Testing of Defense Systems in an Evolutionary Acquisition Environment

Oversight Committee for the Workshop on Testing for Dynamic Acquisition of Defense Systems, National Research Council

ISBN: 0-309-65817-9, 76 pages, 6 x 9, (2006)

Testing of Defense Systems in an Evolutionary Acquisition Environment

Oversight Committee for the Workshop on Testing for Dynamic Acquisition of Defense Systems

Vijay Nair and Michael L. Cohen, Editors

Committee on National Statistics

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

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.

This study is supported by a contract between the National Academy of Sciences and the National Science Foundation (Number SBR-0112521). The work of the Committee on National Statistics is also provided by a consortium of federal agencies through the same grant from the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-10135-2

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America.

Suggested citation: National Research Council. (2006). *Testing of Defense Systems in an Evolutionary Acquisition Environment.* Oversight Committee for the Workshop on Testing for Dynamic Acquisition of Defense Systems. Vijay Nair and Michael L. Cohen, editors. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

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. Ralph J. Cicerone 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. Wm. 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. Harvey V. Fineberg 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. Ralph J. Cicerone and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

OVERSIGHT COMMITTEE FOR THE WORKSHOP ON TESTING FOR DYNAMIC ACQUISITION OF DEFENSE SYSTEMS

VIJAY NAIR *(Chair),* Department of Statistics and Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor
SETH BONDER, The Bonder Group, Ann Arbor, MI
JOHN D. CHRISTIE, Logistics Management Institute, Alexandria, VA
ARTHUR FRIES, Institute for Defense Analyses, Alexandria, VA
STEPHEN POLLOCK, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor
JESSE H. POORE, Department of Computer Science, University of Tennessee, Knoxville

MICHAEL L. COHEN, *Study Director*
MICHAEL SIRI, *Senior Program Assistant*

Preface

The Oversight Committee for the Workshop on Testing for Dynamic Acquisition of Defense Systems, formed under the Committee on National Statistics, is grateful to the many individuals that contributed to our work. First, we are greatly indebted to both Thomas Christie, former director, operation test and evaluation, and to Michael Wynne, secretary of the Air Force and former acting under secretary of defense (acquisition, technology and logistics) for providing the financial support for this study. The panel is also greatly indebted to Ernest Seglie, science advisor to the director of operational test and evaluation, and to Nancy Spruill, director, acquisition resources and analysis, for providing continuous assistance both intellectually and administratively that was invaluable for the success of our work. In addition, Bill Keegan and Bob Williams in the Office of the Secretary of Defense were very helpful in assisting in the framing of our work.

All of the invited speakers to the workshop worked hard and provided very important information on the history of evolutionary acquisition, on best practices in industrial system development, and on the defense system milestone process. The speakers are listed in Appendix D. Of these, we would like to thank in particular Thomas Christie; Philip Coyle, Center for Defense Information; Tim Davis, Ford Motor Company; Donald Gaver, Naval Postgraduate School; David Kelly, Battelle; Art Koehler, Procter & Gamble; Katherine Schinasi, U.S. Government Accountability Office; Steve Vardeman, Iowa State University; Alyson Wilson, Los Alamos National Laboratory; and Michael Wynne.

The participants were also very helpful in generating extremely productive floor discussions. In particular, we would like to thank David Maddox for extremely insightful remarks.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council. The purpose of this independent review is to provide candid and critical comments that will assist the institution 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 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: Donald A. Berry, Department of Biostatistics and Applied Mathematics, The University of Texas, Houston; Paul K. Davis, Defense Strategy, RAND Corporation; Timothy P. Davis, Henry Ford Technical Fellow for Quality Engineering, Ford Motor Company; Michael C. Dieckhoff, Air Force Operational Test and Evaluation Center, Albuquerque; Robert Easterling, Consultant, New Mexico; David M. Maddox, Consultant, Arlington, Virginia; Ronald Manning, Raytheon, Dallas; Daryl Pregibon, Google, Inc., New York; and Stephen B. Vardeman, Statistics and IMSE Departments, Iowa State University.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations nor did they see the final draft of the report before its release. The review of this report was overseen by Thom J. Hodgson, Industrial Engineering Department, North Carolina State University. Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

In the preparation of the report, we would like to thank Michael Siri, for making all the arrangements, formatting the final report, and overall stewardship of the administrative side of the study. Christine McShane expertly served as technical editor, substantially improving the report's readability. Connie Citro, director of the Committee on National Statistics, and Tom Plewes, staff officer for the committee, were helpful as usual in providing guidance and support whenever needed.

Finally, we would like to thank the oversight committee for their collegiality and readiness to help in all phases of the study. It was a great pleasure working together to produce this report.

Vijay Nair, *Chair*
Michael L. Cohen, *Study Director*
Oversight Committee for the Workshop on Testing
for Dynamic Acquisition of Defense Systems

Contents

Executive Summary

Evolutionary acquisition is a Department of Defense (DoD) process for defense system development in which a system is developed in stages as part of a single acquisition program. The different stages can be additional hardware and software capabilities or performance gains due to advances in technological maturity and reliability growth.

DoD has presented evolutionary acquisition as the preferred option for development of key complex defense systems, a leading example being the Army's Future Combat System. While it is quite common to modify a defense system after fielding, the intention in evolutionary acquisition is that these improvements are planned for and accommodated by the choice of system architectures and overall system design, to the extent possible. A further underlying motivation is that some costs and development delays (e.g., due to redesign and retrofitting) that might arise from a single-stage development process could be reduced by (a) giving greater priority to the identification of failure modes early in system development, (b) introducing new technologies only when they are mature, and (c) limiting the introduction of too many new components or subsystems simultaneously. It is argued that this process will shorten the overall system development time, allow rapid insertion of new capability-enhancing technologies, and reduce life-cycle costs.

In evolutionary acquisition, system capabilities are developed and acquired in stages. Hence there is a need for careful reexamination of current testing and evaluation policies and processes, which were designed for

single-stage developments. At the request of DoD, a committee of the National Academies planned and conducted a workshop to discuss the role of testing and evaluation in an evolutionary acquisition environment and to make appropriate conclusions and recommendations.

The specific questions addressed include: What are the appropriate roles and objectives for testing in an evolutionary environment? Can a systematic, disciplined process be developed for testing and evaluation in such a fluid and flexible environment? How can information from the earlier stages of the evolutionary acquisition process be used effectively in developing test designs for subsequent stages? Are there methodologies, either in the academic literature (statistics, operations research, management science, etc.) or best practices in industry that can be adapted for use in the evolutionary acquisition environment in DoD? Are there advantages to data archiving and documenting results from past stages of development? Is there adequate technical expertise within the acquisition community to fully exploit data gathered from previous stages and to effectively combine information from various sources for test design and analysis?

While discussing these questions, it became apparent that there are several broader, contextual issues that must also be addressed if the recommendations on test design are to be effective. Among these issues, the following were considered in the report: Is the meaning and intent of evolutionary acquisition sufficiently clear in DoD, or is there a need for clarity and consistency in the terminology and a need for enforcement of policies and procedures? Can the culture and organization of defense test and acquisition fully support the effective implementation of evolutionary acquisition? If not, what changes are needed in the DoD environment, the acquisition process, and incentives to ensure that the full benefits of testing in the evolutionary environment can be realized? Is the current level of cooperation among the program manager, contractors, and the developmental and operational testing communities adequate for supporting evolutionary acquisition?

While these broader issues are somewhat beyond the study's original scope, the committee concluded that they must be discussed, even if only briefly. The committee draws some conclusions and makes some recommendations on these broader issues. However, the committee could not recommend how to address these problems fully due to the limited scope of this study.

CONCLUSIONS AND RECOMMENDATIONS

Operational testing and evaluation, as portrayed in the current milestone system, supports a decision to pass or fail a defense system before it goes to large-scale procurement. The 1998 National Research Council report *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements* proposed a new paradigm in which testing should be viewed as a "continuous process of information gathering and decision making in which operational testing and evaluation plays an integral role." This new paradigm, originally suggested in the context of the traditional single-stage testing, is even more important in the evolutionary acquisition environment.

Conclusion 1: In evolutionary acquisition, the entire spectrum of testing activities should be viewed as a continuous process of gathering, analyzing, and combining information in order to make effective decisions. The primary goal of test programs should be to experiment, learn about the strengths and weaknesses of newly added capabilities or (sub)systems, and use the results to improve overall system performance. Furthermore, data from previous stages of development, including field data, should be used in design, development, and testing at future stages. Operational testing (testing for verification) of systems still has an important role to play in the evolutionary environment, although it may not be realistic to carry out operational testing comprehensively at each stage of the development process.

Recommendation 1: The under secretary of defense (acquisition, technology and logistics) and the director of operational test and evaluation should revise DoD documents and processes (e.g., DoD Directive 5000.1 and DoD Instruction 5000.2) to explicitly recognize and accommodate a framework in which the primary goal of all acquisition testing and evaluation programs is to experiment, learn about the strengths and weaknesses of system components, and to incorporate these results into system enhancement initiatives.

Under such a continuous learning process, testing activities must go beyond the traditional framework of focusing on estimating the performance of a system under typical test scenarios.

Conclusion 2: Testing early in the development stage should emphasize the detection of design inadequacies and failure modes. This will require testing in more extreme conditions than those typically required by either developmental or operational testing, such as highly accelerated stress environments.

However, the current incentive structure in DoD may discourage testing outside the envelope or identifying limitations and failure modes early in the process.

Conclusion 3: To have a reasonable likelihood of fully implementing the paradigm of testing to learn about and to improve systems prior to production and deployment, the roles of DoD and congressional oversight in the incentive system in defense acquisition and testing must be modified. In particular, incentives need to be put in place to support the process of learning and discovery of design inadequacies and failure modes early and throughout system development.

In evolutionary acquisition, it will be practical to conduct full-scale operational tests only at stages with major upgrades or substantive new capabilities. At other stages, only developmental tests of components and subsystems and some limited tests on functionality, interoperability, etc., will be feasible. Thus, these tests should use operational realism to the extent needed to assess the performance of components and subsystems from an operational perspective.

Recommendation 2: The under secretary of defense (acquisition, technology and logistics) and the director of operational test and evaluation should revise DoD testing procedures to explicitly require that developmental tests have an operational perspective (i.e., are representative of real-world usage conditions) in order to increase the likelihood of early identification of operational failure modes and system deficiencies, so that appropriate actions can be developed and deployed in a timely fashion.

Conclusion 4: In the evolutionary acquisition environment, effective system development and optimization will require a high degree of coordination and communication among system developers, government testers, and system users. In particular, government testers should

have early access to all contractor data sources, including test plans, results of early stage testing and experimentation, and the results of all pertinent modeling and simulation products.

Recommendation 3: The under secretary of defense (acquisition, technology and logistics) should develop and implement policies, procedures, and rules that require contractors to share all relevant data on system performance and the results of modeling and simulation developed under government contracts, including information on their validity, to assist in system evaluation and development.

The traditional single-stage acquisition environment can encourage the adoption of risky, immature technology into an existing system, since it may take a decade or more before a new technology can be incorporated. Evolutionary acquisition provides opportunities to discipline this process by delaying the introduction of risky, immature technology to future stages, and by using advanced technology demonstrations or advanced concept technology demonstrations to further develop and evaluate immature technologies. This will eliminate the possibility of delay of the entire acquisition program due to the use of a single risky technology, or risk using technology that has not been demonstrated to be sufficiently effective or reliable.

Recommendation 4: The under secretary of defense (acquisition, technology and logistics) should require that all technologies to be included in a formal acquisition program have demonstrated sufficient technological maturity before the acquisition program is approved or before the technology is inserted in a later stage of development. The decision about the sufficiency of technological maturity should be based on an independent assessment from the director of defense research and engineering or special reviews by the director of operational test and evaluation (or other designated individuals) of the technological maturity assessments made during the analysis of alternatives and during developmental testing and evaluation.

Regardless of the introduction of evolutionary acquisition, the increasing complexity of defense systems implies that a single all-encompassing,

large-scale operational test, as currently practiced, will not be feasible in many cases. Given the sheer numbers of subsystems, components, materials, software, and resulting system interactions, it is unlikely that the operational evaluation of complex systems can be based primarily on an operational test representing the full spectrum of combat conditions.

Conclusion 5: The DoD testing community should investigate alternative strategies for testing complex defense systems to gain, early in the development process, an understanding of their potential operational failure modes, limitations, and level of performance.

