Reliability Issues for DOD Systems: Report of a Workshop

Committee on National Statistics, Francisco Samaniego and Michael Cohen,Editors, National Research Council

ISBN: 0-309-50515-1, 104 pages, 6 x 9, (2002)

Reliability Issues for DoD Systems

REPORT OF A WORKSHOP

Committee on National Statistics

Francisco Samaniego and Michael Cohen, *Editors*

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.

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

International Standard Book Number 0-309-08606-X

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 (2002) *Reliability Issues for DoD Systems: Report of a Workshop*. Committee on National Statistics. Francisco Samaniego and Michael Cohen, editors. 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. Bruce M. Alberts is president of the National Academy of Sciences.

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

www.national-academies.org

Preface and Acknowledgments

The workshop with which this volume is concerned has a number of important historical antecedents. The Department of Defense (DoD) has long been keenly interested in quantitative analyses assessing the performance of the hardware and software used in its various activities. Its interest in the possibility of revisiting the statistical techniques emphasized in its handbooks and manuals since the early 1960s began in earnest with a DoD-sponsored workshop hosted by the National Academies in September 1992. That workshop served as the stimulus for a study by a panel of the Committee on National Statistics, National Research Council. The panel's report, *Statistics, Testing and Defense Acquisition: New Approaches and Methodological Improvements* (1998) strongly advocates various forms of modernization in the statistical practices used in the various stages of the DoD acquisition process.

Two individuals within DoD played key roles in progressing to the essential next stage: active engagement between the statistical research community and DoD statisticians, engineers, and managers, with a view toward developing a common understanding of the agency's current needs and the statistical research community's most promising methods for addressing those needs. I wish to thank Nancy Spruill, Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, and Ernest Seglie, Science Advisor to the Director of Operational Test and Evaluation, for recognizing the urgency of the problem facing DoD, for raising the level of consciousness about that urgency within the department and among

its own statistical contacts, and for their energetic work in developing the financial support for a new series of workshops on statistical topics of importance to DoD. In the workshop that is the subject of this report—the first in the series—they took a leading role in the identification of potential expert discussants from the defense community. Finally, Drs. Spruill and Seglie put the workshop organizers in touch with other scientists and administrators within DoD who served us as indispensable advisors on the organization and planning of the workshop. Among these, Allen Beckett, Dolores Etter, Hollis Hunter, Robert Nemetz, and Philip Rodgers merit special mention.

I would also like to acknowledge the very special contributions made by a distinguished group of statistical researchers who kindly accepted our invitation to make presentations at the workshop. Each of them took the time to engage in early discussions with DoD personnel working in the areas on which they would speak; this preliminary contact served to enhance both the relevance and the impact of their comments and contributions at the workshop. We gratefully acknowledge the contributions of each of the civilian presenters and discussants at the workshop: Wallace Blischke, University of Southern California; Jane Booker, Los Alamos National Laboratory; Frank Camm, the RAND Corporation; Larry Crow, General Dynamics; Siddhartha Dalal, Telcordia Technologies; Robert Easterling, Sandia National Laboratories; Donald Gaver, Naval Postgraduate School; William Meeker, Iowa State University; William Padgett, University of South Carolina; Stephen Pollock, University of Michigan; Jesse Poore, University of Tennessee; Ananda Sen, Oakland University; Samuel Saunders, Washington State University; Fritz Scholz, the Boeing Company; Duane Steffey, San Diego State University; and Michael Tortorella, Bell Laboratories. I also thank John Rolph and Bill Eddy of the Committee on National Statistics, who served with me as the organizing subcommittee for the workshop.

The workshop also benefitted greatly from extremely informative remarks by its defense community discussants: Allen Beckett, Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics; James Crouch, Wright-Patterson Air Force Base; Paul Ellner, Army Materiel System Analysis Activity; Jack Ferguson, Software Intensive Systems; Arthur Fries, Institute for Defense Analyses; Walter Hollis, Office of the Deputy Under Secretary of the Army; Fred Myers, Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics; Margaret Myers, Office of the Deputy Assistant Secretary of Defense; Ernest Seglie, Office of the Director of Operational Test and Evaluation; Nancy Spruill,

Office of the Under Secretary of Defense for Acquisitions, Resources and Analysis; and Marion Williams, Air Force Operational Test and Evaluation Command.

Several invited guests came to our rescue by serving as session chairs: Asit Basu, University of Missouri-Columbia; Henry Block, University of Pittsburgh; Philip Boland, University College, Dublin; Ronald Glaser, Lawrence Livermore National Laboratory; Patricia Jacobs, Naval Postgraduate School; and Simon Wilson, Trinity College, Dublin.

This workshop 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 its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this report: Luis Escobar, Department of Statistics, Louisiana State University; Hans Mark, Department of Aerospace Engineering, University of Texas; Vijay Nair, Department of Statistics, University of Michigan; and Nozer Singpurwalla, Department of Operations Research, George Washington University.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report nor did they see the final draft of the report before its release. The review of this report was overseen by William F. Eddy, Department of Statistics, Carnegie Mellon 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 authors and the institution.

I would also like to thank Arthur Fries, Institute for Defense Analyses, who contributed greatly to the workshop by helping to identify speakers and to orient their presentations so they were as compatible as possible with the workshop goals. In addition, Art carefully read several drafts of the present report.

Finally, I thank Michael Cohen, Agnes Gaskin, and Julia Kisa, staff of the Committee on National Statistics. Michael Cohen did a yeoman's job

as chief liaison between DoD and the Committee on National Statistics, and more than anyone else provided the impetus for the steady advance of both the scientific and practical issues associated with a successful workshop report. Michael developed the first draft of this report and has been a full partner with me since the beginning of this ambitious project. He is to be commended for his strong dedication to serving simultaneously the highest standards of the discipline and the most pressing needs of DoD. Agnes Gaskin, greatly assisted by Julia Kisa, handled all administrative details with great care and good humor. Finally, Eugenia Grohman managed the production of this report, including the enlistment of Rona Brier as technical editor, who served superbly in that role.

Francisco Samaniego, *Chair*
Workshop on Reliability Issues for
Defense Systems

Contents

1

Introduction and Overview

The final report of the National Research Council's (NRC) Panel on Statistical Methods for Testing and Evaluating Defense Systems (National Research Council, 1998) was intended to provide broad advice to the U.S. Department of Defense (DoD) on current statistical methods and principles that could be applied to the developmental and operational testing and evaluation of defense systems. To that end, the report contained chapters on the use of testing as a tool of system development; current methods of experimental design; evaluation methods; methods for testing and assessing reliability, availability, and maintainability; software development and testing; and validation of modeling and simulation for use in operational test and evaluation. While the examination of such a wide variety of topics was useful in helping DoD understand the breadth of problems for which statistical methods could be applied and providing direction as to how the methods currently used could be improved, there was, quite naturally, a lack of detail in each area.

To address the need for further detail, two DoD agencies—the Office of the Director of Operational Test and Evaluation and the Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics—asked the NRC's Committee on National Statistics to initiate a series of workshops on statistical issues relevant to defense acquisition. The aim of each workshop is to inform DoD about the methods that represent the statistical state of the art and, through interactions of the statisti-

cal and defense communities, explore their relevance for DoD application.

WORKSHOP ORGANIZATION AND GOALS

The issue chosen for this first workshop was statistical methods for assessment of the reliability (including availability and maintainability) of defense systems. A list of potential topics for the workshop evolved from issues raised in Chapter 7 of the above-referenced NRC (1998) report; discussions with the sponsors identified other topics that are a high priority for DoD but were not specifically addressed in that report. Further discussions of the Committee on National Statistics subcommittee responsible for organizing the workshop narrowed these topics down to seven that were selected for focus: (1) global reliability test designs to direct defense system development; (2) recent developments in reliability growth modeling; (3) use of statistical modeling for combining information, with applications to developmental and operational testing; (4) methods for estimating reliability from field performance data; (5) modeling of fatigue; (6) reliability of software-intensive systems; and (7) assessment of life-cycle costs through the use of reliability analyses (see the Appendix for the workshop agenda).

For each topic, two speakers from either academia or industry were identified to provide presentations. In addition, one or more defense specialists with responsibilities directly relevant to each topic were identified. These individuals were asked to interact with the (nondefense) presenters in advance of the workshop to ensure that the presentations would reflect cognizance of the specific problems faced by DoD; the current methods used by DoD to address these problems; and the important, possibly unique, constraints under which DoD operates. They were also asked to serve as discussants for the relevant workshop session.

The Workshop on Reliability Issues for DoD Systems, held June 9-10, 2000, had multiple goals, partly because some of the described techniques are mature, while others are still undergoing active research. In addition, the intended audience for the workshop and this report comprises defense reliability experts, higher-level administrators who could help change the processes used in system development, and defense employees charged with the day-to-day responsibility of assessing the reliability of defense systems. With respect to this last segment of the intended audience, an important consideration for the workshop to address was how to communicate the more readily applied and broadly applicable techniques to those in the DoD

community who have the responsibility of evaluating the reliability of defense systems and who, by nature of other responsibilities and backgrounds, are generally less knowledgeable about statistical techniques than academic researchers. One possible way to facilitate such communication is to upgrade or replace DoD 3235.1-H, *Test and Evaluation of System Reliability Availability and Maintainability: A Primer* (issued in 1962, revised in 1982). This document, referred to henceforth as the RAM Primer, is commonly used (more so in some service test organizations, less in others) by those responsible for the design of tests of system reliability and the associated evaluation in the test service and other defense agencies.

Broadly considered, the goal of the workshop was to foster greater interaction between the academic and defense acquisition communities with regard to both those ideas whose applicability is still uncertain and those that are considered promising. A number of positive impacts for the DoD community were envisioned by the planners of the workshop. First, it was hoped that greater interest would be developed among the academic community in current issues of importance to the defense community involving reliability assessment. Another goal was to inform decision makers about ways of dealing with procedural or other constraints that may hamper the application of statistical techniques in the collection of and access to test and field data, and in the use of testing in the development of reliable systems. Finally, the workshop would acquaint the defense community with state-of-the-art techniques that are applicable to problems in defense system development.

The planned interactions between the statistical research community and DoD were expected to have, in equal measure, strong benefits for the participating researchers. Prominent among these was educating academics about the problems and constraints facing the defense acquisition community, which are often considerably different from those involved in analogous industrial applications.

EIGHT KEY IDEAS

The following eight key ideas represent a useful summary of the workshop sessions: (1) the advantages of methods for reliability growth management, (2) the benefits of broader understanding and use of modern methods for reliability estimation and testing, (3) the need for updating the RAM Primer, (4) gains from the use of alternative modeling approaches, (5) the advantages of state-of-the-art reliability growth models, (6) the po-

tential advantages of the application of methods for combining developmental and operational test information, (7) the development of statistical models of fatigue of materials, and (8) the need for greater use of physics-of-failure models and for modeling some failure sources separately. A short synopsis of each of these eight key ideas is presented below.

Benefits of Methods for Reliability Growth Management

A general issue noted by many workshop participants is that defense systems do not satisfy their operational suitability requirements sufficiently often, and as a result DoD is spending too much for system redesigns, spares management, and maintenance. Speakers stressed that a change in emphasis is needed to address this problem, including greater focus on test and evaluation for suitability, but more important, use of a number of techniques that can help identify design flaws and provide assessments of reliability performance much earlier in system development. This problem relates to the following NRC (1998:105) recommendation:

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

A number of speakers stressed that much progress could result not only from technical changes, but also from management changes that would support a more comprehensive approach to reliability improvement. Such changes would include the following: (1) a change in the function of reliability assessment from that of a statistic used to support promotion decisions to an early and continuing objective measurement (combining a wide variety of types of information) that supports system development by helping to identify components in need of redesign or maturation; (2) the collection of information on system performance in the field to support assessment of life-cycle costs and therefore future decisions on system acquisition; (3) cataloguing of test information and field performance to support feedback loops and thereby improve system performance, as well as the design of future tests; (4) early detection of processes in trouble or "bad actors" (defined below); and (5) development of a better understanding of the relationship between reliability performance in developmental and op-

erational test. Many of these issues relate to the following NRC (1998:120-21) recommendation:

> **Recommendation 7.8: All service-approved reliability, availability, and maintainability data, including vendor-generated data, from technical, developmental, and operational tests, should be properly archived and used in the final preproduction assessment of a prospective system. After procurement, field performance data and associated records should be retained for the system's life, and used to provide continuing assessment of its reliability, availability, and maintainability characteristics.**

Benefits of Broader Understanding and Use of Modern Methods for Reliability Estimation and Testing

Various speakers pointed out areas in which the DoD test and evaluation community could make greater use of modern methods for modeling various aspects of reliability and the advantages of doing so. Examples included (1) methods for combining information across test environments; (2) methods for incorporating subjective assessments; (3) fatigue modeling; (4) statistical methods for software engineering; (5) wider use of nonparametric methods, specifically for reliability growth but also more generally (e.g., for models for combining information and for uncertainty analysis); and (6) alternative methods for modeling reliability growth.

