, discusses the use of constructive simulation to assist in operational test design and then to assist in operational test evaluation, and lastly discusses needed improvements to the current uses of modeling and simulation.

Limits of Simulation

The relevant law (paragraphs 2399(a) and 2399(h)(1), Title 10, U.S. Code) states:

> A major defense acquisition program may not proceed beyond low-rate initial production until initial operational test and evaluation of the program is completed; and the term 'operational test and evaluation' does not include an operational assessment based exclusively on computer modeling, simulation, or an analysis of information contained in program documents.

The panel strongly endorses this policy.

Models and simulations are typically constructed before substantial operational experience is available, either before or during developmental test. They are therefore based on information collected from similar systems, from components used in previous systems, or developmental testing of the current system. As a result (and almost by definition), a simulation often is unable to identify "unanticipated" failure modes—ones that are unique to operational experience. Consequently, it is almost axiomatic that for many systems a simulation can never fully replace an operational test. The challenge is not how to perform a perfectly realistic test, since that is nearly always impossible, but, instead, how to perform a test in which the lack of realism is not a detriment to the evaluation of the system.

There are undoubtedly some characteristics of systems for which modeling is clearly feasible: for example, estimating how quickly a cargo plane can be unloaded or estimating the availability of systems based on failure and repair rates that can be measured in operational conditions. These are situations for which the actions of the typical user are argued, a priori, not to critically affect system performance.

Developmental tests and tests of related systems can obviously provide information about a new system's operational performance. Developmental testing can provide substantial information about failure modes of individual components, including the full prototype. Since many components may have been previously used, possibly in a modified form, in other systems, test results and field data might be available and analyzed in conjunction with the current developmental tests. However, in developmental or laboratory testing, the actions of a typical user are not taken into consideration, and therefore the prototype is not tested as a system-user whole. Given the current lack of operational realism in much of developmental testing, some system deficiencies will not exhibit themselves until the most realistic form of full-system testing is performed. Therefore, failure to conduct some operational testing can result in erroneous effectiveness assessments or in missed failure modes and (possibly) optimistic estimates of operational system effectiveness or suitability. As an example, Table 9-1 shows the mean time between operational mission failures for the command launch unit of the Javelin for several types of testing modes, ranging from pure laboratory or developmental testing (RQT1) to initial operational testing. From the reliability qualification test II (RQTII) to the initial operational test (i.e., as troop handling becomes more typical of field experience), the mean time between operational mission failure decreases from 482 hours to 32 hours. Fries (1994b) discusses a number of similar situations where important system limitations were not observed until field testing.[1]

Of course, developmental tests can be structured to incorporate various aspects of operational use; elsewhere in the report we recommend that this be done more frequently, and this will support simulations that are based on more relevant experience. However, given that developmental tests and experiences on related systems are used to develop simulations, simulations are by nature more limited in the information that they can provide about the operational experience of new systems.

Some operational testing is very difficult. For example, testing multisystem engagements can be extremely costly and dangerous (or impossible, if not enough enemy systems are available). Cost arguments alone for simulations should not be made without some analysis of the tradeoffs. A small number of additional operational test runs may well provide more information, in many situations, than a large number of model runs; the information would be worth the cost. When the interaction of individual systems is well understood, simulations could be

[1] As mentioned previously, even field testing makes several compromises to full realism.

TABLE 9-1 Reliability Assessment of Command Launch Unit of Javelin in Several Testing Situations

Test	Troop Handling	Mean Time Between Operational Mission Failures
RQT I	None	63
RQT II	None	482
RDGT	None	189
PPQT	None	89
DBT	Limited	78
FDTE	Limited	50
IOT	Typical	32

NOTES:
RQT I, Reliability Qualification Test I
RQT II, Reliability Qualification Test II
RDGT, Reliability Development Growth Test
PPQT, Preproduction Qualification Test
DBT, Dirty Battlefield Test
FDTE, Force Development Test and Experimentation
IOT, Initial Operational Test

useful in extrapolating to interactions of larger groups of systems. The use of a small number of operational exercises to help calibrate a simulation is also of potential value and is worth further investigation. The Army Operational Test and Evaluation Command is planning on using this approach for the operational evaluation of the ATACMS/BAT system (see Appendix B).

MODEL VALIDATION

In order to consider using a simulation to augment test design or evaluation, one must understand the extent to which a simulation approximates the real system and in which respects it is more limited, in relation to the intended application. To develop this understanding, model validation is a key. The literature on model validation describes several activities, which can be separated into two broad types: external validation and sensitivity analysis. External validation is the comparison of model output with (operationally relevant) observations from the system being modeled. It is helpful if external validation is accompanied by an uncertainty analysis of a model, which is an analysis of the variability of the model output that results from typical variability in model inputs. An uncertainty analysis provides a yardstick for measuring which differences between model output and observations of the system are capable of being explained by model uncertainty, and, therefore, which are due to model inadequacy.

In stating that external validation is a comparison of model output with field observations, there is an (at least) implicit requirement for a performance criterion that indicates whether the simulation model conforms (adequately) to field experience. Agreement on this performance criterion is a crucial step in an external validation. Consider, for example, the sensor-fuzed weapon. A simulation was used to predict the damage resulting from use of the sensor-fuzed weapon against a tank. The criterion used by the Air Force Operational Test and Evaluation Center to validate the model was whether those subsystems the weapon actually damaged were predicted by the simulation to have a high probability of being damaged. The problem with this criterion is that there is no parallel check that subsystems that were *not* damaged were predicted by the simulation to have a *low* probability of damage. In other words, a simulation that simply predicted that all subsystems would be damaged with high probability would have scored well under this criterion. The stringent test of the simulation model that was not done would be an assessment of its ability to discriminate between subsystems that would and would not be damaged, with high probability. In this case, a laudable attempt to use field data to validate a simulation failed to achieve its intended result due to a poor understanding of appropriate validation criteria.[2]

Sensitivity analysis is an analysis of the response surface of a model, especially the sign and rough magnitude of the relationship between the inputs to and the outputs from a model. The simplest version of a sensitivity analysis is the analysis of the model output from the incremental changes in single inputs from a central scenario, referred to as one-variable-at-a-time sensitivity analysis. A sensitivity analysis can be indicative of model failings if the direction and magnitude of sensitivity of model outputs to various inputs does not agree with subject-matter expertise.

Since external validation of a model that predicts operational performance involves an operational test (or a developmental test with operational features), it can be expensive, and it is not surprising that, in the few examples we reviewed, the number of replications of external validation was fairly small.[3] Also, for the few examples reviewed, sensitivity analysis tended to be one-variable-at-a-time for a limited number of inputs. It is not unusual for constructive simulation models in defense testing applications to have hundreds or thousands of variable inputs and dozens of outputs of interest. While one-variable-at-a-time sensitivity analysis has the benefits of ease of interpretation and ease of estimation of partial

[2] This example also argues for greater statistical expertise in the test community, since this is a mistake that is well described in the statistical literature.

[3] We note, however, that there were certainly more than in other applications of modeling in public policy applications (see, e.g., Citro and Hanushek, 1991).

derivatives at the central scenario, using it will often result in missing important interactions between inputs and missing strong curvatures in the relationship between inputs and outputs. Such analysis is also limited to providing information on the behavior of the simulation in the neighborhood of the central scenario. It is also a very inefficient method for understanding the simulation model's response surface in the same way that experimental designs that vary only one factor between test runs are inefficient. Especially for constructive simulation models that have relatively fast turnarounds, there is no reason to perform such a limited analysis.

Face Validation

External validation and sensitivity analysis are often accompanied by the following activities: (1) model verification, checking to see that the computer code accurately reflects the modeler's specifications; (2) evaluation of model output either for particular inputs or for extreme inputs—where the resulting output can be evaluated using subject-matter expertise; and (3) comparisons of the model output with the output of other models or with simple calculations. Model verification, of course, is an extremely important activity. The quasi-validation activities (2) and (3) can lead to increased face validity—which we define here as the model's output agreeing with prior subject-matter understanding—which is often worthwhile.

However, it is important to stress that face validity is insufficient. If validation is limited to these latter activities, one can be misled by agreement with preconceived notions or with models that are based on a set of commonly held and unverified assumptions. In addition: there will be little or no support for important feedback loops to help indicate areas in the model in need of improvement, there will be little indication of the quality of the model for predicting various outputs of interest, and it will be impossible to construct hypothesis tests that indicate whether discrepancies with the results of other models or with field results are due to natural variation or to real differences in the model(s). Therefore, face validity must be augmented with more rigorous forms of model validation.

Improved Methods for Model Validation

In the last 15-20 years a number of statistical advances have been made that are relevant to the practice of model validation; we have not seen evidence of their use in the validation of constructive simulation models used for defense testing. We note five of the advances that should be considered:

1. Morgan and Henrion (1990), McKay(1995), and others have identified techniques for carrying out an uncertainty analysis[4] in order to assist external validation.

2. One-variable-at-a-time sensitivity analysis assumes a linear response surface, is inefficient in its use of model runs, and is too narrowly focused on a central scenario. These deficiencies can be addressed by using inputs produced by Latin Hypercube sampling (McKay et al., 1979; Iman and Conover, 1980; Owen, 1992) or related fractional factorial techniques. To help analyze the resulting paired input-output data set, McKay's (1995) ANOVA decomposition of outputs from models run on inputs selected by Latin Hypercube sampling is useful for identifying which outputs are sensitive to which inputs. The response surface of these input-output vector pairs can be summarized using various nonparametric regression techniques, such as multivariate adaptive regression splines (Friedman, 1991). For a single output, there are techniques that can identify a small collection of important inputs from a larger collection of candidate inputs. These techniques can be very helpful in simplifying an uncertainty analysis, a sensitivity analysis, and even an external validation, since it indicates which variables are the crucial ones to vary (see Cook, 1994; Morris, 1991; McKay, 1997; and others).

3. One formal approach to the design and analysis of computer experiments has been developed by Sacks et al. (1989), Currin et al. (1991), Morris, Mitchell, and Ylvisaker (1993), and others.

4. Identifying outliers to patterns shown by the great majority of inputs to a constructive simulation can be used to better understand regions of the input-output space in which the behavior of the simulation changes qualitatively (see, e.g., Huber, 1981; Hampel et al., 1986; Atkinson, 1994).

5. The developers of a constructive simulation often make fairly arbitrary choices about model form. Citro and Hanushek (1991), Chatfield (1995), Draper (1995), and others offer ways of addressing this problem of misspecification of the relationship between inputs and outputs in computer and statistical models, wherein the simulation is less precise than one measures by using the usual "analysis of variance" techniques.

Some of the above ideas may not turn out to be directly applicable to defense models, but the broad collection of techniques being developed to analyze nonmilitary simulations are likely to be relevant. Given the importance of operational testing, testing personnel should be familiar with this literature to determine its value in the validation of constructive simulations.

[4] As noted above, a sensitivity analysis is the study of the impact of changes on model outputs from changes in model inputs and assumptions. An uncertainty analysis is the attempt to measure the total variation in model outputs due to *quantified* uncertainty in model inputs and assumptions and the assessment of which inputs contribute more or less to total uncertainty.

In addition to model validation, a careful analysis of the assumptions used in developing constructive simulation models is a necessary condition for determining the value of the simulation. Beyond the usual documentation, which for complicated models can be fairly extensive, an "executive summary" of key assumptions used in the simulation model should be provided to experts to help them determine their reasonableness (and therefore the utility) of the simulation. A full history of model development, especially any modification of model parameters and their justification, should also be made available to those with the responsibility for accrediting a model for use in operational testing.

"Model-test-model" is the use of (constructive) simulation models in conjunction with operational test. In model-test-model, a model is developed, a number of operational test runs are carried out, and the model is modified by adjusting parameters so that it is more in agreement with the operational test results. Such external validation on the basis of operational use is extremely important in informing simulation models used to augment operational testing. However, there is an important difference (one we suspect is not always well understood by the test community) between comparing simulation outputs with test results and using test results to adjust a simulation. Many complex simulations involve a large number of "free" parameters—those that can be set to different values by the analyst running the simulation. In model-test-model some of these parameters can be adjusted to improve the correspondence of simulation outputs with the particular operational test results with which they are being compared. When the number of free parameters is large in relation to the amount of available operational test data, close correspondence between a "tuned" simulation and operational results does not necessarily imply that the simulation would be a good predictor in any scenarios differing from those used to tune it. A large literature is devoted to this problem, known as overfitting.[5]

An alternative that would have real advantage would be "model-test-model-test," in which the final test step, using scenarios outside of the "fitted" ones, would provide validation of the version of the model produced after tuning and would therefore be a guard against overfitting. If there was interest in the model being finalized before any operational testing was performed, this would be an additional reason for developmental testing to incorporate various operationally realistic aspects.

[5] Overfitting is said to occur for a model and data set combination when a simple version of the model (selected from a model hierarchy, formed by setting some parameters to fixed values) is superior in predictive performance to a more complicated version of the model formed by estimating these parameters from the data set. For some types of statistical models, there are commonly accepted measures of the degree of overfitting. An example is the Cp statistic for multiple regression models: a model with high Cp could be defined as being overfit.

Recommendation 9.1: Parameters from modeling and simulation should not be used to fit a simulation to a small number of field events without subsequent validation of the resulting simulation.

A Process of Comprehensive Validation

The panel reviewed several documents that describe the process used to decide whether to use a simulation model to augment an operational test. There are differences across the services, but the general approach is referred to as verification, validation, and accreditation. Verification is "the process of determining that model implementation accurately represents the developer's conceptual description and specifications" (U.S. Department of Defense, 1994a). (For constructive simulations, verification means that the computer code is a proper representation of what the software developer intended; the related software testing issues are discussed in Chapter 8.) Validation is "the process of determining (a) the manner and degree to which a model is an accurate representation of the real-world from the perspective of the intended uses of the model, and (b) the confidence that should be placed on this assessment" (U.S. Department of Defense, 1994a). Accreditation is "the official certification that a model or simulation is acceptable for use for a specific purpose" (U.S. Department of Defense, 1994a). The panel supports the general goals of verification, validation, and accreditation and the emphasis on verification and validation and the need for formal approval, that is, accreditation, of a simulation model for use in operational testing.

Given the crucial importance of model validation in deciding the utility of a simulation for use in operational test, it is surprising that the constituent parts of a comprehensive validation are not provided in the directives concerning verification, validation, and accreditation. A statistical perspective is almost entirely absent in these directives. For example, there is no discussion of what it means to demonstrate that the output from a simulation is "close" to results from an operational test. It is not clear what guidelines model developers or testers use to decide how to validate their simulations for this purpose and how accrediters decide that a validation is sufficiently complete and that the results support use of the simulation. Model validation cannot be algorithmically described, which may be one reason for the lack of specific instruction in the directives. A test manager would greatly benefit from examples, advice on what has worked in the past, what pitfalls to avoid, and most importantly, specific requirements as to what constitutes a comprehensive validation.

This situation is similar to that described in Chapter 1, regarding the statistical training of those in charge of test planning and evaluation. Model validation has an extensive literature, in a variety of disciplines, including statistics and operations research, much of it quite technical, on how to demonstrate that a computer model is an acceptable representation of the system of interest for a specific application. Operational test managers need to become familiar with the general techniques represented in this literature, and have access to experts as needed.[6]

We suggest, then, a set of four activities that can jointly form a comprehensive process of validation: (1) justification of model form, (2) an external validation, (3) an uncertainty analysis including the contribution from model misspecification or alternative specifications, and (4) a thorough sensitivity analysis.

1. All important assumptions should be explicitly communicated to those in a position to evaluate their merit. This could be done in the "executive summary" described above.

2. A model's outputs should be compared with operational experience. The scenarios chosen for external validation of a model must be selected so that the model is tested under extreme as well as typical conditions. The need to compare the simulation with operational experience raises a serious problem for simulations used in operational test design, but it can be overcome by using operationally relevant developmental test results. Although external validation can be expensive, the number of replications should be decided based on a cost-benefit analysis (see the discussion in Chapter 5 on "how much testing is enough"). External validation is a uniquely valuable method for obtaining information about a simulation model's validity for use in operational testing, and is vital for accreditation.

3. An indication of the uncertainty in model outputs as a function of uncertainty in model inputs, including uncertainty due to model form, should be produced. This activity can be extremely complicated, and what is feasible today may be somewhat crude, but DoD experience at this will improve as it is attempted for more models. In addition, exploration of alternative model forms will have benefits in providing further understanding of the advantages and limitations of the current model and in suggesting modifications of its current form.

4. An analysis of which inputs importantly affect which outputs, and the direction of the effect, should be carried out and evaluated by those with knowledge of the system being developed. The literature cited above suggests a number of methods for carrying out a comprehensive sensitivity analysis. It will often be necessary to carry out these steps on the basis of a reduced set of "important" inputs: whatever process is used to focus the analysis on a smaller number of inputs should be described.

Recommendation 9.2: Validation for modeling and simulation, used to assist in the design and evaluation of operational testing of defense sys-

[6] There are tutorials that are provided at conferences, and other settings, and excellent reports in the DoD community (e.g., Wiesenhahn and Dighton, 1993), but they are not sufficient since they do not reflect recent statistical advances.

tems, should include: (1) a complete description of important assumptions, (2) external validation, (3) uncertainty analysis, and (4) sensitivity analysis. A description of any methods used to reduce the number of inputs under analysis should be included in each of the steps. Models and simulations used for operational testing and evaluation must be archived and fully documented, including the objective of the use of the simulation and the results of the validation.

The purpose of a simulation is a crucial factor in validation. For some purposes, the simulation only needs to be weakly predictive, such as being able to rank scenarios by their stress on a system, rather than to predict actual performance. For other purposes, a simulation needs to be strongly predictive. Experience should help indicate, over time, which purposes require what degree and what type of predictive accuracy.

Models and simulations are often written in a general form so that they will have wide applicability for a variety of related systems. An example is a missile fly-out model, which might be used for a variety of missile systems. A model that has been used previously is often referred to as a *legacy* model. In an effort to reduce the costs of simulation, legacy models are sometimes used to represent new systems, based on a complete validation for a similar system. Done to avoid costly development of a de novo simulation, this use of a legacy model presents validation challenges. In particular, new systems by definition have new features. Thus, a legacy model should not be used for a new application unless: a strong argument can be made about the similarity of the applications and an external validation with the new system is conducted. A working presumption should be that the simulation will not be useful for the new application unless proven otherwise.

MODELING AND SIMULATION IN TEST DESIGN

Modeling and simulation may have their greatest contribution to operational test through improving operational test design. Modeling and simulation were used to help plan the operational test for the Longbow Apache (see Appendix B). Constructive simulation models can play at least four key roles.

First, simulation models that properly incorporate both the estimated heterogeneity of system performance as a function of various characteristics (of test scenarios), as well as the size of the remaining unexplained component of the variability of system performance, can be used to help determine the error probabilities of any significance tests used in assessing system effectiveness or suitability. To do this, simulated relationships (based on the various hypotheses of interest) between measures of performance and environmental and other scenario characteristics can be programmed, along with the description of the number and characteristics of the test scenarios, and the results tabulated as in an operational test. Such replications can be repeated, keeping track of the percentage of tests that the system passed. This approach could be a valuable tool in computing error probabilities or operating test characteristics for nonstandard significance tests.

Second, simulation models can help select scenarios for testing. Simulation models can assist in understanding which factors need controlling and which can be safely ignored in deciding which scenarios to choose for testing, and they can help to identify appropriate levels of factors. They can also be used to choose scenarios that would maximally discriminate between a new system and a baseline system. This use requires a simulation model for the baseline system, which presumably would have been archived. For tests for which the objective is to determine system performance in the most stressful scenario(s), a simulation model can help select the most stressful scenario(s). As a feedback tool, assuming that information is to be collected from other than the most stressful scenarios, the ranking of the scenarios with respect to performance from the simulation model can be compared with that from the operational test, thereby providing feedback into the model-building process, to help validate the model and to discover areas in which it is deficient.