There are increased opportunities for combining information to improve test design and analysis in an evolutionary acquisition environment, given data from developmental, operational, and field tests of the system from earlier, fielded versions. Chapter 3 and Appendix B discuss specific examples and methods for combining information for improved test design in a multistage development process that is inherent in evolutionary acquisition. However, the gains from the use of these methods can be assessed only by applying them to defense systems in development. Furthermore, combining information from the various sources, which rely on relevant linkages between previous and current sources of data, requires considerable care and subject-matter expertise. When done correctly, however, it has a tremendous payoff.

Recommendation 5: The Service test agencies should undertake a pilot study, involving a few selected systems developed under the evolutionary acquisition paradigm, in order to identify specific opportunities for incorporating information from previous stages to improve system design and analysis. These case studies will be beneficial in demonstrating both the application of the various techniques and the benefits to be gained from combining data in staged development.

In order to effectively implement evolutionary acquisition, the DoD testing community must have access to greater expertise in various areas. There is also a need for systematically archiving data from various sources and combining them effectively for improved test design.

Recommendation 6:

(a) To support the implementation of evolutionary acquisition, DoD should acquire, either through hiring in-house or through consulting or contractual agreements, greater access to expertise in the following areas: (1) combining information from various sources for efficient multistage design, statistical modeling, and analysis; (2) software engineering; and (3) physics-based and operational-level modeling and simulation.

(b) Test and field data archives should be established to facilitate access to data on test and field performance of systems in development and those fielded. These data archives should be used to support feedback loops to improve the system design, to improve testing methodology over time, and to help validate (and improve) modeling and simulation for operational evaluation.

While evolutionary acquisition is formally defined in DoD Instruction 5000.2, the term has not been used in a consistent manner by DoD leadership or in various other DoD documents. The variety of terms, including *evolutionary acquisition, incremental development*, and *spiral development*, and their inconsistent usage have served to obscure the original goal and intent of the evolutionary acquisition process. Besides clear and consistent definitions, a consistent and definitive articulation of the goals and intention of the evolutionary acquisition process is needed, as well as supporting documentation detailing the specific changes that will need to be implemented throughout the acquisition community.

Recommendation 7: The under secretary of defense (acquisition, technology and logistics) should eliminate inconsistencies in DoD Directive 5000.1 and DoD Instruction 5000.2 and clarify other significant memoranda and documents regarding evolutionary acquisition. All policies and procedures to be used in applying evolutionary acquisition principles to DoD acquisition programs should be strictly enforced. This clarification and enforcement should be applied to program management both in DoD and in all supporting contractor activities.

Evolutionary acquisition is being folded into an acquisition environment that already has a counterproductive incentive system. The flexibilities

inherent in the evolutionary acquisition process present even greater opportunities for these counterproductive incentives to be expressed. Improving the incentive structure is perhaps the most onerous obstacle facing the DoD acquisition process if it is to meet the difficult challenges in acquiring and fielding complex, expensive systems in a timely, efficient manner.

Recommendation 8: The deputy secretary of defense should charge a blue ribbon panel, including experts in organizational behavior, multiobjective decision making, and other relevant areas, to review the Defense Acquisition Performance Assessment panel's proposed changes to DoD acquisition policies and procedures. This review should take place (if possible) before any changes to the policies are implemented in order to assess and improve their likelihood of being successfully implemented in the DoD and defense industry culture.

1

Introduction

EVOLUTIONARY ACQUISITION: DEFINITION AND RATIONALE

The Department of Defense (DoD) recently adopted *evolutionary acquisition*, a dynamic strategy for the development and acquisition of its defense systems. While the term has been used in various ways in DoD documents, the following statement is consistent with its description in the official DoD Instruction 5000.2:

> Evolutionary defense systems are those that are planned, in advance, to be developed in several stages through a single procurement program. Each stage is planned to produce a viable system which, if shown to be superior to existing capabilities, could be fielded. The system requirements for each stage of development may be specified in advance of a given stage or may be decided at the outset of that stage's development.

According to E.C. Aldridge, Jr. (2002), "[evolutionary acquisition] will allow us to reduce our cycle time and speed the delivery of advanced capability to our war fighters." The term "evolutionary acquisition" is used somewhat inconsistently within DoD, and the resulting confusion has served to obscure the original goal and intent of the process. The implica-

tions of this issue are discussed in Chapter 4. In this report, we use only the terms "evolutionary acquisition" and "staged acquisition."

Reducing acquisition time and costs is a problem of great importance to DoD, as it often takes a decade or longer to complete the development and delivery of new major military systems. Even a small reduction in the costs of an ACAT I (acquisition category I) system provides significant savings in program costs. One reason given for delay in the development of complex defense systems is the phenomenon of "excessive requirements." These can be overly optimistic initial expectations of what can be achieved, or they can be add-ons to the initial requirements, revisions of specifications, or changes in designs that occur during development and production. While most of these requirements are well intentioned, the net result is often significant slippage in development schedules (thereby delaying delivery of the capability to users in the battlefield) and substantial increases in procurement costs.

Evolutionary acquisition has been advocated as a way to address these problems. The stages of evolutionary acquisition may correspond to new hardware and software capabilities, increased performance, or to system improvements traced to technological maturity and reliability growth advances. If the acquisition program actually encompasses a "system of systems," then the stages can provide for an increase in the capabilities or performance levels of individual systems, or they can represent new systems that are incrementally added to the system of systems. With evolutionary acquisition, initial and intermediate versions of the system will be released to the field if the decision is made that the system at that stage is superior to those currently in use. For this process to work, however, the initial system architecture must be designed so as to allow for the incorporation of changes dictated by the capabilities or improvements at later stages.

The evolutionary acquisition framework ideally allows for a fielded system to undergo further improvement without initiating a new procurement program. Thus, new technologies (particularly software) can be quickly deployed to enhance capabilities in the field, providing flexibility in responding to changing military missions, threats, and operating environments. By comparison, the traditional acquisition process, with its long lead times, often results in buying systems containing technologies that are outdated when finally fielded.

The idea of modifying fielded defense systems to improve their performance is, of course, not new. The Services have been developing substantial upgrades to existing defense systems for many years. However, many of

these improvements were not planned for in advance in traditional single-stage development programs. As a result, the major block upgrades often have taken substantial amounts of time, redesign work, and cost. In evolutionary acquisition, in contrast, the planning for such major system improvements is supposed to be undertaken in advance to minimize the costs of redesign and retrofitting and to expedite the delivery to the field.

GOALS OF THE REPORT

The Office of the Under Secretary of Defense for Acquisition, Technology and Logistics (USD-AT&L) and the Director of Operational Testing and Evaluation (DOT&E) asked the Committee on National Statistics (CNSTAT) of the National Academies to examine the key issues and implications for defense testing from the introduction of evolutionary acquisition. The specific charge was as follows: "This study will involve the planning and conduct of a workshop to study test strategies for evolutionary acquisition. The study committee will review defense materials defining evolutionary acquisition and will interview test officials from the three major test service agencies to better understand the current approaches used to test systems procured using evolutionary acquisition. Possible alternatives will be examined to identify problems in implementation."

CNSTAT set up an oversight committee to plan and conduct the workshop and write a report based on the presentations and discussions during the workshop and subsequent deliberations among committee members. The members were selected for their expertise in defense acquisition, software system development, industrial system development, operations research, and experimental design. Several teleconferences and meetings at the Pentagon assisted in the planning of the workshop. The committee also examined information in relevant defense documents, the academic literature, reports of the Government Accountability Office (see, e.g., U.S. General Accounting Office, 2002; U.S. Government Accountability Office, 2004), and commercial industrial processes.

The workshop took place on December 13-14, 2004. Participants represented a broad range of expertise from the DoD community, industry, and academia. The workshop agenda and participants are presented in Appendix D. The workshop was organized to learn about the evolutionary acquisition process itself, to examine the role of testing in an evolutionary environment, and to identify relevant best practices in industry and state-of-the-art methodology in the research literature.

Some of the specific questions were: What are the appropriate roles and objectives for testing in an evolutionary environment? Can a systematic, disciplined process be developed for testing and evaluation in such a fluid and flexible environment? How can information from the earlier stages of the evolutionary acquisition process be used effectively in developing test designs for subsequent stages? Are there methodologies, either in the academic literature (statistics, operations research, management science, etc.) or best practices in industry that can be adapted for use in the evolutionary acquisition environment in DoD? Are there advantages to data archiving and documenting results from past stages of development? Is there adequate technical expertise within the acquisition community to fully exploit data gathered from previous stages and to effectively combine information from various sources for test design and analysis?

In the discussion of these questions, it became apparent that there are several broader, contextual issues that must also be addressed if the recommendations on test design are to be effective. Among these broader questions, the committee decided to focus on the following in the report: Is the meaning and intent of evolutionary acquisitions sufficiently clear within DoD, or is there a need for clarity and consistency in the terminology and enforcement of policies and procedures? Can the culture and organization of defense test and acquisition fully support the effective implementation of evolutionary acquisition? If not, what changes are needed in the DoD environment, the acquisition process, and incentives to ensure that the full benefits of testing in the evolutionary environment can be realized? Is the current level of cooperation among the program manager, contractors, and the developmental and operational testing communities adequate for supporting evolutionary acquisition?

Some of these questions raise issues beyond this study's original scope or would have required more resources and expertise than were available. Nevertheless, we have addressed them, since they are critical to carrying out our overall charge. We make a number of general recommendations. Many of these will require substantial changes in the way program managers, test officials, and contractors carry out their responsibilities—changes that will be difficult to implement and institutionalize. Although developing the *details* is beyond the charge and the expertise of the committee, we stress that this organizational redesign will be critical to achieving the potential benefits of the evolutionary acquisition process and therefore requires the serious attention of DoD's top management and technical leadership.

ORGANIZATION OF THE REPORT

The rest of the report is organized as follows. Following this introduction, Chapter 2 addresses the role of testing and evaluation, broadly speaking, in contributing to evolutionary system development. Chapter 3 examines efficient strategies for test design and analysis by combining information in the context of staged system development. Chapter 4 addresses changes in the process and culture and infrastructure changes that are needed to support a more effective implementation of evolutionary acquisition, and in particular the methods proposed in Chapters 2 and 3.

Appendix A is an overview of the current single-stage milestone development process used in DoD. Appendix B provides more technical details in support of Chapter 3. Appendix C examines the special case of software development methods in the context of evolutionary acquisition. Appendix D presents the workshop agenda and the attendees. Appendix E presents biographical sketches of committee members and staff. Box 1-1 is a list of acronyms and abbreviations related to the topics of this report.

Box 1-1
Acronyms and Abbreviations

ACAT I	acquisition category I
ACTDs	advanced concept technology demonstrations
AoA	analysis of alternatives
APBs	acquisition program baselines
ATDs	advanced technology demonstrations
CNA	Center for Naval Analyses
CNSTAT	Committee on National Statistics
DoD	U.S. Department of Defense
DOT&E	Office of the Director of Operational Testing and Evaluation
FOC	full operating capability
FRP	full rate production
IOC	initial operating conditions
IOT&E	initial operational test and evaluation
LRIP	low rate initial production
OSD	Office of the Secretary of Defense
POS	parametric operational situation
RDT&E	research, development, test, and evaluation
USD-AT&L	Under Secretary of Defense for Acquisition, Technology and Logistics

2

Testing and Evaluation in an Evolutionary Acquisition Environment

CONTINUOUS PROCESS OF TESTING, LEARNING, AND IMPROVING SYSTEMS

Operational testing and evaluation, as portrayed in the current milestone system, supports a decision to pass or fail a defense system before it goes to full-scale procurement. However, the U.S. Department of Defense (DoD) and the Services have not been following this consistently, as evidenced by the fact that critical systems have rarely been terminated solely on the basis of testing.

A 1998 report by a National Research Council panel considered many of the themes that are of interest in this report. *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements* (National Research Council, 1998:35) proposed a new paradigm in which testing should be viewed as a "continuous process of information gathering and decision making in which operational testing and evaluation plays an integral role." This new paradigm, suggested in the context of the traditional single-stage development, is even more important in the evolutionary acquisition environment.

With staged development, if a system has gone into full-scale production in Stage I, then a goal in subsequent stages typically will be to identify ways to efficiently incorporate additional capabilities. Furthermore, earlier versions of a fielded system should provide very useful data on strengths and weakness prior to the next and future stages of system development.

Thus, it is even more critical to view testing as a process of experimentation: one that involves continuous data collection and assessment, learning about the strengths and weaknesses of newly added capabilities or (sub)systems, and using all of this information to determine how to improve the overall performance of the system. This should not be viewed as an activity to be carried out solely by contractors near the initiation of system development or by DoD near the end; instead, it should become an intrinsic part of system development with facilitated communication and interaction between the contractor and government testers throughout the developmental process.