Need for Updating the RAM Primer

Given that the RAM Primer has been used to disseminate reliability methods throughout the defense test community for nearly 40 years and that its current version is substantially out of date with respect to a wide variety of currently accepted techniques, a number of speakers strongly suggested that it be updated, possibly in a substantially different format. One possibility mentioned was to have the RAM Primer be a web-based document with embedded software for carrying out the variety of calculations necessitated by modern methods. (This concern about datedness applies to other reliability-related military handbooks and military standards that are of similar vintage.) These suggestions relate to the following NRC (1998:126) recommendation:

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

Gains from the Use of Alternative Modeling Approaches

Many speakers pointed out that the defense test community relies on particular models for specific purposes, the key examples being the power law process in reliability growth estimation and the exponential time-to-failure distribution for a wide variety of reliability questions. In these and other areas, use of alternative models would have the benefits of helping to identify when current models can be improved, to identify situations in which the inference is and is not robust across the different modeling approaches, and to provide an indication of the uncertainty due to model misspecification. Nonparametric models are particularly useful for these purposes since they make relatively few distributional assumptions and are therefore generally applicable. Since it is important that estimates of the uncertainty of reliability estimates take into account as many sources of uncertainty as possible for the benefit of decision makers, it is valuable to use alternative modeling approaches, which can be used to provide an assessment of the uncertainty of estimates due to model misspecification. These suggestions reinforce the following NRC (1998:113) recommendation:

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

the choice of statistical models used in both the design and the analysis of operational suitability tests.

Advantages of State-of-the-Art Reliability Growth Models

Speakers suggested the wider use of reliability growth models that are consistent with the maturation process of test, analyze, and fix. Otherwise, since the power law process and related approaches do not explicitly take into account the results of the process of discovering and fixing the faults found in testing, reliability growth models are liable to produce inaccurate predictions. Two models were proposed that assign probabilities of detection to various system faults. These models use the natural assumption that earlier testing is more likely to find errors that have a higher probability of being discovered, whereas subsequent testing is more likely to discover the less probable errors. Such approaches are consistent with the test, analyze, and fix process. They can therefore can be used to examine the maturation that would result from a specific test design given various characteristics of the system under test, and they can provide an estimate of the reliability of the "matured" system.

Potential Advantages of the Application of Methods for Combining Developmental and Operational Test Information

Speakers described several new approaches based on the use of models for combining information from developmental and operational test when the failure modes in these separate environments of use are well understood (or otherwise satisfy the necessary assumptions underlying the models). In those instances, use of these models can make operational tests much more informative and thereby save test funds in comparison with methods that do not combine information. Use of this approach relates to the following NRC (1998:119) recommendation:

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

proviso that analyses in which this is done be carefully justified and defended in detail.

Development of Statistical Models of Fatigue of Materials

Especially given the recent use of defense systems for longer periods of time, fatigue modeling is clearly of critical importance to DoD. Speakers pointed out that the use of fatigue modeling that derives from a statistical approach provides useful estimates in a variety of important and common DoD contexts. Special emphasis was placed on the need for understanding the science in applications involving fatigue. Presentations during the session on fatigue modeling included both illustrations of the necessary statistics/physical science partnership and descriptions of a variety of models motivated by recent research in materials science. Speakers suggested that fatigue modeling could be one of the techniques included in a revised RAM Primer.

Need for Greater Use of Physics-of-Failure Models and for Modeling Some Failure Sources Separately

Several speakers supported the greater use of physics-of-failure models (i.e., models that directly represent the physical basis for failure) whenever these approaches are applicable. Use of such models would generally provide better estimates of various characteristics of system failure as compared with models not linked to specific failure modes. Discussion of the benefits of physics-of-failure models, along with leading examples, could also be included in a revised RAM Primer.

Further, several speakers proposed dividing failures into those of mature components and those of immature components, the latter through either design or production flaws. For example, the Integrated Reliability Growth Strategy classifies design failures into type A and type B modes, i.e., associated with mature and immature design components. Further, failures in the field are typically overrepresented by poorly produced components, referred to as "bad actors." The identification and separate modeling of failures from type B modes and from bad actors is an approach to reliability estimation that could be effective in providing better estimates of system and fleet reliability. One speaker also mentioned "special-cause" failures—those that are unpredictable and are due, for example, to changes

in manufacturing processes. The study of special-cause failures could also benefit from separate treatment for modeling and prediction.

SCOPE AND ORGANIZATION OF THIS REPORT

As is true with most workshop reports, this report is intended to capture the flavor of the workshop and highlight its primary and most useful ideas and discussion. While the report does not represent a comprehensive transcript of the proceedings, it should certainly serve as a helpful guide to current trends in reliability research and practice that have special relevance to DoD applications. We recognize that much more could be said on statistical modeling, reliability, and system development, perhaps delving into such topics as experimental design for reliability measurement, estimation of reliability for highly reliable systems (so that the probability of not observing a failure during testing is sizable), complications posed by destructive measurements, repeated measure degradation studies, analysis of recurrent events, and use of simulation-based techniques. Perhaps some or all of these topics will be addressed in future workshops in this proposed series.

The next chapter examines methods for measuring and managing reliability growth that were presented at the workshop. This is followed by a chapter on important areas of current research in reliability modeling and inference. The final chapter presents a general discussion of reliability issues and examines the need, expressed on several occasions during the workshop, to disseminate well-understood, broadly applicable methods for reliability test and evaluation, possibly through a comprehensive overhaul of the RAM Primer and other DoD documents that focus on reliability.

2

The Measurement and Management of Reliability Growth

Reliability growth is not a new topic in either engineering or statistics. It has been the subject of intense investigation and spirited application at least since the early 1960s. While the area has been recognized as important in both industrial and military settings for some time, it is featured in this summary, as it was at the workshop, for two important reasons. One is that the area has evolved substantially over the past four decades, yet its utility and power in modern applications do not appear to be widely recognized. A more important reason is that the latest approaches to reliability growth involve a sea change in perspective—from a focus on the measurement or estimation of observed growth to an emphasis on the interdisciplinary collaborations and opportunistic interventions that combine to assist in the identification and understanding of system faults and the creation and attainment of reliability growth goals.

This chapter begins with a brief review of the history of reliability growth estimation. It then proceeds to a discussion of six presentations at the workshop specifically dedicated to the subject—four addressing tools for measuring reliability growth and two reviewing tools for managing reliability growth. The treatment here alters the order in which these presentations were made to roughly parallel the historical review.

HISTORY OF RELIABILITY GROWTH

The historical account provided here of the theory and applications of reliability growth is necessarily brief. The interested reader is directed to

Crow (1984), Jewell (1984), or Ushakov (1994) for a more detailed description of the models and methods mentioned in this chapter, as well as for discussion of related topics that we chose not to feature here.

While there were both formal and informal developments in the area of reliability growth prior to 1964, the field as we know it today had its beginning that year with a highly influential paper by J. T. Duane. The subject of the paper was the observed growth in reliability of specific manufactured items related to the aerospace industry. A simple regression analysis appeared to suggest that the logarithm of the cumulative failure rate of an item at time t was linearly related to the logarithm of t, a relationship that might be expressed as

$$\ln(\lambda_t) = a - b \ln(t).$$

The coefficient b of $\ln(t)$ appeared to vary from one application to the next, depending, for example, on whether the item in question was mechanical or electronic, but the fit of the Duane model in a large collection of quite different applications appeared truly uncanny. Duane's application called for a coefficient of $b = 0.5$, but as applications proliferated, it was observed that b would generally fall in the interval [0, 0.6]. One famous military application of the Duane curve was to the failure rate of the F15-A fighter when its performance was tracked for 4 years in the mid-1970s.

The Duane model gained substantial popularity through the 1960s and 1970s. It appeared to fit reliability growth processes well enough that attempts were initiated to predict the future reliability of an item based on its fitted Duane curve. Such a practice was a bold move indeed, given that the fitted curve offered no explanation of the concomitants of growth, providing no understanding of the growth process itself. Surely the various interventions that occurred as a prototype was developed and improved were somehow linked to the reliability improvement one would experience, but such interventions played no formal role in the Duane model. The model appeared to be saying that it matters little what one does (as long as one does *something*); the improvement seen will follow a Duane curve.

In the early 1980s, Larry Crow, working at the Army Materiel System Analysis Activity (AMSAA), developed a modification of the Duane model that proved to be substantially more flexible. The essence of the AMSAA model was that it was in reality a collection of successive models. It was

recognized that the Duane model tended to apply to the data locally, but that the parameters of the model might well change following a major intervention, giving rise to a new model that would be applicable for the period during which the newly configured prototype was in use. Another extension of note was the use of the Weibull or other well-known parametric distributions for the modeling of failure data. Models with an inherent monotone failure rate, of which the Weibull and the gamma models are the best known, are natural for modeling improvement or deterioration in an item or system of interest and are thus useful tools in modeling reliability growth. Much of the work of this era has been summarized in various military handbooks and codified as military standards. While this work was aimed at allowing for different and unpredicted changes in reliability due to a series of interventions, the focus was still on measuring growth rather than trying to understand its root causes.

The next stage of development of reliability growth modeling involved its application over longer periods, occasionally extending to an item's entire life cycle. Another refinement of interest was the inclusion of covariates that could be used to help predict future performance—covariates that might describe the maintenance process, the level of usage, and the like. An example of the development of models with such features is found in Collas (1991).

TOOLS FOR MEASURING RELIABILITY GROWTH

Four papers were presented in the general area of reliability growth measurement, covering both classical and modern methods in the assessment of the extent to which reliability growth is realized in a developing system, and also exploring models for fault detection and removal.

In one session of the workshop, Ananda Sen was asked to provide a review of the evolution and current state of the "classical theory" of reliability growth. Sen's presentation follows directly from the preceding historical perspective. Following a summary of Ananda Sen's talk, we turn to the presentation made by Donald Gaver. Gaver's assignment was to "think outside of the box," discussing some interesting alternatives to the classical approaches to reliability growth. His presentation, which was based on his investigations into statistical modeling of fault detection and removal, provided a glimpse of a new, interesting, perhaps even radical approach to reliability growth. Gaver studied the performance of a strategy for operational testing in the context of military acquisition in which a prototype is

classified as acceptable when it experiences a suitably specified run of successes. While such methods are not yet in use, Gaver's talk clearly demonstrated that new, different, and promising procedures for acceptance testing are feasible and under development.

Discussion of reliability growth continued in the workshop session dedicated to models, methods, and applications involving the linkage between field performance data and reliability growth. Fritz Scholz described a model applicable to the detection and removal of design flaws in a fielded system and discussed a methodology for estimating and bounding system reliability at each stage of the fault discovery process. William Meeker then presented a series of examples drawn from his experience with field data in the automotive industry—examples that motivate and strongly support the continuous tracking of performance data once an item has been fielded.

After summarizing these four presentations, we turn to the important issue of the management of reliability growth. The presentations of Jane Booker and Larry Crow were both representative of the modern global approach to reliability growth, which incorporates the best features of the classical theory yet goes well beyond it, using ideas that are integrative, inter- and cross-disciplinary, and comprehensive.

The Sen Paper

Ananda Sen provided a review of recent developments in modeling and statistical inference for reliability growth. In typical modern applications of reliability growth theory, a system's reliability is improved through a series of test, analyze, and fix (TAAF) episodes. Reliability growth modeling is a collection of techniques that attempt to model the (hoped-for) increasing reliability of a system as a function of the time spent in development and other relevant variables. Reliability growth modeling has historically played a role in helping to determine whether a system in development is likely to meet reliability requirements in time for graduation to the next development phase, and eventually to operational testing. Sen focused his presentation on systems for which the relevant data input into reliability growth models consists of successive times to failure (that is, total test time). See Sen (1998) for more detail.

Of course, it is not clear whether the process of reliability improvement can be usefully modeled as a function of time alone, since time is an indirect measurement of the workings of the TAAF cycle. However, it is useful to attempt to do so since these models can be used to (1) monitor the

improvement in system reliability throughout the developmental testing stage, which can help in judging when to proceed to subsequent acquisition phases; (2) design more effective operational tests; and (3) potentially predict future reliability and idealized reliability achievement when decisions must be made, for example, between upgrading an existing system and switching to a new one.

A problem that deserves more attention is that system reliability is dependent on the environment of use. The reliability of a system in a cold, wet environment may be substantially different from its reliability in a hot, dry environment. Storage or transport may permanently affect a prototype's reliability. Further, and very important, the reliability of a system in typical developmental test circumstances, with expert users and without counterforces, can be dramatically different from the reliability of the same system in typical operational test circumstances with nonexpert users and deployment in more realistic combat settings. The modeling and measurement of reliability across environments of use is a complicated but important problem.

One class of models that has been used for reliability growth modeling is referred to as learning curve models, especially the power law process. The motivation for the latter model is that, plotted on a log-log scale, the empirical cumulative failure rate in practice has appeared to be linear in time. Equivalently, on the log-log scale, the count of total failures is linearly related to the total time on test. Sen pointed out that the power law process has the following advantages: (1) it is easy to work with analytically, (2) well-developed inferential procedures are associated with it, (3) it has an easily interpretable reliability growth (or decay) scenario, and (4) its validity can be tested with readily available goodness-of-fit procedures. On the other hand, it has the following disadvantages: (1) the error rate is assumed to decrease to zero as time increases (which may not be a problem if either the produced system is extremely reliable or the development process is short relative to the time the system is predicted to be extremely reliable); (2) given the use of a continuous representation of system reliability as a function of time, the failure rate after fixes have been incorporated is assumed to be essentially identical to the failure rate beforehand; and, most important, (3) the TAAF cycle is not explicitly represented in the model.