Third, there may be an advantage in using simulation models as a living repository of information collected about a system's operational performance. This repository could be used for test planning and also to chart progress towards development, since each important measure of performance or effectiveness would have a target value from the Operational Requirements Document, along with the values estimated at any time, using either early operational assessments or, for requirements that did not have a strong operational aspect, the results from developmental testing. The Air Force Operational Test and Evaluation Center is in the process of testing this concept for the B-1B defensive system upgrade.

Fourth, every instance in which a simulation model is used to design an operational test, and the test is then carried out, presents an opportunity for model validation. The assumptions used in the simulation model can then be checked against test experience. Such an analysis will improve the simulation model under question, a necessary step if the simulation model is to be used in further operational tests or to assess the performance of the system as a baseline when the next innovation is introduced. Feedback of this type will also help provide general experience to model developers as to which approaches work and which do not. (Of course, this kind of feedback will not be possible without the data archive recommended in Chapter 3. Also mentioned in Chapters 3, 6, and 8, inclusion of field use data in such an archive provides great opportunities for validation of methods used in operational test design.)

Recommendation 9.3: Test agencies in the military services should increase their use of modeling and simulation to help plan operational

tests. The results of such tests, in turn, should be used to calibrate and validate all relevant models and simulations.

Recommendation 9.4: Simulation should be used throughout system development as a repository of accumulated information about past and current performance of a system under development to track the degree of satisfaction of various requirements. The repository would include use of data from all relevant sources of information, including experience with similar systems, developmental testing, early operational assessments, operational testing, training exercises, and field use.

A final note is that validation for test design, although necessary, does not need to be as comprehensive as validation for simulation that is to be used for augmenting operational test evaluation. One can design an effective test for a system without understanding precisely how a system behaves. For example, simulation can be used to identify the most stressful environment without knowing what the precise impact of that environment will be on system performance.

MODELING AND SIMULATION IN TEST EVALUATION

The use of modeling and simulation to assist in the operational evaluation of defense systems is relatively contentious. On one side, modeling and simulation is used in this way in industrial (e.g. automobile) applications. Simulation can save money, is safer, does not have the environmental problems of operational test, is not constrained in defense applications by the availability of enemy systems, and is always feasible in some form. On the other side, information obtained from modeling and simulation may at times be limited in comparison with that from operational testing. Its exclusive use may lead to unreliable or ineffective systems passing into full-rate production before major defects are discovered.

An important example of a system for which the estimated levels for measures of effectiveness changed due to the type of simulation used is the M1A2 tank. In a briefing for then Secretary of Defense William Perry (see Wright, 1993), detailing work performed by the Army Operational Test and Evaluation Command, three simulation environments were compared: constructive simulation, virtual simulation, and live simulation (essentially, an operational test). The purpose was to “respond to Joint Staff request to explore the utility of the Virtual Simulation Environment in defining and understanding requirements.” In the test, the constructive model indicated that the M1A2 was better than the M1A1. The virtual simulation indicated that M1A2 was not better, which was confirmed by the field test. (The problems with the M1A2 had to do, in part, with immature software.) The specific limitations of the constructive simulation were that the various assumptions underlying the engagements resulted in the M1A2 detecting and killing more targets. Even though the overall results agreed with the field test, the virtual simulation was found to have problems as well. The primary problem was the lack of fidelity of the simulated terrain, which resulted in units not being able to use the terrain to mask movements or to emulate having dug-in defensive positions. In addition, insufficient uncertainty was represented in the scenarios. In this section we discuss some issues concerning how to use validated simulations to supplement operational test evaluation.

(The use of statistical models to assist in operational evaluation—possibly in conjunction with the use of simulation models—is touched on in Chapter 6. An area with great promise is the use of a small number of field events, modeling and simulation, *and* statistical modeling, to jointly evaluate a defense system under development. Unfortunately, the appropriate combination of the first two information sources with statistical modeling is extremely specific to the situation. It is, therefore, difficult to make a general statement about such an approach, except to note that it is clearly the direction of the future, and research should be conducted to help understand the techniques that work.)

Modeling and simulation have been suggested as ways of extrapolating or interpolating to untested situations or scenarios. There are two general types of interpolation or extrapolation that modeling and simulation might be used to support. First, in *horizontal* extrapolation, the operational performance of a defense system is first estimated at several scenarios—combinations of weather, day or night, tactic, terrain, etc. Simulation is then used to predict performance of the system in untested scenarios. The extent to which the untested scenarios are related to the tested scenarios typically determines the degree to which the simulation can predict performance. This extrapolation implies that the tested scenarios need to be selected (to the extent possible) so that the modeled scenarios of interest have characteristics in common with the tested scenarios (see discussion in Chapter 5 on Dubin's challenge). One way to ensure this commonality is to use factor levels to define the modeled scenarios that are less extreme than those used in the tested scenarios. In other words, extrapolation to an entirely different sort of environment would be risky, as would extrapolation to a similar environment, but at a more extreme level. For example, if a system was tested in a cold (32°F) and rainy (0.25″ per hour) environment and in a hot (85°F) and dry (0.00″ per hour) environment, there would be some reason to hope, possibly using statistical modeling, for a simulation to provide information about the performance of the system in a moderately hot (65°F) and somewhat rainy (.10″ per hour) environment. (The closer the untested environment is to the tested one, the closer one is to interpolation than extrapolation.) However, if no tested environments included any rain, it would be risky to use a simulation to extrapolate to rainy conditions based on the system performance in dry conditions. (Accelerated life testing, discussed in Chapter 7, is one way to extrapolate with respect to level.)

Second, *vertical* extrapolation is either from the performance of a single system (against a single system) to the performance of multiple systems in a team, against a team of enemy systems; or from the performance of individual subsystem components to performance of the full system. The first type of vertical extrapolation involves an empirical question: whether the operational performance estimated for a single system can be used in a simulation to provide information about multiple system engagements. Experiments should be carried out in situations in which one can test the multiple system engagement to see whether this type of extrapolation is warranted. This kind of extrapolation should often be successful, and given the safety, cost, and environmental issues raised by multisystem engagements, it is often necessary. The second type of vertical extrapolation depends on whether information about the performance of components is sufficient for understanding performance of the full system. There are systems for which a good deal of operational understanding can be gained by testing portions of the system, for example, by using hardware-in-the-loop simulations. This is, again, an empirical question, and tests can be carried out to help identify when this type of extrapolation is warranted. This question is one for experts in the system under test rather than a statistical question.

The ATACMS/BAT calibration described in Appendix B represents both horizontal and vertical extrapolation. First, extrapolation is made to different types of weather, terrain, and tactics. Second, the extrapolation is made from several tanks to a larger number of tanks. The first extrapolation requires more justification than the second. In such situations, it might be helpful to keep the degree of true extrapolation to a minimum through choice of the test scenarios.

A third possible type of extrapolation is *time* extrapolation; the best example is reliability growth testing (see Chapter 7).

In conferences devoted to modeling and simulation for testing of military systems, most of the presentations have been concerned with potential and future uses of simulations for operational test evaluation. We have found few examples of constructive simulations that have been clearly useful for identifying operational deficiencies of a defense system. This lack may be due to the limitations of modeling and simulation for this purpose, to its lack of application, or to the lack of feedback to demonstrate the utility of such a simulation. To make a strong case for the increased use of modeling and simulation in operational testing and evaluation, examples of simulation models that have successfully identified operational deficiencies that were missed in developmental test need to be collected, and the simulations analyzed to understand the reasons for their success.

We are reluctant to make general pronouncements about which type of simulations would be effective or ineffective for operational assessment of a military system. Everything else being equal, the order of preference from most preferred to least preferred should be live, virtual, and then constructive simulation. For constructive simulations and the software aspects of virtual simulation, the more "physics-based" the better: the actions of the system (and any enemy systems) should be based on well-understood and well-validated physical representations of the process. The use of computer-aided design and computer-aided manufacturing (CAD/CAM) representations of a system, for just this reason, is clearly worth further exploration for use in modeling and simulation for defense testing. In all cases, however, comprehensive model validation and feedback from real experience will indicate which approaches work and which do not in given situations.

IMPROVED ADMINISTRATION OF MODELING AND SIMULATION

While DoD directives are complete in their discussion of the need for validation and documentation, it is not clear that testers understand how to implement the directives. Not much seems to have changed since the following was noted over 10 years ago (U.S. General Accounting Office, 1987a:46, 61):

> In general, the efforts to validate simulation results by direct comparison to data on weapon effectiveness derived by other means were weak, and it would require substantial work to increase their credibility. Credibility would also have been helped by . . . establishing that the simulation results were statistically representative. Probably the strongest contribution to credibility came from efforts to test the parameters of models and to run the models with alternative scenarios.
>
> The office of the secretary of the Department of Defense has issued no formal guidance specifically for the management of simulations or how to conduct them and assess their credibility. Although several directives and at least one military standard have some bearing on simulations, we found no documented evidence that the secretary's office has sought to develop and implement appropriate quality controls that could be expected to directly improve the credibility of simulations.

A similar, more recent, observation was made by Giadrosich (1990):

> The problem of how to keep track of the many assumptions, input data sources, multiple configurations, and so forth associated with digital simulation models, and how to defend the resulting validity and accreditation of the models has not been adequately addressed. These issues are further complicated by the fact that a widely accepted formal process for model validation and accreditation does not exist. The lack of transparency in most complex modeling analyses, the inability to get agreement on the basic assumptions, and the ability of these assumptions to drive the analysis results usually lead to poor model credibility.

And independent contractors have observed (Computer Sciences Corporation, 1994:ii,3,4):[7]

> The major disadvantages noted are: 1) potential inconsistencies in approach implementation; 2) potential biases of the analysts coupled with resource limi-

[7] Some of these comments refer to more general use of modeling and simulation and not specifically to modeling and simulation for operational test augmentation.

> tations that might cause incomplete data gathering and/or improper comparisons; and 3) a possible lack of uniformity and experience in establishing acceptance criteria for each specific application.
>
> Other interviewees pointed out that there is seldom enough time to carry out a formal model accreditation for a particular application since the study results are often required in the matter of a few weeks. . . . Occasionally, the lack of time forces analysts to use models that are not really suited to the study but are either the only ones available or the only ones with which the study analysts are familiar.
>
> Several V&V [validation and verification] techniques were mentioned as being valuable in supporting accreditation. Although almost all interviewees recognized the desirability of performing extensive verification code checks and in depth comparisons between model results and real world data, funding and time constraints frequently preclude such extensive V&V. Instead, many model users turn to other less costly methods that are also less beneficial. Such methods typically include reviews of past usage and VV&A [validation, verification, and accreditation] results, face validation, and some comparisons between models.
>
> Many organizations, faced with a lack of time and resources to perform in depth V&V, have relied on prior V&V, face validation, and benchmarking as the basis for informal model accreditation. Among those who use these methods are the Army Aviation and Troop Command (ATCOM), the Air Force Operational Test and Evaluation Center (AFOTEC), AMSAA [U.S. Army Materiel Systems Analysis Activity], OPTEC [U.S. Army Operational Test and Evaluation Command], ASC [Aeronautical Systems Command], and NAVAIR [Naval Air Systems Command]. A common view is that past usage, coupled with reviews by subject matter experts (SMEs) or comparisons between models, are sufficient to justify model selection and use.

All of these comments suggest an operational testing and evaluation community that needs guidance. A number of organizational improvements might remedy the cited problems. The key to the following suggestions are independence, expertise, and resources: validation must be carried out by those without a vested interest in the outcome of the validation process, with relevant expertise, and with sufficient resources committed to ensure comprehensive validation. Without these criteria, simulations in operational test design and evaluation should not be used.

Validation needs to be carried out by individuals with expertise. This means in-depth knowledge of the system being modeled, as well as facility with the technical, computer, and statistical tools needed to assess the model. The validation needs to be carried out by individuals that do not have a preference as to whether the model receives accreditation. System developers obviously have in-depth knowledge, and they usually have modeling (and possibly model validation) expertise, but they have a strong preference that the model be accredited. The operational tester has no preference with respect to accreditation, except insofar as it permits information to be gathered when field testing is too difficult, but the tester may have limited modeling or model validation skills. Competitors to the system developers—those that bid on the same contract—will have the expertise, but they may be biased against accreditation. Therefore, it may be difficult to satisfy the joint desires of expertise and independence, but the closer DoD can come to this, the better. The Navy Operational Test and Evaluation Force generally uses the Center for Naval Analysis to perform much of its model validation, which seems an excellent combination of independence and expertise. The other services might investigate whether other federally financed research and development centers, or possibly teams of academics might be used for this purpose.

As for resources, a substantial fraction of the money awarded for production of a simulation model should be earmarked for validation. The precise amount should be related to such considerations as the expense of external validation and the complexity of the simulation model. These funds should be protected from "raiding" by the developers of the simulation model if shortfalls in the simulation development budget occur.

As a final check on the comprehensiveness of validation and accreditation, the Office of the Director, Operational Test and Evaluation (DOT&E) should have the final say as to whether a particular simulation model can be used in conjunction with an operational test. The service test agency should provide all documentation concerning validation activities to DOT&E, along with the plan of how the simulation is to be used, *in advance of its use*. DOT&E would then examine the materials to determine whether the service test agency should be allowed to go forward with its plan to use the simulation model in operational test design or evaluation. Alternatively, each service could have its own modeling and simulation validation center, with a major function of the overall center being to oversee and support the individual service centers.

Finally, there is a need to develop a center of expertise of modeling and simulation for test design and evaluation in DoD. Such a unit could provide clear descriptions of what constitutes a comprehensive validation, examples of successful and unsuccessful experience in using simulation models for various purposes, and expertise as to the statistical issues that arise in the use of simulation for operational test design and evaluation. Either the recently established Defense Modeling and Simulation Office should add these responsibilities to its mission, or a new center for simulation; validation, verification, and accreditation; and testing should be established.

> **Recommendation 9.5: A center on use of modeling and simulation for testing and acquisition should be established, with a primary focus on validation, verification, and accreditation. This center could be included in the charter of the Defense Modeling and Simulation Office or established by another relevant organization in the defense community.**

The notion of a casebook of examples is especially appealing. It could help educate test managers about the approaches to simulation of the operational performance of a defense system that work (and which do not), based on field use data, how these simulations were designed, and what statistical issues arise in relating information from the simulation to the field testing.

> **Recommendation 9.6: Modeling and simulation successes and failures for use in operational test design and evaluation should be collected in a casebook so that information on the methods, benefits, risks, and limitations of modeling and simulation for operational test can be developed over time.**

The effectiveness of modeling and simulation for operational testing and evaluation is idiosyncratic. Some methods that work in one setting might not work in another. Indeed, it is unclear whether a simulation model could be declared to be working well when there is in fact limited information on operational performance prior to operational testing. Experiments need to be run for systems for which operational testing is relatively inexpensive and simulation models (which in this case would be redundant) developed to see if they agree with the field experience. Problems that are identified in this way can help analysts understand the use of models and simulations in situations in which they do not have an opportunity to collect full information on operational performance. Even in situations when operational testing is limited, it will be important to reserve some test resources to evaluate any extrapolations. For many types of systems, the state of the art for modeling and simulation is lacking, and field testing must stand on its own.

10

Increasing Access to Statistical Expertise for Operational Testing

As we argue throughout this report, operational testing and, more broadly, the process of system development are activities that can greatly benefit from the (further) application of statistical methods and statistical principles. The panel repeatedly learned of operational tests for which additional statistical knowledge and expertise could have been used to improve test design and evaluation. Examples of missing statistical knowledge include:

- applying principles of optimal experimental design to nonstandard situations,
- methods for combining information from developmental test and tests of other systems with the results of operational test;
- state-of-the-art validation of simulation models for augmenting operational testing;
- use of Markov chain methods for software testing; and
- use of such techniques as nonhomogeneous Poisson processes to model time between failures for a defense system.

As detailed in Chapter 3, the acquisition process itself would benefit from the changes that statistically based principles have brought about in many industrial applications of system development.

The panel is aware of some efforts in the DoD test community to make greater use of the statistical knowledge and tools available to it and to increase the statistical resources that are at hand. Courses in statistics are made available to staff at the service test agencies, particularly the Air Force Operational Test and

Evaluation Center, in relevant areas such as experimental design and reliability theory. The service test agencies, particularly the Army Operational Test and Evaluation Command, make use of statistical consultants. Also, DOT&E has access to expert statistical assistance at the Institute for Defense Analyses. All the service test agencies have military staff who operate with both dedication and professionalism, but they operate in the context of statistical training that prepares them to apply standard methodology, rather than to produce customized solutions as needed. Finally, the service test agencies make use of and occasionally develop statistical software to help in test design and evaluation. RAPTOR, developed at the Air Force Operational Test and Evaluation Center and used to evaluate the reliability of component systems, is a particularly relevant and impressive example of this.

However, the DoD test community generally has limited access to, and makes little use of, individuals who have highly advanced training in statistics, specifically, the level of training that is typical of a doctorate from a graduate program in statistics. The panel knows of only one Ph.D.-level statistician who is a full-time employee in any of the three largest service test agencies at this time. The service test agencies also do not make enough use of the statistical expertise at the Naval Postgraduate School, the Center for Naval Analysis, the Institute for Defense Analyses (through DOT&E), Aerospace, RAND, and other similar institutions. It appears that the Army, Navy, and Air Force test agencies rarely consult with academic statisticians, even for test design and evaluation issues concerning multibillion dollar systems.

CONSTRAINTS ON ACCESS TO STATISTICAL EXPERTISE

Some of the limited interaction with statistical experts is understandable. First, as stated above, the problems in design and evaluation are heavily constrained by various factors, including test designs that are constrained by budgets to have small sample sizes, test facility scheduling constraints, and test facility limitations. Navy tests seem particularly constrained. Evaluations are often limited by time, and they are focused on the calculation of means and percentages and, sometimes, significance tests for identified measures to be used in the decision regarding whether to enter into full-rate production; this focus reduces incentives for more thorough and sophisticated analyses. These constraints can at times limit the utility gained through interaction with statistical experts. (We note, however, that many of the constraints would disappear with adoption of a test and acquisition strategy recommended in Chapter 3.) Yet these constraints can also increase the value of interactions with statistical experts, since constraints present nonstandard design problems; budgetary limitations make efficient test design even more important; and evaluations can be expedited by making use of effective diagnostics that quickly identify data features worth investigating.

Second, military staff rotate from one assignment to another as they are promoted, and it is not reasonable for them to have comprehensive or advanced training in statistics, which would necessarily limit their education for other areas important to their success in later assignments. Therefore, we believe the military staff of service test agencies cannot have appreciably greater statistical training than they currently do; however, the civilian work force has no such limitation.

Third, to be of assistance in any applied setting and work with system specialists, particularly in defense testing, a consulting statistician would have to be knowledgeable about DoD systems, how these systems are used, and DoD procedures. A key example of when the interaction between statistician and subject-matter specialist would be most useful is in identifying design flaws in a system under development. Statistics can be extremely useful towards this goal, though significant progress would depend on close interaction of a statistician, system specialists, and operational testers. Since knowledge of defense systems and procedures could take quite a while to acquire, a long-term commitment would be needed on the part of, say, a service test agency, and a statistical consultant. Given the uneven demand for statistical expertise in any given program, such a commitment would need to be carefully considered.

Fourth, there is the widespread belief in the DoD test community that statisticians will only recommend unaffordably large operational tests, based on arguments related to tests with sufficient power, regardless of the costs of test units; that they will ask for too much time for evaluation, regardless of the need for a timely decision; and, more broadly, that the discipline of statistics is generally useful only for large sample size questions. The last point, in particular, does not reflect awareness of the most up-to-date statistical advances. One of the themes of this report is that, through careful modeling and efficient experimental design or both, it is possible to achieve substantial resources savings and, in particular, a reduction in the sample size required for a fixed level of precision.