In the evolutionary acquisition context, experimentation in early stages can be used to identify system flaws and to understand the limitations of system design. The focus in later stages should be on problems identified in the field and/or unresolved from earlier testing, evaluating the most recent modifications to the system, and assessing the maturity of a new component or subsystem design. This experimentation can be at the component level, at the subsystem level, or at the system level, with varying degrees of operational realism, depending on the goals.

Operational testing and evaluation (or testing for verification) will still have a major role to play, since it is the only way to verify that the systems are in fact operationally effective and suitable. In fact, it is critical that there is adequate oversight and accountability in this flexible environment. However, it is not realistic to undertake comprehensive operational tests at each stage of the development process. These should be undertaken only at stages encompassing major upgrades, the introduction of new, major capabilities, or new major (sub)systems. At other stages, a combination of data and insights from component or subsystem testing and developmental tests (reflecting operational realism as feasible) can be used instead, along with engineering and operational user judgment.

Conclusion 1: In evolutionary acquisition, the entire spectrum of testing activities should be viewed as a continuous process of gathering and analyzing data and combining information in order to make effective decisions. The primary goal of test programs should be to experiment, learn about the strengths and weaknesses of newly added capabilities or (sub)systems, and use the results to improve overall system performance. Furthermore, data from previous stages of development, including field data, should be used in design, development, and testing at future stages. Operational testing (testing for verification) of

systems still has an important role to play in the evolutionary environment, although it may not be realistic to carry out operational testing comprehensively at each stage of the development process.

The use of the term "testing" to describe this continuous process is unfortunate, and we do not expect that it will be changed. Nevertheless, it is important for DoD to recognize explicitly that the primary goal of test and evaluation programs in evolutionary acquisition will be to experiment, learn, and use the results to improve system performance.

Recommendation 1: The under secretary of defense (acquisition, technology and logistics) and the director of operational test and evaluation should revise DoD documents and processes (e.g., DoD Directive 5000.1 and DoD Instruction 5000.2) to explicitly recognize and accommodate a framework in which the primary goal of all acquisition testing and evaluation programs is to experiment, learn about the strengths and weaknesses of system components, and to incorporate these results into system enhancement initiatives.

TESTING OUTSIDE THE ENVELOPE

In a learning environment, testing should go beyond estimating typical performance of a new or modified system to also emphasize the early discovery of failure modes,[1] deficiencies, and weakness in the system design. This will require adding more operational realism to developmental tests as well as testing with demanding and possibly even accelerated stress environments. This is consistent with the practices of some companies (for example, Ford Motor Company), in which reliability improvement is based primarily on failure-mode avoidance, that is, finding and eliminating failure modes early in development rather than merely demonstrating compliance with prescribed reliability metrics.

The General Accounting Office, on the basis of a study of industrial best practices, advocates a knowledge-based approach to system develop-

[1]Failure modes are present and past circumstances of system use that result in the system either failing or not performing satisfactorily.

ment (see U.S. General Accounting Office, 2004). Katherine Schinasi described this approach in a presentation at the workshop. Failures in early-stage testing are viewed as rich sources of information about design inadequacies. The objective is to "break it big early," rather than to utilize test events that concentrate on performance under typical stresses. This is consistent with suggestions in *Statistics, Testing, and Defense Acquisition* (National Research Council, 1998) to test outside the envelope and to include characteristics of operational realism early in developmental testing. This will assist in discovering both reliability failure modes and deficiencies in operational use as early as possible—when it is least costly to explore redesign and improvement options, and to implement specific modifications and enhancements (these are sometimes referred to as "countermeasures" in industry, although the term has a different meaning in DoD).[2]

This early stage experimentation is particularly important for the development of reliable defense systems. Tom Christie's introduction to the FY 2004 Director of Operational Test and Evaluation's annual report states:

> The Defense Science Board in 2000 pointed out that 80 percent of recent U.S. Army defense systems brought to operational test failed to achieve even half of their reliability requirement. This was followed later by data showing that with all the streamlining of the acquisition process, the number of systems failing to meet the reliability requirement had increased. As stated earlier, this trend is evident in the reports DOT&E sends to Congress.

The experience of the acquisition community is that system reliability in operational tests is generally substantially worse than that in developmental tests. For example, operational test mean times to first failure typically are a factor of roughly two to four times shorter than those for developmental tests. Clearly, developmental testing gives an incomplete assessment of operational system quality. In addition, if failure modes iden-

[2]In many industrial system development programs, when countermeasures are identified to address system shortcomings, a verification test is often carried out. This is generally not feasible in the defense environment, given the cost of replications in operational test. The system resulting from a design change should be viewed as another stage in the development of an evolutionary system, with the type of testing used dictated by the nature of the design change.

tified in operational tests (or when fielded) had been discovered earlier in the development process, they would have been much less expensive to fix and would have contributed to more rapid system maturation. This emphasizes the need to test defense systems more stressfully and as early as possible.

We think that for test purposes, "edge of the envelope" can be defined fairly rigorously. The space of conceivable military scenarios for operational testing includes a number of uncontrollable dimensions (e.g., environmental characteristics, potential missions, threat objectives and characteristics, etc.), and these dimensions can be usefully parameterized to identify the edge of the envelope. For example, in Bonder (2000), with the parametric operational situation (POS) space formulation,[3] each point in this space represents an operational situation that U.S. forces might have to be deployed to and operate in. Some of these situations are more stressful than others. Operational testing should be performed in the most stressful situations in which U.S. forces and/or system performance can barely be successful (i.e., the edge of the envelope). This will provide information about system performance/force capability under very stressful conditions and facilitate estimating performance under less stressful conditions (i.e., inside the envelope) as *interpolations* rather than *extrapolations.*

The use of a POS space to design a testing strategy is another reason to include modeling and simulation as an integral part of the operational testing process. The POS space is developed by identifying: (1) likely geographic areas around the world that U.S. forces might have to be deployed to and operate in, (2) the set of uncontrollable "operational parameters" that affect U.S. military/system performance in those areas (terrain, weather, threat objectives, levels, equipment, tactics, countermeasures, etc.), and (3) the likely range of each parameter in the POS space. Using this space, parametric simulation-based analyses should be used to identify feasible stressful situations for operational testing. As noted elsewhere in this re-

[3]The parametric operational situation space was originally developed in a capability-based planning study to structure a versatile rapid reaction force for the North Atlantic Treaty Organization's Supreme Allied Commander for Europe in 1992. This approach to capability-based planning was then used in a number of other studies in the 1990s. Although different in concept and technical details, the POS approach has similar objectives to the capability-based planning approach currently used by the Joint Chiefs of Staff.

port, modeling and simulation-based analyses should also be used to identify issues and hypotheses for testing, to guide the development of testing strategies and designs, and to analyze and extend test results.

> **Conclusion 2:** Testing early in the development stage should emphasize the detection of design inadequacies and failure modes. This will require testing in more extreme conditions than those typically required by either developmental or operational testing, such as highly accelerated stress environments.

The current incentive structure in DoD may discourage testing outside the envelope or identifying component or system limitations and failure modes early in the process. Test strategies by both contractors and DoD apparently are designed to emphasize situations in which a defense system can "do well." This may be one of the reasons that reliabilities assessed during developmental tests are much higher than those assessed during operational tests. For example, the ineffectiveness in the early identification of failure modes may be due to the fact that the current acquisition system uses the funding mechanism to punish early failures in programs compared with programs that do not appear to have initial problems. Pressures on program officials to meet budgets and deadlines, due to congressional and other oversight, result in test strategies geared toward demonstrating "successful" performance. Thus, testing is often carried out under benign or typical stresses and operating conditions, rather than striving to determine failure modes and system limitations by testing under more extreme circumstances. Comprehensive testing targeted toward identifying system failure modes and system limitations—a cornerstone of commercial system development—does not appear to have a high priority in DoD.

> **Conclusion 3:** To have a reasonable likelihood of fully implementing the paradigm of testing to learn about and to improve systems prior to production and deployment, the roles of DoD and congressional oversight in the incentive system in defense acquisition and testing must be modified. In particular, incentives need to be put in place to support the process of learning and discovery of design inadequacies and failure modes early and throughout system development.

DEVELOPMENTAL AND OPERATIONAL TESTING

As noted already, there is a need for more operational realism in early test events to help identify failure modes and system performance shortcomings that will appear under operational circumstances. In evolutionary acquisition, it will be practical to conduct full-scale operational tests only at stages with major upgrades or substantive new capabilities. At other stages, only developmental tests of components and subsystems and some limited tests on functionality, interoperability, etc., will be feasible. These tests, while not full-scale operational tests, should use operational realism to the extent needed to assess the performance of these components and subsystems from an operational perspective.

Thus, the current distinction between operational and developmental testing will have to be reconsidered. A new paradigm, better coordinated between the two approaches, would allow system development to benefit from a more continuous, strategic approach to testing—efficiently supporting the evaluation of competing designs at the beginning of system development, and then the evaluations of new components, subsystems, and complementary systems prior to integration with the existing overall system of interest. Testing at the component or subsystem levels would also emphasize the identification of operationally relevant failure modes and design limitations.

Recommendation 2: The under secretary of defense (acquisition, technology and logistics) and the director of operational test and evaluation should revise DoD testing procedures to explicitly require that developmental tests have an operational perspective (i.e., are representative of real-world usage conditions) in order to increase the likelihood of early identification of operational failure modes and system deficiencies, so that appropriate actions can be developed and deployed in a timely fashion.

DoD'S ROLE IN SYSTEM DESIGN AND DEVELOPMENT AND INTERACTION WITH CONTRACTORS

In commercial enterprises, techniques for experimentation are most effectively applied at the product design and development stages. In DoD, however, most of the testing efforts (including elements of developmental testing) are performed too late in the process to detect deficiencies, improve

the design, or enhance system performance. To quote *Statistics, Testing, and Defense Acquisition* (National Research Council, 1998:38) again, "It is now generally acknowledged [in industry] that quality cannot be 'inspected' into a product. . . . Rather, quality must be designed into any complex new product at virtually every stage of its development. This is accomplished by creating acquisition processes that make use of continuous monitoring and testing."

Testing in the design and development stage will aid critical decisions on the choice of materials, design layout, selecting parameter values for optimizing system performance, and exploring robustness to the impact of various uncontrollable "noise" factors in manufacturing and field-use conditions. In the commercial sector, a considerable part of the design and development takes place within the company, with direct control over these decisions. In DoD, in contrast, there appears to be no mechanism for the relevant DoD test and acquisition officials to be closely involved with the contractors in the design and development phase. Under current contractual procedures, the contractor's experiments, processes, simulation models, test results, and decisions often are proprietary.

Admittedly, even in private industry, some manufacturers rely on specifications, testing, and inspection to enforce the quality and reliability of suppliers' products. But leading companies like Toyota maintain very close relationships with suppliers and work hand-in-hand with them to improve their processes and products (see Box et al., 1988). This has allowed them to implement leading-edge quality practices, such as just-in-time and lean manufacturing, with cooperation and information throughout the value chain. Current rules and regulations do not allow DoD to take advantage of such practices.

Effective implementation of evolutionary acquisition requires closer interaction and a high degree of coordination and communication among system developers, testers, and system users. A critical component of such interaction is the need for DoD testers to have full and early access to all contractor data sources in order to support a mutually flexible and iterative approach for design, development, and testing.

Conclusion 4: In the evolutionary acquisition environment, effective system development and optimization will require a high degree of coordination and communication among system developers, government testers, and system users. In particular, government testers should have early access to all contractor data sources, including test plans,

results of early stage testing and experimentation, and the results of all pertinent modeling and simulation products.

Recommendation 3: The under secretary of defense (acquisition, technology and logistics) should develop and implement policies, procedures, and rules that require contractors to share all relevant data on system performance and the results of modeling and simulation developed under government contracts, including information on their validity, to assist in system evaluation and development.

The committee does not underestimate the difficulty of getting contractors to share their test data and in separating information that is proprietary from data that should properly be shared with the program officials and DoD test community. Clearly, this will require considerable discussion and effort. Nevertheless, the committee thinks that these should be carefully negotiated up front as part of the system development contract.

Cooperation between DoD and contractors may have been discouraged in the past due to concerns over the possible impact it would have on their independence and the need for objective assessment of system performance. While such independence is needed, it is also important to learn from industrial best practices in order to work closely with the contractors and suppliers. DoD must recognize the value of the information on upstream design problems in guiding test design and in supporting the continuous assessment and improvement of system performance.