Sen then described an alternative model form that was developed to address the above deficiencies in the power law process model. This class of alternatives, represented as a step-function approximation to the power law process, was first proposed by Benton and Crow (1989) and is referred to

as the step-intensity model. It was developed to directly represent the TAAF cycle. These alternatives assume that the time to failure is exponentially distributed, but these distributions have mean times to failure that vary following some prescribed formulation linked to the TAAF process. Sen and Bhattacharya (1993) retained the power law form but gave it an interpretation that was linked more directly to the TAAF process—referred to as exponential reliability growth.

Other alternatives to the power law process exist to handle situations such as upward trends in the failure rate, situations in which the time to first failure can be infinite with positive probability, and failure-rate functions that have a bathtub shape (failure rate initially decreasing and subsequently increasing). A second distinct class of models is derived as solutions to differential equations. The defining equations represent the relationship between cumulative expected time between failures and nonlinear functions of time. Unfortunately, these procedures are complicated to use for purposes of statistical inference. A third distinct class of models makes use of a Bayesian formulation through which the subjective inputs of experts in appropriate disciplines can be elicited, quantified, and included in the analysis. Finally, there are nonparametric approaches to modeling reliability growth that are straightforward applications and generalizations of the Kaplan-Meier estimates used in survival analysis.

Clearly there is a wide variety of reliability growth models from which to choose. No single model is best for all purposes. Parametric models permit extrapolation to areas in which few or no failures have been observed, but they are based on assumptions that need to be validated or evaluated for robustness. Fully nonparametric models are essentially always valid but, for a fixed desired precision, typically require a substantial number of replications; they can also be inefficient relative to parametric alternatives when the relevant assumptions of the latter are approximately true.

Sen argued that it is important for decision makers to be provided with a full representation of the complete evolution of the bottom-line result, instead of a simplistic presentation of a single point estimate or an elementary pass/fail pronouncement. A full representation of the results should include some physical justification or validation of the model's assumptions (given that a parametric approach is used), together with use of a nonparametric approach for purposes of validation and comparison and, at times, results from leading parametric alternative representations. Agreement of alternative modeling approaches offers important assurance of the stability of the results presented. Disagreement is often indicative of the

failure of one or more of a model's assumptions and necessitates examination of each assumption concerning the failure process. The degree of disagreement can be used to measure the potential for model misspecification, which in turn can aid in informing decision makers as to the quality of the reliability growth estimation.

The Gaver Paper

Donald Gaver has led a research team in developing a new methodology that explicitly represents testing as part of system development for systems consisting of separate stages linked in a series structure. (The obvious example is a one-shot system, such as a missile or torpedo.) With this methodology (see Gaver et al., 2000), the failure mode discovery process that derives from testing is directly represented—i.e., system failures are activated based on an explicit random process—and the resulting impact on reliability growth is estimated. (Testing is assumed to be carried out at specific time periods, with no explicit representation of the actual length of time between tests.) The approach assumes that the only testing is full-system testing, with components operating during the test in the sequence natural to system use. Therefore, the later stages of the system are not tested if an earlier stage fails beforehand. This relative lack of testing of later stages of the process for staged systems is often ignored using current approaches for modeling reliability growth.

The simulation framework that has been developed to represent reliability growth explicitly can be used to answer a variety of important questions concerning system and test performance. Also, some of these questions can be addressed analytically. (The solution involves the use of various recursive identities.) Some questions that can be answered analytically concern the properties of stopping rules of the form "accept a system when it runs successfully r times in a row" for various values of r. This class of stopping rules is considered easy to apply and ensures that each stage of the system passes a test r times. Requiring r successes in a row helps control the "false acceptance rate," with the desired rate being achievable through the appropriate setting of r. Further, the system developers have an incentive to create as reliable a system as possible as early as possible to achieve a high probability of passing the test (either developmental or operational). The test design has the added advantage of focusing on success rather than failure, with poorer-performing systems being eliminated because of their inability to accumulate the requisite success run within a specified time frame

or within a fixed, predetermined budget. A specific measure of interest for these stopping rules is the expected number of test replications needed to pass a system. The following are some other questions of interest that this simulation structure can address:

- After a specified number of operational test events of the system (and associated fixes), what is the probability that the system will meet its reliability requirements when fielded?
- How many operational tests are likely to be required to achieve the *r*th successful test?
- How many operational tests are likely to be needed to achieve *r consecutive* successful tests? Other stopping rules besides *r* consecutive successful tests can also be examined.

The framework also provides the ability to address a wide variety of additional what-if questions.

To carry out an analysis of a test design, one selects parameters describing the number and probability of failure modes in each component of a system. One then inputs the test parameters (e.g., test size) and runs simulated test replications to estimate the operating characteristics of the proposed test design (i.e., the probability of improperly rejecting a good system and of improperly failing to reject a bad one).

Some of the mathematical details are as follows. This approach assumes that there are an unknown number of initial design faults $d_s(0)$ at each of the s sequential stages of the system and that the undiscovered (and therefore unfixed) faults are revealed (and removed through a redesign) with some unknown probability $1\text{-}\theta_s$. Further, at time t (that is, after the tth test), each of the s stages has $d_s(t)$ remaining faults, given the discovery and treatment of faults in earlier tests.

The model makes use of some simplifying assumptions: (1) when a failure occurs, the design source of the failure is always identified and removed, and no new failures of this type are introduced; (2) the process that exhibits faults in each component follows an independent binomial distribution with parameters $1\text{-}\theta_s$ and $d_s(t)$; (3) all of the faults at a given stage have equal probabilities of discovery; (4) the probabilities of fault discovery are not dependent on any environmental conditions, aging, and so on; and (5) if two faults for a given stage simultaneously express themselves, only one is identified and removed. While these assumptions are acknowledged to be

an oversimplification, more realistic versions of this approach can be (and are being) developed using straightforward generalizations of the above model.

One interesting implication of this research is that reliability growth under the assumed circumstances will not necessarily have the general pattern identified by Duane (1964). Consideration of subsystems tested in series with this framework could certainly lead to other patterns of reliability growth.

One possible generalization of this model is to place a Bayesian prior on the $d_s(0)$'s. Doing so would (1) allow the introduction of expert judgment, (2) reduce the assumptions concerning the $d_s(0)$'s to a small number of hyperparameters, and (3) allow some borrowing of information across components. Another generalization that could be explored would be to assume that the θ_s's were draws from some distribution, instead of assuming fixed parameters $1\text{-}\theta_s$. Doing so would (1) help account for overdispersion, (2) reduce the number of parameters to be estimated, and (3) remove the homogeneous failure rate assumption.

The Scholz Paper

Nonhomogeneous Poisson processes (Poisson processes in which the failure rate changes as a function of time) are commonly used for modeling reliability growth. As mentioned above, an extremely popular model is the Duane power law model, a particular nonhomogeneous Poisson process in which the failure rate is assumed to be a power function of time. A chief deficiency of the Duane model is that it is not based on a physical cause-and-effect connection between an observed pattern of system failures and reliability growth (as was noted in the previous section). To address this concern, Fritz Scholz proposed the following model for a system of defect detection and reliability growth. (This idea was originally developed in the context of software testing, but it can be applied to any system that satisfies the cited assumptions. The description provided here is for the continuous case; Scholz, 1986, provides more detail and also addresses the discrete case.)

Assume one wants to measure reliability for a system that is suspected of having a number of design flaws. To measure reliability, the system is subjected to a series of test events. The system is assumed to be a deterministic function of the inputs to the system. A test is the exercising of the system using a selected subset of the set of possible inputs to the system. For a software system, the inputs are user-supplied fields, such as keystrokes

and mouse movements. For a hardware system, the inputs can include the environment of use and the actions of friendly and enemy soldiers. A few of the design flaws are assumed to be easy to find in that many inputs are likely to expose them, while many of the flaws are assumed to be relatively difficult to find in that very few inputs will disclose their presence. That is, assuming that inputs are selected uniformly from the space of all possible inputs, a few design flaws will be discovered with high probability, and many more will be discovered with low probability.

Some mathematical details follow. The system is assumed to have N faults. The assumption is also made that the waiting time to discovery of fault i (i = 1, ... , N) can be modeled using independent, exponential random variables Z_1, Z_2 , ... Z_N with respective failure rates λ_i. (Here fault i means the fault with label i, not the ith fault discovered.) The results of the testing are the first k waiting times (or cycles of operation) between the discovery of successive faults (again, not faults with successive labels), which are denoted D_1,D_2, ... , D_k. Conditional on the unobservable fault labels, the distribution of the D_i's is that of independent, exponential random variables with decreasing failure rates (decreasing since, of course, each time a fault is discovered, the system becomes more reliable). This conditional distribution is used to derive the unconditional marginal distribution of the first k D_i's, which in turn can be used to derive useful estimates concerning system reliability.

While the mathematics underlying inference for this model are complicated given that faults identified previously have an impact on the probability of discovering future faults, Scholz has derived the maximum-likelihood estimates of the residual failure rate at each stage in the fault discovery process using tools from the field of isotonic regression. Scholz has also provided upper bounds for confidence intervals for the residual failure rate. In other words, Scholz's method estimates and bounds system reliability at each stage of the fault discovery process.

The Meeker Paper

The Department of Defense collects considerable information on the performance of its systems while in development—especially test results—as well as when fielded. However, since test results are currently collected mainly to support decisions on whether to promote systems to the next stage in the milestone process, test data are often not saved and archived in a manner that facilitates their further use. Further, while data on field

failures and field performance are often retained, they are rarely archived in a manner that facilitates analysis of improvements in the system development process. In particular, such information could be used to improve the future design of developmental and operational tests by helping to explore how system flaws were missed during previous developmental and operational testing and how this can be remedied in the future.

In contrast, in industrial applications, test and field use data are often employed for these and other purposes and are frequently archived in a manner that facilitates analysis in support of these uses. Specifically, field use data are employed for prediction of future warranty or maintenance costs, as well as for early detection of reliability problems in fielded systems. Albeit less frequently, these data are also used to provide information on the discovery of failure modes and their frequency of occurrence—information that is in turn used to improve developmental and operational test procedures. Further, this information supports comparisons of system performance (failure modes and their frequencies) in developmental or laboratory tests, versus performance in operational tests, versus performance in the field. Understanding how system performance is related in tests with various degrees of operational realism is extremely valuable for performing reliability growth modeling and for learning how to design laboratory and operational testing with greater operational realism. Finally, field performance data are used to feed component-level reliability information back to design engineers so they can improve current or future component or system designs.

While field performance data have many potential uses in industry, they also have disadvantages. Some disadvantages stem from the primary reason for the collection of field performance data in industry—to support administrative action such as warranty management. Therefore, the data often are not as suitable for the analyses outlined above as would be the case for data from a structured experiment. Deficiencies include the following. First, a sizable fraction of the data is missing, and there are reporting errors and delays. Second, while collecting time of actual use would be optimal for measuring system life, what is commonly available is only calendar time. Third, the environment of use is commonly known only partially or totally unknown. Fourth, in warranty situations, failures are reported only for units that are under warranty. As a result, data are reliable only until the warranty period is exceeded, and the status of units that are not reported is unknown (including retired units and units that were never put into service). Finally, most field performance data are collected only for repairable systems.

A further reason that test and field performance data are not fully utilized in industrial applications is that easy access to these data has an associated cost. While the collection of field performance data is effectively free since they are often required for other purposes, field performance (and test) data must be catalogued in a database structure in a way that facilitates the above uses. The construction and maintenance of such a database is time- and resource-intensive. This point can be illustrated if one considers the need to create a (living) cross- referencing system that identifies all (current and future) systems having components in common with a given system, all test event results, the conditions underlying each test event, the performance of components when fielded, and the conditions underlying field use. The benefits from the use of such data must be sufficient to offset this substantial cost. Making this argument was one goal of some of the presentations related to this topic at the workshop.

Bill Meeker provided an overview, from an industrial perspective (in particular, automobile warranty data), of the many opportunities to learn from the analysis of field performance data. He focused on features of such data that would be expected for defense systems: (1) data are collected until the system is a certain age or until it has covered a fixed number of miles, (2) there is only limited information on the exact cause of failure, (3) there is good information on the date of manufacture, (4) there is often useful information on the rate of use for each system, and (5) there are potential biases in estimation resulting from various homogeneity assumptions (e.g., high-speed drivers may have a different miles-per-failure distribution).

Field performance data have the following key applications. A primary use is to support early detection of production processes in trouble. A common approach used for the purpose is to graph the observed percentage of system failures by months in service alongside a graph of the upper bound for an estimate of the same based on a quantile of a standard cumulative distribution function used to model failure rates (e.g., the Poisson distribution) with its parameters estimated using historical data. Two detection rules are used to signal the need for corrective action: (1) if the observed failure rate at a point in time exceeds a particular quantile based on the historical data, or if some function of the observed number of failures (usually chosen to approximate some standard distribution) exceeds the historical estimate plus a critical value times an estimate of the standard deviation of the historical estimate; and (2) if the difference between some function of the observed number of failures at time t and at time $t - 1$ is greater than the historical estimate for the same plus a critical value times

an estimate of the standard deviation of the historical estimate (of this difference). The critical values are chosen, using historical data, to balance errors of identifying a process in need of correction when it is functioning fine against the cost of letting a process pass that is in need of correction. (For details, see Kalbfleisch et al., 1991 and Wu and Meeker, 2002.)

A second important application of field performance data is the prediction of future warranty or total maintenance costs (the second possibility currently being more relevant to DoD systems). Clearly, information on the rate of field failures of various types could be extremely useful for estimating field maintenance, repair, and component replacement costs.