While the panel accepts the first three arguments above as possible constraints on the availability and utility of statistical expertise, it remains that test design, test evaluation, and system development are all activities that are informed through the application of statistical techniques and principles. We therefore argue that the DoD test and acquisition community needs to develop increased interaction with statistical expertise at the doctorate level. Although some of the technical issues discussed in this report are fairly routine, many require methods that are either not presented until late in graduate work or are current research problems. Therefore, the need for interaction with expert statisticians is clear and currently unmet. This lack of high-level statistical advice and interaction in the DoD test community also reduces the chances that when this type of expertise is needed in new application areas, it will be recognized.

Conclusion 10.1: The level of statistical expertise in the service test agencies and at the Director, Operational Test and Evaluation is currently inadequate to effectively carry out their missions.

RECOMMENDATIONS

There are a variety of ways to enhance access to statistical expertise in the defense testing community, considering the constraints discussed above. More statistical expertise can come in at least four forms: (1) increased training of current staff, (2) increased hiring of expert (master's- and doctorate-level) statisticians, (3) increased use of statistical consultants including those available through federally financed research and development centers, government laboratories, and supporting institutions, and (4) increased hiring of temporary staff through interagency professional agreements, use of temporary openings for academic statisticians on sabbaticals, and formation of fellowships in conjunction with groups such as the American Statistical Association.

Increased access to statistical expertise and more staff expertise in statistics can be instituted at several places: DOT&E, the service test agencies, and institutions such as the Institute for Defense Analyses, the Naval Postgraduate School, Aerospace, the Center for Naval Analysis, Lincoln Laboratories, and universities. Since there are four methods for enhancing statistical expertise and three types of places where it can occur, there are a number of possibilities for new approaches. As discussed above, when considering anything other than training existing staff, a key problem is that for individuals to be useful they must know more than statistics; they must also have some knowledge of the defense acquisition system and the system under test. They should also be somewhat familiar with the physics of combat systems and the realities of military operations. It would also be extremely valuable for the individuals to have some training in physics and engineering. (While this need for some subject-matter expertise is true in most applied settings of statistics—e.g., biostatisticians usually need to be familiar with medicine—it is extremely important in this situation.) We are confident that creative solutions to this problem can be developed by the defense testing and academic communities.

The development of statistical expertise for the defense testing agencies need not be focused exclusively on test and evaluation. The interaction of the defense and statistical communities, in general, is less than it could be; one approach could be the establishment of a defense statistical consulting unit within the Department of Defense, where test design and evaluation would be one among several of its primary responsibilities. Another approach may come from the academic community: for example, the Georgia Institute of Technology has recently established the Test and Evaluation Research and Education Center, a program that will be conducive to various kinds of interactions with the defense testing community, such as sabbaticals for visiting faculty at federally financed research and development centers and sabbaticals for members of the defense testing community at the university. In addition, graduates of this program could be provided with substantial statistical training and other expertise to prepare them for employment at service test agencies. One possible use of academic statisticians would be to assist in developing and advancing educational programs on test design and evaluation for professional staffs of service test agencies and monitoring their implementation.

Finally, other things being equal, the closer the statistical expertise is to those involved with test and evaluation, the more likely it is that they will be used.

> **Recommendation 10.1: The service test agencies should place greater emphasis on the statistical training of their nonmilitary staff, especially in the areas of experimental design, reliability theory, data analysis, use of statistical software, and total quality management.**

In addition to increasing the availability of statistical expertise in the defense testing community, it is also desirable to increase the familiarity of defense decision makers, particularly those in the acquisition process, with statistical concepts and the benefits that advanced statistical applications can provide to improving the process of defense acquisition.

There is some evidence that improving the decision makers' understanding of the application of statistical principles could reduce the likelihood that they would unintentionally approve goals that almost guarantee poor operational test and evaluation results. One straightforward example was the approval of an original objective for a jamming system that would improve aircraft survivability by X percent over the performance without the jamming system: The objective was impossible to meet because the application of the criterion would have required achieving a survivability greater than 100 percent. A second example was the setting of a requirement that a replacement item be "at least as reliable as the item it was to replace and that this requirement be met with a confidence factor of Y percent." The unintended result was that the developers of the replacement system had to strive for an extremely high design reliability, potentially at an unreasonably high cost, in an attempt to achieve the combination of the reliability and confidence specified. If the approval authorities of the requirements for the replacement item had better understood the statistical implications of what they were requiring the designers to achieve, in all likelihood they would have made a different decision. In the second example above, the performance standard was inadvertently set too high; in the first example it was impossible to meet. Developing a better appreciation of statistical principles could help DoD acquisition decision makers both avoid making bad decisions and setting unachievable goals and improve the quality of test plans and evaluations provided to them.

Over two decades ago, Deputy Secretary of Defense David Packard expressed major concern about the growth of cost estimates during the system development process and with the poor initial estimates. As a result he required the use of parametric cost estimates[1] to test the reasonableness of estimates provided when the Department was making major commitments of funds for development and initial production (Packard, 1971). This led Secretary of Defense Melvin Laird to establish the Cost Analysis Improvement Group (CAIG) as part of his office (Laird, 1972); it continues to function well today. The initial CAIG objectives were to review the cost estimates provided to the Defense Systems Acquisition Review Council (the predecessor to the current Defense Acquisition Board) and to develop uniform criteria to be used by all DoD units making such cost estimates. By the time DoD Directive 5000.4 provided a permanent charter for the CAIG, the membership had been expanded beyond the five members from the Office of the Secretary to include service members appointed by the secretaries of the individual military departments (U.S. Department of Defense, 1973). We mention this story because we believe that some of the concepts applied to improving cost analysis can be applied to improving statistical analysis for acquisition programs.

Specifically, we believe it would be desirable for the Department of Defense to form a small group of the best statistically trained individuals available in the Office of the Secretary and in the individual services who are involved in the acquisition process, to:

- review, comment, and provide advice on key system specific acquisition proposals and documents that can affect the outcome of tests and evaluations for major defense acquisition programs; and
- develop criteria and provide advice on statistical criteria to be used by all DoD components responsible for setting goals and testing systems against such goals in the DoD acquisition process.

It is suggested that such a group be given a permanent charter and that it be responsible primarily to the Undersecretary of Defense for Acquisition and Technology, who can direct the group's support to any part of the acquisition community. Such a group needs to be large enough and stable enough to maintain continuity over time in methods and advice. The specifics of the charter for such a Statistical Analysis Improvement Group should be developed by key personnel in the office of the Secretary of Defense:

Recommendation 10.2: DoD should create a Statistical Analysis Improvement Group in the Department of Defense (using existing person-

[1] Parametric cost estimates are developed by relating the actual historical costs of earlier similar weapon systems to their performance characteristics to make statistical projections of the most likely costs of new weapons. The parametric approach is intended to capture the costs of setbacks and design changes encountered by almost all programs—costs that are not usually anticipated in an industrial engineering or "grass roots" approach (see Margolis, 1975).

nel on a part-time basis) to support the Under Secretary of Defense (Acquisition and Technology) in applying the best statistical principles in the acquisition of defense systems.

A FINAL ISSUE: SPONSORED RESEARCH

There are a number of military organizations with responsibility for sponsoring research on technical areas with potential military applications. Among the most active and visible of these are the Army Research Office (ARO), the Office of Naval Research (ONR), and the Air Force Office of Scientific Research (AFOSR). Over many years, these agencies have played important leadership roles in the initiation of relevant research in the mathematical and physical sciences and in several key areas in the engineering sciences. Of particular interest to the panel are the policies, practices, and priorities of these agencies in sponsoring statistical research.

These agencies have, over the past decade, given little emphasis to the development of appropriate statistical methods relevant to the developmental and operational testing associated with the Department of Defense's acquisitions programs. In fact, there seems to have been a substantial change in their priorities in this regard. The Army Design of Experiments Conference, instigated by Samuel Wilks in the early 1950s and sustained by ARO for some 40 years, has been discontinued. Similarly, AFOSR, which sponsored a number of major research initiatives in the fields of reliability and stochastic processes in the 1970s and 1980s, now makes fairly modest investments in research in these areas. ONR announced in the mid-1980s that it would no longer designate the area of reliability as an area of emphasis; this was soon followed by a sharp reduction in ONR funding for research in reliability. The other agencies have since followed suit. There is a clear need to modernize statistical practices in the OT&E community, and the research that might facilitate the needed advances is today happening by accident rather than by design.

As is evident to readers of this report, there are a number of statistical subfields that are central to modern test and evaluation activity in military acquisitions. They include:

- the design of experiments;
- reliability theory, and more generally, the "quality sciences," including standard suitability issues such as availability and maintainability and areas such as statistical process control;
- sequential analysis;
- software testing;
- validation of modeling and simulation;
- Bayesian methods; and
- data visualization.

The panel has noted that there are deficiencies, some quite serious in magnitude, in current knowledge and practices in these subject areas vis-à-vis military testing and evaluation. These are some areas that could be emphasized by agencies such as the ARO, ONR, and AFOSR.

It is not appropriate to call for specific work in areas like experimental design, reliability, or software testing without having a clear notion about the potential value of different contributions in these areas. An effective program of sponsored research would require prior input from both the potential users of research findings and the research community. It would require specifying the innovations that practitioners need and the theoretical developments required to address the applications of interest.

There is a fundamental need for a collaborative effort between the military operational test and evaluation community and the statistical research community directed at defining new research initiatives. Developing such a research program, including preferred providers in government, industry, and academia would be a good continuing task for the Statistical Analysis Improvement Group that we recommend be established. One of its missions should be to advise the Undersecretary of Defense for Acquisition and Technology on the priorities of research on technical issues raised by the developmental and operational testing of defense systems and the potential sources for such research. At the same time, we suggest that the ARO, ONR, and AFOSR, consider increasing the priority of research on technical issues raised by the developmental and operational testing of defense systems. These applications are extremely important and might greatly benefit by new advances in statistical and related methodology. Sponsored research can be an extremely useful tool, for generating progress on many of the important issues described here, and it has not been used particularly effectively in recent years. Support of both external sponsored research and internal statistical methods research is critical if the services are to embrace improved statistical methods in designing and interpreting test and evaluation activities.

APPENDICES

APPENDIX A

Case Studies and System Descriptions

In this appendix we present several examples, or case studies, that we refer to in the body of the report. We intend this appendix to help orient readers in the statistical community who are unfamiliar with the testing and evaluation of defense systems, as well as to provide additional background and support for recommendations addressed to those in the defense acquisition community.

LONGBOW APACHE HELICOPTER

The Longbow Apache (AH-64D) is a modified version of the Army's existing attack helicopter, the Apache (AH-64A). A key distinguishing feature of the new Longbow helicopter is a fire control radar (FCR) system that allows the helicopter to engage targets with radar-guided Hellfire missiles. Because the missiles can be used without visual or optical acquisition of the target, the FCR system is expected to increase the operational effectiveness of the helicopter in conditions of adverse weather and reduced visibility (e.g., due to smoke or fog) when the performance of laser, optical, and infrared sensors is degraded. In addition, the Longbow AH-64D includes an improved data modem that is intended to enable it to transfer digital target data to other attack helicopters, including the Apache AH-64A, thus enhancing the capabilities of integrated attack helicopter teams.

The Longbow is considered a major acquisition program (ACAT I). In constant fiscal 1994 dollars, the total procurement cost of the system is estimated at $5.3 billion, and the total 20-year life-cycle cost is estimated at $14.3 billion.[1] The total research, test, development, and evaluation costs are expected to be approximately $567 million. Because the Longbow is an ACAT I program, its operational testing and evaluation is subject to oversight by DOT&E, in the Department of Defense.

Operational testing of the Longbow comprised a series of gunnery tests with live ammunition at the Naval Air Warfare Center (in China Lake, California) in January and February 1995, and a series of force-on-force tests with simulated ammunition (at Fort Hunter Liggett, California) in February and March 1995. Separate air transportability and FCR conversion exercises were conducted as part of the force-on-force testing. These operational test events followed a period of developmental testing, training, and force development testing and experimentation. Both the gunnery and force-on-force tests involved a comparison of the performance of the Longbow AH-64D against the existing Apache AH-64A system.

The initial operational test and evaluation plan for the Longbow Apache involved a total of 46 measures of performance (MOPs). Of these, 21 were related to operational effectiveness and 26 were related to operational suitability. Of the 26 suitability MOPs, 11 concerned the Longbow's reliability, availability, and maintainability. Operational requirements in the initial plan were specified for 6 of the 11 measures of reliability, availability, and maintainability; see Table A-1. Because the Longbow AH-64D is a modification of an existing system, many of its RAM requirements were specified on the basis of the comparable Apache AH-64A requirements and observed operating experience. For example, the availability of the Longbow AH-64D, excluding the FCR system, is required to at least match the projected baseline availability value of 0.916, which was computed using Army logistics records for the Apache AH-64A. The availability requirement for the Longbow with the FCR system is 0.824, a value that allows for a 10 percent degradation in availability to maintain the new FCR system.

The primary purpose of the gunnery test was to assess the effectiveness of the Hellfire radar-guided missile when used as part of the Longbow Apache system. The gunnery test collected data on the proportion of targets designated by the helicopter's FCR that were acquired by the missile's radar. The test also provided information about the range in which the missile could effectively engage targets. Such tests involving a small number of live-ammunition firings are needed to establish credible data on system performance. Live ammunition is not used in force-on-force tests for obvious reasons. The gunnery phase consisted of live missile firings from both the Longbow AH-64D and the Apache AH-64A

[1] The source of the estimates is Annex D of the Longbow Apache Test and Evaluation Master Plan (U.S. Department of Defense, 1993), which cites December 1993 estimates from the Longbow Program Office and the President's fiscal 1995 budget as the original sources.

TABLE A-1 Summary of Longbow RAM Requirements and Demonstrated Performance in Initial Operational Testing and Evaluation (in hours)

Measure	Required	Demonstrated	Objective
Reliability			
Mean Time Between Mission Failure	15.3	22.18	21.0
Mean Time Between FCR System Failure	102.0	136.0	102.0
Availability			
With FCR	0.824	0.914	
Without FCR	0.916	0.925	
Maintainability			
FCR Mean Time to Repair	0.5	3.85	
FCR Maintenance Ratio	0.023	0.173	

teams against an array of enemy targets. Firings occurred under three distinct conditions: day mission with smoke obscuration, night mission with clear visibility, and night mission with smoke obscuration. The test and evaluation plan listed the required number of missile shots in each of the three scenarios as 12, 8, and 4, respectively, for the Longbow team and 5, 3, and 3, respectively, for the Apache baseline team. The panel does not know the basis for determining these gunnery test sample sizes.

Sample size for the force-on-force test was determined from considering the criterion that the modernized Apache AH-64D team demonstrate increased targets hit when compared with the AH-64A team. This criterion was interpreted as requiring a sample of sufficient size to detect a (one-sided) difference of 10 percentage points in the probabilities of hits of the Longbow team and the Apache team against the Red forces. Assuming a baseline hit probability of 0.5 yields a conservative sample size estimate that 325 missile shots by both the Longbow AH-64D and the baseline AH-64A are needed to estimate the difference in hit probabilities at 0.10 levels of both consumer's and producer's risk. If a higher risk level of 0.15 is accepted by producer and consumer, then the sample size requirement can be reduced to 212 missile shots. The conclusion in the test and evaluation report is stated as follows (U.S. Department of Defense, 1993):

> A sample size of 212 to 325 missile shots by the Longbow team, as well as by the AH-64A team in each cell of the event matrix, will detect a difference of 10 percent or more and give 85-90 percent confidence in the criteria conclusions. Based on the expected number of shots per mission, the number of missions [sic] under each condition can be estimated.

The force-on-force test was conducted in three major scenarios: close battle at day, close battle at night, and deep battle at night. (Close battles occur at the military front; deep battles involve strikes inside enemy lines.) A total of 320 missile shots were required in both night battle scenarios, and 256 missile shots were required in the day battle scenario.

The Longbow Apache test appears to be typical in that the size of the force-on-force test was primarily determined by considerations of effectiveness rather than suitability. A common approach in operational testing seems to be: Do the testing necessary to assess effectiveness, and accumulate as much of the requisite RAM data as feasible.

The primary data sources for measuring the reliability, availability, and maintainability performance of the Longbow system were the gunnery and force-on-force tests conducted as part of the initial operational test and evaluation plan. Secondary data sources included development testing, logistical demonstrations, and force development test and experimentation. According to the test and evaluation plan (U.S. Department of Defense, 1993:2-92): "Secondary data will be used to supplement [operational test] results by demonstrating historical performance, helping identify trends in anomalies found in the primary MOPs, and characterizing performance in conditions not encountered in the IOT." The test and evaluation plan also stated that, because of the different conditions obtained under gunnery and force-on-force testing, results from the two phases would not be merged; instead, they would be presented separately in the evaluation report. For the RAM-related MOPs, however, the operating hours appear to reflect the accumulated hours under both the gunnery and force-on-force phases.

The demonstrated RAM-related performance of the Longbow Apache AH-64D helicopter in comparison with its operational requirements is summarized in Table B-1. The mean time between mission failure (MTBF[M]) value of 22.18 hours is based on 11 observed mission failures of the Longbow AH-64D helicopter in a total of 244 operating hours. The mean time between FCR system failure (MTBF[S]) value of 136 hours is based on 2 failures in 272 system operating hours. The maintainability of the FCR system is summarized by two measures: (1) its mean time to repair of 3.85 hours, comprising six corrective maintenance actions and a total of 23.07 hours in repair time; (2) its maintenance ratio of 0.173, reflecting a total of 47.05 person-hours spent on system maintenance divided by the total of 272 system operating hours.

Computation of the availability measure is a bit more involved. The observed availability values (A_0) of 0.914 and 0.925 for the Longbow AH-64D—including and excluding the FCR system, respectively—are computed according to the formula:

$$\begin{aligned} A_o &= 1 - \left(\frac{\text{number of EMAs} \times \text{mean downtime}}{\text{total time}} \right) \\ &= 1 - \left[\left(\frac{OT}{MTBEMA} \right) \times (MTTRe + ALDAT) \times \left(\frac{1}{TT} \right) \right] \\ &= 1 - \left[\left(\frac{OT}{TT} \right) \times \left(\frac{MTTRe + ALDT}{MTBEMA} \right) \right], \end{aligned}$$

where *OT* denotes the total operating (flight) time of the Longbow system, *TT* denotes the total calendar time during the test period, *MTTRe* denotes the mean time expended on essential maintenance actions, *ALDT* denotes the total administrative and logistics downtime (e.g., time spent waiting for parts, maintenance personnel, or transportation), and *MTBEMA* denotes the mean time between essential maintenance actions. The sample values of these constituent parts observed during the initial operational testing and evaluation are described in the Longbow Apache operational evaluation report.

On the basis of the results in Table B-1, the Longbow Apache was judged to have met its reliability and availability requirements, but the FCR system did not meet its maintainability requirements. However, because the evaluator judged the FCR maintainability measures of secondary importance and because the realism of the 0.5-hour mean repair time requirement for the FCR was questioned, the Longbow system was determined to have demonstrated adequate RAM performance in its initial operational test.

C-130H HERCULES AIRCRAFT

The primary mission of the C-130H Hercules aircraft is to provide intratheater airlift—particularly, the tasks of normal airland, tactical airland, and tactical airdrop—in worldwide environments. The C-130H is an upgraded version of the older C-130, with five previously untested modifications: a low-power color radar; rearranged cockpit instruments; an electrical system upgrade; new, flat-screen electronic flight instruments; and a mode advisory caution and warning system.

The purpose of the qualification operational test and evaluation was to evaluate the effectiveness and suitability of the modified C-130H. Instead of focusing on such parameters as range and payload (that are typically of primary interest for cargo aircraft in initial operational testing and evaluation), the qualifying testing and evaluation focused on the operation of the new modifications. The effectiveness evaluations essentially involved comparisons of the new modifications with the old subsystems to be replaced. Effectiveness data were gathered during two weeks of flight testing (approximately 60 flying hours) in March 1994 (at Little Rock Air Force Base).