ENSURING MATURITY OF NEW TECHNOLOGY

In addition to overly optimistic operational requirements (discussed in Chapter 4), incorrect assessment of the maturity of new technology and complications in converting technological advances into producible and reliable products are major causes of slippage in an acquisition schedule and associated cost growth. For example, the U.S. General Accounting Office (1992:51) found that "successful programs have tended to pursue reasonable performance objectives and avoid the cascading effects of design instability." In 2001, the Center for Naval Analyses (CNA) addressed the instability of requirements and performance characteristics for major DoD acquisition programs. On the basis of interviews, it reported that senior

DoD officials believed the Army's acquisition process had too many sources of requirements and too little coordination or corporate direction. As a result, "too many requirements get 'approved' without adequate consideration of resource availability or near-term feasibility." CNA also reviewed the acquisition program baselines (APBs) for 70 DoD acquisition category I (ACAT I) programs to assess changes in performance characteristics (APBs are the "contracts" between DoD/Service acquisition executives and their program managers containing the performance goals and thresholds for system performance). For 70 Army, Navy, and Air Force programs, CNA found that there were 20 to 30 performance characteristics per program and that 10 to 30 performance characteristics changed for each program (i.e., about 50 to 100 percent changed). They also noted that the frequency of APB changes was roughly one per every 2.5 to 2.8 years.

In the traditional single-stage acquisition environment, it may take a decade or more before a new but possibly risky technology can be incorporated into an existing system. This may have encouraged the incorporation of risky, immature technology into acquisition programs. There are also currently no management penalties for using such immature technologies.

Evolutionary acquisition provides increased opportunities to better discipline the management of acquisition programs. Decision makers can weed out unwarranted optimism in draft "requirements" and choose only mature technologies to include in the early stages of acquisition programs. They can delay the introduction of risky, immature technology to future stages. Engineers can demonstrate a satisfactory level of technological maturity in separate advanced technology demonstrations (ATDs) or advanced concept technology demonstrations (ACTDs). Demonstrating technological maturity before including the technology in a formal acquisition program will eliminate the need to delay the entire acquisition program because of one risky technology area or risk using technology that may not be sufficiently effective or reliable. (This separation of the roles of technology development and product development is supported in U.S. General Accounting Office, 2004.)

The under secretary of defense (acquisition, technology and logistics) or the relevant service acquisition executive could request the director of defense research and engineering to certify (or refute) sufficient technological maturity for critical components to be included in any particular stage of the system's development. The relevant acquisition authority could also provide for special reviews of the analysis of alternatives (AoA) for assess-

ment of technology risk and maturity called for in Section 3.5.3 of DoD Instruction 5000.2 or the materiel developer's assessment, during developmental test and evaluation, of technical progress and maturity against critical technical parameters (as documented in the test and evaluation master plan) called for in Section E5.5.4 of DoD Instruction 5000.2. By taking such steps to achieve early confirmation of maturity or to delay introduction of an immature technology to a later stage of development, the acquisition executive can avoid putting the entire acquisition program in an unacceptable high-risk position.

Recommendation 4: The under secretary of defense (acquisition, technology and logistics) should require that all technologies to be included in a formal acquisition program have demonstrated sufficient technological maturity before the acquisition program is approved or before the technology is inserted in a later stage of development. The decision about the sufficiency of technological maturity should be based on an independent assessment from the director of defense research and engineering or special reviews by the director of operational test and evaluation (or other designated individuals) of the technological maturity assessments made during the analysis of alternatives and during developmental testing and evaluation.

COMPARISON WITH BASELINE SYSTEMS IN EVOLUTIONARY ACQUISITION

In traditional acquisition programs, operational tests and evaluations often test both the system under development and the baseline system (i.e., the control) that is scheduled to be replaced (or the system or family of systems that currently performs comparable missions). Baseline testing provides direct contrasts—both relative to specific required performance characteristics and on a broader mission performance scale—that can establish the degree to which the new system is an improvement over the old one. The ability to make such comparisons is especially important when the integrity of prescribed performance requirements for the new system can be questioned (e.g., lacking a solid basis, sensitive to test scenario and test execution specifics, subject to data measurement uncertainties), when those requirements are limited in scope (e.g., focused solely on technical performance characteristics rather than operational mission accomplishment), or when the new system's performance in the operational test and evaluation

appears to come up short relative to one or more specific established requirements.[4]

In an evolutionary acquisition framework, baseline testing should retain its essential role. The current system provides the baseline in Stage I, while the new system from the previous stage will serve that role in subsequent stages. When there are major changes to the system, the additional system(s) or capabilities (or both) that distinguish the new stage should be subjected to extensive technical testing and operational assessments that explore functionality, interoperability, safety, etc., over the spectrum of relevant environmental conditions.

In a staged development process, the system at the initial stage will have advanced through its own test and evaluation. In addition, there will be field data on actual field performance. These sources of data will provide very useful comparative information on the performance of the baseline system, much more so than in the current acquisition process, in which an existing system is compared with a completely new prototype.

OPERATIONAL TESTING OF VERY COMPLEX DEFENSE SYSTEMS

There is another consideration that is independent of the issue of evolutionary acquisition, and that is the increasing complexity of major defense systems. The complexity includes: (1) components, subsystems, and systems of systems and their interactions, (2) network centricity, (3) leading-edge materials technology, (4) leading-edge guidance and control technology, (5) unknown human factors, (6) unknown vulnerability to countermeasures, and (7) software demands on the order of tens of millions of lines of code. As a result, DoD fast is approaching a period in which a single all-encompassing large-scale operational test, as currently practiced, will cease to be feasible.[5]

---

[4]Recommendations 2.1 and 3.2 in *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements* (National Research Council, 1998) recognize the importance of operationally relevant and mission-focused requirements, acknowledging some of the related difficulties inherent in current DoD acquisition processes.

[5]It has been suggested that with evolutionary acquisition substantially more funds will be needed for the testing of each of the systems produced at each stage of development—funds that may not be available. We are not convinced that the test budget needs to be increased to support widespread adoption of evolutionary acquisition. However, even if that

The sheer number of subsystems, components, materials, software, and resulting system interactions may become overwhelming. Operational testing that is typically limited to at most a few dozen test scenarios cannot hope to exercise the systems in enough ways to discover all of the important design deficiencies with a reasonable degree of confidence. As discussed above, extensive operational testing will be reserved for stages that have the most substantive changes involving the most components and subsystems, and will be limited to a relatively small number of test scenarios, even across stages of development. For other stages, testing will be limited to component, subsystem, and software testing.

Thus, it is unlikely that the operational evaluation of complex systems can be based primarily on an operational test representing the full spectrum of applicable (simulated) combat conditions in the field at an affordable cost, within a reasonable time, and with meaningful conclusions. The number of all relevant combinations of scenarios for testing the system will simply be too large for any traditional operational testing.

This concern about the ability to operationally test very complex systems is not new. A good example is the Peacekeeper missile, which was tested in the early 1980s. It could not be comprehensively tested using standard methods, since, for example, only 20 operational replications were feasible. As a result, assessments of performance over a wide range of operationally realistic trajectories could not be definitive. Given these circumstances, the operational evaluation needed to be based largely on a model of the system, with the model's validity established as well as possible given the test replications. Various "tricks" were used so that each test replication provided as much information as possible. For example, a flight might have several accelerometers and instrumentation might be very detailed. The model depended heavily on all of the information that went into the original design, developmental testing, and other sources. This example demonstrates that even with very limited testing, one could ultimately be confident in one's assessment of the system's performance capabilities.

---

were the case, the general discussion of the proper size of the test budget needs to be considered in the broader context of estimates of life-cycle costs and the effectiveness of the fielded systems given more continuous testing. It is the general experience of industrial systems development that identification of failure modes early enough to identify superior system designs not only reduces the frequency of retrofitting and redesign work, but also improves the performance of the system in the field. Therefore, greater use of testing that identifies failure modes and system limitations early can pay for itself in reduced maintenance, repair, and replacement costs, as well as in protocols for how the systems are to be effectively used.

An alternative approach to testing is to get experts knowledgeable in the system together with experts in statistical experimentation design to (a) model and analyze the system design at important levels of abstraction for important scenarios of use; (b) identify the most likely critical or vulnerable components, interfaces, or interactions and design and conduct relatively cheap, focused tests of these components or interfaces; (c) present the results and case analysis to key people in the acquisition processes; and (d) refocus the development to the critical or vulnerable components.

Traditional operational test and evaluation can and should still be conducted for simpler systems to demonstrate that these systems are likely to work as intended in the field. But as the percentage of defense systems that are extremely complex appears to be increasing, the operational test and evaluation strategy must focus on the essential components, subsystems, and interfaces. Although such an approach will not provide a guarantee that the comprehensive and complex system will work, it will provide useful evidence that certain scenarios and zones of use will work reliably. Analogous to mapping a mine field, testing can show "safe paths" of important use. For example, there are techniques for identifying small subsets of all possible combinations of components, subsystems, and environments, so that critical pair-wise or higher order combinations are tested; Cohen et al. (1994) discuss an application to software testing. Also, aspects of the "exploratory analysis under uncertainty" paradigm (e.g., Davis et al., 1999) may be applicable.

In summary, with unlimited time and test budgets, the preferred approach would be to test the system or its components in every operationally relevant scenario. However, given the large number of all possible combinations of the factors and test scenarios, testing will have to make use of clever strategies. These include testing only in selected scenarios to examine the performance of those components or interfaces that seem to be most problematic, testing only a subset of all possible interactions or inter-operability features, testing those scenarios that correspond to the most frequent types of use, or some combinations of these strategies. Clearly, there is a need to investigate various alternatives and develop specific proposals.

Conclusion 5: The DoD testing community should investigate alternate strategies for testing complex defense systems to gain, early in the development process, an understanding of their potential operational failure modes, limitations, and level of performance.

3

Combining Information in Staged Development

There are many opportunities for combining information to improve test design and analysis even in traditional single-stage acquisition (see, for example, National Research Council, 1998). These include data from developmental tests, operational tests, and even training exercises, as well as test results of earlier systems composed of similar components. Such opportunities are considerably greater in an evolutionary setting, in which there is substantial information about the operational performance of the system from earlier, fielded versions.

The similarity of the systems at different stages of development can be exploited so that information from past stages can be used to support efficient test design for later stages. In particular, statistical models can be used to link the performance of a system in developmental tests with that of previous versions of the system in developmental tests, between different stages of the system in operational tests, between developmental and operational test results for various versions of the same system, and linking test data with field performance data. There are both informal and formal methods for combining information from various sources (see, for example, National Research Council, 1992, 2004). See also Berry (2005) for related results in the area of drug development.

TEST DESIGN

Among the various techniques in the experimental design literature, sequential design and analysis are the most relevant to evolutionary acquisi-

tion. However, there are also important differences with traditional use of sequential designs. Evolutionary acquisition represents a "block" sequential case, in which experiments corresponding to a particular stage are carried out in blocks. Furthermore, the system under study changes from stage to stage; unlike many other sequential studies, the decision on what to do in the next stage does not necessarily depend on results from the previous stage. Rather, this decision is based on strategic consideration about the new additional capabilities that are needed for the system in the field.

In each block there are several types of experimental strategies that could be considered, such as screening experiments for identifying important factors, response-surface designs for determining the optimal factor combinations, and so on.[1] Their usefulness for developmental and operational testing is discussed in *Statistics, Testing and Defense Acquisition: New Approaches and Methodological Improvements* (National Research Council, 1998; see also National Research Council, 2004; Box, Hunter, and Hunter, 1978; Box and Draper, 1987; Myers and Montgomery, 2001; Wu and Hamada, 2000). This section provides a qualitative description of how one could use information from past stages to improve test design in the current stage. A more technical description of some of these ideas is given in Appendix B.

One can use test results and field performance of the system from the previous stage in both qualitative and quantitative ways in test design. These involve obtaining information on (a) factors that were found to be unimportant in previous tests; (b) how the performance (response surface) varied as a function of changes to key inputs or test scenarios; (c) hazard rate behavior of failure data, such as increasing hazard rate, infant mortality, etc., for reliability test design; (d) estimates of variability that are needed for allocating the test resources to different test scenarios in the current stage; and so on. Results from the earlier stages can also suggest whether to oversample or undersample certain test scenarios in the current stage of development or to push the testing envelope in certain directions. The system in a subsequent stage in development could be tested in problematic circumstances to see whether reliability growth or the addition of new com-

[1]The naïve use of both screening experiments and response-surface models will almost always be inferior to the use of techniques that are developed in collaboration with system experts, who can help guide the choice of variables and model forms.

ponents is having a beneficial impact on performance. Poor performance can be analyzed in subsequent stages of system development to understand why certain prototypes were poorly manufactured, if variables can be found that discriminate between the good actors and poor actors, and so on.