A third use of field performance data is to establish a "transfer function" between developmental and operational tests and between operational tests and field performance. Knowledge of the ways in which developmental and operational tests are unreliable predictors of field performance has great value for reliability growth estimation, and could be useful both in linking developmental and operational test results and in providing information on how to design developmental and operational tests with greater operational realism. Meeker described the following possibility for addressing a linkage between developmental and operational testing.

Developmental tests are often accelerated, meaning that stresses are frequently increased in an effort to simulate the greater passage of time and greater use. To make accelerated testing informative for decisions concerning operational or field performance, a model (e.g., a degradation model) is used to relate accelerated test time to actual use time. This model must describe the effects of acceleration, the impact and distribution of environmental conditions, and the distribution of use rates in actual use of the system. (This type of model is often related to physics-of-failure models, discussed below.) A successful model of this form could be used to link developmental test data on system reliability to operational test and field performance. Meeker gave an example concerning the use of washing machines. Here the failure probability was expressed as a function of the number of cycles of use, and users were divided into categories based on their rate of use in cycles per unit time. Within these categories, the rate of use was assumed to be constant. Use of this assumption made it possible to translate the failure probability, initially expressed as a function of the number of cycles (which could be experimented on for high use) into a failure probability expressed as a function of time (see Meeker et al., 2002). Agreement between this mixture distribution and field performance can be used as a validation tool, validating, for example, that the percentages of people

in the various categories of use rate do not change over time. Divergences from these assumptions and identification of remedies are the subject of ongoing research.

Once these and other uses of field performance data have been institutionalized with the accompanying benefits, the quality of such data is likely to improve. One important application in which industrial field performance data have recently been improving in quality is sensors that can collect the entire history of use of, say, an automobile, including stresses, speeds of use, temperature, etc. That information can be linked to information on system reliability or performance to support much richer statistical reliability modeling. Use of such sensors could be valuable in operational testing for similar reasons.[1] Several companies are undertaking efforts to save additional data in their warranty databases so the data can be used not only for financial purposes, but also for reliability assessment and estimation. Such efforts represent a cultural change. A hurdle is that development or expansion of such a database sometimes requires innovative funding approaches.

Discussion of Gaver, Sen, Scholz, and Meeker Papers

In his discussion of the papers by Sen and Gaver, Paul Ellner addressed the complication involved in reliability growth modeling of translating reliability estimates from developmental test to predictions of reliability in the field from operational test. At present, analysts may use a reliability growth model to extrapolate from developmental test results to operational test results. This approach can be severely biased, producing overly optimistic reliability estimates since the failure modes can be substantially different in the two situations (actually three—developmental test, operational test, and field performance). Efforts to perform this translation face the following challenges: (1) determining the (approximate) relationship between failure modes that occur jointly in developmental and operational environments, and (2) identifying a function expressing the probability of failure in operational test as a function of the probability of failure in operational test due to failure modes not present in developmental test. This translation probably cannot be done at the system level; it must be carried out at the com-

[1]These sensors could be used to monitor reliability degradation during field use and to support efficient logistics management.

ponent level. Clearly, a direct way to address this issue is to use developmental testing that is more representative of operational use, to the extent possible.

Ellner pointed out some assumptions that limit the applicability of the exponential reliability growth model, though he was relatively confident that these could be addressed. For Gaver's model, the greatest challenge is that of initial input, that is, the number of faults in each stage of the system and the probability of discovering a fault during a test. Ellner suggested that the number of faults can be assumed to be quite large for complicated systems, and that giving the probabilities of discovering a fault a hierarchical structure is also a promising generalization of the model.

Ellner also strongly supported Sen's proposal regarding the use of many alternative models that are consistent with the data to assess model misspecification. If these models agree with respect to decisions, one can be confident; if they disagree, the discrepancy will have to be analyzed.

Ellner remarked that an AMSAA website[2] and an Institute of Electrical and Electronics Engineers working group are both concerned with updating Military Handbook 189 on reliability growth management (U.S. Department of Defense, 1981). He suggested that efforts to update this handbook would be more successful if the responsibility were assigned to a specific organization.

In his discussion of the papers by Scholz and Meeker, James Crouch pointed out that DoD already makes considerable use of operational test and field performance data, at least in the area of reliability testing of jet engines. Performance data are used to manage and control various aspects of turbine engine reliability, specifically (1) the engine in-flight shutdown rate; (2) the rate at which the engine needs to be repaired; and (3) the line replaceable unit rate, the maintenance rate for replacement of the components that surround the engine. The use of operational test data is complicated by engine-to-engine variations (it is typical to develop only three or four prototypes for operational test), and the use of both operational test and field performance data is complicated by variations in operational use on which data are not easily collected. These problems are currently being addressed using modeling and simulation.

The Air Force uses Pareto charts (histograms of the number of failure occurrence reports by root-cause category) based on field performance data

[2]http://amsaa-www.arl.mil/AR/rel_growth_guide.htm

to examine different causes of engine removal. In the long term, redesigns are often based on these analyses. The Air Force also has a deficiency reporting system that initiates an engineering investigation of a problem. (At times, unfortunately, the reports are incomplete, and improper malfunction codes are entered.) In addition, the Air Force has a warranty program that reports component failures and conducts warranty investigations. As mentioned above, while the time to failure is generally known, the cycles at failure or other characteristics are sometimes unknown. Also as mentioned above, once the parts and engines have outlived the warranty, problems in collecting data can become more prevalent.

The Air Force has been giving reliability issues much higher priority of late. One successful program addressing high-cycle fatigue is referred to as reliability-centered maintenance. Reliability-centered maintenance is a systematic approach to preventative maintenance in which optimal maintenance processes are employed at the component or subsystem level. The Air Force is also using highly accelerated life testing and highly accelerated stress screening to identify failure modes. Analysis of the common and unique failure modes from accelerated developmental testing and operational testing may make it possible to better understand the distinctions between these two types of testing.

In the floor discussion, Dan Willard questioned the benefit of going beyond the understanding of failure modes from such activities as accelerated testing. Scholz responded that the failure modes discovered in accelerated testing may differ from those found in the field. While it is valuable to discover and correct as many faults as possible before fielding, there is additional value in comparing those faults identified through accelerated testing and those discovered after fielding.

TOOLS FOR MANAGING RELIABILITY GROWTH

Reliability growth management consists of procedures and infrastructure used during system development to track and expedite reliability growth—especially including use of various feedback mechanisms to improve system design. (The hope is that these feedback mechanisms can also be used to improve the process of reliability growth management itself over time.) A key goal is early assessment of the likelihood of meeting the targeted operational reliability. Generally speaking, DoD has accorded subordinate priority to system reliability as compared with system effectiveness as a result of the primary conflict scenarios on which the agency has, until

recently, focused its attention. In recent years, the types of military engagements faced and anticipated have changed quite dramatically. As a consequence, the importance of placing increased emphasis on the development of highly reliable systems has grown. Speakers and discussants strongly confirmed the need for improved reliability growth management through frequent and thorough testing and inspection and through the application of global, cross-disciplinary strategies for achieving and surpassing reliability growth targets.

It was pointed out by more than one speaker that at present, defense systems regularly fail to satisfy their operational suitability requirements in the field. (Suitability encompasses reliability and related measures, including maintainability and availability.) As a result, DoD is spending far too much for system redesigns late in system development, and for spares management and system maintenance (and also system redesign) after the system has been fielded. DoD systems also are frequently submitted for operational testing before they are sufficiently mature with respect to system reliability. For example, it was pointed out that 80 percent of Army systems failed to achieve even half of their requirements for mean time to failure in operational test (Defense Science Board, 2000).

Mention was made of a number of methods currently used in industrial applications of system development and reliability growth management that are not being used in defense system development, but appear to be relevant to the latter systems. First, early assessment of (operational) reliability could play an important role in improving system design, as opposed to current primary use in supporting the milestone decision process. Second, little or no use is currently made of test or field failure data to (1) support a better understanding of system life-cycle costs; (2) help determine how failure modes escaped detection during developmental or operational testing; or (3) relate the reliability of systems and failure modes in operational test to the reliability of systems and failure modes in developmental test, which could support methods for combining information from developmental and operational testing (as discussed earlier). The estimation of system life-cycle costs was noted as a particularly important use of field performance data. (As mentioned above, the increased accessibility of such data to support this type of analysis would require the institution of a data archive.) Third, it was noted that it is typical in reliability growth modeling for defense systems—used to predict system reliability in the future (e.g., to ascertain when it would be appropriate to enter operational test)—to make no attempt to model

the system defect discovery process. Ignoring this information will likely produce much less predictive models than the approaches available today that model this process. Fourth, methods exist and are currently used for developing early assessments of system reliability that make full use of the disparate information available in industrial applications (e.g., information derived from maintenance records, computer simulations, expert knowledge, historical data, and test data, and information from similar systems, or systems with similar parts, components, or processes). However, these methods are not currently applied to defense systems (with a few notable exceptions). Finally, when determining total time on test (in operational test) or other aspects governing an operational test design, requisites for testing system effectiveness typically have been the dominant consideration, while the requisites for producing assessments of operational suitability have received considerably less weight.

The above five areas are ones in which greater attention to reliability measurement, reliability modeling and data collection, and the management of defense system reliability could prove beneficial. Two specific approaches to reliability growth management were considered at the workshop, as summarized below.

The Booker Paper

It is currently typical for reliability assessment of defense systems to be used primarily as input into the DoD acquisition milestone process, for deciding whether a system in development can proceed to the next milestone. Since operational testing is carried out near the end of the second phase of system development (known as engineering and manufacturing development), there is little or no opportunity for reliability assessment of a system's operational performance to inform system design during its early stages. In contrast, in various industrial applications of system development, reliability assessment has an earlier and more continuous influence on system design. A major benefit of this early influence is that, generally speaking, the earlier modifications are made to system design, the less costly those modifications are. Further, the better a system design is, the more likely it is that the system will pass operational test on its first attempt. Finally, and most important, the better the system design is, the more likely it is that the system ultimately approved for full-rate production will perform better and be less costly to operate in the field, since it will likely require less maintenance and repair. Changing the role of reliability assess-

ment from one of confirming that a system or its components meet specific performance requirements to one of understanding as much as possible, as early as possible, about the (operational) strengths and weaknesses of the current system design requires planning and carrying out additional, targeted testing of the system. In that testing, previous assessments are used in determining the test timing, test size, test scenarios, and number of replications at each scenario for the various test events that are needed. Two processes were presented at the workshop that, to different degrees, (1) provide early reliability assessments; (2) assist in the design of these early, additional, operationally relevant tests; and (3) use early reliability assessments for improving system design.

The first such process was presented by Jane Booker, representing a team at Los Alamos National Laboratory that has developed the Performance and Reliability Evaluation with Diverse Information Combination and Tracking (PREDICT) system of early reliability assessment. PREDICT (now known as Information Integration Technology) is a comprehensive framework that facilitates the use of disparate sources of information—such as expert opinion, simulations, historical data, test data, and maintenance data—for the system in question, or information on similar systems or systems with similar components to produce reliability assessments through use of a combination of information models. PREDICT also provides estimates of the uncertainty of these assessments. Both the estimates and their estimated uncertainties can be displayed graphically for easier understanding by decision makers.

PREDICT uses this information for a variety of purposes. The first is to identify which components, if improved, would most improve overall system reliability. Also, these assessments and their associated uncertainty can help in designing system tests that are more informative by targeting test events to areas of lowest reliability or of greatest uncertainty. These test results can be used to propose system design changes. Further, this framework can be used to carry out "what-if" analyses. For example, to gauge test size, one could ask what the result would be if another test run were carried out and were successful. One could also input a different maintenance schedule or a redesign of a subsystem to examine the impact on total system performance.

For an initial assessment of system reliability, PREDICT uses the following inputs: (1) system requirements or performance measures; (2) system structure, including information on subsystems and components and

on failure modes for components that is input through use of a variety of representations, including logic models/diagrams, event/fault trees, directed graphs, Bayesian networks, process trees, and reliability block diagrams; (3) system process, using inputs from physics and chemistry, mechanical engineering, quality control tests, assembly, and testing at various levels (system, subsystem, and component); and (4) inputs concerning the reliability of components in analogous systems, and expert opinion on the reliability of the components. All of these inputs are documented in a knowledge base that provides information at customized levels for various queries. Inputs for a given system are also available to provide information concerning the performance of similar or related systems in the future. These initial assessments are updated in accordance with the receipt of new information and test results. (Updates are also based on refinements to system structure or changes to requirements or performance measures.)

PREDICT tracks performance as system development proceeds. Once a system has been fielded, PREDICT can be used to track performance in the field; that is, it can continuously update reliability assessments on the basis of new information (e.g., on the aging of the system).

PREDICT also provides a platform that facilitates consideration of various action items, such as whether one can support a system in the field or how the number of maintenance actions can be reduced once the system has been fielded. PREDICT can also support decisions involving either the development of a new system or a choice among several system designs through balancing of the costs of development and the costs of fielding to arrive at a system that minimizes life-cycle costs.

As an example, consider an air-to-air heat-seeking missile. The major subsystems are the warhead, the missile, the aircraft, command and control, and logistics and maintenance. Taking the command and control subsystem in more detail, the aircraft has power, avionics, environmental, acquisition and fire control, flight structure, launching, flight control, and missile interface elements, as well as human intervention. There are also complex interactions between subsystems that act across major subsystems. PREDICT attempts to represent all of this structure using various forms of sensitivity analysis.

PREDICT has been used successfully by Delphi Automotive Systems and in the nuclear weapons program at Los Alamos National Laboratory. PREDICT can also be implemented in dynamic environments where testing is not feasible, such as in the nuclear weapons program.