Minimum reliability, availability, and maintainability test times were statistically determined according to the standard formula assuming exponentiality (see Milstar example, below). Required test times for the five new subsystems ranged from approximately 200 hours to 17,000 hours. A compromise test period of 5 months was selected with two justifications: the schedule would permit the issuance of a final report in time to influence fiscal year production decisions; and approximately 1,000 flying hours were expected to be logged during this 5-month period, which would meet the statistical test requirements for three of the five new subsystems.

To evaluate the availability and reliability of the C-130H, maintenance data from February to June 1994 were entered into the Core Automated Maintenance System at the Little Rock Air Force Base and at the Wyoming Air National Guard base. These data were subsequently downloaded into the Micro-Omnivore database of the Air Force Operational Test and Evaluation Center (AFOTEC). A test data scoring board maintained quality control of the maintenance events and checked the data for completeness and accuracy. The combined data represented approximately 2,200 flying hours.

The C-130H was found to be effective but not suitable. In some cases, observed values of the reliability, availability, and maintainability measures differed from operational requirements by as much as two orders of magnitude. AFOTEC's recommendation was to delay deployment of the C-130H until adequate technical orders, support equipment, and spares were available and training was complete. AFOTEC also recommended testing of the C-130H in extreme cold, desert, and tropical climates to stress the aircraft in these conditions.

B-1B BOMBER

The B-1B aircraft is a long-range, supersonic bomber with after-burning engines, capable of conducting high-speed, low-level flight profiles. An operational readiness assessment was mandated by Congress to determine whether the B-1B can achieve and sustain a desired readiness rate of 75 percent over a 6-month period. In the 2 years preceding the assessment, mission capable rates in the B-1 fleet averaged approximately 57 percent, primarily because of inadequate funding for repair of parts through interim contractor support and for stocking of repairable and new spare parts. In effect, the assessment was a test of the adequacy of the Air Force's planned upgrades in the levels of B-1 spare parts, logistics support equipment, and maintenance personnel.

The test was conducted at Ellsworth Air Force Base (South Dakota) from June 1 to November 30, 1994. The 28th Bomb Wing was selected as the military test unit on the basis of several considerations, including stability of force structure during the assessment period, impact of testing on field training operations, programmed aircraft maintenance and modification schedules, and overall unit experience in conventional operations. Aircraft operations during the 6-month evaluation period included planned peacetime training and a 2-week deployment to a remote location for exercises consistent with the B-1B's expected use in a conventional conflict.

The term "75 percent operational readiness rate" was defined as an average mission capable rate of 75 percent of possessed aircraft hours. An aircraft is considered mission capable if it is capable of performing at least one of its designated missions. Note that the mission capable rate is a measure of the performance of the entire B-1B support structure, encompassing more than the aircraft itself. In particular, the mission capable rate is not a direct measure of the

B-1B's inherent reliability or its ability to successfully complete assigned missions.

Data were collected for a total of 12 reliability, availability, and maintainability measures. The three reliability measures were break rate of system failures during assigned missions; mean time between corrective maintenance events; and rate of aircraft-to-aircraft or engine-to-aircraft cannibalization actions per 100 sorties flown. The four maintainability measures were the 12-hour fix rate, expressed as the percentage of inoperable aircraft that are returned to flyable status within 12 hours; maintenance person-hours per flying hour; mean man-hours to repair; and mean repair time, not including maintenance or supply delays. The five availability measures were the mission capable rate; the total not mission capable rate; two submeasures intended to isolate maintenance and supply problems; and the utilization rate, expressed as the average number of hours flown per aircraft per month.

With improved reliability, availability, and maintainability support, the 28th Bomb Wing achieved a readiness rate of 75 percent by the start of the operational readiness assessment and demonstrated a cumulative readiness rate of 84.3 percent during the 6-month evaluation period. Test results were regarded as evidence that, with 100 percent staffing, sufficient spare parts, and a robust base repair capability, a B-1B wing can achieve and sustain the required 75 percent readiness rate.

(According to AFOTEC presentation materials, data analyses for the 12 reliability, availability, and maintainability parameters involved trend analysis and analysis of variance—in addition to reporting mean values—but the final assessment report does not mention them.)

MILSTAR SATELLITE COMMUNICATION SYSTEM

Milstar is a constellation of four satellites and associated ground support, designed to provide secure, high-priority, strategic communication. The initial operational test and evaluation was an 18-month test, from February 1995 to July 1996, with oversight from DOT&E. This was a joint (or multiservice) test and evaluation conducted by the Air Force with involvement of other military services. Its purpose was to evaluate the operational effectiveness and suitability of the Milstar Low Data Rate System.

The testing and evaluation consisted of limited field exercises to test the endurance of the mobile constellation control station—defined as the Milstar terminal and satellite control station aligned in series configuration. The user requirement is that the mobile station system "endure" for X days (a classified value) at a deployed location without outside support. The required number of test hours per platform (T) was determined statistically—assuming exponentially distributed failure times. Statistical means were to be used in reporting test data.

Field testing resources alone were judged insufficient to support evaluation, and the use of a complementary model was proposed. For modeling purposes, the system was defined as two mobile constellations and four satellites. Modeling efforts were to be concentrated on the effect of reliability and sparing levels on the operation of the stations and satellites. Lack of a rigorous model validation, however, precluded the formal incorporation of modeling results into the system evaluation.

THE ATACMS/BAT SYSTEM

In the Army Tactical Missile System/Brilliant Anti-Tank (ATACMS/BAT) System, the Army Operational Test and Evaluation Command proposes to use a relatively novel test design in which a simulation model, when calibrated by a small number of operational field tests, will provide an overall assessment of the effectiveness of the system under test.

BAT submunitions use acoustic sensors to guide themselves toward moving vehicles and an infrared seeker to home terminally on targets. The submunitions are delivered to the approximate target area by the ATACMS, which releases the submunitions from its main missile. The ATACMS/BAT system is very expensive, costing several million dollars per missile.

The experimental design issue is how to choose the small number of operational field tests such that the calibrated simulation will be as informative as possible. Here we provide a description of the ATACMS/BAT operational testing program to illustrate one context in which to think about alternative approaches to operational test design.

Plans for Operational Testing

As noted above, the ATACMS/BAT operational test will not be a traditional Army operational test involving force-on-force trials, but will be similar to a demonstration test, using a model to simulate what might happen in a real engagement. To calibrate the simulation, various kinds of data will be collected, for example, from individual submunition flights and other types of trials, including operational test trials. This relatively novel approach has been taken because budgetary limitations on the sample size and the limited availability of equipment such as radio-controlled tanks for testing make it infeasible to develop a program of field testing that could answer the key questions about the performance of this system in a real operational testing environment.

Operational testing of the ATACMS/BAT system is scheduled to take place in 2000. According to the last approved Test and Evaluation Master Plan (U.S. Department of Defense, 1995b):

> This portion of the system evaluation includes the Army ATACMS/BAT launch, missile flight, dispense of the BAT submunitions, the transition to independent flight, acoustic and infrared homing, and final impact on targets. Evaluation of

> this discrete event also includes assessment of support system/subsystem RAM requirements, software, terminal accuracy, dispense effectiveness, kills per launcher load, and BAT effectiveness in the presence of countermeasures.

Initial operational test and evaluation is the primary source of data for assessing these system capabilities. There is no baseline system for comparison.

The number of armored vehicle kills (against a battalion of tanks) is the bottom-line measure of the system's success. Tank battalions vary in size, but typically involve about 150 vehicles moving in formation. (Unfortunately, every country moves its tanks somewhat differently.) Under the test scoring rules, no credit is given if the submunition hits the tank treads or a truck or if two submunitions hit the same tank.

There is one operational test site, and the Army has spent several million dollars developing it. There will be materiel constraints on the operational test. Only seven test events, with eight missiles, each of which has a full set of 13 BAT submunitions, are available for testing. Also, the test battalion will involve only 21 remotely controlled vehicles. Thus, the Army plans to use simulation as an extrapolation device, particularly in generalizing from 21 tanks to a full-size battalion.

Important Test Factors and Conditions

All stages of missile operation must be considered in an operational test, particularly acoustic detection, infrared detection, and target impact. Factors that may affect acoustic detection of vehicles include distance from target (location, delivery error), weather (wind, air, density, rain), vehicle signature (type, speed, formation), and terrain. For example, the submunitions are not independently targeted; they are programmed with logic to go to different targets. Their success at picking different targets can be affected by such factors as wind, rain, temperature, and cloud layers. Obviously, one cannot learn about system performance during bad weather if testing is conducted only on dry days. However, it is difficult to conduct operational tests in rain because the test instrumentation does not function well, and much data can be lost. Such factors as weather (rain, snow) and environment (dust, smoke) can also affect infrared detection of vehicles. Factors that affect the conditional probability of a vehicle kill, given a hit, include the hardness of the vehicle and the location of the hit. Possible countermeasures must also be considered: for example, the tanks may disperse at some point, instead of advancing in a straight-line formation or may try to use decoys or smoke obscuration.

The operational test design, or shot matrix, in the Test and Evaluation Master Plan lists eight test events that vary according to such factors as range of engagement, target location error, logic of targeting software, type of tank formation, aimpoint, time of day, tank speed and spacing, and threat environment; see Table

A-2. Three levels are specified for the range of engagement: near minimum, near maximum, and a medium range specified as either "2/3s" or "ROC," the required operational capability. Target location errors vary with the method of control (either centralized or decentralized) and are quantified in terms of either a median level or one standard deviation (sigma) above the median. The logic of targeting software (primary or alternate) and type of tank formation (linear or dispersed) are combined into a single, two-level factor. Aimpoint distance is either short or long, and aimpoint direction is either left or right. The aimpoint factors are not expected to be important in affecting system performance. (The payload inventory is also unimportant in this context.) Tanks are expected to travel at lower speeds and in denser formations during night operations; therefore, tank speed and spacing are combined with time of day into a single two-level factor (day or night). Three different threat environments are possible: benign, Level 1, and Level 2 (most severe). Clearly, in view of the limited sample size, many potentially influential factors are not represented in the shot matrix.

Statistical Methods for Further Consideration

One approach to operational testing for the ATACMS/BAT would be to design a large fractional factorial experiment for those factors thought to have the greatest influence on system performance. The number of effective replications can be increased if the assumption that all of the included design factors are influential turns out to be incorrect. Assuming that the aimpoint factors are inactive, a *complete factorial* experiment for the ATACMS/BAT system would require $2^3 \times 3^2 = 72$ design points. However, *fractional factorial* designs with two- and three-level factors could provide much information while using substantially fewer replications than a complete factorial design. Of course, these designs are less useful when higher-order interactions among factors are significant. (For a further discussion of factorial designs, see National Research Council, 1995:Appendix B; Box and Hunter, 1961.)

Another complication is that environment (or scenario) is a factor with more than two settings (levels). In the extreme, the ATACMS/BAT operational test results might be regarded as samples from several different populations representing test results from each environment. Since it will not be possible to evaluate the test in several unrelated settings, some consolidation of scenarios is needed. It is necessary to understand how to consolidate scenarios by identifying the underlying physical characteristics that have an impact on the performance measures and to relate the performance of the system, possibly through use of a parametric model, to the underlying characteristics of those environments; see discussion in Appendix C of the panel's interim report (National Research Council, 1995).

Although the above fractional factorial approach has advantages with respect

TABLE A-2 Army Tactical Missile System Block II/Brilliant Anti-Tank Operational Test Shot Matrix

Test[a]	Range	Target Location Error (Method of Control)	Target/Logic	Aimpoint	Payload	Environment[b]
DT/OT 1	ROC	ROC	Primary/linear	Long/right	9 Tt, 4 Tsw	Benign ROC
DT/OT 2	2/3s	Median centralized	Primary/linear	Short/right	13 Tt	Level 1 night
OT 1	Near max	1 sigma centralized	Primary/linear	Short/left	13 Tt	Level 1 and 2 day
OT 2	Near min	Median decentralized	Alternate/dispersed	Long/left	7 Tt, 6 Tsw	Level 1 day
OT 3[c]	2/3s	Median dentralized AMC	Primary/linear	C3 system determined	10 Tt, 3 Tsw	Level 1 night
OT 4[c]	2/3s	Median centralized AMC	Primary/linear	C3 system determined	10 Tt, 3 Tsw	Level 1 night
OT 5[d]	2/3s	1 sigma decentralized	Primary/linear	Long/on line	13 Tt	Level 1 and 2 night
OT 6[d]	2/3s	Median centralized	To be determined	To be determined	13 Tt	Level 1 day

NOTES:
AMC, Army Materiel Command
C3, command, control and communications
DT, developmental testing
OT, operational testing
ROC, required operational capability
See text for discussion.

[a] This shot matrix reflects a two-flight developmental testing/operational testing program of engineering and manufacturing development assets conducted by the Test and Evaluation Command, and a six-flight operational testing program of low-rate initial production Brilliant anti-tank assets conducted by the Test and Experimentation Command. This shot matrix is subject to threat accreditation.

[b] Flights will be conducted during daylight hours for safety and data collection. Night/day environments pertain to vehicle speed and spacing.

[c] OT 3 and OT 4 are ripple firings.

[d] Flight objectives may be revised based on unforeseen data shortfalls.

to understanding system performance equally in each scenario, there are some benefits of the current Army approach if one assumes that the majority of interest is focused on the "central" scenario, or the scenario of most interest. In the current approach, the largest number of test units are allocated to this scenario, while the others are used to study one-factor-at-a-time perturbations around this scenario, such as going from day to night or from linear to dispersed formation. This approach could be well suited to gathering information on such issues while not losing too much efficiency at the scenario of most interest. If it turns out that changing one or more factors has no effect, the information from these settings can be pooled to gain further efficiency at the scenario of greatest interest.

THE COMMON MISSILE WARNING SYSTEM

The common missile warning system (CMWS)[2] is a warning system that alerts aircraft crew that missiles have been launched at the aircraft. The CMWS is intended to be effective against short-range infrared surface-to-air missile (SAM) threats. The system is composed of multiple sensors, coupled with a processing and control system, which examines the threat missile "plume" to determine whether the missile is a threat. If so, the CMWS triggers countermeasures and warns the aircrew. A key factor for the success of this system is the time between the reception of the warning and the missile strike. The basic requirement for the operational test of the system is the demonstration, at typical levels of statistical confidence,[3] that the CMWS provides adequate warning *across all threats and scenarios*. However, there is no requirement that the test provide statistical confidence for assertions about the performance of the system at the threat- or scenario-level.

During test planning of the CMWS, it was initially assumed that the key factors related to warning time were altitude of the aircraft, horizontal distance (range) from the ground-based missile threat, aspect angle (the profile of the plane in the direction of the flight path of the missile), and the particular threat. (This assumption is currently being reviewed and is probably going to be revised once detailed simulation models of the relevant physical processes are developed.) Furthermore, using engineering and threat information, AFOTEC reduced the number of essential threats to four given the similar plumes and range and altitude of targets (threat envelopes) against which various threats are effective. Further, it is assumed that each of these four threats has equal priority, i.e., that their potential for use by various opponents is similar.

While the overall goal was the estimation of average warning time across threats and scenarios, it was also desired to estimate the average warning time for

[2] Much of the following is taken from Sheedy and Stolle (1996). In addition, a very useful informal summary was provided by Steve Ordonia.

[3] Unspecified in the documentation but ranging from 99% to 80%.

each of these individual threats across scenarios by producing a mathematical model that would predict warning time as a function of the above three factors. The hope was to support development of both an estimate of overall performance and these mathematical models for each threat for the least cost, i.e., for the fewest missile firings.

Additional information indicated that missile warning time was likely to be linearly related to altitude, and at most quadratically related to range and aspect angle. Therefore, missile warning time was measured at two different altitudes (low and high), three different ranges (short, medium, and long) and three different levels of aspect angle (rear, beam, and front), which would permit estimating the assumed relationship between these factors and warning time. The result was a total of $2 \times 3 \times 3 = 18$ test scenarios of interest for each of the four threats, representing 72 threat-scenario combinations.

Fixed Grid Approach

The first approach to this problem defined the three test factors in terms of a "fixed" grid, i.e., identical across all threats. So, e.g., the two altitudes, low and high, were defined across threats as:

Low:	$\text{Alt} \le 5{,}000$ ft
High:	$5{,}000 \text{ ft} < \text{Alt} \le 10{,}000$ ft.

The three ranges, in nautical miles (NM) were defined, again across threats, as:

Short:	$\text{Range} \le 1$ NM
Medium:	$1 \text{ NM} < \text{Range} \le 3$ NM
Long:	$3 \text{ NM} < \text{Range} \le 6$ NM.

Finally, the three aspect angles in degrees were defined, across threats, as:

Front:	$315^\circ \le \theta < 45^\circ$
Beam:	$45^\circ \le \theta < 135^\circ$; $225^\circ \le \theta < 315^\circ$
Rear:	$135^\circ \le \theta < 225^\circ$.

This fixed grid approach to the definition of test scenarios caused a problem, since some of the test scenarios in the resulting factorial design were not testable. For example, threat 2 could not be tested at long range. In addition, the low altitude, short range, beam scenario was testable for threat 1 but not testable for threat 2. The result was that only 32 of the 72 scenario/threat combinations were testable. (Note that other complications have been omitted in this description). The hope of an early design using this fixed grid was also to have estimates of performance in each scenario for each threat, which resulted in a suggested

LOW ALTITUDE (<5 K)								
SHORT RANGE (<1 NM)			MEDIUM RANGE (1-3 NM)			LONG RANGE (<3 NM)		
FRONT	BEAM	REAR	FRONT	BEAM	REAR	FRONT	BEAM	REAR
■ 5	■ 5	▒ 0	■ 15	■ 5	■ 5	▒ 0	▒ 0	▒ 0
▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	■ 15	▒ 0	▒ 0	░ 0
░ 0	░ 0	▒ 0	■ 5	■ 15	▒ 0	■ 5	■ 5	■ 5
■ 5	■ 5	▒ 0	■ 15	■ 5	■ 5	■ 5	■ 5	░ 0
■ 5	■ 5	▒ 0	■ 15	■ 5	■ 5	■ 5	■ 5	■ 5
MEDIUM ALTITUDE (5K - 15K)								
SHORT RANGE (<1 NM)			MEDIUM RANGE (1-3 NM)			LONG RANGE (<3 NM)		
FRONT	BEAM	REAR	FRONT	BEAM	REAR	FRONT	BEAM	REAR
■ 0	■ 0	▒ 0	■ 5	■ 0	■ 0	▒ 0	▒ 0	▒ 0
▒ 0	▒ 0	░ 0	▒ 0	▒ 0	■ 0	▒ 0	▒ 0	■ 5
░ 0	░ 0	▒ 0	■ 5	■ 0	▒ 0	░ 0	■ 0	▒ 0
■ 0	■ 0	■ 0	■ 0	■ 0	■ 0	░ 0	░ 0	░ 0
■ 0	■ 0	▒ 0	■ 0	■ 0	░ 0	■ 0	■ 0	■ 0
HIGH ALTITUDE (> 15K)								
SHORT RANGE (<1 NM)			MEDIUM RANGE (1-3 NM)			LONG RANGE (<3 NM)		
FRONT	BEAM	REAR	FRONT	BEAM	REAR	FRONT	BEAM	REAR
▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0
▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0
▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0	▒ 0
■ 0	■ 0	▒ 0	■ 5	■ 0	▒ 0	▒ 0	▒ 0	▒ 0
░ 0	░ 0	▒ 0	■ 5	■ 0	▒ 0	▒ 0	▒ 0	▒ 0

Legend:
■ Indicates threat engagement capability
░ Indicates partial threat engagement capability
▒ Indicates no threat engagement capability

FIGURE A-1 Threat Categorization: experimental design for fixed grid approach. NOTE: In the partial threat and no threat cells, no shots could be taken, as indicated, justifying the switch from a fixed grid approach to the scalable grid approach; see text for discussion. NM: Nautical miles. SOURCE: "Common Missile Warning System Sample Size Strategy," 2-11-96 presentation, Air Force Operational Test and Evaluation Center

design of 5 shots per scenario and threat combination. (Assuming an extremely successful performance by the CMWS, this design would have had marginally acceptable power with respect to the assumption that the CMWS did not meet its required level of performance.) Given that some cells were designed to have an additional 10 shots, this design required 205 missile shots (see Figure B-1). It was decided that this was more shots than could be afforded. (Figure B-1 displays three altitutdes, which were later collapsed to two.)