Clearly, the extent to which information from previous stages can be used effectively depends on the similarities and differences between the stages. This is explored in the context of a regression model in Appendix B for several cases. In situations in which there is good prior information, one can use a Bayesian approach effectively to develop good test designs (Chaloner and Verdinelli, 1995). It is also related to the work done by Los Alamos and Procter & Gamble reported on at the workshop.

One can also develop and use a more formal decision-theoretic approach for combining information from developmental, operational, and field tests (as well as combining information from tests at different stages of development). Appendix B includes a discussion of some preliminary ideas. A key problem with this approach is that it requires inputs that cannot be realistically quantified, for example, the price of late fielding for a system, the price of having a fielded system perform poorly in operations, the benefit of deploying a good system earlier, the benefit of winning a battle faster as a result of a fielded new system, the deterrence value of fielding a new system (even if they are not effective), and so on.

ANALYSIS

Many techniques exist for combining data from various sources to improve the quality (efficiency) of the analysis and conclusions (see, for example, National Research Council, 1992, 2004). These are particularly relevant in evolutionary acquisition, in which data from developmental tests, operational tests, and field performance from previous stages of the system are available. We do not provide a detailed discussion of these methods, as the issues are very similar to those in National Research Council (2004) and references therein.

Combining information from the various sources, which rely on relevant linkages between previous and current sources of data, requires considerable care and subject-matter expertise. When done correctly, however, it has a tremendous payoff. For example, ACAT I defense systems typically have dozens of primary measures of performance and measures of effectiveness that are functions of system requirements. These measures must be evaluated in a wide variety of operational scenarios and environments. It is

true that measures of performance and effectiveness often are interpreted as average performance across these scenarios, but one cannot focus entirely on these averages, since it is also important to identify any extreme heterogeneity of system performance across scenarios. This is an area in which considerable expertise in modeling and statistics is needed. In addition, it should be noted that, while Bayesian methods can be used to formally incorporate information from previous stages, they also provide considerable opportunities for abuse in the use of prior information. Thus, in addition to the right expertise, there is also need for oversight and proper documentation of the analysis and conclusions. These issues are discussed in the next chapter.

Current practice in DoD does in fact employ many of the above suggestions, albeit informally and not always consistently, and it is subject to the infrastructure limitations described elsewhere in this report. For example, deficiencies discovered in the field are often the focus of subsequent testing, and new capabilities are often the drivers of operational test design. However, this use of prior information is not routine and is not nearly as effective as it could be. The techniques discussed in the appendixes and similar ones need to be formalized and more broadly adopted and institutionalized in the documentation and casebooks of best practices.

Recommendation 5: The Service test agencies should undertake a pilot study, involving a few selected systems developed under the evolutionary acquisition paradigm, in order to identify specific opportunities for incorporating information from previous stages to improve system design and analysis. These case studies will be beneficial in demonstrating both the application of the various techniques and the benefits to be gained from combining data in staged development.

MODELING AND SIMULATION

Evolutionary acquisition will bring about changes in the role and value of modeling and simulation for operational evaluation and operational test design. The most substantial change is that staged development provides the opportunity to refine models and simulations via the validation provided by the collection of field performance data of earlier stage versions of the system. This, in turn, supports feedback loops for model (and system) improvement. Through this process, the utility of a model or simulation to

better mimic the crucial determinants of effectiveness and suitability becomes greatly enhanced. In other words, the process of *model-test-model* becomes much more effective through the "testing" that field use (or training exercises) provides. The recalibrated models and simulations can then be used to assist in next-stage test design and operational evaluation.

In the mid- to late 1990s, the Air Force Operational Test Center had a pilot project in which it attempted to employ modeling and simulation to track development of the B1B Defensive System Upgrade. The idea was to incorporate the information gained in developmental testing, operational testing, live-fire testing, training exercises, etc., into one software representation. The model would then serve as a repository of information on current system performance that could be queried in various ways. For example, if successfully formulated, this software representation of the system could be used to track reliability growth, to identify which components needed either additional work or further testing, to identify problematic scenarios of use, and so on. Because of the complexity and risk with electronic warfare systems, modeling and simulation were to be major evaluation tools to evaluate jamming effectiveness against radio frequency threats. Although this effort was the first of its kind (as far as we know) and the program was ultimately discontinued, evolutionary acquisition programs, particularly their intrinsic feedback loop mechanisms, naturally support the construction of similar virtual system representations.

Scarcity of time and resources is the usual reason offered for expanding the role of engineering-level modeling and simulation within operational test and evaluation. These science-based models that directly incorporate detailed information on the mechanisms of system functionality and operation, have been proposed as a means for expanding the scope of evaluations beyond the sets of experimental circumstances actually tested. This is a plausible approach only when the relevant technical components of system performance can be modeled sufficiently well for the problem at hand, via physical and/or chemical modeling, and when the testing yields direct validations of modeling and simulation constructs and assumptions. Moreover, extrapolations to nontested regions may be supportable to the degree that their salient characterizations are well understood and encompassed by validated modeling and simulation depictions. Typically (e.g., in the ballistic missile defense arena), such extremely detailed modeling and simulation representations entail considerable complexity and comprehensiveness, and their development often requires extensive dedicated resources and substantial time commitment.

In addition to their potential utility for expanding the scope of operational performance evaluations, modeling and simulation can also serve as valuable test planning tools—exploring test scenario options, addressing test sizing issues, identifying critical performance issues, and so on. Depending on the application, this can be accomplished by detailed physics-based models or by lower fidelity models that capture primary system performance without attempting to replicate the underlying physical processes that define system performance. One advantage of simpler models is increased responsiveness and flexibility, both in terms of initial availability to support operational testing and evaluation planning as well as turnaround time to complete detailed studies (e.g., comprehensive sensitivity analyses). However, without a physical justification, extrapolation away from conducted testing conditions and circumstances is generally not warranted.

As Art Koehler described in a presentation at the workshop, a particular application of modeling and simulation that is directly relevant to evolutionary acquisition is being developed by Procter & Gamble and Los Alamos National Laboratories. They have designed a simulation and analysis system that can examine the impacts on system performance and system reliability from changing a major component of an existing complex system. This simulation and analysis system therefore provides a key component of what one would need to know in an evolutionary context in order to plan for and accommodate changes resulting from component upgrades, possibly through redesign of other parts of the existing system.

SOFTWARE DEVELOPMENT IN AN EVOLUTIONARY CONTEXT

Software development is one area in which evolutionary acquisition is expected to play a major role. Given the increasing role of software in complex defense systems, staged development should take place consistent with current best practices. The general approach described in Appendix C, sometimes referred to as the cleanroom process, is one of several similar approaches representative of current best practices. The software engineering process described in Appendix C, however, is designed to be carried out by contractors, and the extent to which it could (or should) be mandated in government contracts is unclear. Nevertheless, there is a need to explore how this might be done.

4

Changes to Infrastructure, Process, and Culture in Support of Evolutionary Acquisition

A number of changes to the infrastructure of the test and acquisition community are needed to take full advantage of the opportunities present in the evolutionary acquisition process and to confront the formidable challenges posed.

TECHNICAL EXPERTISE

The DoD testing community will need greater access to, and use of, expertise in statistical methods, particularly in the areas of experimental design, modern analysis methods, and, more generally, eliciting and combining information. Test designs for even ACAT I systems currently make use of relatively standard "cookbook" designs, often modifying a design used for a similar system in the past. This approach to operational test design is not designed to fully exploit the increased information available in an evolutionary context, and it will often be inadequate to effectively test the highly complicated systems-of-systems that are becoming increasingly common. Formulating effective test designs for systems developed in stages will often involve questions that are of the level and complexity of publishable research. Much greater access to and use of the highest levels of statistical expertise will therefore be needed. Several ways of developing a more collaborative relationship with the statistical community were outlined in Chapter 10 of *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements* (National Research Council, 1998).

The increasing role that software-based components are playing in complicated defense systems and the evidence that a substantial percentage of cost increases, schedule delays, and performance problems are due to software problems require greater access to expertise in software engineering and software testing methods. Although a number of extremely effective relevant procedures have been developed in the past decade, there is little evidence that these techniques have been applied to defense systems (see, for example, National Research Council, 2003).

The greater role that modeling and simulation will play in testing and evaluation of increasingly complex systems-of-systems will also require greater access to expertise in the use of physics-based modeling and simulation and modeling at the operational level.

DATA ARCHIVING

A test and field data archive is absolutely necessary to provide facilitated access to information on the test designs and outcomes of tests from previous stages of development. Methods of combining information clearly require information from previous stages of development. By a test (and field) data archive, we mean the following (as discussed in National Research Council, 2004:59-60):

> a rich set of variables to adequately represent the test environment, the system under test, and the performance of the system. . . .
>
> In order to accurately represent system performance, including the appearance of various failure modes and their associated failure frequencies, the circumstances of the test must be understood well enough that the test, training exercise, or field use can be effectively replicated, including the environment of use (e.g., weather, terrain, foliage, and time of day) and type of use (e.g., mission, intensity, and threat).
>
> While a system is under development, the system design is often under constant modification. Given the need to be able to replicate a test event in the database, it is crucial to represent with fidelity the system that was in operation during the event so that proper inference is possible.

> In addition to storing the length of time between system failures, it is also important to identify which hardware or software component malfunctioned; the maintenance (including repair) record of the system; the time of previous failures; the number of cycles of use between failures; the degree of failure; and any other variables that indicate the stresses and strains to which the system was subjected, such as speed and payload. It is also useful to include the environments and stresses to which individual system prototypes have been exposed historically (e.g., in transport, storage, and repeated on/off cycling), in order to support comprehensive failure mode analysis, especially if an apparent declining trend in system reliability appears.

If the above data are available, they can be used to help design efficient testing in the system's evolutionary development process; they can greatly facilitate the effective combining of data from different sources over time; they can be used to interpolate the results of testing in a limited number of operational situations in order to assess capabilities of the system in situations not tested; they can provide a "hot plant" for reliability assessments; and they can assist in developing performance correlates (e.g., reliability) for design of the system's evolutionary upgrades.

For programs of the size and complexity that are typical of recent ACAT I systems, and especially for those developed using evolutionary acquisition, there will be an enormous amount of data from contractor testing, early and late government testing, results from training exercises, results from modeling and simulation, and results from field use for those systems that have gone into production. Being able to find and utilize these diverse sources of data for various purposes requires that they be documented and arrayed in a way that facilitates a variety of analyses. This will also strongly support the operation of the various feedback loops needed to improve system design, to identify the sources of field performance failures, to improve modeling and simulation, and to improve test planning and conduct. Evolutionary acquisition is tailor-made to support the operation of these feedback loops, given the continuity of the systems developed, and given that there is an opportunity to examine their performance once fielded. While there are existing data archives maintained by some of the Service test agencies, none of them contain all the information that is listed here and that is essential for fully understanding the system that was tested

or fielded, the governing operating conditions, and the associated performance records.

This information is not easy to collect in controlled settings, such as operational tests, and it is considerably more difficult to collect in less controlled types of use, such as training exercises and field use. However, much in this direction can be accomplished. For example, in many commercial industries, sensors are attached to fielded systems to collect much of the information automatically.

Recommendation 6:

(a) To support the implementation of evolutionary acquisition, DoD should acquire, either through hiring in-house or through consulting or contractual agreements, greater access to expertise in the following areas: (1) combining information from various sources for efficient multistage design, statistical modeling, and analysis; (2) software engineering; and (3) physics-based and operational-level modeling and simulation.

(b) Test and field data archives should be established to facilitate access to data on test and field performance of systems in development and those fielded. These data archives should be used to support feedback loops to improve the system design, to improve testing methodology over time, and to help validate (and improve) modeling and simulation for operational evaluation.

To help implement this recommendation, ideally, a data plan could be established early in the program that would be connected to a modeling plan, so that at every stage the needed data could be obtained. Various performance parameters and related statistics could therefore have a common format that would facilitate using those data.

TERMINOLOGY AND DOCUMENTATION

Evolutionary acquisition is defined in DoD Instruction 5000.2, although the term has not been used in a consistent manner by DoD leadership or in various other DoD documents. The variety of terms, including *evolutionary acquisition, incremental development*, and *spiral development*, and their inconsistent usage have served to obscure the original goal and intent of the evolutionary acquisition process.