The Crow Paper

A second system that also uses early reliability assessments to improve system design and development is the Integrated Reliability Growth Strategy (IRGS), currently in use at General Dynamics Advanced Technology Systems and several other institutions. IRGS, which was outlined at the workshop by its primary developer, Larry Crow, is a process that generates early and substantial reliability growth through continuous testing and assessment to determine which of a system's components have a mature or immature design. On this basis, IRGS directs design modifications of the immature components. The result is a reliability growth program, iterating between design modification and testing, that tends to reduce substantially the time needed to achieve reliability goals, for example, to attain a required reliability level before entering operational test.

Permitting a system to enter late-stage developmental test with a substantial number of reliability flaws places too heavy a burden on developmental and operational test to discover the remaining problems. This is also an expensive way of discovering defects since it is likely that the system will experience difficulties in operational test, and it may have to undergo design modifications and later repeat some operational test events. Today, it is not uncommon for some DoD systems to enter into late-stage developmental test when their reliability is at 30 percent of the ultimate goal, whereas the goal for industrial applications is for a system to be at 75 percent of its eventual reliability before entering into formal testing. The latter is accomplished by identifying design flaws in earlier stages of the development process, thereby producing a mature system design much earlier. Again, the overall change in strategy is based on modifying the function of reliability assessment from that of a statistic used to support promotion decisions to that of an early and continuing objective measurement (combining a wide variety of types of information) that is used to support system development by helping to identify components in need of redesign or maturation.

IRGS takes as input a system design that supports prototypes with approximately 25–30 percent of the final required reliability. The complete system undergoes a requirements review, including performance requirements and requirements involving the environment, reliability, safety, maintainability, and support. IRGS then categorizes failure modes for complex systems into type A and type B failure modes. Type A modes correspond to components that have mature designs and are unlikely to be

capable of substantial improvement. These modes are typically associated with off-the-shelf components with demonstrated high reliability and a proven, inherited design. Type B modes correspond to components that are candidates for improvement. These components involve unproven new technology or a new design, or may be off-the-shelf components that require improvement before use. System maturity can be measured as the mean time to failure that is due to A components as a percentage of total system mean time to failure. (See Crow [1998] for some related estimation issues.)

The foundation of IRGS is a process that identifies and mitigates type B failure modes by converting them to type A modes. This conversion is accomplished through an iterative process of testing and analysis. Testing steps include evaluation; qualification; reliability growth modeling; and application of the Failure Reporting, Analysis, and Corrective Action System (FRACAS). Analysis steps include understanding of failure modes and fault tree analysis, analysis of reliability design trade-offs, safety, maintainability, design-stress reliability, material and supplies analysis, and analysis of manufacturing for reliability. Analysis and testing are used to identify which components are likely involved in any problems. This is accomplished by applying the broad concept of the type A–B mode approach to component reliability described above. A process that tracks the reliability of components is fed information from this analysis and testing scheme, and the sources of any problems are identified and appropriate corrective actions sought through use of design reviews and consultation with project teams. The result is a reliability growth program iterating between design modification and testing that tends to reduce the time to achieve reliability goals. As the reliability of components improves, their designation changes from type B to type A.

IRGS has been applied successfully for various purposes. For example, it has been used to demonstrate that a system in development would be highly likely to exceed its required reliability of 8 years between failures. It has also been used to show that design upgrades are improving system reliability. Finally, it has been used to monitor a wide variety of system integration tests and hardware and software upgrades and their impact on system reliability.

In all of these applications, it is important to define specific metrics that can be used for tracking performance over time so that reliability growth can be concretely documented. Metrics such as the proportion of failure modes that are classified at any point in time as type B, as well as the

performance of components, subsystems, and systems before and after interventions and/or design improvements, serve not only to measure progress toward meeting a project's reliability goals, but also to inform participants at all levels of the constructive contributions made by the reliability program. Since the success of IRGS relies on the collaboration of a wide array of scientists, engineers, and management personnel, it is imperative that improvements in system performance be documented and widely communicated.

Discussion of Booker and Crow Papers

The two systems for using reliability assessment as an input into system development described by Booker and Crow make credible the claim that comprehensive reliability improvement programs can be both feasible and effective. The cornerstone of both methodologies is their interactive character, with iterations of the traditional test-fix-test cycle leading to interventions that improve components and subsystems. Both methods seek to utilize input from experts, with the PREDICT methodology doing so in a more formal way.

The discussants for these papers, Walt Hollis and Arthur Fries, expressed their optimism that these methods could be implemented in defense system development and would provide substantial benefits for many types of systems. (In the floor discussion, Dan Willard mentioned that his agency had developed a tool that appears to have some similarities to PREDICT, and they were going to compare the two to see whether there are advantages that could be shared.) One promising idea would be to use IRGS as the process for managing reliability growth, with PREDICT being used for reliability assessment.

Hollis mentioned three Army systems for which measurement of system reliability is extremely difficult, for different reasons: the National Missile Defense System, the Theater High-Altitude Air Defense System, and the Patriot Missile System. For these systems, reliability growth must necessarily rely on component testing and simulations. Since achievement of high operational reliability cannot be tested in, it must be designed into the system.

Fries pointed out that there is less and less time available for evaluation, and there is a growing need to begin operational evaluations earlier. To do so, one is obligated to use other information sources from the development process. IRGS and PREDICT are both very worthy attempts to accomplish this. Both are both highly structured (fully documented) ways

of breaking a system up into subsystems and subprocesses and having experts examine each carefully, suggesting additional testing when necessary to direct improvements. These tools can also help guide final operational tests.

Hollis and Fries pointed out that the adoption of these methods does require up-front investment, and DoD program managers need to be willing to expend those funds. Support for this investment will come with expected positive experiences, which IRGS has already demonstrated for defense systems.

Finally, the discussants pointed out that the key to the successful use of both of these processes is that there must be an early and constant emphasis on the performance of the system under operational conditions, as opposed to meeting a required reliability level that is based on laboratory performance. It is still the case that operational sources of reliability problems appear very late in system development. These problems are typically ones that could have been identified much earlier. Both approaches can address this problem if necessary change in emphasis is achieved.

CONCLUDING REMARKS ON THE MEASUREMENT AND MANAGEMENT OF RELIABILITY GROWTH

The reliability growth management processes outlined above, and reliability growth management more generally, require a variety of sources of information on system reliability as key inputs. Especially important are data from developmental and operational testing and from the field performance of related systems.

Because operational testing is costly (and occurs late in the budget cycle when there is little possibility of a large reallocation of funds for operational test), only a limited amount of information is typically collected in terms of both the number of replications and the number of separate test environments and scenarios that can be examined. Given this limited information, it is typically the case that operational testing data alone cannot confirm, with the usual levels of statistical confidence, that a defense system's suitability requirements are met. It would be generally useful, therefore, to combine operational test data with appropriate portions of developmental test data on the same system, and data from field use and developmental and operational testing of related systems to provide less variable estimates of system reliability to inform decisions about system promotion. Further, as mentioned previously, early assessment of a system's

reliability is extremely useful to help guide early design changes, and any such early assessment must be based on a combination of information from a variety of sources given the scarcity of direct assessment early in system development. Therefore, combining information, some of it possibly subjective, is likely to prove beneficial in some situations.

It is becoming increasingly common—though by no means widespread as yet—for reliability assessment for industrial systems late in development to make effective use of information on the reliability of related systems (e.g., systems with identical or similar components) and information for the same system from laboratory testing. Even earlier in system development, some industries have demonstrated the utility of information on related systems and expert judgment to help make initial assessments of system reliability that are useful for developmental test planning and for tracking of reliability growth.

In the field of statistics, combining information models are currently being developed primarily from a Bayesian perspective. Much progress is occurring in this area as a result of the development of simulation methods that have greatly facilitated the calculation of Bayesian estimates. This rapid progress increases expectations that more and more types of applications will be addressed using these new methods.

Certainly such methods cannot be used without some scrutiny, and the benefits of use of these models for defense systems will almost undoubtedly vary with the specific application. The linkage between failure modes and failure frequencies across systems and across environments of testing or field use must be well understood before these models are applied. Aggressive efforts are needed to validate the assumptions made. The expectation is that over time, estimates for some measures for some types of systems will be found to benefit greatly from use of these models, whereas for other systems, these models will be much less useful. Chapter 3 provides a description of some specific methods that were suggested for use at the workshop.

3

Current Research in Reliability Modeling and Inference

Four of the seven sessions at the workshop addressed reliability-related areas (other than reliability growth) in which recent advances and ongoing research could especially benefit the DoD test and evaluation community in its current activities and applications. This chapter presents the issues, methods, and approaches that were raised in these sessions. The topics to be discussed here include: (1) approaches to combining information from disparate sources that are aimed at achieving improvements in the accuracy and precision of the estimation of a system's reliability; (2) model-based approaches to selecting inputs for software testing; (3) current models for estimating the fatigue of materials; and (4) reliability management to support estimates of system life-cycle costs. Before proceeding to fairly detailed coverage of the sessions on each of these topics, we briefly describe the motivation for the selection of these as topics worthy of special attention at the workshop and give a brief overview of each.

The combination of information from disparate sources (the value of which is discussed in Chapter 2) is a problem that has interested statisticians for many years. Indeed, the fields of Bayesian statistics, empirical Bayes methods, and meta-analysis all emerged to address this problem. The idea of exploiting "related information" in the process of interpreting the outcome of a given experiment arises in many different forms. In the DoD acquisition process, data are collected during the various stages of developmental testing, and these data may well be of use in the process of analyzing

the outcomes of the subsequent operational test. In the workshop session on combining information, Duane Steffey described use of a parametric hierarchical Bayes framework for combining data from related experiments, and Francisco Samaniego followed with a description of nonparametric methods for handling the same problem. These presenters argued that existing methods and others under current investigation constitute promising ways of modeling the data-combination challenges that arise in developmental and operational testing.

Some of the earliest work on fatigue modeling occurred in the context of addressing problems that were common in the aircraft industry during and following World War II. While some of the early attempts at modeling fatigue in the materials used in aircraft construction were primarily mathematical in nature, the field has evolved and seen some notable advances and achievements through the collaboration of mathematical/statistical workers, materials engineers, and other scientists. Sam Saunders described some of the early work in this area at Boeing, including the development of the widely used Birnbaum-Saunders model, and underscored the importance of understanding the science involved in a particular application before attempting to model the problem statistically. Joe Padgett's presentation was focused on a class of models for fatigue of materials or systems due to cumulative damage and the modeling of crack growth due to fatigue. This work, which combines current thinking in materials science and sophisticated statistical modeling, provides a broad collection of models on which to base estimation and prediction in this area.

Modern military systems have become increasingly dependent on computer software for their successful and reliable operation. Given that the area of software reliability is broad enough to merit a workshop of its own, the goal of the session was scaled down to providing the flavor of two particular lines of research in the area. Siddhartha Dalal's presentation focused on efficient methods of selecting factorial experiments with attractive coverage possibilities. He described approaches to experimental design that allow the experimenter to sample a reasonably broad array of combinations of factors while controlling the scale of the overall experiment. Jesse Poore's presentation focused on methods of testing software that take special account of anticipated usage patterns, thereby enhancing the likelihood of good performance in the software's intended domains of application.

APPROACHES TO COMBINING INFORMATION FROM DISPARATE SOURCES

A variety of sources of information on the reliability of a defense system under development are available at the different stages of system development. Data from developmental and operational tests and from field performance for systems with similar or identical components are typically available at the beginning of system development. There are also data from the developmental tests (contractor and government) of the system in question. Finally, there are often field use and training exercise data, as well as "data" from modeling and simulation.

Attempts to combine data from tests or field experience for a related system with those for a given system must be made with caution since large changes in reliability can result from what would ordinarily be considered relatively minor changes to a system, and even identical components can have importantly different impacts on system reliability when used in different systems. Data from field and training exercises must be carefully considered since field use and training exercises are not well-controlled experiments. Further, the utility of modeling and simulation results depends heavily on the validity of the models in question.

Even identical systems can have dramatically different reliabilities in developmental and operational testing as a result of the different conditions involved. In developmental testing, the system operators are typically fully acquainted with the system, the test conditions are carefully controlled, and the test is often at the component level (e.g., hardware-in-the-loop testing). On the other hand, operational testing involves using the full system in operational conditions as realistic as possible, with the actions of the participants relatively unscripted and the system being operated by personnel more typical of real use (with the anticipated amount of training). Clearly, these are distinctly different conditions of use.

On the other hand, the cost of operational testing (and the need for expeditious decision making) necessarily limits the number of operational test replications. Given the importance of assessing the reliability of defense systems in development, including how this assessment factors into the ultimate decision on whether to proceed to full-rate production, it is extremely important to base reliability assessments of defense systems in development on as much relevant information as possible. As a result, it has been suggested, especially of late, that the various sources of information be combined, if possible, to provide the best possible estimates of sys-

tem reliability in an operational setting (see also Chapter 2). Given the differences in conditions of use, however, the combination of developmental and operational test data for identical systems (and data from test and field use for similar systems) must be considered carefully. It was stressed repeatedly at the workshop that any attempt to combine information from disparate sources should be preceded by close scrutiny of the degree of "relatedness" of the systems under consideration and the conditions of use, and the appropriateness of modeling these relationships. It is clear that without this care, use of these additional sources of information could result in assessments that are less accurate or precise than those relying exclusively on operational test data. Combining of information is therefore an important opportunity, but one that must be explored with caution.