The objectives of the test design were then revisited, and it was understood that no scenario- or threat- level analysis was requested, besides the development of mathematical models at the threat level. Also, as was recognized by AFOTEC, there were real advantages for the test design to be a full factorial experiment (possibly with replications), i.e., to have each combination of levels of each test factor appear the same number of times in the test. This would facilitate development of the threat-level models and test evaluation.

Scaleable Grid Approach

Therefore, AFOTEC replaced the fixed grid approach with a scaleable grid for each threat. For example, the low and medium levels of altitude were shifted from 0–5,000 feet (low) and 5,000–10,000 feet (medium) to instead fluctuate based on the effective target altitude for the threat. (Details on the precise definition of the scaleable grid are not provided here.) This was also done for range and aspect angle. This approach is quite sensible in that it provides test scenarios that are well-distributed inside the feasible test region for each threat, which would not have been as true using the fixed grid approach.

As a result of using the scaleable grid, each of the 18 scenarios for each threat (which were defined differently for each threat) became testable, and therefore average missile warning time across scenarios and threats could be estimated for the CMWS (which now had a different interpretation than in the fixed grid situation), and a mathematical model of average missile warning time for each threat as a function of altitude, range, and aspect angle, and some two and three-level interactions of these variables, could be easily developed. Finally, these models could be used to provide predictions and associated confidence intervals for additional scenarios that were within the test domain for each threat.

Test Sample Size Argument

The remaining question was whether one missile shot per scenario - threat combination would be sufficient to produce an estimate of average missile warning time across scenarios and threats that would pass a significance test at typical levels with high probability, thereby supporting that the CMWS satisfied its required level of performance. The argument used to justify test size was as follows. Assume that one missile shot is taken in each of the 18 scenarios for each of the four threats as a full factorial experiment over the three test factors. That would result in 72 missile shots. In addition, given that there was a requirement for investigation of the performance of the CMWS against multiple-threats in a given quadrant, the number of missile shots was increased by two for each threat. The resulting test would therefore have 20 shots per threat, or 80 overall shots.[4] Characterizing a single missile test shot as being successful for the CMWS if the

[4] There were actually 96 test missile allocated, a 20 percent margin in case there were problems with the missiles.

warning provided was less than or equal to x seconds, where x was the required average warning time,[5] the argument for the number of missile shots was that the test needed to demonstrate that the success rate was significantly greater than .80 across scenarios and threats. Using 80 total missile shots, the displayed number of failures result in the associated statistical confidence that the real success rate was greater than .8:

Total Missiles: 80

Number of Failures	Confidence
8	99%
9	97%
10	94%
11	90%
12	84%

(Note that this argument assumes a constant success rate across threats and scenarios, which is unlikely to obtain.)

Of course, this does not directly answer the question of how reliable estimates of the average warning time across scenarios and threats will be or address power considerations. However, the probability of a warning time less than x is certainly of interest, and there may have been little a priori information on what standard deviation might have been expected for the CMWS to permit inference about the ability to estimate average warning time. Therefore, this approach has the advantage of being a quantifiable answer to an important and related question.

Note that although there is no user requirement for estimates of average warning time at the threat level at typical levels of statistical confidence, this test design can produce estimates with reasonable levels of statistical confidence at that level. The analogous table for 20 missile shots is:

By Each Threat: Total Missiles 20

Number of Failures	Confidence
0	99%
1	93%
2	79%

[5] The value for x was not provided to enable the above documents to be unclassified.

Once the experiment is carried out, regression models will be fit for three purposes (note that it is unlikely, given the scaleable grid, that one could develop a single mathematical model that would be useful across threats):

1. to compute confidence bounds on the average missile warning time for a particular threat,
2. to investigate performance of the CMWS in alternative scenarios for a given threat within the test domain for that threat, and
3. as input to a digital model to examine the impact of missile warning time on the success of the total mission.

An open question is whether the test size could be reduced by making use of preliminary versions of the regression models that will be developed, since these models will produce estimated average warning times that, with some reasonable assumptions, will have lower variances than one could get through direct measurement for an individual threat. Or, one could instead maintain the variance of the estimated average warning time while reducing the number of required test shots, through use of a design that would have permitted estimation of the four regression models.

SELECTED SYSTEM DESCRIPTIONS

This section contains brief descriptions of several of the remaining systems mentioned in the report. The panel's objective is to provide readers not actively involved in the defense community a basic understanding of these systems that can be used when interpreting the discussion and related examples in the report.

Air Defense Antitank System
Don Richardson, Institute for Defense Analyses

The Air Defense Antitank System (ADATS) is a short-range air defense weapon that was part of the Army's follow-on program to the Sgt. York division air defense gun. ADATS was intended to provide air defense protection to the Army's forward maneuver forces of tanks and armored vehicles. The ADATS system consisted of eight laser-guided missiles, a search and tracking radar, and infrared and optical sensors.

Advanced Amphibious Assault Vehicle
U.S. Department of Defense, 1998b

The Advanced Amphibious Assault Vehicle (AAAV) [a USMC ACAT ID program] is a high water-speed amphibious armored personnel carrier to replace the current family of Marine Corps assault amphibians, the AAV7A1 series. . . . Armed with a medium-caliber machine gun and a cannon of 25-35 mm, the

AAAV will use GPS, a forward-looking infrared radar, and a night vision system for navigation, targeting, and intelligence gathering.

The AAAV must operate in all climates, over all terrain, and in all weather and lighting conditions. At sea, it must achieve a water speed of 20 knots in 3-ft significant wave height and cross a surf zone characterized by up to 8-ft plunging surf. The 20-knot water speed is significantly greater than the 7-knot speed provided by the current AAV7A1 amphibious tractor. . . .

The AAAV's land mobility characteristics must be comparable to the Marine Corps' M1A1 Abrams main battle tank. This requires a top speed of approximately 45 miles per hour, the capability to traverse the same terrain at the same speed as the tank during cross-country operations, and the capability to cross the same obstacles and terrain features (for example, trenches, hills, walls, and soft soils) as the tank.

AH-1 Cobra

Arthur Fries, Institute for Defense Analyses

The Bell AH-1 Cobra attack helicopter was first delivered to the Army in the summer of 1967. This single engine, twin-bladed aircraft was a modification of the UH-1 Iroquois. To reduce its susceptibility to detection and engagement by enemy forces, the width of the airframe was reduced. Hence the two man crew occupied tandem seats, vice the side-by-side seating arrangement in the UH-1. Two stub-wing pods were capable of carrying combinations of 2.75 inch rockets, mini-guns, and TOW anti-tank missiles. It was the Army's primary attack helicopter during the Vietnam conflict.

AQUILA

Arthur Fries, Institute for Defense Analyses

The AQUILA was a remotely piloted air vehicle (RPV) system developed by the Army in the 1980s and terminated late in that decade. It was designed to perform reconnaissance, target acquisition, artillery fire adjustment, and target designation for laser-guided munitions such as Copperhead artillery rounds and HELLFIRE missiles. The concept of operations was for the RPV to penetrate enemy territory 20 to 30 kilometers, where it might be acquired and engaged, or countered, by enemy systems such as air defense units or enemy radio-frequency jammers.

AN/ALQ-165 Airborne Self Protection Jammer

U.S. Department of Defense, 1998b

The AN/ALQ-165 Airborne Self Protection Jammer (ASPJ) [a canceled Navy ACAT I program] is an automated modular reprogrammable active radar frequency (RF) deception jammer designed to contribute to the electronic self protection of the host tactical aircraft from a variety of air-to-air and surface-to-air RF threats. The ASPJ was designed to accomplish threat sorting, threat identification, and jamming management in a dense signal environment to counter multiple threats. The modular architecture supports internal integration with other avionics/weapons systems in a variety of aircraft. The basic system consists of five weapons replaceable assemblies/line replaceable units (WRAs/LRUs) which include two receivers, two transmitters, and one processor. Each WRA is interchangeable among different tactical aircraft. Additional transmitters can be installed on aircraft with larger radar cross sections to increase the effective radiated power.

C-17 Airlift Aircraft

U.S. Department of Defense, 1998b

The C-17 [an Air Force ACAT ID program] is a four engine turbofan aircraft capable of airlifting large payloads over intercontinental ranges without refueling. Its design is intended to allow delivery of outsize combat cargo and equipment directly into austere airfields. The C-17 will deliver passengers and cargo over intercontinental distances, provide theater and strategic airlift in both airland and airdrop modes, and augment aeromedical evacuation and special operations missions.

CH-46

Navy Fact File

http://www.chinfo.navy.mil/navpalib/factfile/aircraft/air-ch46.html

The CH-46 *Sea Knight* was first procured in 1964 to meet the medium-lift requirements of the Marine Corps in all combat and peacetime environments since that time.

[It is a] medium lift assault helicopter, primarily used to move cargo and troops. The CH-46D *Sea Knight* helicopter is used by the Navy for shipboard delivery of cargo and personnel. The CH-46E is used by the Marine Corps to provide all-weather, day-or-night assault transport of combat troops, supplies, and equipment. Troop assault is the primary function and the movement of supplies and equipment is secondary.

CH-47 Chinook

Arthur Fries, Institute for Defense Analyses

The Boeing CH-47 is a twin engine, medium lift helicopter. Each engine drives a three-bladed rotor. The CH-47A was first introduced into the Army in late 1962. Subsequently there have been three major modifications resulting in the CH-47B, C and D. It is used to lift both cargo and personnel. Cargo loads of

8 to 10,000 pounds can be carried internally or externally. CH-47s are normally found in corps level aviation support units, and provide most of the tactical cargo airlift within the corps and division areas.

CH-54 Skycrane

Arthur Fries, Institute for Defense Analyses

The Sikorsky CH-54 is the Army's heavy lift helicopter. Initial deliveries of this aircraft were made in late 1964. It is capable of lifting loads of 15 to 20,000 pounds. In Vietnam it was used to lift heavy equipment, such as dozers and lightly armored vehicles. Very few CH-54s were fielded, and none currently reside in the active forces.

Dragon

Rosser Bobbitt, Institute for Defense Analyses

The Dragon is a man-portable medium anti-tank missile system with command to line-of-sight guidance (gunner holds cross-hairs on target). Its maximum range is about 1,000 meters against the *last*, but not *current*, generation of threat main battle tanks. Recent modifications have extended its range to 1,500 meters and increased its penetration of the current generation of main battle tanks. It is being replaced in the active component of the U.S. Army by the Javelin anti-tank missile system with longer range and a fire-and-forget capability.

Family of Medium Tactical Vehicles

U.S. Department of Defense, 1998b

The Family of Medium Tactical Vehicles (FMTV) consists of fourteen wheeled tactical vehicles based on a common truck cab, chassis, and internal components and two tactical trailers. The components are primarily non-developmental items integrated in rugged tactical configurations. The light-medium tactical vehicles (LMTV) are 2.5-ton payload capacity models consisting of cargo, air drop cargo, and van variants. The medium tactical vehicles (MTV) are 5-ton payload capacity models consisting of cargo (with and without material handling crane), air drop cargo, tractor, wrecker, dump, air drop dump, fuel tanker, and expansible van variants.

FMTV supports Joint Vision objectives: focuses logistics through the transport of troops, water and ammunition distribution, and general cargo transport; information superiority through the provision of mobility to the new generation of automated systems, sophisticated management information systems, and communications links; and precision engagement as the prime mover for towed artillery and Patriot and as the chassis for the High Mobility Artillery Rocket System (HIMARS).

Javelin Antitank Missile
U.S. Department of Defense, 1998b

The Javelin [an Army ACAT ID program] is a manportable, fire-and-forget, antitank missile employed by dismounted infantry to defeat current and future threat armored combat vehicles. Javelin is intended to replace the Dragon system in the Army and the Marine Corps.

The Javelin consists of a missile in a disposable launch tube and a reusable Command Launch Unit with a trigger mechanism and day/night sighting device for surveillance, and target acquisition and built-in test capabilities. The missile locks on to the target before launch using an infrared focal plane array and on-board processing, which also maintains target track and guides the missile to the target after launch.

Joint Surveillance Target Attack Radar System
U.S. Department of Defense, 1998b

The Joint Surveillance Target Attack Radar System (JSTARS) [an Air Force ACAT ID program, contributes] a synoptic battlefield view to operational maneuver commanders. The system [is required] to perform battlefield surveillance, battle management for both air and land component forces, and indications and warning functions. . . . The JSTARS system is intended to meet the operational need for locating, classifying, and supporting precision engagement of time-sensitive moving and stationary targets.

The JSTARS system consists of the Air Force E-8C aircraft, an Army ground station, and the data link that connects the two elements. . . . The JSTARS system brings to the battlefield the technical capability to perform surveillance through interleaved high resolution synthetic aperture radar (SAR), moving target indicator (MTI), and the computer capability to integrate battlefield and geographic information into a near real-time picture of the ground battle.

M1 Abrams Tank
Bernard Kempinski, Institute for Defense Analyses

M1, M1A1 and M1A2 are versions of the Army's Abrams main battle tank. The Abrams is [a] tracked armored vehicle mounting either a 105mm or 120mm cannon, three machine guns and four crew men. Its mission is to close with and destroy enemy armor and troops. It is heavily protected and designed to absorb a high level of damage.

OH-58D Kiowa Warrior

U.S. Department of Defense, 1997

The OH-58D Kiowa Warrior [an Army ACAT IC program] is a two-[pilot,] single engine armed reconnaissance helicopter. . . . The principal difference between the Kiowa Warrior and its immediate OH-58D predecessor is a universal weapons pylon on both sides of the aircraft capable of accepting combinations of the semi-active laser Hellfire missile, the Air-to-Air Stinger (ATAS) missile, 2.75" Folding Fin Aerial Rocket (FFAR) pods, and a 0.50 caliber machine gun. In addition to these weapons, the Kiowa Warrior upgrade includes changes designed to provide improvements in air-to-air and air-to-ground communications, mission planning and management, available power, survivability, night flying, and reductions in crew workload through the use of on-board automation and cockpit integration.

The primary mission of the Kiowa Warrior is armed reconnaissance in air cavalry troops and light attack companies. In addition, the Kiowa Warrior may be called upon to participate in the following missions or tasks: (1) Joint Air Attack (JAAT) operations; (2) air combat; (3) limited attack operations; and (4) artillery target designation.

Sensor Fuzed Weapon

U.S. Department of Defense, 1998b

The CBU-97/B Sensor Fuzed Weapon (SFW) [an Air Force ACAT ID program] is an anti-armor cluster munition to be employed by fighter/attack and bomber aircraft to provide . . . multiple kills per pass against armored and support vehicle combat formations. . . . SFW is currently delivered as an unguided, gravity weapon. After release, the TMD opens and dispenses the ten submunitions which are parachute stabilized. At a preset altitude sensed by a radar altimeter, a rocket motor fires to spin the submunition and initiate an ascent. The submunition then releases its four projectiles, which are lofted over the target area. The projectile's sensor detects a vehicle's infrared signature, and an explosively formed penetrator fires at the heat source.

Sergeant York

Elliot Parkin, Institute for Defense Analyses

The Sergeant York, a radar controlled twin 40 mm air defense gun for defense of armored forces, was evaluated by the Army during an April 1985 operational test at Fort Hunter Liggett, California. The operational test consisted of 30-minute force-on-force battles in which the Sergeant York defended an armored unit, conducting offensive or defensive operations, from enemy air attack. During these mock-combat battles, the Sergeant York could not effectively engage the standoff helicopted threat that enemy force successfully employed beyond the range of the Sergeant York's guns. The $4 billion program was subsequently canceled with the government writing off the $1.8 billion already spent before testing. The 65 systems that had already been delivered to the Army were scrapped.

Stinger

U.S. Department of Defense, 1998b

The Stinger missile . . . is the Army's system for short-range air defense that provides the ground maneuver commander force protection against low-altitude airborne targets such as fixed-wing aircraft, helicopters, unmanned aerial vehicles, and cruise missiles. The Stinger is launched from a number of platforms: Bradley Stinger Fighting Vehicle, Bradley Linebacker, Avenger (HMMWV), and helicopters as well as Man Portable Air Defense (MANPADS).

TOW

Rosser Bobbit, Institute of Defense Analyses

The TOW is an optically tracked wire-guided heavy anti-tank missile system introduced into the U.S. Army during the Vietnam War. It is fired from a ground mount, the High Mobility Multi-Purpose Vehicle (Hummer), Bradley Fighting Vehicle, and Cobra Attack Helicopters. Through continuous modification it has been kept up to date against current threat main battle tanks. It is being replaced by the Follow-on-to-TOW missile system, very similar in concept but with longer range and a fire-and-forget capability.

UH-1 Iroquois (Huey)

Arthur Fries, Institute of Defense Analyses

The UH-1 is a single engine, twin-bladed transport helicopter built by Bell Helicopter. It was first fielded in the Army in 1959, and served as its primary troop and light cargo lift helicopter until the subsequent introduction of the UH-60 Blackhawk. The UH-1 had a crew of three, including the crew chief, and seats for approximately six passengers. Thus, two Huey's could lift a full strength rifle squad. It could carry either internal or external cargo loads. It was the primary tactical lift carrier for the Army during the Vietnam conflict. It has also served as an airborne command and control aircraft.

UH-60 Black Hawk

U.S. Department of Defense, 1998b

The UH-60 Black Hawk [an Army ACAT IC program] is a single rotor medium-lift helicopter powered by twin General Electric T700-GE-701C turboshaft engines rated at 1,700 shp each. The Black Hawk helicopter provides utility and assault lift capability across a wide range of missions. The Black Hawk is the primary helicopter for air assault, general support, and aeromedical evacuation units. In addition, modified Black Hawks operate as command and control, electronic warfare, and special operations platforms.

APPENDIX B

Abstracts of Background Papers

Although the panel was not charged with developing or executing technical analyses related to operational testing and evaluation, we found that exploring certain technical issues in depth contributed to our deliberations. We present here abstracts of three studies, which are published separately (Cohen, Rolph, and Steffey, 1998).

STRATEGIC INFORMATION GENERATION AND TRANSMISSION: THE EVOLUTION OF INSTITUTIONS IN DoD OPERATIONAL TESTING

***Eric M. Gaier,* Logistics Management Institute, and**
***Robert C. Marshall,* Pennsylvania State University**

This paper presents a model that extends the information transmission literature to consider the question of strategic information generation. We analyze a principal-agent game in which the agent strategically chooses the probability with which s/he can distinguish a given state from its complement. We call this stage *test design*. After observing an information partition, the agent reports to the principal. We analyze several ways in which the principal might choose to extend oversight authority over the process. As the main result of the paper we show that oversight of the test design stage always improves the welfare of the principal while oversight of the reporting stage may not. The model is used to examine the historical evolution of operational testing in the Department of Defense.

ON THE PERFORMANCE OF WEIBULL LIFE TESTS BASED ON EXPONENTIAL LIFE TESTING DESIGNS

***Francisco J. Samaniego* and *Yun Sam Chong*,**
University Of California, Davis

It is common to plan a life test based on the assumption of exponentiality of observed lifetimes or lives between failures. Analysts are then able to calculate specifically how many items should be placed on test (or the number of observed failures it takes to terminate the test) and the maximum total time on test required to resolve the hypothesis test of interest. Once the test data are in hand, one has the opportunity to confirm the exponentiality assumption or to decide that an alternative modeling assumption is preferable. This paper pursues the question: "What if the data point toward a nonexponential Weibull model?" We identify circumstances in which the available data permit testing the original hypotheses with better performance characteristics (that is, smaller error probabilities) than the test originally planned; a complementary analysis of situations leading to poorer performance is also given. We give indications of the potential savings in the number of systems and the time on test that would accrue from having modeled the experiment correctly in the first place.