For example, it has been put forward in some documents (Aldridge, 2002) that spiral development is the process used in an individual stage of (what is referred to as) incremental development to determine appropriate requirements for that stage. However, in DoD Instruction 5000.2, which represents official DoD policy, spiral and incremental development are represented instead as two parallel forms of evolutionary acquisition—with incremental development having fixed requirements in advance, and spiral development having requirements that are developed when specific stages of development have been initiated. Such conflicting use of terminology by DoD managers has created some confusion.

Another area of confusion is whether the requirements are defined clearly up front for each stage of the development, or if it is only decided at the beginning of that stage. There was widespread concern among the industrial participants at the workshop that this flexibility in the identification of requirements can lead to serious problems in evaluating the performance of the systems and potential for "gaming" the development process.

Besides clear and consistent definitions, a consistent and definitive articulation of the goals and intention of the evolutionary acquisition process is needed, as well as supporting documentation detailing the specific changes that will need to be implemented throughout the acquisition community (e.g., how the milestone system is intended to operate in conjunction with evolutionary acquisition). There also is a need for more direction as to how evolutionary acquisition will affect testing, both development and operational, and what the process should be for deciding how to use test results to help determine whether to field a system having attained a given stage of development. This would assist all of those in the acquisition community, including contractors, program officials, and testers, in adjusting to this new acquisition methodology.

At first glance, it might seem attractive to develop a formal taxonomy of programs with separate guidelines in each cell. However, this runs the risk of making the process too rigid and cumbersome and negating the flexibilities inherent in evolutionary acquisition. The best alternative would be to work with the relevant technical personnel in DoD to develop a set of simple, clear, and coherent guidelines.

Recommendation 7: The under secretary of defense (acquisition, technology and logistics) should eliminate inconsistencies in DoD Directive 5000.1 and DoD Instruction 5000.2 and clarify other significant memoranda and documents regarding evolutionary ac-

quisition. All policies and procedures to be used in applying evolutionary acquisition principles to DoD acquisition programs should be strictly enforced. This clarification and enforcement should be applied to program management both in DoD and in all supporting contractor activities.

PROCESS AND CULTURE IN DoD

Evolutionary acquisition is being folded into an acquisition environment that already has a counterproductive incentive system (see, e.g., National Research Council, 1998:23-29). The flexibilities inherent in the evolutionary acquisition process present greater opportunities for these counterproductive incentives to be expressed. For example, evolutionary acquisition allows requirements to be set at the beginning of each stage of development, rather than stating them for all stages at the outset. While there are obvious advantages to having fluid requirements for a system developed in stages (e.g., addressing unpredictable changes in threats to the system, being able to incorporate new and unanticipated technological advances in capabilities), there is also great potential for abusing this fluidity. It has been shown several times that even in the single-stage context, initial system requirements used to justify an acquisition program often are revised downward during system development (see, for example, Christle et al., 2001). This suggests that the initial requirements (and cost estimates) for a defense system are sometimes used as a sales brochure to support the decision to initiate an acquisition program. Then, once a program has become more established, modifications are made to adjust the requirements to lower levels of performance.

The interplay, or lack thereof, between user, developer, and tester in setting requirements for initial and intermediate stages of system development (often mentioned as part of evolutionary acquisition) is currently a factor in setting unrealistic initial requirements. The more flexible approach of evolutionary acquisition may encourage this practice, since requirements that are difficult to meet might be pushed back to later stages of development.

Clearly, there is a trade-off between gaining efficiencies inherent in a flexible acquisition process and the pressures for greater oversight and systematic documentation of the overall development process. To date, few if any program managers for defense systems have requested, or been approved for, evolutionary acquisition classification. It is reasonable to sup-

pose that, in addition to the overall confusion surrounding evolutionary acquisition, some of the associated activities listed in DoD Instruction 5000.2 are considered burdensome. In particular, Congress has imposed strict reporting requirements on any program officially designated in DoD as an evolutionary system, and this may discourage program officials from requesting this designation. While there are trade-offs between flexibility and oversight, the need for clear accountability is even more critical in the flexible evolutionary environment, in which there is also a desire to reduce the oversight burden.

It is worth comparing some key aspects of the product development process in the commercial industrial sector with that in DoD. In industry, the broad spectrum of activities for the development of a major new product is typically overseen by a chief engineer. The chief engineer is accountable for the entire product line, from concept design, through product development, to manufacturing and product launch. This is intended to ensure overall accountability, from original justification of the product and projection of market share to quality and costs of design and development and ultimately to reliability and warranty costs. By contrast, in DoD, mission capabilities are often overstated and costs understated at the initiation of an acquisition program, with transference or recognition of added costs downstream along with delays in product development. Moreover, the typical tenure of program managers is much shorter in DoD, often on the order of three years, so there is generally no full accountability for the decisions made over the life of an acquisition program.

The incentive structure in industry encourages a "survival of the fittest" behavior, so that poor quality or high-cost operations (or both) will fail in the long run. The reporting and management structure in industry is well defined. In DoD, the existence of multiple players, reporting structures, and layers of oversight groups complicates the decision-making process substantially. The current acquisition process, environment, and culture have significant built-in inefficiencies.

These problems will become even more critical under evolutionary acquisition and can seriously hinder the process and even lead to abuse. Similar comments were made in the past in the context of the current acquisition process (see National Research Council, 1998). However, there is little evidence of any systematic effort to address the underlying issues. Improving the incentive structure (i.e., aligning the incentives of the major players) is perhaps the most onerous obstacle facing the DoD acquisition process if it is to better meet the difficult challenges in acquiring and fielding complex, expensive systems in a timely, efficient manner.

Recommendation 8: The deputy secretary of defense should charge a blue ribbon panel, including experts in organizational behavior, multiobjective decision making, and other relevant areas, to review the Defense Acquisition Performance Assessment panel's proposed changes to DoD acquisition policies and procedures. This review should take place (if possible) before any changes to the policies are implemented in order to assess and improve their likelihood of being successfully implemented in the DoD and defense industry culture.

References

Aldridge, E.C., Jr.
2002 Evolutionary acquisition and spiral development. *Crosstalk: Journal of Defense Software Engineering* (August).

Berry, D.A.
2005 Bayesian clinical trials. *Nature Reviews* (December).

Bonder, S.
2000 Versatility Planning: An Idea Whose Time Has Come—Again! Steinhardt lecture presented at the Institute for Operations Research and the Management Sciences Conference, Salt Lake City, May 9.

Box, G.E.P., and N.R. Draper
1987 *Empirical Model-Building and Response Surfaces*. New York: John Wiley and Sons.

Box, G.E.P., W.G. Hunter, and J.S. Hunter
1978 *Statistics for Experimenters: An Introduction to Design, Analysis, and Model-Building*. New York: John Wiley and Sons.

Box, G.E.P., R. Kackar, V. Nair, M. Phadke, A. Shoemaker, and C.F.J. Wu
1988 Quality practices in Japan. *Quality Progress* 37-41.

Chaloner, K., and I. Verdinelli
1995 Bayesian experimental design: A review. *Statistical Science* 10:273-304.

Christie, T.
2002 Testers shouldn't be blamed for defense program setbacks. *National Defense Magazine* (May). Available: http://www.nationaldefensemagazine.org/issues/2002/May/Testers_Shouldnt.htm [accessed March 2006].

Christle, G.E., A.R. DiTrapani, K.D. Marmaud, M.P. Sullivan, J.A. Thomas, M.A. Miller, J.A. Ferara, and R.V. Johnson
2001 *The Army Acquisition Management Study: Congressional Mandate for Change*. Alexandria, VA: Center for Naval Analysis.

Cohen, D.M., S.R. Dalal, A. Kajla, and G.C. Patton
1994 The automatic efficient tests generator. *Proceedings of the Fifth International Symposium on Software Reliability Engineering*, pp. 303-309, Institute of Electrical and Electronics Engineers, Inc.

Davis, P.K., Bigelow, J.H., and McEver, J.
1999 *Analytical Methods for Studies and Experiments on "Transforming the Force."* Santa Monica, CA: RAND.

Mills, H.D.
1971 Top-down programming in large systems. In R. Rustin ed., *Debugging Techniques in Large Systems*. Upper Saddle River, NJ: Prentice-Hall.

Myers, R.H., and D.C. Montgomery
2001 *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York: John Wiley and Sons.

National Research Council
1992 *Combining Information: Statistical Issues and Opportunities for Research.* Washington, DC: National Academy Press.
1998 *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements*. Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, Michael L. Cohen, John E. Rolph, and Duane L. Steffey, editors. Washington, DC: National Academy Press.
2003 *Innovations in Software Engineering for Defense Systems*. Oversight Committee for the Workshop on Statistical Methods in Software Engineering for Defense Systems. S.R. Dalal, J.H. Poore, and M.L. Cohen, eds. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC.: The National Academies Press.
2004 *Improved Operational Testing and Evaluation and Methods of Combining Test Information for the Stryker Family of Vehicles and Related Army Systems.* Phase II Report, Panel on Operational Test Design and Evaluation of the Interim Armored Vehicle. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Prowell, S., C. Trammell, R. Linger, and J. Poore
1999 *Cleanroom Software Engineering: Technology and Process*. Boston: Addison-Wesley.

Trammell, C.J., M.G. Pleszkoch, R.C. Linger, and A.R. Hevner
1996 The incremental development process in cleanroom software engineering. *Decision Support Systems* 17(1):55-71.

U.S. Department of Defense
2003 Department of Defense Directive 5000.1, May 12. Available: http://www.dtic.mil/whs/directives/corres/pdf2/d50001p.pdf [accessed March 2006].
2003 Department of Defense Instruction 5000.2, May 12. Available: http://www.dtic.mil/whs/directives/corres/pdf/i50002_051203/i50002p.pdf [accessed March 2006].

U.S. General Accounting Office
1992 *Weapons Acquisition: A Rare Opportunity for Lasting Change*. (GAO/NSIAD-93-15/December.) Washington, DC: U.S. General Accounting Office.

2002 *Best Practices: Capturing Design and Manufacturing Knowledge Early Improves Acquisition Outcomes*. (GAO-02-701, July 15.) Washington, DC: U.S. General Accounting Office.

2004 *Using a Knowledge-Based Approach to Improve Weapon Acquisition*. (GAO-04-386SP, January.) Washington, DC: U.S. General Accounting Office.

Wu, C.F., and M. Hamada

2000 *Experiments, Planning, Analysis, and Parameter Design Optimization*. New York: John Wiley and Sons.

APPENDIX A
Overview of the Current Milestone Development Process

The current process used for the development of a single-stage defense system consists of several well-defined steps, each ending with a milestone, that is, a decision point (see Figure A-1). These represent established criteria as to whether the system design is ready to be promoted to the next step in the process. Two primary types of defense testing are associated with this process: developmental testing and operational testing (discussed in detail below). We first describe the milestone development process.

Milestone Development: The first phase of system development is concept refinement, and it culminates in Milestone A: new acquisition program approval. The second phase, technology development, ends in Milestone B, at which stage a well-defined system design has been agreed on for addressing an identified need. The third phase of development is system development and demonstration. Once the system design is verified using developmental testing and appropriate operational assessments, Milestone C authorizes production of prototypes that are likely to satisfy the operational requirements. During this phase, production verification testing as well as other tests are carried out. It concludes with operational testing and evaluation in support of the decision to enter into full-rate production. Passing the operational test generally promotes the system to full-rate production and deployment (fielding). The last phase of acquisition is operations and support, which includes maintenance of production, system support in the field, and any follow-on product improvements and any additional operational testing and evaluation that is needed.

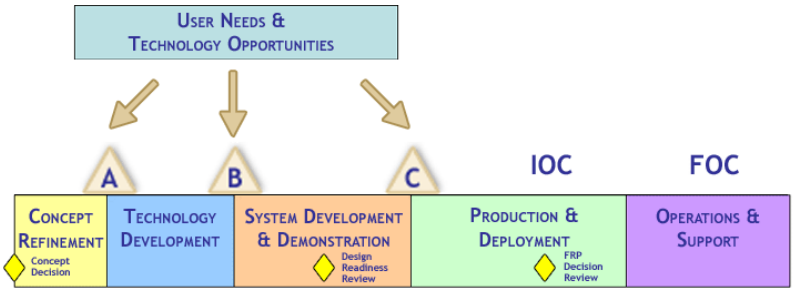

FIGURE A-1 The defense milestone development process.
SOURCE: U.S. Department of Defense Instruction Number 5000.2. May 12, 2003 Under Secretary of Defense for Acquisition, Technology and Logistics. Available: <http://akss.dau.mil/dag/DoD5002/Figure1.asp>.