One session of the workshop focused specifically on the use of models for combining information from developmental and operational tests when the failure modes in these separate environments of use are well understood (or otherwise approximately satisfy the necessary assumptions). It was argued that in those instances, use of the proposed models can provide improved estimates and thereby support better decision making. Two specific approaches to combining information were proposed at the workshop, as described below.

The Steffey Paper

Duane Steffey reported on recent research on the estimation of mean time to failure under specified conditions of use, given information about the performance of the same system under different test conditions. (For details, see Samaniego et al., 2001). Of course, a key application for which such extrapolation would be needed is one in which the former conditions of use were developmental test conditions and the latter operational test conditions, with the hope of combining developmental and operational test information to support an operational evaluation. There are two questions of interest: (1) How can a meaningful notion of relatedness be characterized in a statistical model? (2) What method or methods of estimation are most appropriate for this problem? The approach relies on the following assumption: that the complexity of and difference between the two sets of experimental conditions make it impossible to link the information derived under those sets of conditions using parameters that define the test conditions. In other words, a trustworthy parametric model of reliability as a function of the test conditions (that is, covariates such as amount of

TABLE 1 Fictitious Developmental and Operational Test Data

Developmental Test Data		Operational Test Data
28.73	18.01	13.48
21.76	1.55	18.63
6.01	35.54	4.54
46.68	22.06	23.51
7.58	2.58	5.34
11.27	20.89	8.39
16.08	7.15	39.97
8.06	10.19	7.79
9.97	67.03	33.14
41.66	7.79	6.14

training) cannot be developed. The estimation approach described by Steffey is (Bayesian) hierarchical modeling using a relatively simple characterization of relatedness of conditions of use.

A dataset motivated the discussion. Consider the following (fictitious) lifetimes of experimental units (hours to failure) from developmental and operational testing (DT and OT) as displayed in Table 1. For developmental testing, the mean time to failure is 19.53, whereas for operational testing, it is 16.09.

The statistical model used assumes that there exists a probability distribution with mean μ^D that generates DT mean times to failure. Likewise, there also exists a probability distribution that generates OT mean times to failure. These means of the distributions that generate mean times to failure (μ^D, μ^O) are referred to as grand means. Then, to obtain the observed time to failure for a given system for either developmental or operational test, one draws a random waiting time from a distribution with the appropriate mean. This can be considered a staged process in which the second and final stage represents the variability of an individual system's waiting times to failure about each individual system's mean, and the initial stage represents the variability between the mean times to failure for individual systems (from the same manufacturing process) about a grand mean time to failure. It makes sense to assume that the DT grand mean is some factor larger than the OT grand mean, since operational test exposes a system to more opportunities for failure. This multiplicative factor is designated λ. (There are non-Bayesian approaches in which a λ factor is used to convert

operational test hours into "developmental test hours" for purposes of weighting as combined estimates.)

The goal for combining information in this framework is estimation of μ^O. To this end, three alternative estimators were considered: (1) the unpooled estimator, here the average time to failure relying solely on data collected during operational test; (2) a specific weighted average of the observed OT individual mean time to failure and the observed DT individual mean time to failure, referred to as the linear Bayes estimator; and (3) an estimator that makes full use of the hierarchical Bayes approach. To compare the performance of these (and other potential) estimators, the natural metric is Bayes risk relative to a true prior representing the true state of nature, which is the average squared error (averaged over the process described above that first draws a mean reliability for a specific system, in either developmental or operational test, and then draws a time to failure from the assumed probability distribution centered at those means). The reduction in average squared error that results from switching from an unpooled to a pooled estimator measures the gain from the use of developmental test data.

The linear Bayes estimator is considered since (1) it can be computed explicitly and can serve as an approximation of the full hierarchical approach (one simply chooses the weights to minimize Bayes risk), (2) it makes explicit the use of developmental test data, and (3) it is possible to characterize the circumstances under which this estimator is preferred to the unpooled estimator. For discussion of linear Bayes methods, see: Hartigan (1969), Ericson (1969, 1970), Samaniego and Reneau (1994), and Samaniego and Vestrup (1999).

Returning to the above dataset, the unpooled estimator is the mean operational test waiting time to failure, or 16.09. Assuming that $\lambda = 0.75$—which of course would not be known in practice—and some additional but reasonable assumptions about the developmental and operational test experiments, the optimal coefficients for the linear Bayes estimator are $c_1^* = .4$ and $c_2^* = .43$ (see Samaniego et al. [2001] for details), producing the linear Bayes estimate of .4 (19.53) + .43 (16.09) = 14.73, which is considerably lower than the unpooled estimator of 16.09.

Steffey demonstrated analytically that the Bayes risk for the linear Bayes estimator, when λ is known, is necessarily smaller than that for the unpooled estimator. This begs the question of what happens in the case when λ is not known.

To proceed it is necessary to place a prior distribution around λ, de-

noted $\pi(\lambda)$. Two possible approaches can now be used. First, one can construct a different linear Bayes estimator that makes use of the mean and variance of the probability distribution. Second, one can make use of a hierarchical Bayes estimator that assumes a joint prior distribution for the means of the distributions of the operational test and developmental test mean waiting times to failure.

In the earlier linear Bayes approach, the optimal coefficients for the developmental and operational test (observed) mean times to failure were selected to minimize the Bayes risk. Now, the optimal coefficients of the developmental and operational test mean times to failure are selected to minimize the *expected* Bayes risk, given that one must average over the uncertainty in λ. This makes the resulting optimal coefficients slightly more complicated than when λ was assumed to be a known constant. Analytic results show that the expected Bayes risk for the unpooled estimator is greater than that for the optimal linear estimator, as in the case for fixed λ, when the assumed $\pi(\lambda)$ has nearly the same center as the true prior distribution. When the assumed prior distribution is substantially incorrect, the unpooled estimator can be preferable to the optimal linear Bayes estimator. Therefore, the benefits of pooling depend on the information available about the relationship between the two testing environments.

Steffey also examined the less analytically tractable hierarchical Bayes estimator, providing some information on the differences between its performance and that of the linear Bayes estimator in simulation studies. Generally speaking, use of the more complicated hierarchical estimator results in additional benefits relative to use of the optimal linear Bayes estimator, although much of the improvement over the unpooled estimator is realized at the linear Bayes stage.

The Samaniego Paper

Nonparametric estimation and testing avoids the use of parametric assumptions and instead uses quantities such as empirical distributions or the relative ranks of observations to support estimation and inference.[1]

[1]Some nonparametric methods may assume that the data are generated by broad distributional families or may have other parametric aspects. Therefore, we use the term "nonparametric" to indicate methods that *avoid* the use of parametric assumptions. We use the terms "distribution-free" and "fully nonparametric" to indicate methods that make no assumptions about the data-generating mechanism.

Because parametric models describing a probability distribution that is assumed to generate the data are avoided, nonparametric approaches are much more likely to be valid. This greater validity comes with the disadvantage that nonparametric methods are typically outperformed by parametric alternatives when the assumptions used by the parametric approach are approximately correct. It is generally understood, however, that the loss in efficiency sustained by nonparametric methods when parametric assumptions hold exactly is often quite modest, and is thus a small price to pay for the broad protection these methods offer against model misspecification. Francisco Samaniego offered a brief review of nonparametric methods in reliability, and then suggested some nonparametric approaches to combining information from "related" experiments.

Parametric models that are often used to describe the distribution of times to failure include the exponential, Weibull, gamma, and lognormal distributions. Selection of parameters (e.g., the mean and variance) identifies specific members from these distributional families. In shifting from one family of distributions to another, say, from the lognormal to the gamma, different shapes for failure time distributions are obtained, though they are all typically skewed distributions with long right-hand tails. Samaniego demonstrated a phenomenon often encountered in applied work—the futility of performing goodness-of-fit tests based on small samples. He displayed a simulated dataset that appeared to be reasonably well fit by all four of the aforementioned parametric models on the basis of sample sizes of 20, but were clearly poorly fit by these models when sample sizes of 100 were available. Generally speaking, the use of goodness-of-fit tests to test for a specific parametric form should be preceded by use of graphical and other exploratory tools, and consideration of applicable physical principles, to help identify reasonable parametric distributional models. However, for small sample sizes, these techniques typically will not provide sufficient information to identify good parametric models. This inability to distinguish among various parametric families for the smaller datasets that are typical of defense operational testing motivates the use of nonparametric estimation for which no parametric form is assumed.

Nonparametric reliability models are typically based on certain distributional assumptions, such as notions of aging or wear-out, that are motivated by experience with the application of interest. One notion of aging is "increasing failure rate" (IFR). Systems having time-to-failure distributions with this property are increasingly less likely to function for t additional units of time as they grow older. (A related characteristic is "increas-

ing failure rate average" [IFRA].) Another model of aging is "new is better than used" (NBU). For systems with time-to-failure distributions having this characteristic, the probability of lasting t units of time when the system is new is greater than the probability of lasting an additional t units of time given that one such system has already lasted Δ units. (This notion is slightly distinct from IFR since it links the performance of older systems to that of a new system and not to the performance of intermediate aged systems.) Another widely used modeling assumption is that of decreasing mean residual life (DMRL). This assumption characterizes a time-to-failure distribution in which the expected additional or residual lifetime of a system of age t is a decreasing function of t. This concept is distinguished from IFR since it relates mean lifetimes rather than lifetime probabilities.

Samaniego argued that instead of assuming a specific distributional form for the time-to-failure distribution and estimating parameters to identify a particular member of these distributional families, one could estimate the lifetime distribution under one of the above nonparametric assumptions. For example, under the assumption that the time-to-failure distribution is IFR, the nonparametric maximum-likelihood estimate of the hazard rate (the instantaneous failure rate conditional on the event that the system has lasted until time t, which is essentially equivalent to estimation of the time-to-failure distribution) at time t is a nondecreasing step function whose computation involves the well-understood framework of isotonic regression. Similar constraints from assumptions such as NBU produce alternative nonparametric estimators.

These are one-sample techniques for the problem of estimating the properties of a single time-to-failure distribution. With respect to the problem of combining information, the natural situation is that of comparing samples from two related experiments. Rather than make the linked-parameter assumption of Samaniego et al. (2001) (i.e., the λ factor), Samaniego instead used nonparametric assumptions about the relationship between the time-to-failure distributions for developmental and operational testing of a system. Three well-known formulations of the notion that a sampled quantity (failure time) from one distribution tends to be smaller than a sampled quantity from another distribution are as follows (see Shaked and Shanthikumar [1994] for further details): (1) stochastic ordering, when the probability that the next failure will be t time units or greater for a system in developmental test is greater than the probability that the next failure will be t time units or greater for the same system in operational test, for all t; (2) hazard-rate ordering, when the instantaneous failure rate,

given that a system has lasted until time *t*, for the system in developmental test is smaller than that for the system in operational test, for all *t*; (3) likelihood ratio ordering, when the ratio of the time-to-failure density for developmental test to that for the time-to-failure density for operational test is a decreasing function of time.

Samaniego then discussed a new type of ordering of distributions, referred to as "stochastic precedence." Distribution A stochastically precedes distribution B if the probability is greater than .5 that a random variable from distribution A is less than a random variable from distribution B. The assumption that operational test failure times stochastically precede developmental test failure times has repeatedly been verified empirically in a wide array of applications. When the assumption is warranted, relying on it and using the associated inference substantially improves estimation of the cumulative time-to-failure distribution function for operational test data.

Attention was then turned to the process of estimating the lifetime distributions from two experiments under the assumption that one experiment (for example, OT) stochastically precedes the other (for example, DT). The estimation process is accomplished as follows. Should the standard estimates of the empirical cumulative distribution functions (ecdf) for failure times from operational and developmental testing satisfy the property of stochastic precedence, those ecdf's are used, unchanged, to estimate the operational and developmental test time-to-failure distributions. However, if the ecdf's fail to satisfy stochastic precedence, the ecdf's can be "adjusted" in one of several ways to arrive at a pair of estimators that do satisfy stochastic precedence. Samaniego discussed two specific approaches to such adjustment—the first involving a rescaling of the data from both samples to minimally achieve stochastic precedence, and the second involving data translation (that is, a change of location). Under the assumption that stochastic precedence holds, both methods were shown to offer improvement over estimators that rely exclusively on data from just one of the experiments. Asymptotic results show improvement in the integrated mean squared error of the competing estimators, and simulations demonstrate their efficacy in small-sample problems as well (see Arcones et al. [2002] for details).

In summary, this research demonstrates that developmental test data can be used to improve an estimator of the time-to-failure distribution of operational test data under quite minimal assumptions. Such an approach might also be used in gauging the robustness of parametric approaches to estimation. As research advances, more nonparametric models and infer-

ential methods will be available and will constitute an increasingly comprehensive collection of tools for the analysis of life-testing data.

Discussion of Steffey and Samaniego Papers

The discussion focused on the ability to capture the degree to which the reliabilities of different systems tested in different environments are related. The argument was made that developmental testing conditions are by nature quite different from those for operational testing, in part because they have somewhat different objectives. The goal of developmental test is to identify key areas of risk and then determine how to mitigate that risk. For this reason, much of the effort in developmental test focuses on the working of individual components. On the other hand, the goal of operational testing and evaluation is to examine whether the entire system is consistently effective and suitable in an operationally realistic environment with its intended users. Clearly many operational problems that may not arise in a laboratory setting cause system redesigns late in the development process, when they are more costly. As a result, there are now increasing efforts to make greater use of conditions in developmental testing that approximate more closely the most realistic operational test conditions. Such efforts will increase the opportunities for combining information since they will lessen the differences between developmental and operational testing.