Various approaches to testing hypotheses concerning Weibull means are discussed. The first two sections of the paper are expository and review the main issues in exponential life testing and some properties and procedures associated with the Weibull distribution. In Sections three and four we develop the mechanics of Weibull life testing, and carefully examine the performance of Weibull life tests based on exponential life test plans.

APPLICATION OF STATISTICAL SCIENCE TO TESTING AND EVALUATING SOFTWARE INTENSIVE SYSTEMS

***Jesse H. Poore*, University of Tennessee, and**
***Carmen J. Trammel*, Software Engineering Technology, Inc.**

Defense systems are becoming increasingly software intensive. While software enhances the effectiveness and flexibility of these systems, it also introduces vulnerabilities related to inadequacies in software design, maintenance, and configuration control. Effective testing of these systems must take into account the special vulnerabilities introduced by software. The software testing problem is complex because of the astronomical number of scenarios and states of use. The domain of testing is large and complex beyond human intuition. Because the software testing problem is so complex, statistical principles must be used to guide testing strategy in order to get the best information for the resources invested in testing.

In general, the concept of "testing in" quality is costly and ineffectual; software quality is achieved in the requirements, architecture, specification, design, and coding activities. The problem of doing just enough testing to remove uncertainty regarding critical performance issues, and to support the decisions that must be made in the software life cycle is a problem amenable to solution by statistical science. The question is not whether to test, but when to test, what to test, and how much to test.

Statistical testing enables efficient collection of empirical data that will remove uncertainty about the behavior of the software-intensive system and support economic decisions regarding further testing, deployment, maintenance, and evolution. A statistical principle of fundamental importance is that a population to be studied must first be characterized, and that characterization must include the infrequent and exceptional as well as the common and typical. It must be possible to represent all questions of interest and all decisions to be made in terms of this characterization. When applied to software testing, the population is the set of all possible scenarios of use with each accurately represented as to frequency of occurrence. The operational usage model is a formalism presented in this paper that enables the application of many statistical principles to software testing and forms the basis for efficient testing in support of decision making.

Most usage modeling and related statistical testing experience to date is with embedded real-time systems, application program interfaces, and graphical user interfaces. One very advanced industrial user of this technology is the mass storage devices business. Use of this technology has led to extensive test automation, significant reduction in the time these software-intensive products are in testing, improved feedback to the developers regarding product deficiencies or quality, improved advice to management regarding suitability for deployment, and greatly improved field reliability of products shipped.

From a statistical point of view, all the topics in this paper follow sound problem-solving principles and are direct applications of well-established theory and methodology. From a software testing point of view, the application of statistical science is relatively new and rapidly evolving, as an increasing range of statistical principles is applied to a growing variety of systems. Statistical testing is used in pockets of industry and agencies of government, including DoD, on both experimental and routine bases. This paper is a composite of what is in hand and within reasonable reach in the application of statistical science to software testing.

APPENDIX C

A Potential Taxonomy for Operational Tests

INTRODUCTION

Over the course of this study, the panel has heard presentations on the following types of defense systems: airport runway repair systems, software systems for personnel administration, radar-assisted combat helicopters, shoulder-fired anti-tank missiles, radar jammers, systems for prioritizing strategic threats, and operating systems designed to run other software. Systems can make use of entirely new technology, or well-tested technology that has worked well on similar systems. They can be slight upgrades or embody entirely new capabilities. They can be 100 percent software or 100 percent hardware (though all recent ACAT I systems contain software components). There are systems whose failure could result in loss of life. One would expect that operational tests for systems this varied should take on very different forms. It is necessary to consider a new system's characteristics to select an appropriate operational test design.

Certainly, when a member of the relevant service test agency is called upon to design a test for a system, if a similar system could be discovered that was considered to have been well-tested in the past, it would make sense to borrow certain features of that previous test to use on the new system, especially if the previous system was tested relatively recently. Unfortunately, it would seem that a taxonomy that could encompass such diversity would necessarily have a large number of cells, and that would require a large number of well-designed, fully analyzed tests that covered the range of DoD operational test experience. On the other hand, taxonomies that tried to avoid this difficulty by collapsing cells to form aggregate cells would run the risk of recommending tests that were not appropriate since there would be too much heterogeneity within cells. These two requirements for few cells and homogeneous cells run against each other. The panel presents an attempt at the creation of a practical taxonomy, but this is only a preliminary attempt at such a construct. It is important to keep in mind that the kinds of military systems under development is an extremely dynamic process. The systems that are under development today are tremendously different from those developed in the 1950s. Therefore, any taxonomy would have to be reevaluated on a regular basis, and flexibility of the implementation of such a taxonomy would be very important. Further, there is no reason why a taxonomy is the preferable construct. Another possibility is a type of checklist, where answers to a succession of questions about system characteristics add different types of features to the recommended operational test.

Following is a discussion of characteristics that may affect the nature of the operational test of a military system. Boxes C-1 and C-2 list some of the many characteristics considered for inclusion in our taxonomy. These potential taxonomy dimensions have been divided into two broad categories, (1) those which have a broader application but probably do not have a direct bearing on the operational test, and (2) those dimensions which may have a direct bearing on how an operational test is designed and how preparations are made for it.

The panel selected three characteristics that seem important for test design. The taxonomy is designed to serve the following purposes:

- Reflect the prevalence of various types of systems;
- Highlight attributes that might call for different statistical approaches, affect decision tradeoffs, or involve qualitatively different consequences; and
- Provide a framework for developing the test scenarios.

A USEFUL TAXONOMY

The panel has developed a taxonomy for defense systems for the purpose of classifying systems by their operational test design needs rather than the uses to which the system will be put. For example, if the test issues and the type of data collected to address those issues are similar for a missile and a combat aircraft, then the taxonomy should put these two systems together in the same category. On the other hand, if say, a telephone for administrative use and a similar telephone for transmissions of intelligence information have quite different test issues, then the taxonomy should put them in different categories.

The panel's taxonomy has a broader conceptual base and utility than test design, however. A proper taxonomy entails notions of the loss function underlying the decision of whether to enter into full-rate production—since that is the purpose of operational test—and as discussed above, this decision directly involves the issue of test benefit versus test cost. To understand the benefit gained from the use of various test scenarios, one must consider the likely variability of

Box C-1 Dimensions with Broad System Implications

1. Cost. What is the cost of a test item? What is the number of items to be procured?

2. New versus evolutionary. Is the system a de novo development? Is it an upgrade? Is it a modification? Is it a derived design? Is it a replacement for another system?

3. Testing consequences. What are the consequences of not achieving a successful replacement? What are the consequences of achieving a replacement at a much higher cost than anticipated? What are the consequences of receiving it at a much later date than planned? What are the consequences of receiving it at a much lower level of performance than promised?

4. What are the range of possible decisions that might be appropriate, given the outcome of this operational test?

5. What is the cost of repairing and supporting the system?

6. What are the logistics costs of fielding the system?

7. Roles and missions. Could this system perform in roles and missions different from those which are tested? Could this system provide a backup for other systems, in case they perform badly or are seriously attrited?

8. Force flexibility. Would this system significantly improve the flexibility inherent in our fielded portfolio of such systems? Would it allow us to perform new missions, or to perform existing missions in more than one way? Would this system free up other systems for more valuable uses?

9. Is it plausible that the system opens up opportunities for radical approaches that are not yet well understood?

10. To what extent could this system create fear among potential adversaries about our capabilities?

11. Segmentation. To what extent is it possible to segment the system into subsystems and components, particularly ones that can be tested independently? To what extent do systems integration problems and subsystem interactions interfere with the validity of “segmented tests”?

12. Architecture. To what extent does the design or nature of this system allow for improvements on a subsystem-by-subsystem basis? To what extent is system performance affected by the current state of development of the individual subsystems?

the system across environments, threats, tactics, and doctrine, across prototypes, across levels of training of the users involved, etc. It forces the developer of the taxonomy to think statistically, that is where is the variability of system performance likely to lie? So the goal of a taxonomy is a way of formalizing a decomposition of variance of the system as a function of various characteristics of the situations in which the system might be used.

Box C-2 Dimensions with a Direct Bearing on Test Design and Execution

1. Role of software. Is the system a software product? Does it have significant software content? Does it use a dedicated computer or require the development of new computer hardware?

2. Environment of use. How stressful is the environment within which the system must operate?

3. Environment of test and evaluation. How close are test environments to actual-use (combat) environments? What is the relevance of simulation? To what extent are performance evaluations dependent upon indirect measurement and inference?

4. Scenario dependence. To what extent is the value of the weapon system, or its operational performance, affected by the testing scenario? For example, does the scenario correspond to operations on the first day of the war or after air superiority has been achieved? Does it correspond to a scenario in which we have ample warning or are caught by surprise? Is it assumed that air bases are available nearly, or that operations must be adapted for primitive air strips?

5. Have the potential users and testers had adequate time to develop tactics that will utilize this system most effectively?

6. Training. To what extent does the performance of the system depend on the training of those who will operate it? Of those, who will operate any “enemy systems” during the testing? To what extent may the assessed performance of the system be affected by the training of those who will collect, reduce, and interpret the data collected during the operational test?

7. Instrumentation. To what extent is test range instrumentation adequate for assessing the system performance during the operational tests? To what extent might the act of instrumenting the test articles interfere with their performance?

8. Maturation. To what extent is this system matured? To what extent is its performance likely to improve markedly as it is better understood? After it has been fielded? After test and operational data have accumulated?

9. Sources of information. To what extent will there be continued reliance on data collected through operational tests or ongoing use (with proper documentation) in understanding the performance, reliability, and nature of this weapon system?

10. Process perfection. To what extent will the performance of this system gradually improve as production, testing, repair, and support processes are perfected?

11. Heterogeneity. To what extent will differences among produced items be testable, or recognizable, before the items are used? To what extent will military commanders be able to either manipulate or hedge against apparent heterogeneity among fielded units of this weapon system?

12. What is the effect of a system failure?

Characteristics that are considered important in some systems are represented by dimensions of the taxonomy. For example, there are systems for which the environment presents real stress, and systems for which it does not. Therefore, sensitivity to environment is a potential dimension of the taxonomy. Likewise, characteristics that are considered less important for most systems are ignored in the taxonomy. There is finally the fact that the population size of each cell in the taxonomy will differ considerably. So, to take an extreme example, if all systems but one are insensitive to direct sunlight, it is unreasonable to expect the taxonomy to have a sunlight dimension to accommodate a single system. Any taxonomy has to be used with a view towards its limitations. These notions of what factors are related to system response, and how many systems are affected by a given factor, governed the panel in its attempt to create the taxonomy, and assisted in the goal of creating a taxonomy with few cells, but enough to create relatively homogeneous cells.

The panel used the following criteria to develop its taxonomy. These three criteria should be used in planning, conducting, and evaluating a test.

1. Skill level required of the people part of the system. Skill levels may be:

a. Highly skilled with extensive, multifaceted training. Such systems include aircraft systems, naval combat vessels, and armored ground combat systems.

b. Highly skilled in a single or limited number of actions. Such systems include individual weapons, ground and sea transportation systems, communication systems, and radar systems.

c. Little skill required. In this case, training is not a significant issue. Such systems may be clothing, rations, or temporary shelter systems.

2. Nature of opposition. Often in the use of a military system there is an opposition force whose mission is to degrade or prevent the effective performance of the system under test. We will divide the nature of this opposition into three categories: active, passive, and none.

a. Active opposition. This is used in an operational test of a system designed for combat operations. The nature of the test is usually but not always a force-on-force combat test in which there is a manned active opposing force.

b. Passive opposition. This category is for scripted unmanned opposition which is under total control of the test designer. Included among this type of opposition is scripted jamming of a communication network and the emplacement of an enemy minefield.

c. No opposition. Included are items which are tested with no opposition and include such things as clothing, rations, and most information management and transportation systems.

3. Effect of a system failure. The nature of the seriousness of a system failure depends upon several factors. If the failure occurred during a mission, was it a (1) total system failure (loss of system forever), (2) critical component failure (causes immediate withdrawal), (3) component failure with ability to continue mission in a degraded mode, or (4) component failure which will not degrade mission capability. If the failure occurred while not on a mission, can the system embark upon the next mission? If so, will it be in a degraded mode or can it continue to function without degradation?

Also included in this criterion is the effect of a failure on the mission. For example, if a system failed during a mission, the total loss of that system for that mission may have a significant effect while the ability of the system to continue after a component failure would probably have little or no effect. Combining these ideas, we can divide system failures into the following categories:

a. Catastrophic. A failure would probably cause a total failure of the mission; or would represent the loss of that system forever.

b. Serious. A failure would seriously degrade the chance of mission success or would represent the loss of that system for the duration of that mission.

c. Significant. A failure would probably degrade the mission, or would cause a delay in the accomplishment of the mission.

d. Minor. Would have little or no effect on mission success. May cause inconvenience or create additional cost.

The Effect on Test Design

The Effect of Training on an Operational Test

One of the major factors to be considered in the design of a test is the control and evaluation of the variability of the dependent variables; i.e., the measures which indicate how a system performed. In addition to the hardware and software, a "system" also includes the people and how the system is employed. A great source of variation in the performance of most systems is the variability in human behavior. The degree to which human variability can be understood and minimized will have a large effect on the reliability of the overall test data.

At the very outset of test planning, provision should be made to minimize the effect of human variability. This can best be done by planning for appropriate training of the people operating the systems. The higher the skill level required, the greater the importance of adequate training. For those systems requiring highly skilled operators, testing should be conducted to assess whether the training is adequate. Force development tests and experiments are excellent tools for the assessment of the adequacy of training. There are times when even a "golden crew," an extremely competent crew, is useful to examine the upper limit of the performance of a system. The data generated by such crews should be used carefully and not be advertised as the "expected" performance of a system.

Finally, when there is an active opposing force, these crews should also be well trained, and they should present an opposition of as near as possible equal quality in each trial. Variability in opposing force performance causes unwanted variability in the measured performance in the system under test. This leads us to the next criterion.

The Effect of Opposing Force on an Operational Test

The sources of variability in the outcome of a force-on-force battle are legion—so much so that in most operational tests, trial based data (such as exchange ratios) seldom behave in a manner that lend themselves to reliable statistical measures. Some sources of variation can be controlled (time of day, mission, terrain, etc.), others cannot be controlled but can be measured (temperature, accidents, failures, etc.), and still others are neither controlled nor measurable (interpersonal relations, inattention, etc.). All of these affect the active opposition in a similar manner to how they affect the system under test. In such tests, the designer should concentrate on three things: control unwanted variability to the extent possible, design for event based data rather than trial based data,[1] and design for evaluation of the system using qualitative measures.

In cases of passive opposition, the nature of the opposition should be designed into the test and scripted exactly to assure no unwanted variability is produced by the opposition. Often in such tests, sufficient data can be generated to do an adequate statistical analysis.

When there is no opposition, the variability of the results comes solely from the variability of the system under test and its environment. Most measures from such tests lend themselves readily to statistical analysis.

The Effect of Failures in an Operational Test

Simply put, a failure is not necessarily a failure. The effect of each failure should be considered in the evaluation of reliability, availability, maintainability (RAM) data. In the scoring conference (the evaluation of RAM data) this effect should be considered. In an aircraft system should a failure of the heating system, a failure of the RADAR warning device, a failure of a missile to launch when it should have, and an engine failure (in a single engine aircraft!) all be treated

[1] Trial based data are such things as how many of the enemy were killed or who won the battle. Event based data are such things as "given an engagement opportunity, did an engagement take place" and "did a message from the commander get to the intended recipient"?

alike? Provision should be made in the test plan to evaluate RAM failures according to some taxonomy similar to that given above.

Proposed Taxonomy

Figure C-1 is a graphic representation of a proposed taxonomy for operational test that the panel believes represents a good first step towards the development of a taxonomy that categorizes defense systems by the type of operational test that is needed. If such a taxonomy is found useful, to operationalize this taxonomy an example system in each cell would have to be identified that had gone through operational test recently, and a well-designed test would have to be drawn up through a collaboration of test experts in the relevant test service agency, DOT&E, IDA and possibly statistical experts from academia.[2] Then, new entrants to each cell could make use of minor adjustments to operational test designs of the leading case in the cell. Certainly, a feedback system could modify the test design of the leading cases as deficiencies in tests of other cell members were discovered.

To repeat some cautionary statements from above, this attempt at a taxonomy is only a preliminary step. Again, the types of military systems under development is an extremely dynamic process and therefore, any taxonomy would have to be re-evaluated on a regular basis. Finally, there is no reason why a taxonomy must be the preferable construct: the formation of checklists for this purpose should also be investigated.

[2] Figure C-2 places a few defense systems in some of the categories. The intent is to demonstrate the kinds of systems that would fit into some cells. We are not proposing that they be seen as the systems on which one would base other operational tests.

Mission Disruption	Opposition	Skill		
		HM	HL	L
Catastrophic	None	1	2	3
	Active	4	5	6
	Passive	7	8	9
Serious	None	10	11	12
	Active	13	14	15
	Passive	16	17	18
Significant	None	19	20	21
	Active	22	23	24
	Passive	25	26	27
Minor	None	28	29	30
	Active	31	32	33
	Passive	34	35	36

FIGURE C-1 A taxonomy for operational tesing. NOTES: The numbers in this figure are keyed to Figure C-2. HM: High; multifaceted training; HL: High; limited training; L: Low

1 Parachutes (People) Parachutes (Things) C-17 CH-46 CH-47 CH-54	**4** Apache AH-1 Cobra OH-58D	**9** Mine Detection Equipment Mine Clearing Devices
10 UH-1 Iroquois (Huey) UH-60	**13** JSTARS M1 M1A2 SGT. YORK LOSFH	**14** Bradley Amphibious
18 Jammers	**20** Road Building Equipment Trucks HETs HMMWVs Management Info Systems	**23** Rifles Machine Guns Pistols Dragon Tow Stinger
25 RPV (Air)	**26** Mortars Artillery (Tracked) Artillery (Stationary) Cannons	**29** Defensive Emplacements
30 Clothing Food Tentage	**36** RPV (Ground) Radios Telephones Mines (Defensive) Mines (Scatterable) Mines (Emplaced)	

FIGURE C-2 Examples of defense systems in the proposed taxonomy.

APPENDIX D

Elements ISO 9000

The International Organization for Standardization (ISO) is a federation of organizations that represent 92 member countries established in 1946 to "promote the development of international standards and related activities to facilitate the exchange of goods and services" (Breitenberg, 1993). The United States representative to the ISO is the American National Standards Institute (ANSI).

In 1987 ISO issued the ISO 9000 Standard Series, a set of five individual but related standards (9000 through 9004) that provide guidance on requirements for quality systems; see Box D-1. The most comprehensive set of standards are ISO 9001; its various elements or specific requirements are listed in Box D-2. Significant changes were made to the ISO 9000 standards in 1994. ISO 9004-1 is now the overall guidance standard in the basic set of quality management standards; however, it also retains its original role of a separate standard for organizations to use internally to improve their quality management practices. Further revisions are anticipated in 1998 to move the ISO 9000 family toward "total quality management." The ISO's Technical Committee 176 has already started developing the revisions, based on quality management principles and extending beyond the product realization process.

Initially designed to be used in contractual arrangements between two parties or for internal auditing, use of ISO 9000 standards has greatly expanded over the last 10 years. More than 96 countries have adopted the ISO quality standards, and most industries throughout the world, including many in the United States, recognize them. Today, many government agencies and private industries worldwide require the organizations with which they do business to comply with one of the standards in the series. The standards have been applied to manufacturing,

Box D-1 Quality Systems Standards Descriptions

ISO 9000 — Explains fundamental concepts and terms, and is a guide to the use of the other four standards (9001-9004). The two relevant subsections are:

- ISO 9000-3 — Guidelines for applying ISO 9001 to the development, supply, and maintenance of software.
- ISO 9000-4 — Guidelines for dependability (RAM) and program management. Covers the essential features of a comprehensive dependability program for the planning, organization, direction, and control of resources to produce products that will be reliable and maintainable.