We describe next the two types of testing—developmental and operational testing—that take place during the development process. Developmental testing includes testing of components, subsystems, and software to ensure that performance capability and reliability are designed into the system early. Developmental testing then proceeds to system-level and system-of-systems–level testing to ensure that the system has matured to the point at which it can be expected to meet initial operational test and evaluation and operational employment requirements. Operational testing is a test of the full production representative system in as operationally realistic a setting as possible, with representation of opposition forces, countermeasures, and participation by users with training equivalent to what would be available in the field under circumstances of typical use.

These two types of defense testing have different objectives. Developmental testing is used to determine the capabilities of components and of the full system against individual stress factors, to help identify failure modes. The intent is to discover changes that can greatly improve the overall design. In addition, developmental testing is intended to help produce a system that would be expected to pass operational testing in proceeding to full-rate production. To examine whether a system design is mature, it is vital to learn about the performance of each component and, to some extent, their interactions. Therefore, during developmental testing, testing under more strenuous circumstances, that is, testing at the edge of the envelope, may be conducted.

In operational testing, an implicit assumption that is often made is that the system design is basically mature and that, barring unusual circumstances, the system will soon be purchased and fielded. Therefore, the system is typically not stressed at excessive levels. Instead, the testing roughly is guided by the typical scenarios of use (e.g., as promulgated by the U.S. Army's formal Operational Mode Summary/Mission Profile document). The objective of operational testing is usually stated as to confirm that the system satisfies its operational requirements. This means that significance tests at standard statistical levels would fail to reject the null hypothesis that the system's performance satisfied its requirements. (The requirements are often stated in terms of averages over scenarios of use, in which the scenarios and their frequencies are governed by the Operational Mode Summary/Mission Profile document.)

Early operational assessments occupy a middle ground between developmental and operational testing. They may involve component testing with typical users or other means of gleaning operationally relevant insights.

APPENDIX B
Combining Information for Test Design with Staged Development

The extent to which one can use information from the previous stages to design experiments or tests in the current stage depends clearly on the similarities in the system at the different stages. This section considers several simple situations and describes some technical ideas on how information from previous stages can be exploited for efficient test design.

A SIMPLE EXAMPLE WITH TWO FACTORS

Suppose that at stage $(k - 1)$, the test design for the system had two scenarios, each at two settings: factor 1 corresponded to sunny versus rainy conditions with x_1 being the indicator variable for sunny/rainy; similarly, factor 2 corresponded to day versus night, and x_2 is the indicator variable for day/night. Let Y be some performance of interest, and suppose the performance at the four combinations of test scenarios can be captured by the following simple model:

$$Y = \beta_{00} + \beta_{10}x_1 + \beta_{20}x_2 + \varepsilon$$

where ε is a random variable with mean 0 and constant variance. Note that there is no interaction term in this model.

Assume that the current stage (stage k) involves addition of a new night vision capability. Based on the subject-matter knowledge, we think that there will not be much change in performance for rainy-versus-sunny con-

ditions and that the only major change will be for day-versus-night test scenarios. Then, it is intuitively clear that most of the test resources in stage k should be used to assess the improvement in night-versus-day scenarios, with substantially fewer resources allocated to testing the effect of sunny-versus-rainy conditions for the new system. The actual allocation can be decided informally, depending on the level of confidence in the knowledge that the addition of the new night vision capability will not affect the rainy-versus-sunny comparison.

A more formal way of allocating test resources can be based on a Bayesian framework for design of experiments (Chaloner and Verdinelli, 1995). The estimates $\hat{\beta}^{k-1}$ and $\mathrm{Var}\left(\hat{\beta}^{k-1}\right)$ from stage $(k-1)$ can be viewed as prior information for the unknown parameters in stage k. Since we do not expect the new capability to affect the sunny-versus-rainy test comparison, we can expect the parameter β_{10} to remain relatively unchanged in stage k. So the prior information from stage $(k-1)$ for β_{10} is quite reliable, and $\hat{\beta}_{10}^{k-1}$ and $\mathrm{Var}\left(\hat{\beta}_{10}^{k-1}\right)$ will serve as good estimates of the mean and variance of the prior distribution. However, we expect β_{20} to change considerably, so we have to place a lot less weight on the prior information from stage $(k-1)$ for this parameter. The exact decision will depend on our belief on how relevant the data from the previous stage are. For example, if we believe that the data from the previous stage still provides unbiased information, we can use $\hat{\beta}_{20}^{k-1}$ as the mean and an inflated value of $\mathrm{Var}\left(\hat{\beta}_{20}^{k-1}\right)$, with the inflation factor depending on our judgment of the relevance. In the extreme case in which the previous data provide no useful information, we can use a "noninformative" prior for β_{20}. Alternatively, if the performance is expected to improve with the new night vision capability, the comparison of night-versus-day from the previous stage can be used as a lower bound for the prior distribution. The Bayesian theory for optimal design can now be used to obtain appropriate allocation of test resources to the various scenarios. There are several optimality criteria, such as D-, A-, and G-criteria. A good review can be found in Chaloner and Verdinelli (1995).

The same ideas can be used if we have additional interaction terms of the form

$$Y = \beta_{00} + \beta_{10}x_1 + \beta_{20}x_2 + \beta_{12}x_1x_2 + \varepsilon$$

or a model with factors or test scenarios with more than two levels and quadratic terms:

$$Y = \beta_{00} + \beta_{10}x_1 + \beta_{11}x_1^2 + \beta_{20}x_2 + \beta_{21}x_2^2 + \beta_{12}x_1x_2 + \varepsilon.$$

EXTENSIONS TO MORE REALISTIC SCENARIOS

The above discussion can be generalized to address a variety of interesting, more complex cases. For example:

1. Suppose that we have new factors or test scenarios (new x's) in stage k. In other words, the system at the new stage will be asked to perform well in entirely new environments, for new missions, or against new threats. The latter could arise in the above example, for instance, if there was no night vision capability in stage $(k - 1)$ so that it was not a factor in the previous operational test. With the new capability added to the requirements, it becomes necessary to test for it. In this case, one will have to examine how the parameters in the model from stage $(k - 1)$ map into those for stage k since only one level (daylight) was tested before. One can use a noninformative prior for the parameters to represent this state of ignorance.

2. Suppose that the functional form of $Y = f(x;\beta)$ changes. If the change is arbitrary, then there is no linkage to borrow information from earlier stages of development. Often, however, there will be some common elements that we can usefully exploit. For example, suppose we have an additive model of the form

$$f^{(k)}(x;\beta) = \lambda_1 f^{(k-1)}(x_1;\beta_1) + \lambda_2 g^{(k)}(x_2;\beta_2),$$

where $g^{(k)}(x_2;\beta_2)$represents the effect of the new factors in the current stage. Then the problem decouples into two independent ones, and we can focus more resources in the current stage on estimating the second component. This is just a more complex version of the earlier, simple example with two factors. If there is, in addition, interaction between the two terms, then resources must also be allocated to estimating that term.

3. Suppose now that new measures of performance or new measures of effectiveness become of interest in the intermediate stages of system development. For example, suppose the amount of collateral damage becomes a new metric used to evaluate a system. Again, the relationship between the new measures and those from the previous stages can be exploited for improved efficiency. This will require careful modeling of the relationships.

4. In many situations, additional capabilities acquired in later stages

can be viewed as new subsystems. If the effect of these new subsystems on system performance can be assumed to be *additive* (as in point 2 above), then the majority of the experimentation resources in the *k*-th stage should be spent on the new subsystems with limited resources devoted to system integration (interactions). More specifically, the following strategy should be useful in such situations:

- Do full operational testing and integration testing only after substantial stages.
- Do limited integration testing at intermediate stages at which modifications are small to moderate.
- Build in realism in developmental tests and carry out full component testing in developmental test.

As pointed out at the workshop by Steve Vardeman, one can use more formal decision-theoretic methods to combine both costs and data from test results at different stages of development, including results from developmental, operational, and field tests. Examples of such methods were described at the workshop. (See, for example, Gaver et al., 2005). These techniques require inputs (typically costs) that are often difficult to estimate or quantify; examples include cost of fielding a system late, cost of having a fielded system perform poorly in operations, the benefit of deploying a good system earlier, benefit of winning a battle faster as a result of a fielded new system, and so on. Nevertheless, analyses based on such approaches can provide useful insights into the trade-offs involved, especially when they are coupled with sensitivity analyses on the robustness of the conclusions to the inputs.

REFERENCES

Chaloner, K., and I. Verdinelli
1995 Bayesian experimental design: A review. *Statistical Science* 10:273-304.

Gaver, D., P. Jacobs, and E. Seglie
2005 *Modern Military Evolutionary Acquisition and the Ramifications of "RAMS."* Technical Report. Monterey, CA: Naval Postgraduate School.

APPENDIX C
Software Development in Evolutionary Acquisition

Software development is a natural area for evolutionary acquisition; in fact, it has been noted that the concept came from this area. Given the increasing and important role of software in complex defense systems, evolutionary acquisition should take advantage of and be consistent with current best practices in this area. This section provides a brief description of one approach, sometimes referred to as the cleanroom process. It is similar to several other approaches that are representative of best current practices.

CLEANROOM REFERENCE MODEL

The Cleanroom Reference Model developed by the Software Engineering Institute at Carnegie Mellon University defines a set of 14 integrated processes for software management, specification, development, and certification. Prowell et al. (1999) provides a detailed discussion of these individual processes and their role in the software development process. Here we focus on one of the processes: incremental development under statistical quality control, the cleanroom approach to establishing and maintaining management control of a software system that is developed in stages. In the incremental development process (Mills, 1971), at every stage of development one has a workable system to which additional features are added and verified sequentially. The process permits systematic incorporation of changes to requirements throughout the development cycle, which is crucial in an evolutionary context.

INCREMENTAL DEVELOPMENT

Large software systems are organized collections of parts. The way a software system is put together from these parts has a critical impact on project success. Incremental development is a top-down approach to development in which a software system is developed and tested as a succession of cumulative subsets of function. A minimal system is developed in the first increment, and function is added in each successive increment until the system is complete.

Each increment implements one or more end-user functions. Each increment contains all previously developed capability plus this new capability, so the system is "grown" in cumulative increments. Incremental development enables intellectual control over system development through referential transparency. (Referential transparency essentially implies the use of pluggable components with a fixed interface.) When referential transparency holds, a system part can be implemented from its subspecification with no need for backtracking; there is no rework of previous increments. This strategy enables correctness verification of each increment in a complete system context. This referential transparency clarifies one of the advantages of having system requirements for an evolutionary system known in advance of development.

Cleanroom incremental development permits continual integration of referentially transparent user-function increments over the entire development life cycle. Because the design of each increment is based on a verified subspecification and tested interface in a prior increment, deep design and interface errors should be rare. The system evolves in well-defined increments throughout the development. Testing and certification activities begin early in the development cycle.

Incremental development as practiced in cleanroom also provides a basis for statistical process control. Each cleanroom increment is a complete cycle of the process, involving specification, development, and verification of new user function plus testing of all work completed to date. This allows for measures of performance in each iteration of the process to be compared with performance targets to determine whether or not the process is "in control" (i.e., performing as expected). If standards are not met, the team can examine performance data from the increment to identify problems, adjust project plans if necessary, and modify the software development process to prevent recurrence of the problems identified. For example, if testing of an increment reveals that the process is "out of control"

(i.e., quality standards are not met), testing ceases and developers return to the design stage. If the process is in control, work on the next increment can continue. Feedback produced in each increment is also used to improve the process in the next increment.

The incremental development process enables early and continual feedback by customers on the executing functionality of an evolving system, to permit changes if necessary. Because the increments execute in a system environment and represent subsets of user function, early increments can be exercised by users for feedback on system functionality and usability. Such feedback helps avoid developing the wrong system and builds user acceptance of the eventual product. (This technique is consistent with some of the early interest in having requirements developed through the interaction of system developers and system users.)

Most crucially, in the context of evolutionary acquisition, incremental development allows systematic accommodation of inevitable changes in system requirements and the project environment. At the completion of each increment, the impact of accumulated changes in system requirements can be assessed in terms of current specifications and increment designs. If changes are isolated to future increments, they can often be incorporated into the existing incremental development plan, with possible adjustments to schedules and resources. If changes affect completed increments, a modified system development can be started from the top down, usually with substantial (often total) reuse of code from existing increments, with adjustments to schedules and resources as required.

The overall objective is to develop a system with each new increment as an elaboration of the functions implemented in prior increments. That is, new functions in an increment should "plug in" to the previous increment at predefined points in its structure, and they should satisfy the subspecifications associated with the processing requirements at those points. Within the framework of functional dependencies exhibited by a system, incremental planning is also influenced by a wide range of management and technical factors in a project. For example, the user may wish to place certain system functions into operational use prior to system completion. Such functions are likely candidates for early increments. (For more details on incremental development, see Trammell et al., 2005.)