One of the discussants, Fred Myers, argued further that if combining data is to be part of the operational test evaluation, it must be factored into the entire testing strategy and planning. It should be described in the test and evaluation master plan so that the developmental and operational test environments can be linked in some manner. This is a natural assignment for an integrated test team. Further, if contractor (as opposed to governmental) developmental test data are to be used, a better understanding is needed of the specific test conditions used and what the results represent, and there must be full access to all of the test data.

Myers added that some caution is needed because of the requirement of Title 10 U.S.C. 2399 for the independence of the operational tester. To effect this combination, operational evaluation data must be validated by the operational tester independently of the developmental tester. Another caution is that for combining information models to have a good chance of success, it must be determined that the prototypes used in developmental testing are production-representative of the system. If not, this complicates

the relationship between the reliability of the system in developmental and operational test events.

The second discussant, Ernest Seglie, expressed his belief that experts in operational test agencies would be able to guess the ratio of the mean waiting time to failure for operational test to that for developmental test (λ in Steffey's notation). Presumably, program managers would not allow a system for which they were responsible to enter operational test unless they were relatively sure it could meet the reliability requirement. In data just presented to the Defense Science Board (2000), one finds that 80 percent of operationally tested Army systems failed to achieve even half of their requirement for mean time between failures. This finding reveals that operational test is demonstrably different from developmental test, and that perhaps as a result, priors for λ should not be too narrow since it appears that the information on system performance is not that easy to interpret.

Seglie stated that there are two possible approaches for improving the process of system development. First, one can combine information from the two types of testing. Seglie admitted that he had trouble with this approach. The environments of developmental and operational testing are very different with very different failure modes. In addition, combining information focuses too much attention on the estimate one obtains instead of on the overall information about the system that one would like to give to the user from the separate test situations. It is extremely important to know about the system's failure modes and how to fix them. Therefore, one might instead focus on the size of λ and use this information to help diagnose the potential for unique failure modes to occur in operational test and not in developmental test. Doing so might demonstrate the benefits of broadening the exposure of the system during developmental test to include operational failure modes. Until a better understanding is developed of why developmental and operational tests are so different, it appears dangerous to combine data. A better understanding of why λ differs from 1 should allow one to incorporate (operational test-specific) stresses into developmental testing to direct design improvements in advance of operational test.

Samaniego commented on Seglie's concerns as follows. First, it should be recognized that Bayesian modeling has been shown to be remarkably robust to the variability in prior specifications. The Achilles heel of the Bayesian approach tends to be overstatement of the confidence one has in one's prior mean (that is, in one's best guess, a priori, of the value of the parameters of interest). With prior modeling that is sufficiently diffuse,

accompanied by a sensitivity analysis comparing a collection of plausible priors in a given problem, what some think of as the "arbitrariness" of Bayesian analysis can be minimized (see, e.g., Samaniego and Reneau [1994], for results and discussion on this issue). Second, the use of methods for data combination by no means precludes the investigation of failure sources and the various steps that are associated with reliability growth management. Samaniego suggested that it is critical for theoretical and applied research to proceed simultaneously and interactively on both of these fronts.

In the floor discussion, Larry Crow asked whether anyone had compared failure modes in developmental versus operational test. Specifically, he wanted to know how much of the increased failure rate in operational test is due to new failure modes and how much is due to higher failure rates of known failure modes. Jim Streilein responded that his agency has an idea of what λ is for some systems. He added that efforts are being made to carry out failure modes and effects analyses, but the problem is that these estimates are never fully cognizant of the environment of use. Further, for most components, there is no physical model that can be used to provide reliability estimates for a given material or manufacturing process. Streilein is therefore not sanguine about combining information models until more information is available.

TWO MODEL-BASED APPROACHES TO SELECTING INPUTS FOR SOFTWARE TESTING

It is well known that software is essentially a ubiquitous component of today's complex defense systems and that software deficiencies are a primary cause of problems in defense system development (see Mosemann, 1994). Software reliability, while sharing some aspects of hardware reliability, is clearly distinct in critical ways. For example, the smallest change to a software system can have a dramatic impact on the reliability of a software system. Further, there are no analogous notions to burn-in or fatigue for software components. Given the distinct nature of problems involving software and the broad aspects of the subject area, the hope during the planning stages of this workshop was that an entire subsequent workshop would be devoted to this issue, for which two presentations on software engineering at this workshop would serve as a preview. (This workshop was held July 19–20, 2001.)

This preview session outlined two approaches to the selection of in-

puts to software systems for testing purposes. The first selects a very small set of inputs with the property that all pairwise (more generally *k*-wise) choices for fields (which comprise the inputs) are represented in the test set. It has been demonstrated empirically that a large majority of the errors can be discovered in such a test set. The second approach uses a graphical model of software usage, along with a Markov chain representation of the probability of selection of inputs, to choose test inputs so that the high-probability inputs are selected for the testing set. The session's goal was to demonstrate that a number of recently developed statistical methods in software engineering are proving useful in industrial applications.

The Dalal Paper

The number of potential inputs to a software system is often astronomically high. In testing a software system, therefore, there is much to be gained by carefully selecting inputs for testing. Consider the interoperability problem in which a number of component systems must interact smoothly, and each of the components has a separate schedule for release of updated versions. To test the combined system, one must consider the (potentially) large number of possible configurations, with each configuration representing the joint use of specific releases. Given the time required to put together a specific configuration, the value of techniques that can reduce the number of tests needed to examine the reliability of all *k*-wise (for some integer *k*) combinations is clear. More generally, for any software system, input fields play the same role as system components in the interoperability problem, except that the combinatorial complexity is typically much greater. An empirically supported assumption is that the large majority of software faults are typically due to the interaction of a small number of components or fields, often just two or three. Thus one approach to software testing in this situation is to test all two-way (and possibly three-way) combinations of configurations or fields. That is, for two-way combinations, each version of the first software component is used with each version of the second software component, and so on.

Siddhartha Dalal discussed new test designs that include inputs, for a very modest number of test runs, covering all possible two-way (or generally *k*-way) combinations of fields. For even moderate-sized problems, these designs include dramatically fewer test cases than standard designs having the given coverage property. Consider, for example, a user interface with 13 entry fields and 3 possible values per field. In this case, the number of

possible test cases is extremely large—1,594,323. (In more realistic applications, it is common to have 75 or 80 different fields; in that case, even with dichotomous field inputs, one would have billions of potential inputs that might require testing.)

One possible approach for efficiently choosing inputs that cover all two- or three-way interactions involves the use of orthogonal arrays. An example of an orthogonal array with seven fields, each with two possible input values, is shown below in Table 2. In this example, eight test cases can be identified that provide coverage of all pairwise field values for each of the seven fields. For instance, the possible input pairs of field 1 and field 2 are (1,1), (1,2), (2,1), and (2,2). All of these possible pairs are represented in the Field 1 and Field 2 columns of Table 2; the same is true for any combination of two fields. Unfortunately, there are substantial problems with the use of orthogonal arrays. First, they do not exist for all combinations of fields and numbers of values per field. For an orthogonal array to exist, each field must have the identical number of values per field. More constraining, orthogonal arrays exist only for the case of pairwise changes, leaving open which approach to use for examining simultaneous changes to three, four, or more fields.

Dalal and others have shown that there are designs offering the same advantages as orthogonal designs but having substantially fewer test cases to examine. The reason for the inefficiency in orthogonal designs is that they are constrained to cover all pairs (or triplets, and so on) of field values an equal number of times. The new approach to this problem is referred to

TABLE 2 Example of an Orthogonal Array for Seven Fields, Each with Two Possible Values

Test	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7
# 1	1	1	1	1	1	1	1
# 2	1	1	1	2	2	2	2
# 3	1	2	2	1	1	2	2
# 4	1	2	2	2	2	1	1
# 5	2	1	2	1	2	1	2
# 6	2	1	2	2	1	2	1
# 7	2	2	1	1	2	2	1
# 8	2	2	1	2	1	1	2

as Automatic Efficient Test Generator (AETG) (see, e.g., Cohen et al., 1996; Dalal et al., 1999). With AETG, one ignores the requirement of balance, and as a result one can identify designs that select fewer test cases that maintain coverage of the same pairwise (or *k*-wise) combinations of field values. In the test outlined above with seven fields, eight test cases are required. AETG, in contrast, generates the matrix of test inputs shown in Table 3 for the problem of 10 dichotomous fields. Here, even with an additional two fields, one is able to test all pairwise field values for all combinations of two fields with only six test inputs. In the case of 126 dichotomous fields, one needs only 10 test cases. There is a great deal of interesting mathematics associated with this new area of combinatorial design, with much more work remaining to be carried out.

For the problem mentioned at the outset of the presentation—13 fields, each with 3 possible values—AETG produces the array of test inputs shown in Table 4, with the given input values for the 13 fields. It was necessary to have 1.5 million inputs to cover all of the possible combinations of field values. With AETG, however, if one requires only test cases that cover all pairwise input values, one needs only the 15 test cases shown in Table 4. (It is, of course, important to consider the consequences of using a model of such modest size for problems in which the natural parameter space is of very high dimension. Therefore, sensitivity analysis is recommended to validate such an approach.)

It should be stressed that in real applications, the problems are typically more challenging since various complicating constraints operate when one is linking fields of inputs. Such complexity also can be accommodated with this methodology.

More broadly, AETG represents a game plan for efficiently generating test cases, running these cases to identify failures, analyzing the results, and making improvements to the software system, and then iterating this entire process to attain productivity and quality gains.

Jerry Huller at Raytheon has used this procedure and obtained a 67 percent cost savings and a 68 percent savings in development time. The effort required to carry out AETG is typically longer than is generally allocated to testing because the approach requires careful attention to the constraints, the various fields, and so on. Therefore, a cost-benefit argument must be made to support its use.

TABLE 3 AETG Test Design for Ten Dichotomous Fields Covering All Pairwise Input Choices

Test	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7	Field 8	Field 9	Field 10
# 1	1	1	1	1	1	1	1	1	1	1
# 2	1	1	1	1	2	2	2	2	2	2
# 3	1	2	2	2	1	1	1	2	2	2
# 4	2	1	2	2	1	2	2	1	1	2
# 5	2	2	1	2	2	1	2	1	2	1
# 6	2	2	2	1	2	2	1	2	1	1

TABLE 4 AETG Test Design for 13 Trichotomous Fields Covering All Pairwise Input Choices

Test	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7	Field 8	Field 9	Field 10	Field 11	Field 12	Field 13
#1	1	1	1	1	1	1	1	1	1	1	1	1	1
#2	1	2	2	2	2	2	2	2	2	2	1	1	1
#3	1	3	3	3	3	3	3	3	3	3	1	1	1
#4	2	1	1	2	2	2	3	3	3	1	2	2	1
#5	2	2	2	3	3	3	1	1	1	2	2	2	1
#6	2	3	3	1	1	1	2	2	2	3	2	2	1
#7	3	1	1	3	3	3	2	2	2	1	3	3	1
#8	3	2	2	1	1	1	3	3	3	2	3	3	1
#9	3	3	3	2	2	2	1	1	1	3	3	3	1
#10	1	2	3	1	2	3	1	2	3	1	2	3	2
#11	2	3	1	2	3	1	2	3	1	2	3	1	2
#12	3	1	2	3	1	2	3	1	2	3	1	2	2
#13	1	3	2	1	3	2	1	3	2	1	3	2	3
#14	2	1	3	2	1	3	2	1	3	2	1	3	3
#15	3	2	1	3	2	1	3	2	1	3	2	1	3

The Poore Paper

Jesse Poore described a model for software testing that has as its basis a Markov chain representation of the transition from one state of use to another as a software system executes, where the transitions are generated by the user taking various actions. (See, e.g., Whittaker and Thomason, 1994; Walton et al., 1995). As mentioned above, there is a potentially astronomical number of scenarios of use of a software system. It is therefore natural to consider using statistical principles as a basis for the selection of the inputs to use in testing.

The particular approach taken makes use of an operational usage model, which is a formal statistical representation of all possible states of use of a system. The specific structure of the model is a directed graph, where the nodes are the states of use and the arcs are possible transitions from one state of use to another. (The states of use should not be confused with the state of the software system operating in a computer, that is, which line of code is being executed.) The states of use can be defined at any desired level of the software's natural hierarchical structure. Thus, for example, if a specific module is well known to be error free, one might model only the entering and exiting of that module. The stimuli that cause the current state to change are represented on these arcs connecting the states of use. For example, a human being hitting a key on a keyboard or pointing and clicking a mouse is a stimulus that can result in a state change. On the directed graph are a starting state and a terminal state, and use of the system is a path from the starting to the terminal state across the arcs. If one is interested in generating random test cases, one can use information on how the software is used to provide probabilities for these paths.

Given this structure, it is natural to model the working of the software system as a Markov chain, which assumes that the probabilities of moving from one state to another (the transition probabilities) are independent of the history of prior movement, conditioned on knowledge of the current state of the system. These transition probabilities can be set to values based on knowledge of specific environments of use, or in the absence of this knowledge, can be set uniformly across all states with which the given state "communicates." (That is, some movements from one state to another may be forbidden given the assumed functioning of the system, and these are the states with which a given state does not communicate.)