ISO 9001 — Provides quality system requirements in design, development, production, installation, and servicing. This is the most comprehensive set of specific standards.

ISO 9002 — Provides quality system requirements in production and installation. Identical to 9001 except for the deletion of requirements for design and development.

ISO 9003 — Provides quality requirements for product inspection and test.

ISO 9004 — Provides guidance on designing and improving quality systems. It has several subsections, including 9004-2 which deals with service organizations.

service industries, software development organizations, and the process industries. Currently, the standards are most often applied to business processes that directly affect the products and service provided by an organization. In the future, the scope of application will be expanded to all areas of enterprises, including finance, accounting, human resources, market research, and marketing.

BASIC PRINCIPLES

The ISO 9000 standards are based on three basic principles: processes affecting the quality of products and services must be documented; records of important decisions and related data must be archived; and when these first two steps are complied with, the product or service will enjoy continuing improvement. We now discuss each of these principles and discuss how adherences results in product and service improvement.

Documenting Processes Documentation often helps to discover important ad hoc processes that are inconsistently applied. This discovery then leads to changes in the process that results in their more consistent application, which, in turn, reduces the variability in the output of these processes.

Documenting processes also provides a better understanding of how an orga-

Box D-2 Elements or Specific Requirements of ISO 9001

1. Management Responsibility. Establish quality policy, define organizational responsibilities and empowerment of management representatives, provide verification resources, and review the quality system periodically, maintaining complete review records.

2. Quality System. Provide and maintain an effective, up-to-date, documented system that defines quality procedures and that ensures consistent product conformance to specified requirements.

3. Contract Review. Establish and maintain procedures for review of proposals and contracts, assuring that requirements are adequately defined and documented and that the supplier is capable of meeting contractual requirements.

4. Design Control. Define and maintain procedures to control and verify product design and design changes to ensure the outputs of design meet specified requirements. The items defined include the planning of activity assignments; definition of interfaces, design input, and design output; verification; and control of design changes.

5. Document Control. Create and maintain procedures to control all documents and data relating to the requirements of this standard, to make readily available both process or methodology documents and product documents, and to control changes to documents.

6. Purchasing. Select subcontractors based on their ability to meet subcontract requirements, assure purchasing documents clearly describe the products and services ordered, maintain complete and accurate purchasing information, assure adequacy of vendors' quality systems, and verify that purchased product conforms to specified requirements.

7. Purchaser Supplied Product. Establish and maintain procedures for verification, storage, and maintenance of purchaser-supplied product provided for incorporation by the supplier; and protect and store such product, just as any product manufactured or purchased by the business unit, and notify the customer when the product is unfit for use.

8. Product Identification and Traceability. Create and maintain procedures for identifying, recording, and tracing the product during all stages of production, testing, delivery, and installation so that the product can be identified in the event a recall or other activity is required.

9. Process Control. Identify and plan the production and installation processes that directly affect quality and ensure that these processes are carried out under controlled conditions, including documented work instructions or methodology, suitable work environments and resources, process and product monitoring, and availability of appropriate workmanship standards.

10. Inspection and Testing. Provide for the inspection or verification of incoming product before use, the in-process inspection and testing of product, and the final inspection and testing of product, including the provision for maintaining inspection and test records.

11. Inspection, Measuring, and Test Equipment. Control, calibrate, and maintain inspection, measuring, and test equipment (whether owned, leased on loan, or purchaser-provided) to demonstrate the conformance of product to the specified requirements, and maintain associated records and procedures to determine actions necessary when test equipment is found to be out of calibration.

12. Inspection and Test Status. Put in place a system to identify the inspection and test status of product (awaiting, passed, or failed inspection/test), prevent use of defective product, maintain associated records, and identify the entity accepting conforming product for release.

13. Control of Nonconforming Product. Establish and maintain procedures and records to ensure that product not conforming to specified requirements is prevented from inadvertent use or installation and that repaired product is reinspected and/or tested according to the quality plan or procedures.

14. Corrective Action. Create, document, and maintain procedures for investigating the causes of nonconforming product, analyzing processes and all associated activities and data to detect and eliminate potential causes of nonconforming product, assuring that process changes are effective, and implementing and recording changes in the process.

15. Handling, Storage, Packaging, and Delivery. Establish, document, and maintain procedures to protect product from damage during handling, storage, packaging, and delivery.

16. Quality Records. Create and maintain procedures for managing the identification, collection, indexing, filing, storage, protection, maintenance, access, and disposition of quality records.

17. Internal Quality Audits. Carry out a comprehensive system of planned and documented internal quality audits, by trained personnel independent of the organization being audited, to verify that activities are conducted with documented arrangements and to determine the effectiveness of the quality system.

18. Training. Establish and maintain procedures to identify the training needs of each person whose work affects the quality of product delivered to customers, to train new people quickly in the quality standards, to populate all jobs with people qualified on the basis of education, experience, and/or training, and to maintain training records.

19. Servicing. As specified in the contract, create and maintain procedures for providing and verifying that service and ongoing support meet specified requirements.

20. Statistical Techniques. Where needed, establish procedures for identifying adequate statistical techniques required for verifying the acceptability of process capability and product characteristics, and demonstrate that such techniques are applied appropriately.

SOURCE: American Telephone & Telegraph, 1994.

nization does business. First, customer requirements are better understood. In addition, since the process of designing and producing products and services is well understood, the customer needs can be more directly addressed. Finally, there is a clear process for exploring the sources of and resolving customer complaints.

Retaining Records and Data Supporting Decisions Decisions, only made by those with the authority to make them, that affect the quality of a product or service, need to be documented, along with all supporting data and records. These decisions need to be periodically reviewed for their basis and authorization. Data obtained during the life of the product or service must be archived for a specified period of time if the data provide information about the product or service quality. These data are then used for product or service improvement, or for development of new products or services.

Continuing Improvement The three elements of continuing product improvement are: audit, corrective and preventive action, and management review.

Internal auditing provides important information as to the potential sources of ineffectiveness in a business process. In conjunction with surveillance by ISO 9000 registrars, internal auditing and external surveillance produce corrective action requests that lead to an understanding of the cause of the ineffectiveness and its resolution. The final step of management review assures that continuing product or service improvement is in place and the integration of these improvements into the overall goals of the business.

BENEFITS OF APPLICATION TO DoD

DoD might draw constructively on industrial practices, particularly in such areas as documentation, uniform standards, and the pooling of information on operational suitability, following the direction of the ISO 9000 series.[1] Documentation of processes and retention of RAM-related records (for important decisions and valuable data) are practices now greatly emphasized in industry. The same should be true for DoD, especially for the purposes of assessing operational suitability in support of major production decisions.

DoD should document, consistently across the services, all relevant sources

[1] Interestingly, many aspects of the ISO 9000 standards were originally developed from DoD's Quality Management Program, MIL-Q-9858. It was revised in 1963, and in 1968 NATO adopted its provisions as the Allied Quality Assurance Publication (AQAP-1). In 1970 the Ministry of Defence of the United Kingdom adopted AQAP-1, and in 1979 the British Standards Institution developed BS 5750, the first commercial quality management system. The ISO 9000 standard series was created from these predecessors.

of test data and information. Development teams would not have to "reinvent the wheel" when beginning a project. Instead, they would identify and learn from previous studies and findings. Setting multiservice operational test and evaluation standards similar to those in ISO 9000 is critical to the creation of an environment of more systematic data collection, analysis, and documentation.

The panel recommends (see Chapter 3) that DoD and the services develop a centralized test and evaluation data archive and standardize test data archival practices based on ISO 9000. This archive should include both developmental and operational test data; use of uniform terminology in data collection across services; and careful documentation of development and test plans, development and test budgets, test evaluation processes, and the justification for all test-related decisions, including decisions concerning resource allocation.

Standardization, documentation, and data archiving facilitate use of all available information for efficient decision making. Routinely taking these steps will:

- provide the information needed to validate models and simulations, which in turn can be used to plan for (or reduce the amount of) experimentation needed to reach specified operational test and evaluation goals;
- allow the "borrowing" of information from past studies (if they are clearly documented and there is consistent usage of terminology and data) to inform the assessment of a system's performance based upon limited testing, by means of formal and informal statistical methods and other approaches;
- make possible the use of data from developmental testing for efficient operational test design;
- allow learning from best current practices across the services; and
- lead to an organized accumulation of knowledge within the Department of Defense.

A key benefit of documentation and archival of test planning, test evaluation, and in-use system performance is the creation of feedback loops to identify system flaws for system redesign, and to identify where tests or models have missed important deficiencies in system performance.

Retention of records may involve additional costs, but it is clearly necessary for accountability in the decision-making process. The trend in industry is to empower employees by giving them more responsibility in the decision-making process; along with this responsibility comes the need to make people accountable for their decisions. This consideration is likely to be an important organizational aspect of the operational testing of defense systems.

APPENDIX E

Glossary and Acronyms

ACAT: Acquisition Category; a designation for each program based on cost that determines both the level of review that is required by law and the level at which milestone (see below) decision authority rests in DoD. There are four acquisition categories, ACAT I through ACAT IV; the most expensive systems are designated ACAT I. ACAT I programs have two sub-categories: ACAT ID (milestone decision authority is the USD[A&T]); and ACAT IC (milestone decision authority is the DoD Component Head.

AFOTEC: Air Force Operational Test and Evaluation Center

AMSAA: Army Materiel Systems Analysis Activity

Analysis of Alternatives: (formerly Cost and Operational Effectiveness Analysis — COEA); this is a cost-benefit analysis tool that provides justification for the selection of one procurement option over an alternative.

Acquisition Program Baseline (APB): can be viewed as a contract between the milestone decision authority and the relevant service, including the program manager and his/her supervisors.

ATACMS: Army Tactical Missile System

BAT: Brilliant Anti-Tank

CAIG: Cost Analysis Improvement Group

DAB: Defense Acquisition Board

Developmental Testing (DT): responsible for performing specification-based testing, verifies that the system meets all specifications, and certifies that the system is ready to enter operational testing; determines *if* and *how* the system works.

DIS: Distributed Interactive Simulation

DoD: Department of Defense

DOT&E: Director, Operational Test and Evaluation

DT&E: Developmental Testing and Evaluation

EADSIM: Force-on-Force Air Defense Model

EMD: Engineering and Manufacturing Development

Evolutionary Procurement: pertains to continuous changes to a system and the associated stage-wise testing and development, so that what is operationally tested is not necessarily what is deployed; particularly relevant to software and software-intensive systems.

IDA: Institute for Defense Analyses; a federally funded research and development center established to assist the Office of the Secretary of Defense, the Joint Staff, the Unified Commands and Defense Agencies in addressing important national security issues, particularly those requiring scientific and technical expertise.

IOT&E: Initial Operational Test and Evaluation

ISO: Organization for International Standardization; a worldwide federation of 92 member countries established to promote the development of international standards and related activities to facilitate the exchange of goods and services.

JROC: Joint Requirements Oversight Council; serves to support milestone review, validate the Operational Requirements Document, and validate mission need.

LOSFH: Line of Sight—Forward Heavy; LOSFH was the generic name given early on to the conceptual ADATS system, while ADATS was the particular piece of hardware/software that was chosen to fill that role.

M&S: Modeling and Simulation

MCOTEA: Marine Corps Operational Test and Evaluation Activity

MDAPs: Major Defense Acquisition Program; ACAT I programs are MDAPs

Measure of Effectiveness (MOE): A measure of how well an operational task or an assigned task is accomplished; can be measured directly or may require the aggregation of MOPs.

Measure of Performance (MOP): A measure of how well a system performs its function or how well a design characteristic meets an operational requirement.

Milestone: One of five steps in the procurement of a weapons system; Milestone 0 is the concept studies approval, Milestone I is the concept demonstration approval, Milestone II is the development approval, Milestone III is the production approval, and Milestone IV is the major modification approval.

Mission Needs Statement (MNS): a conceptual document, prepared by the relevant military service in response to a perceived threat, that is supposed to identify a broadly stated operational need (not a specific solution to counter the perceived threat).

MTBOMF: Mean Time Between Operational Mission Failure

MTTF: Mean Time To Failure

Operational Effectiveness: the capability of a system to perform its mission in the operational environment, and in the face of the expected threat, including countermeasures.

OMS/MP: Operational Mode Summary and Mission Profiles; defines the environment and stress levels the system is expected to encounter in the field. They include the overall length of the scenarios of use, the sequence of missions, and the maintenance opportunities.

Operational Requirements Document (ORD): a document that describes in some detail the translation from the broadly stated mission need to the system performance parameters that the users and the program manager believe the system must have to justify its eventual procurement.

Operational Suitability: the degree that a system can be placed satisfactorily in field use with consideration given to availability, compatibility, transportability, interoperability, wartime usage rates, maintainability, safety, human factors, manpower supportability, logistics supportability, natural environmental effects and impacts, documentation, and training requirements.

Operational Testing and Evaluation (OT&E): pertains to field tests, under realistic conditions, to determine system effectiveness and suitability for use in combat by typical military users; assesses *when* and *where* the system will work.

OPTEC: Army Operational Test and Evaluation Command

OPTEVFOR: Navy Operational Test and Evaluation Force

OSD: Office of the Secretary of Defense

PA&E: Office of Program Analysis and Evaluation

PEO: Program Evaluation Office

PM: Program Manager; the "champion" in the Department of Defense of a military system in development.

RAM: Reliability, Availability, and Maintainability

ROC: Required Operational Capability

Test and Evaluation Master Plan (TEMP): documents the overall structure and objectives of the test and evaluation program, provides a framework for generating detailed test and evaluation plans, and documents associated schedule and resource implications.

USD(A&T): Under Secretary of Defense (Acquisition and Technology)

VV&A: Verification, Validation, and Accreditation

References

American Telephone & Telegraph
1994 *Using ISO 9000 to Improve Business Processes*. Murray Hill, N.J.: AT&T Corporate Quality Office.

Ascher, H., and H. Feingold
1984 *Repairable Systems Reliability: Modeling, Inference, Misconceptions and Their Causes*. New York: Marcel Dekker.

Atkinson, A.C.
1994 Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* 89(428):1329-1339.

Bain, L.J., and M. Engelhardt
1991 *Statistical Analysis of Reliability and Life Testing Models, Theory and Methods*. Second edition. New York: Dekker.

Barlow, R.E., and F. Proschan
1975 *Statistical Theory of Reliability and Life Testing; Probability Models*. New York: Holt, Rinehart and Winston, Inc.

Blue Ribbon Defense Panel
1970 *Report to the President and the Secretary of Defense on the Department of Defense: Appendix F—Staff Report on OT&E*. Washington, D.C.: U.S. Department of Defense.

Box, G.E.P., and J.S. Hunter
1961 The 2^{K-p} fractional factorial designs, Part 1. *Technometrics* 3(3):311-351.

Breitenberg, M.
1993 *More Questions and Answers on the ISO 9000 Standard Series and Related Issues*. Standards Code and Information Program, Office of Standards Services. Report #NISTIR 5122. Gaithersburg, MD: National Institute of Standards and Technology.

Bridgman, M.S., and D.V. Glass
1992 *Better Assessment of Operational Suitability*. Volume 1: Final Report. McLean, VA: Logistics Management Institute.

Carlin, B.P., and T.A. Louis
1996 *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman and Hall.

Chatfield, C.
1995 Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society* 158(3):419-444.

Citro, C.F., and E.A. Hanushek, eds.
1991 *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*. Volume I: Review and Recommendations, Volume II: The Uses of Microsimulation Modeling. Panel to Evaluate Microsimulation Models for Social Welfare Programs, Committee on National Statistics, National Research Council. Washington, D.C.: National Academy Press.

Clemen, R.T.
1991 *Making Hard Decisions: An Introduction to Decision Analysis*. Boston: PWS Kent.

Cochran, W.G.
1977 *Sampling Techniques*. 3rd edition. New York: Wiley.

Cohen, M.L., J.E. Rolph, D.L. Steffey, eds.
1998 *Statistics, Testing, and Defense Acquisition: Background Papers*. Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics, National Research Council. Washington, D.C.: National Academy Press.

Coleman, D.E., and D.C. Montgomery
1993 A systematic approach to planning for a designed industrial experiment. *Technometrics* 35(1):1-12.

Computer Sciences Corporation
1994 *A Comparative Analysis of Tri-Service Accreditation Policies and Practices*. Volume 1 of the Accreditation Requirements Study Report. Report JTCG/AS-93-SM-20. Camarillo, Calif.

Cook, R.D.
1994 Using dimension-reduction subspaces to identify important inputs in models of physical systems. *American Statistical Association Proceedings of the Section on Physical and Engineering Sciences*. Washington, D.C.: American Statistical Association.

Cox, D.R., and P.A.W. Lewis
1972 *The Statistical Analysis of Series of Events*. London: Chapman and Hall.

Currin, C., T. Mitchell, M. Morris, and D. Ylvisaker
1991 Bayesian prediction of deterministic functions, with applicability to use design and analysis of computer experiments. *Journal of the American Statistical Association* 86:953-963.

Dalal, S.R., E.B. Fowlkes, and B. Hoadley
1989 Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *Journal of the American Statistical Association* 84(408):945-957.

Daniel, C.
1976 *Applications of Statistics to Industrial Experimentation*. New York: Wiley.

Deming, W.E.
1986 *Out of the Crisis*. Cambridge, Mass.: Center for Advanced Engineering Study.

Draper, D.
1995 Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society* Series B, 57(1):45-97.

Escobar, L., and W. Meeker
1994 Fisher information matrix for the extreme-value, normal and logistic distributions and censored data. *Applied Statistics* 43:533-540.

Friedman, J.
1991 Multivariate adaptive regression splines (with discussion). *Annals of Statistics* 19:1-141.

Fries, A.
1994a Design of experiments in operational test and evaluation: Where should we go next? *The ITEA Journal of Test and Evaluation* 14(4):20-33.

1994b Recent Demonstrations of the Inherent Value of Operational Test and Evaluation (OT&E). Unpublished white paper prepared for the Director, Operational Test and Evaluation. Alexandria, Va.: The Institute for Defense Analyses.

Gaier, E.M., and R.C. Marshall

1998 Strategic information generation and transmission: The evolution of institutions in DoD operational testing. In M.L. Cohen, J.E. Rolph, and D.L. Steffey, eds., *Statistics, Testing, and Defense Acquisition: Background Papers*. Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics, National Research Council. Washington, D.C.: National Academy Press.

Galway, L.A., and C.H. Hanks

1996 *Data Quality Problems in Army Logistics: Classification, Examples, and Solutions*. Santa Monica, Calif.: RAND.

Gaver, D.P., and I.G. O'Muircheartaigh

1987 Robust empirical Bayes analyses of event rates. *Technometrics* 29:1-15.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin

1995 *Bayesian Data Analysis*. New York: Chapman and Hall.

Giadrosich, D.L.

1990 Modeling and simulation in operational test and evaluation: A dinosaur's viewpoint. *The ITEA Journal of Test and Evaluation* XI(1):45-52.

Hahn, G.J.

1977 Some things engineers should know about experimental design. *Journal of Quality Technology* 9(1):13-20.

1984 Experimental design in the complex world. *Technometrics* 26(1):19-31.

Hall, M.J., L. Weidell, S.S. Nair, and M.O. Gonzales

1994 *Test and Evaluation Plan for the Initial Operational Test and Evaluation of the Longbow Apache (AH-64D)*. Alexandria, Va.: United States Army.

Hampel, F., E.M. Ronchetti, P.J. Rousseuw, and W.A. Stabel

1986 *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hill, H.M.

1960 Experimental designs to adjust for time trends. *Technometrics* 2:67-82.

Hoadley, B.