An important motivation of the use of incremental development methods is the fact that requirements can rarely be established with certainty at the outset of a project. Under incremental development, customers provide feedback on an evolving system by direct operation of user-executable in-

crements. The relative clarity of requirements may influence an increment plan in two ways. Volatile requirements may be implemented in an early increment, so they can be clarified. Alternatively, unstable requirements may be planned for later implementation, when questions affecting the requirements have been settled.

Increasingly, customers in the commercial sector are specifying formal software reliability requirements. There are methods that can be used to compute the reliability needed for each subsystem to achieve a system-wide reliability. Subsystems with the highest reliability requirements may be candidates for early increments. A functional usage distribution is developed as part of a top-level cleanroom specification. Expected usage probabilities of system functions are established from historical data and estimates provided by customers. System functions with high expected usage probabilities will receive greatest exposure in the field, and they should therefore benefit from the greatest exposure to testing. Since increments are cumulative, the functions developed in early increments will be tested every time a new increment enters the testing process. System functions expected to receive the greatest operational usage by customers are therefore candidates for early increments. Some functions expected to receive low usage may even be regarded as optional and scheduled for development in the final increment if time permits.

Systems that involve both hardware and software must be developed as a coordinated effort between hardware and software engineers, and incremental development is an ideal framework for this coordination.

APPENDIX

D

Workshop Agenda and Attendees

AGENDA

Workshop on Testing for Dynamic Acquisition of Defense Systems
December 13-14, 2004
National Academy of Sciences Building
2100 C Street, NW
Washington, DC

Monday, December 13, 2004

Lecture Room

8:30 a.m.	Introduction and Overview *Vijay Nair (University of Michigan)*
8:40	Setting the Stage: Definitions and Intentions *Michael W. Wynne (Acting Under Secretary of Defense for Acquisitions, Technology and Logistics)*
9:00	Floor Discussion

9:10	The Role of Testing in the Proposed New Evolutionary Paradigm: DOT&E's Perspective *Thomas P. Christie (Director, Operational Test and Evaluation)*
9:30	Floor Discussion
9:40	Relevant Best Practices in Industry GAO Studies on Commercial Approaches to Incremental Acquisition: What Works and What Doesn't *Katherine Schinasi (Managing Director, Acquisition and Sourcing Management, GAO)*
10:00	Break
10:15	Speakers from Industry: *Art Koehler and Charley Eberhard (Procter & Gamble)*
10:35	*Tim Davis (Ford Motor Company)*
10:55	*Stan Young (NISS)*
11:15	*David Kelly (Battelle)*
11:35	Floor Discussion
11:45	Comments on Evolutionary Acquisition and Testing *Philip Coyle (Defense Consultant and Senior Advisor, Center for Defense Information)*
12:05 p.m.	Floor Discussion
12:15	Lunch
1:15	Techniques for Test Design in the Context of Staged Acquisition *Steve Vardeman (Iowa State University)*
1:35	*Donald Gaver (NPS)*
1:55	*Jesse Poore (University of Tennessee-Knoxville)*
2:15	*Alyson Wilson (LANL)*
2:35	Floor Discussion

2:45	Infrastructure Issues: Combining Information, Data Warehousing, and Summary of Previous Panel Recommendations *Steve Pollock (University of Michigan)*
3:05	Floor Discussion
3:15	Break
3:30	Panel Presentation: Perspectives on Testing in Evolutionary Acquisition from the Service Test Agencies *Steven K. Whitehead* (Technical Director, OPTEVFOR)*
3:40	*Frank J. Apicella (Technical Director, AEC)*
3:50	*Doug Marlowe (Technical Director, AFOTEC)*
4:00	Floor Discussion
4:10	Panel Discussion—Potential Recommendations and Further Activities *Ernest Seglie (DOT&E)* *Seth Bonder (The Bonder Group)* *Katherine Schinasi (GAO)* *David Kelly (Batelle)* (Moderator: *Vijay Nair*)
4:50	Floor Discussion
5:15	Adjourn

Tuesday, December 14, 2004

Room 250

9:00 a.m.	Participant Presentations and Structured Discussion
12:00 p.m.	Lunch
1:00	Executive Session (closed)
4:00	Adjourn

*Could not attend.

WORKSHOP ATTENDEES

Christine Anderson-Cooke, Los Alamos National Laboratory
Frank Apicella, Army Evaluation Command
Sheila Bahner, Office of the Secretary of Defense for Acquisition, Technology and Logistics
Charles Bedard, Army Test and Evaluation Command
Don Berry, M.D. Anderson Cancer Center
Seth Bonder, The Bonder Group
Bruce Braun, National Research Council
Herman Chernoff, Harvard University
John Christie, Logistics Management Institute
Tom Christie, Office of the Secretary of Defense–Operational Test and Evaluation Directorate
Jerry Clark, Government Accountability Office
Mike Clarke, National Research Council
Michael Cohen, National Research Council
Philip Coyle, Center for Defense Information
Steve Daly, Office of the Secretary of Defense–Operational Test and Evaluation Directorate
Paul Davis, RAND
Tim Davis, Ford Motor Company
Charley Eberhard, Procter & Gamble
John Foulkes, Office of the Secretary of Defense for Acquisition, Technology and Logistics
Arthur Fries, Institute for Defense Analyses
Don Gaver, Naval Postgraduate School
Lou Gordon, Susquehanna
Alan Karr, National Institute of Statistical Sciences
David Kelly, Battelle
Jerry Kitchen, Air Force Operational Test and Evaluation Center
Art Koehler, Procter & Gamble
David Maddox, United States Army (ret)
Doug Marlowe, Air Force Operational Test and Evaluation Center
Bill Meeker, Iowa St. University
Jim Morrison, Government Accountability Office
Vijay Nair, University of Michigan
Lloyd Pickering, Air Force Operational Test and Evaluation Center
Tom Plewes, National Research Council

Steve Pollock, University of Michigan
Jesse Poore, University of Tennessee
Rick Sayre, Office of the Deputy Under Secretary of the Army for Operations Research
Katherine Schinasi, Government Accountability Office
Ernest Seglie, Office of the Secretary of Defense–Operational Test and Evaluation Directorate
Arun Seraphin, Senate Armored Services Committee
Michael Siri, National Research Council
Nancy Spruill, Office of the Secretary of Defense for Acquisition, Technology and Logistics
Brian Taylor, Army Test and Evaluation Command
Steve Vardeman, Iowa State University
Alyson Wilson, Los Alamos National Laboratory
Michael Wynne, Office of the Secretary of Defense for Acquisition, Technology and Logistics
Stan Young, National Institute of Statistical Sciences

APPENDIX E
Biographical Sketches of Oversight Committee and Staff

VIJAY NAIR (*Chair*) is the Donald A. Darling professor of statistics and professor of industrial and operations engineering at the University of Michigan. He has been chair of the Department of Statistics since 1998. He was a research scientist at Bell Laboratories for 15 years before joining the faculty at Michigan. His area of expertise is engineering statistics, including quality and productivity improvement, experimental design, reliability, and process control. He is a fellow of the American Association for the Advancement of Science, the American Statistical Association, and the Institute of Mathematical Statistics, as well as an elected member of the International Statistical Institute. He is a former editor of *Technometrics* and *International Statistical Review* and has served on many other editorial boards. He is currently the chair of the board of trustees of the National Institute of Statistical Sciences and a member of the Committee on National Statistics at the National Research Council (NRC). He has been a member of several NRC panels, including the Panel on Statistical Methods for Testing and Evaluating Defense Systems and the Assessment Panel on NIST's Information Technology Center. He has a Ph.D. in statistics from the University of California, Berkeley.

SETH BONDER is the founder of Vector Research, Incorporated, and was chairman/chief executive officer until 2001. The firm provides analysis and information technology services to national security, health care delivery, and financial enterprises in the public and private sectors. His area of

expertise in the field of systems, policy, and operations analysis is the development of new procedures and their application to defense and health care enterprises. He has advised senior management in the Department of Defense and defense industries, including various secretariats, service chiefs of staff, commanders of unified commands, and the Joint Chiefs of Staff. He has recently worked on studies to enhance the protection of U.S. military forces overseas against terrorist attacks and to restructure Army forces to improve their capability and versatility to perform a broader spectrum of missions over the next two decades. He is a member of the U.S. Army Science Board and a past president of the Military Operations Research Society. He is an INFORMS fellow and a member of the National Academy of Engineering. He has a Ph.D. in industrial engineering (operations research) from Ohio State University.

JOHN D. CHRISTIE is a senior fellow at the Logistics Management Institute with an extensive background in defense acquisition policy and program analysis. From 1989 to 1993, he served as director of acquisition policy and program integration for the under secretary of defense (acquisition), directing the preparation of a comprehensive revision to all defense acquisition policies and procedures. While there he also prepared comprehensive acquisition program alternatives for the secretary of defense that resulted in multibillion dollar budget reductions. As a member of the Army Science Board in the 1980s, he directed a review of all Army analytical community and operations research activities that supported the overall Army acquisition process and its integration with the programming and budgeting process. In the 1990s he served on the staff of the Commission on Roles and Missions of the Armed Forces that provided recommendations to improve defense management. He was a member of the Panel on Statistical Methods for Testing and Evaluating Defense Systems of the Committee on National Statistics. He has S.B., S.M., E.M.E., and Sc.D. degrees from the Massachusetts Institute of Technology, all in mechanical engineering.

MICHAEL L. COHEN (*Study Director*) is a senior program officer for the Committee on National Statistics. He previously assisted the Panel on Estimates of Poverty for Small Geographic Areas and directed the Panel on Statistical Methods for Testing and Evaluating Defense Systems. Formerly, he was a mathematical statistician at the Energy Information Administration, an assistant professor in the School of Public Affairs at the University

of Maryland, and a visiting lecturer in statistics at Princeton University. His general area of research is the use of statistics in public policy, with particular interest in the census undercount, model validation, and robust estimation. A fellow of the American Statistical Association, he has a B.S. in mathematics from the University of Michigan and M.S. and Ph.D. degrees in statistics from Stanford University.

ARTHUR FRIES is a staff member and project leader at the Institute for Defense Analyses (IDA), where he has been employed since 1982. His recent work experience there includes leading studies on such topics as operational test and evaluation of antiarmor defense systems for the director of operational test and evaluation, validation of operational test and evaluation simulation models, counternarcotics strategies and data trends for the Department of Defense and the Office of National Drug Control Policy, and counterterrorism strategies and technology assessments for the Department of Homeland Security. His statistical research interests include minimum aberration designs, properties of the inverse Gaussian distribution in application to fatigue modeling and experimental design, and reliability growth methodologies. He served as a member of the NRC's Committee on Commercial Aviation Security. He has published widely in numerous journals, including *Technometrics*, the *Journal of the American Statistical Association*, *Statistics & Probability Letters*, and the *IEEE Transactions on Reliability*. In 1999, he was elected a fellow of the American Statistical Association. He has an M.A. in mathematics and a Ph.D. in statistics, both from the University of Wisconsin-Madison.

STEPHEN POLLOCK is Herrick emeritus professor of manufacturing and professor in the Department of Industrial and Operations Engineering at the University of Michigan. His research interests are in mathematical modeling, operations research, and Bayesian decision theoretic methods. A member of the National Academy of Engineering and a former member of the U.S. Army Science Board (1994-1999), he was a member of the National Research Council's Committee on Applied and Theoretical Statistics, the Panel on Statistical Methods for Testing and Evaluating Defense Systems of the Committee on National Statistics and he also chaired the Panel on Operational Test Design and Evaluation of the Interim Armored Vehicle. In addition to his career at the University of Michigan, he spent four years at the U.S. Naval Postgraduate School. He has S.M. and Ph.D.

degrees in physics and operations research from the Massachusetts Institute of Technology.

JESSE H. POORE holds the Ericsson/Harlan D. Mills chair in software engineering in the Department of Computer Science at the University of Tennessee. He is also director of the University of Tennessee–Oak Ridge National Labs Science Alliance, a program to promote and stimulate joint research between the University of Tennessee and Oak Ridge National Labs, in order to manage joint programs and encourage interdisciplinary collaborations. He conducts research in cleanroom software engineering and teaches software engineering courses. He has held academic appointments at Florida State University and Georgia Tech; he 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 the Association for Computing Machinery and the Institute of Electrical and Electronics Engineers and a fellow of the American Association for the Advancement of Science. He was a member of NRC's Panel on Statistical Methods for Testing and Evaluating Defense Systems. He has a Ph.D. in information and computer science from Georgia Tech.