1981 Quality management plan (QMP). *Bell System Technical Journal* 60:215-273.

Huber, P.J.

1981 *Robust Statistics*. New York: Wiley.

Iman, R.L., and W.J. Conover

1980 Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Communications in Statistics—Theory and Methods* 9(17):1749-1842.

Ishikawa, K.

1985 *What is Total Quality Control? The Japanese Way*. Englewood Cliffs, NJ: Prentice-Hall.

Kaminski, P.G.

1995 *Reinventing DoD Test & Evaluation*. Address to the International Test and Evaluation Association Symposium, Huntsville, Alabama, October.

Laird, M.

1972 Cost Estimating for Major Defense Systems. Memorandum from the Secretary of Defense for the Secretaries of the Military Departments.

Lall, P., M. Pecht, and M.J. Cushing

1994 A physics-of-failure (POF) approach to addressing device reliability in accelerated testing. *Proceedings of the 5th European Symposium on Reliability of Electronic Devices, Failure Physics and Analysis*. Glasgow, UK, October 4-7.

Lawless, J.F.
1982 *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons.

Margolis, M.A.
1975 The CAIG: In pursuit of improved cost estimates. *Defense Management Journal* 11:20-24.

McKay, M.D.
1995 *Evaluating Prediction Uncertainty*. Technical report (NUREG/CR-6311) prepared for the Division of Systems Technology of the Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission. Los Alamos, NM: Los Alamos National Laboratory.

McKay, M.D.
1997 Nonparametric variance-based methods of assessing uncertainty importance. *Reliability Engineering and System Safety* 57:267-279.

McKay, M.D., W.J. Conover, and R.J. Beckman
1979 A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21:239-245.

Meeker, W.Q., and Escobar, L.A.
1993 A review of recent research and current issues in accelerated testing. *International Statistical Review* 6:147-161.

Meeker, W.Q., L.A. Escobar, and D.A. Hill
1992 Sample sizes for estimating the Weibull hazard function from censored samples. *IEEE Transactions on Reliability* 41(1):133-138.

Meeker, W.Q., and W. Nelson
1976 Weibull percentile estimates and confidence limits from singly censored data by maximum likelihood. *IEEE Transactions on Reliability* R-25(1):20-24.
1977 Weibull variances and confidence limits by maximum likelihood for singly censored data. *Technometrics* 19(4):473-476.

Mizuno, S., ed.
1988 *Management for Quality Improvement: The Seven New QC Tools*. Portland, Ore.: Productivity Press.

Morgan, M.G., and M. Henrion
1990 *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge, Mass.: Cambridge University Press.

Morris, C.N.
1983 *Journal of the American Statistical Association* 78(381):47-65.

Morris, M.D.
1991 Factorial sampling plans for preliminary computational experiments. *Technometrics* 33(2):161-174.

Morris, M., T. Mitchell, and D. Ylvisaker
1993 Bayesian design and analysis of computer experiments. *Technometrics* 35(3):243-255.

Nair, V.N., D.A. James, W.K. Ehrhich, and J. Zevallos
1998 Statistical issues in assessing software testing strategies. *Statistica Sinica* 8(1):165-184.

National Research Council
1995 *Statistical Methods for Testing and Evaluating Defense Systems: Interim Report*. Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics. Washington, D.C.: National Academy Press.
1996 *Statistical Software Engineering*. Panel on Statistical Methods in Software Engineering, Committee on Applied and Theoretical Statistics. Washington, D.C.: National Academy Press.

Nelson, W.
1990 *Accelerated Testing: Statistical Models, Test Plans and Data Analyses*. New York: John Wiley and Sons.

Nelson, W., and W.Q. Meeker
1978 Theory for optimum accelerated censored life tests for Weibull and extreme value distributions. *Technometrics* 20(2):171-177.

Oshana, R.
1997 Software testing with statistical usage based models. *Embedded Systems Programming* 10(1):40-55.

Owen, A.B.
1992 Randomized orthogonal arrays for computer experiments, integration, and visualization. *Statistica Sinica* 2:439-457.

Packard, D.
1971 Use of Parametric Cost Estimates. Memorandum from the Deputy Secretary of Defense for the Secretaries of Military Departments. December 7.

Pollock, S.M.
1997 A Bayesian Approach to “Dubin’s Challenge”: Optimally Selecting a Small Number of Operational Test Environments. Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor.

Poore, J.H., and C.J. Trammell
1996 *Cleanroom Software Engineering: A Reader*. Cambrdige, Mass.: Blackwell Publishers.

Proschan, F.
1963 Theoretical explanation of observed decreasing failure rate. *Technometrics* 5:375-383.

Rolph, J.E., and D.L. Steffey, eds.
1994 *Statistical Issues in Defense Analysis and Testing: Summary of a Workshop*. Committee on National Statistics and Committee on Applied and Theoretical Statistics, National Research Council. Washington, D.C.: National Academy Press.

Ross, S.M.
1996 *Simulation: Statistical Modeling and Decision Science*. San Diego, Calif.: Academic Press.

Sacks, J., W.J. Welch, T.J. Mitchell, and H.P. Wynn
1989 Design and analysis of computer experiments. *Statistical Science* 4(4):409-435.

Samaniego, F.J., and Y.S. Chong
1998 On the performance of Weibull life tests based on exponential life testing designs. In M.L. Cohen, J.E. Rolph, and D.L. Steffey, eds., *Statistics, Testing, and Defense Acquisition: Background Papers*. Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics, National Research Council. Washington, D.C.: National Academy Press.

Seglie, E.
1992 How Much Testing is Enough? Background paper prepared for The Workshop on Statistical Issues in Defense Analysis and Testing, Committee on National Statistics, National Research Council. September 24-25.

Sheedy, J., and M. Stolle
1996 Mathematical Modeling to Support Operational Test & Evaluation. Prepared for the Second Test & Evaluation International Aerospace Forum, The Royal Aeronautical Society, London, UK.

Steinberg, D., and W. Hunter
1984 Experimental design: Review and comment. *Technometrics* 26(2):71-130.

Taub, A.E., and M.A. Thomas
1983 *Confidence Intervals for CEP When the Errors are Elliptical Normal.* Dahlgren, Va.: Naval Surface Weapons Center.

U.S. Department of Defense
1960 *Handbook H108: Sampling Procedures and Tables for Life and Reliability Testing (Based on the Exponential Distribution).* New York: John Wiley and Sons.

1973 *OSD Cost Analysis Improvement Group*. DoD Directive No. 5000.4. June 13.

1981 *Military Handbook 189, Reliability Growth Management*. MIL-HDBK-189. Philadelphia: Naval Publications and Form Center.

1982 *Test and Evaluation of System Reliability, Availability, and Maintainability: A Primer*. Third edition, Report No. DoD 3235.1-H.

1992 *COMPOTEVFOR Test Plan for Project 1376-OT-IIC1*.

1993 *Longbow Apache Test and Evaluation Master Plan — Annex D: Mission Profiles and Operational Mode Summary for AH-64 Apache Longbow*. Prepared by Scenarios and Simulations Branch.

1994a *DoD Modeling and Simulation (M&S) Management*. Directive No. 5000.59. Undersecretary of Defense for Acquisition.

1994b *Reliability, Availability, and Maintainability (RAM) Rationale Report for the Longbow (AH-64D)*. Fort Rucker, Ala.: U.S. Army Aviation Center and Fort Rucker.

1995a *Family of Medium Tactical Vehicles (FMTV)*. Office of the Director, Operational Test and Evaluation, The Pentagon, Washington, DC, 18 August. (unclassified version), [Online]. Available: http://www.dote.osd.mil/.

1995b *Test and Evaluation Master Plan for the ATACMS/BAT*. U.S. Army Operational Test and Evaluation Command.

1996 *DoD 5000 Series*. Washington, D.C.: U.S. Department of Defense.

1997 *Director, Operational Test and Evaluation FY 1996 Annual Report*. (unclassified version), [Online]. Available: http://www.dote.osd.mil/.

1998a *Army Performance Improvement Criteria*. Strategic Management and Innovations Division, Management Directorate. Headquarters, Department of the Army.

1998b *Director, Operational Test and Evaluation FY 1997 Annual Report*. (unclassified version), [Online]. Available: http://www.dote.osd.mil/reports/FY97/ 97tocmain.html

U.S. General Accounting Office

1986 *Weapon Performance: Operational Test and Evaluation Can Contribute More to Decisionmaking*. Washington, D.C.: U.S. Government Printing Office.

1987a *DoD Simulations: Improved Assessment Procedures Would Increase the Credibility of Results*. Washington, D.C.: U.S. Government Printing Office.

1987b *Testing Oversight*. Washington, D.C.: U.S. Government Printing Office.

1996 *Best Practices: Commercial Quality Assurance Practices Offer Improvements for DoD*. Washington, D.C.: U.S. Government Printing Office.

1997 *Test and Evaluation: Impact of DoD's Office of the Director of Operational Test and Evaluation*. Washington, D.C.: U.S. Government Printing Office.

Vardeman, S.

1992 Response to How Much Testing is Enough? a background paper prepared for The Workshop on Statistical Issues in Defense Analysis and Testing, Committee on National Statistics, National Research Council. September 24-25.

von Winterfeldt, and W. Edwards

1986 *Decision Analysis and Behavioral Research*. Cambridge: Cambridge University Press.

Wiesenhahn, D.F., and R.D. Dighton

1993 A Framework for Using Advanced Distributed Simulation in Operational Test. Alexandria, Va.: The Institute for Defense Analyses.

Wright, S.J.

1993 Assessment of the Utility of Simulation Environments in Support of Operational Test & Evaluation. Briefing for Dr. William J. Perry, Deputy Secretary of Defense, September 30. U.S. Operational Test and Evaluation Command.

Zelen, M., and M. Dannemiller

1961 The robustness of life testing procedures derived from the exponential distribution. *Technometrics* 3:29-49.

Biographical Sketches of Panel Members and Staff

JOHN E. ROLPH *(Chair)* is professor of statistics and chair of the Department of Information and Operations Management in the University of Southern California Marshall School of Business. He previously was on the research staff of the RAND Corporation and has held faculty positions at University College London, Columbia University, the RAND Graduate School for Policy Studies, and the Health Policy Center of RAND/University of California at Los Angeles. His research interests include empirical Bayes methods and the application of statistics to the law and to public policy. He has served as editor of the American Statistical Association magazine *Chance*, and he currently is incoming chair of the National Research Council's Committee on National Statistics. He is a fellow of the American Statistical Association, the Institute of Mathematical Statistics, the American Association for the Advancement of Science and is an elected member of the International Statistical Institute. He received A.B. and Ph.D. degrees in statistics from the University of California at Berkeley.

MARION R. BRYSON is the Director of Research and Development for North Tree Management, an entrepreneurial activity in Monterey, California. He has held many positions in the federal government, 22 years primarily in the operational test arena. He served as scientific advisor at CDEC, director of CDEC, and technical director of the Test and Experimentation Command. Prior to his government service, he taught in several colleges and universities, including Duke University. He is a past president and fellow of the Military Operations Research Society. He is the recipient of the Vance Wanner Memorial Award in Military

Operations Research and the Samuel S. Wilks Award in Army Experimental Design. He holds a Ph.D. degree in statistics from Iowa State University.

HERMAN CHERNOFF is professor of statistics in the Department of Statistics at Harvard University. He previously held professorships at the Massachusetts Institute of Technology, Stanford University, and the University of Illinois at Urbana. His current research centers on applications of statistics to genetics and molecular biology, and his past work specialized in large sample theory, sequential analysis, and optimal design of experiments. He is a member of the National Academy of Sciences and the American Academy of Arts and Sciences, and has served as president of the Institute of Mathematical Statistics, and associate editor of several statistical journals. He is a fellow of the Institute of Mathematical Statistics and the American Statistical Association. He received a B.S. degree in mathematics from City College of New York, Sc. M. and Ph.D. degrees in applied mathematics from Brown University, an honorary A.M. degree from Harvard University, and honorary Sc.D. degrees from the Ohio State University and the Technion.

JOHN D. CHRISTIE is a senior fellow and assistant to the president at the Logistics Management Institute, a nonprofit institution in McLean, Virginia. Before joining the institute he was the Director, Acquisition Policy & Program Integration for the Undersecretary of Defense (Acquisition) in the U.S. Department of Defense. Prior to that he was vice president of two different professional service firms, while also serving for 7 years as a member of the Army Science Board. During an earlier period of government service he held various positions at the Federal Energy Administration and the Defense Department. Previously, he was a member of the Bell Labs technical staff. He holds S.B., S.M., E.M.E., and Sc.D. degrees from the Massachusetts Institute of Technology, all in mechanical engineering.

MICHAEL L. COHEN is a senior program officer for the Committee on National Statistics. Previously, he was a mathematical statistician at the Energy Information Administration, an assistant professor in the School of Public Affairs at the University of Maryland, a research associate at the Committee on National Statistics, and a visiting lecturer at the Department of Statistics, Princeton University. His general area of research is the use of statistics in public policy, with particular interest in census undercount and model validation. He is also interested in robust estimation. He received a B.S. degree in mathematics from the University of Michigan and M.S. and Ph.D. degrees in statistics from Stanford University.

ANURADHA DAS served as research assistant with the Committee on National Statistics, National Research Council. In addition to the Panel on Statistical

Methods for Testing and Evaluating Defense Systems, she concurrently worked on projects related to integrated environmental and economic accounting and longitudinal research on children. She previously worked on studies related to health, aging, disability, and the census. She holds a bachelor's degree in mathematics and a master's degree in computer and information science.

CANDICE S. EVANS is a senior project assistant with the Committee on National Statistics. She is also currently working with the Panel on Estimates of Poverty for Small Geographic Areas. Previously she has worked with the Panel on Retirement Income Modeling and the Panel on International Capital Transactions.

ERIC M. GAIER is a Research Fellow with Logistics Management Institute of McLean, Va. Currently, his research at the Institute focuses on the economic impact of implementing advanced NASA technologies in the integrated aviation community. Previous research has focused on the role of information in principal agent relationships. He holds a B.S. in Economics from Florida State University and a Ph.D. in Economics from Duke University.

LOUIS GORDON is a consultant in Palo Alto, California. Previously, he was a statistician at the Filoli Information Systems Corporation. He has held academic appointments at the University of Southern California and at Stanford University. He has also worked as a statistician in industry and in the federal government. He has held J.S. Guggenheim and Fulbright fellowships. His research interests are in nonparametric statistics.

KATHRYN BLACKMOND LASKEY is an associate professor of systems engineering at George Mason University. She was previously a principal scientist at Decision Science Consortium, Inc. Her primary research interest is the study of decision theoretically based knowledge representation and inference strategies for automated reasoning under uncertainty. She has worked on methods for automated construction of Bayesian belief networks and for recognizing when a system's current problem model is inadequate. She has worked with domain experts to develop Bayesian belief network models to be used in automated reasoning. She received a B.S. degree in mathematics from the University of Pittsburgh, an M.S. degree in mathematics from the University of Michigan, and a joint Ph.D. in statistics and public affairs from Carnegie Mellon University.

ROBERT C. MARSHALL is a professor and head of the Department of Economics at Penn State University. Previously, he taught at Duke University. His research—using theoretical, empirical, and numerical methods of analysis—has included a broad range of topics—housing, labor, the expected utility paradigm, and measurements of mobility. He is best known for his work on auctions and procurements, which has focused on collusion by bidders. He received an A.B. degree from Princeton University and a Ph.D. degree from the University of California at San Diego.

VIJAYAN N. NAIR is professor of statistics and professor of industrial and operations engineering and incoming chair of the Department of Statistics at the University of Michigan, Ann Arbor. Previously, he was a research scientist at Bell Laboratories. His research interests include statistics in advanced manufacturing, quality and reliability improvement, design and analysis of industrial experiments, and process control. He is a fellow of the American Statistical Association and of the Institute for Mathematical Statistics, an elected member of the International Statistical Institute, and a senior member of the American Society for Quality. He is currently the joint editor of the International Statistical Review, is a past editor of *Technometrics* and has served on the editorial boards of several other statistical journals. He has a B. Econs (Hons.) degree from the University of Malaya and a Ph. D. in statistics from the University of California, Berkeley.

ROBERT T. O'NEILL is director of the Office of Epidemiology and Biostatistics and Acting Director of the Division of Epidemiology and Surveillance in the Center for Drug Evaluation and Research (CDER) of the Food and Drug Administration. He is responsible for postmarketing surveillance and safety of new drugs, and for providing statistical support to all programs of CDER, which include advice in all drug/disease areas on the design, analysis, and evaluation of clinical trials performed by sponsors seeking approval to market new drugs. He is a fellow of the American Statistical Association, and a former member of the board of directors of the Society for Clinical Trials and he is active in several professional societies. He received a B.A. degree from the College of the Holy Cross and a Ph.D. degree in mathematical statistics and biometry from Catholic University of America.

STEPHEN M. POLLOCK is professor of industrial and operations engineering at the University of Michigan, Ann Arbor. Previously, he served as a consultant at Arthur D. Little, Inc., and a member of the faculty at the Naval Postgraduate School. He teaches courses in stochastic processes, decision analysis, and reliability and mathematical modeling and has engaged in a variety of research areas and methods, including search theory, sequential detection of change, queuing systems, criminal recidivism, police patrol, and filling processes. He also serves as a consultant to more than 30 companies and other organizations. He is a fellow of the American Association for the Advancement of Science, and has been senior editor of *IIE Transactions*, area editor of *Operations Research*, and president of the Operations Research Society of America. He holds a B. Eng. Phys. from Cornell and S.M. and Ph.D. degrees in physics and operations research from the Massachusetts Institute of Technology.

JESSE H. POORE is professor of computer science at the University of Tennessee and president of Software Engineering Technology, Inc. He conducts research in cleanroom software engineering and teaches software engineering courses. He has held academic appointments at Florida State University and Georgia Tech; has served as a National Science Foundation rotator, worked in the Executive Office of the President, and was executive director of the Committee on Science and Technology in the U.S. House of Representatives. He is a member of ACM, IEEE and a fellow of the American Association for the Advancement of Science. He holds a Ph.D. in information and computer science from Georgia Tech.

FRANCISCO J. SAMANIEGO is professor in the Intercollege Division of Statistics and Director of the Teaching Resources Center at the University of California at Davis. He has held visiting appointments in the Department of Statistics at Florida State University and in the Department of Biostatistics at the University of Washington. His research interests include mathematical statistics, decision theory, reliability theory and survival analysis, and statistical applications, primarily in the fields of education, engineering and public health. He is a fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the Royal Statistical Society and is a member of the International Statistical Institute. He received a B.S. degree from Loyola University of Los Angeles, an M.S. degree from Ohio State University, and a Ph.D. from the University of California, Los Angeles, all in mathematics.

DENNIS E. SMALLWOOD is the Roger's Professor in the Department of Social Sciences at the United States Military Academy. Previously he was a senior economist with RAND where he conducted research related to national security, including defense acquisition, industrial base and costing issues. He has held previous positions at the Pentagon, working on strategic arms control issues; he also served as head of the Economic Analysis and Resource Planning Division, Assistant Secretary of Defense for Program Analysis and Evaluation. He was previously an associate professor of economics at the University of California at San Diego where he worked on issues related to the economics of health and of law. He received B.A. and M.A. degrees in mathematics from the University of Michigan and a Ph.D. degree in economics from Yale University.

DUANE L. STEFFEY is senior program officer with the Committee on National Statistics, and he served as the panel study director until July 1995. Concurrently, he is associate professor of mathematical sciences at San Diego State

University, where he teaches courses in Bayesian statistics, statistical computing, and categorical data analysis. He previously worked at Westinghouse and was involved in conducting probabilistic risk assessment of commercial nuclear energy facilities. He engages broadly in interdisciplinary research and consulting, and current professional interests include applications of statistics in environmental monitoring, transportation demand modeling, and census methodology. He received a B.S. degree and M.S. and Ph.D. degrees in statistics, all from Carnegie Mellon University.