While a system has only one structure of states and arcs, it may be applied in many different environments. For example, a word processor

may be used by a novice writing a letter or by a technical writer, and the two will use the system differently. The usage environment, which provides the probability of a transition from each state to any other state in the system, is a modeling structure that can and should be implemented separately in each intended environment of use, therefore requiring different transition probabilities. The selection of test inputs based on a model of user behavior permits the software test to focus on those inputs that are most likely to appear in operational use

"Testing scripts" are also attached to each arc. By this is meant not only that the stimuli or the inputs that would cause movement along each arc are attached, but also that, when the input is applied and this arc is followed, there is an anticipated change in some output that accompanies the state change. Failure to observe this change will indicate an error in the code. For manual testing, the expected result given a stimulus can be checked against the actual result given written instructions. Using automated testing, one can add commands to the automated test equipment on each arc so that software commands are communicated to the test equipment when a test case is generated; these commands can be used to automatically measure the degree of agreement with the software requirements.

Development of these models is a top-down activity. There is generally a model initially at some high level of aggregation, with more detail being added incrementally as more of the detail of submodules is required. As submodules are incorporated into other software products, the associated models can be transplanted as well.

This Markov chain representation of the functioning of a software system supports the following standard analyses (and others) that derive from a test: (1) the probability that at a random time the system is in a given state; (2) the expected number of transitions in a test (and the variance in the number of transitions); (3) the expected number of transitions until a state first appears (and the associated variance); and (4) importantly and in a rigorous sense, how much testing is enough. These outputs can be compared with external understanding of the system to judge the validity of the model. (Small changes in transition probabilities can have unanticipated large impacts on the probabilities of paths through the system.) Various measures can also be developed to estimate test resources required to provide a given level of understanding of system performance. Some of this analysis can be carried out before coding. Therefore, one could use this analysis to argue against development of an untestable system.

In addition to these more standard analyses, one can obtain, through simulation, estimates of answers to any additional well-framed questions concerning states, arcs, paths, and so on. Also, one can specify transition probabilities by developing a system of constraints, that is, linear equalities and inequalities involving the transition probabilities, and then optimizing an objective function using linear programming.

Discussion of Dalal and Poore Papers

In the discussion of these papers, Margaret Myers asked software experts to examine four questions: (1) How does one estimate the reliability of a software system that is structured as a series of component software systems? (2) How does one address the integration of commercial off-the-shelf software systems, and how does one estimate the reliability of the resulting system? (3) How does one estimate reliability for a system that is being acquired in a spiral environment? (4) What sort of regression testing is useful in evolutionary acquisition?

Jack Ferguson pointed out that software methods are extremely important to DoD. DoD spends approximately $38 billion a year on research, development, testing, and evaluation of new defense systems, and it is estimated that approximately 40 percent of that cost is for software. Any method that can help DoD to make even a small improvement in software development can represent a large amount of savings. However, it should be kept in mind that the software is not always the problem. DoD is using software to do more and more, mainly to provide the flexibility required to meet new environments of use. Therefore, the problem is often fundamentally a domain-analysis problem.

Ferguson added that given the widespread use of commercial off-the-shelf systems, it is important to consider the use of black-box testing, as discussed in Poore's (and Dalal's) presentation. The workshop devoted to software reliability should devote a good deal of time to these methods.

Ferguson remarked that the traditional way of determining software reliability required a great deal of inspection, which is no longer workable. It was stressed that models such as those used by Dalal and Poore need to be developed as early as possible in system development, since, as with hardware systems, it is generally much less costly to fix a problem discovered early in the design phase.

CURRENT MODELS FOR ESTIMATING THE FATIGUE OF MATERIALS

Fatigue is likely the leading cause of failure of military hardware in the field. Thus it is extremely important for DoD to develop a better understanding of the sources of fatigue, to control the rate of fatigue, and to measure and predict fatigue in deployed hardware. Therefore, a high priority at the workshop was a session on the evolution and current status of statistical work in fatigue modeling.

The field of fatigue modeling lies at the interface of the disciplines of statistics and materials science, and success stories in this field invariably involve close collaboration between both disciplines. Materials scientists understand the structure and properties of the relevant materials while statisticians can model the behavior of these materials and analyze experimental or observational data that help refine these models. The products of this collaboration form the basis for replacement and repair policies for fatigue-prone systems and for the general management of hardware subject to fatigue. It is important to pursue efforts to enhance the statistics/materials science collaboration.

The Saunders Paper

Sam Saunders provided a historical perspective on fatigue modeling. Attention to this problem stems from analysis of the Comet, a post–World War II commercial jet aircraft. In the mid-1960s, around two dozen separate deterministic fatigue decision rules had been published, but none of them was very successful. The most accurate on average was found to be Miner's rule: that the damage after n service cycles at a stress level that produces an expected lifetime of N cycles is proportional to n/N. Subsequently, it was proven that Miner's rule made use of (apparently unknowingly) the expected value for fatigue life assuming that damage increments were generated from a specific class of distribution functions. However, the distribution of fatigue life *about* its expectation was either not considered or ignored.

A useful stochastic approach to the problem of fatigue modeling was provided by the development and application of the inverse Gaussian distribution (see, e.g., Folks and Chhikara, 1989). (The Birnbaum-Saunders distribution was developed first, but it is a close approximation to the inverse Gaussian distribution, which is easier to work with analytically.) Some

mathematical details are as follows. Let $X(t)$ denote cumulative damage until time t, and assume that $X(t) \geq 0$. Assuming that as t grows, $X(t)$ becomes approximately normal, and assuming that $\mu\ E(X(t)) = \mu t$ and $\mathrm{Var}(X((t)) = \sigma^2 t$, the distribution of the first t at which $X(t)$ exceeds Δ can be shown to be:

$$F(t) = \Phi\left(\left(\frac{\sigma}{\sqrt{\mu\Delta}}\right)^{-1}\left[\left(t\frac{\mu}{\Delta}\right)^{1/2} - \left(t\frac{\mu}{\Delta}\right)^{-1/2}\right]\right)$$

There are many generalizations of this argument, including those (1) for means and variances that are other functions of t, (2) for means and variances that are functions of other factors, and (3) where the distribution of $X(t)$is substantially non-normal.

Saunders described a current application derived from the generalized inverse Gaussian distribution applied to waiting times to failure for polymer coatings. These coatings often have requirements that they last for 30–40 years. Modeling this requires some understanding of the chemical process of degradation, which in turn entails understanding how sun, rain, ultraviolet exposure, temperature, and humidity combine to affect coatings. Further, the chemistry must be linked to observable degradation, such as loss of gloss, fading, and discoloration. (For more details, see Saunders [2001].)

The Padgett Paper

Joe Padgett outlined several currently used approaches he has been researching that can be applied to either the modeling of the failure of material or systems due to cumulative damage or the modeling of crack growth due to fatigue. A good motivating example is the modeling of the tensile strength of carbon fibers and composite materials. In fibrous composite materials, it is essentially the brittle fibers that determine the material's properties. To design better composites, it is important to obtain good estimates of fibers' tensile strength. Numerous experiments have been conducted to provide information on the tensile strength of various single-filament and composite fibers. Failure of the fibers due to cumulative damage can be related to times to "first passage," that is, times at which a sto-

chastic process exceeds a threshold, which naturally suggests use of the inverse Gaussian distribution. (There is also empirical evidence to support this model.)

In his research, Padgett has used fatigue modeling in conjunction with accelerated testing. The overall strategy is to use fatigue modeling to draw inferences about system failure during normal use based on observations from accelerated use. A key example is the case in which the length (L) of a carbon fiber serves as the accelerating variable, since the longer a carbon fiber is, the weaker it is. A key assumption of accelerated testing is that there is a functional relationship between the acceleration variable, L here, and the parameters (mean and variance) of the failure distribution.

To carry out such a program, the system is first observed for various values of the accelerating variable. Then a model of system reliability as a function of the accelerating variable is developed. Finally, the model is used to translate back to estimate the reliability at levels of typical use. Of course, one is extrapolating a model to a region in which little or no data are collected, making the inference somewhat risky. For this reason, efforts to validate models and to derive models using relevant physical principles (physics-of-failure models) are of critical importance.

The above represents the general approach that is currently being explored. A specific model that can be applied to the situation of discrete damage is as follows (for details, see Durham and Padgett, 1997, 1991; a related method is discussed in Castillo and Hadi, 1995). Consider a load being placed on a system as increasing in discrete increments. The system is assumed to have an unknown, fixed strength Ψ. The initial damage to the system is caused by its manufacture and is assumed to be due to the most severe flaw in the system. This flaw is quantified by a random variable X_0, which can be the associated initial crack size or "flaw size." The system is then placed under increasing loads, with each additional increment of stress causing some further damage $D \sim F_D$. The amount of damage at each step i, D_i, is a random variable, which can be viewed as crack extension due to the added load. The initial strength of the system is denoted by W, which is the result of a reduction of Ψ by X_0. The approach assumes that additional increments of stress are loaded on the system until failure, and the model provides estimates of the mean number of increments of load to system failure, the mean critical crack size, and the full distribution of the number of increments until failure. The model is completed by the specification of the distribution of W, which should be based on knowledge of the physics of failure of the system of interest. Durham and Padgett (1991)

apply this methodology to the modeling of crack formation in gun barrels, using linear programming to estimate the various parameters.

In some situations, it may be more appropriate to model the damage process (e.g., crack size) as a continuous rather than a discrete process. Here, the "system" of size L is placed under continuously increasing loads (e.g., tensile stress) until failure. (Or analogously, a stress is increased until a crack extends, resulting in failure.) Various estimated acceleration functions are used to provide estimated parameters for the inverse Gaussian distribution. Examples include the power law model, the Gauss-Weibull additive model, or the Bhattacharyya-Fries inverse linear acceleration law. One can use various goodness-of-fit tests to determine which model fits the data best. Finally, approximate confidence intervals from maximum-likelihood considerations can be constructed.

Discussion of Saunders and Padgett Papers

In the discussion of these papers, Ted Nicholas outlined two methods for modeling fatigue currently used by the Air Force. First is a typical functional form that models expected lifetime (before a crack initiates) as a function of stress, which in the Air Force's case is the number of cycles an engine is operational. There is natural variability among individual systems about this mean; thus systems are designed so that the lower bound of, say, a 99.5 percent tolerance interval lies above a required level given a certain rate of cycles of use. This is referred to as the "design-allowable curve." An unsolved problem is that the functional form often must be fit on the basis of limited data, especially at the tail end with respect to amount of use.

Another approach to fatigue modeling used by the Air Force is the damage tolerance methodology. With the above approach, all of the items subject to fatigue are thrown out as soon as they have been operated the number of cycles determined by the design-allowable curve. Of course, a number of the items thus discarded have residual lifetimes many times longer than the number of cycles for which they were operated. This process can therefore be extremely expensive. The damage tolerance methodology assumes that one is sophisticated at predicting how cracks grow. If there is an inspection interval at a given number of cycles, and one can be sure by examination that there are no cracks larger than the inspection limit, one can institute a process whereby any crack will be discovered during the next inspection period before it can grow to a critical size. To accomplish this, the intervals between inspections are set at

one-half or one-third of the minimum time it takes a crack to grow from the inspection limit to the critical size. This approach can play a role in engine design.

Nicholas then discussed a substantial remaining problem—high-cycle fatigue—which is due to low-amplitude, high-frequency vibrations. With low-cycle fatigue, a crack typically develops early in an item's life, and gradually propagates until it can be discovered when it grows to be larger than the inspection limit. With high-cycle fatigue, one typically has no indication of any fatigue damage until it is almost too late. As a result, there is as yet no reliable method for detecting high-cycle fatigue damage in the field. The current idea is to stay below a statistically significant stress level so the item will never fail. (This is not a notion of accumulated stress, but a notion of current stress levels.) However, there is no guidance on what to do when there are transient events during which the stress level exceeds the limit. If the perspective of accumulated stress is taken, should cycles during these transient events count more than cycles within the stress limit? The model that underlies this approach is that there are distributions of stresses and of material strength, and one does not want to have a pairing in which the individual stress received from the stress distribution exceeds the individual strength received from the strength distribution. The important complication is that the strength distribution for an aging system is moving toward lower values during service, and it can decrease substantially as a result, for example, of damage from a foreign object.

Fighting the accumulation of stress is highly complicated. Both vibratory and steady stresses must be considered, along with the statistics of material behavior. In addition, computational fluid dynamics plays a role. One must also take into account the effects of friction, damping, and mistuning. Finally, one can have certain types of fatigue failures, only say, when an aircraft is flown under particular operational conditions. Linking a vibrational problem with flight conditions is important, but can be extremely difficult. Nicholas agreed that enhanced communication between materials scientists and statisticians is needed to continue work on these issues.

RELIABILITY MANAGEMENT TO SUPPORT ESTIMATES OF SYSTEM LIFE-CYCLE COSTS

The workshop sponsors were extremely interested in exploring the issue of how early reliability assessments of defense systems in development

might be used to address issues involving the life-cycle costs of proposed systems and systems under development. Defense systems incur costs during the development process (including testing costs), costs in production, costs through use and repair, and sometimes redesign costs. Maintenance, repair, and redesign costs increase with the decreased reliability of a system and its components. Today there is a widespread perception within DoD that the percentage of the costs of defense systems that is incurred after production is too large, and thus that greater resources should perhaps be expended during the design and development stages to reduce postproduction costs, thereby reducing life-cycle costs. Estimating life-cycle costs and their contributing components can help in evaluating whether this perception is true and what specific actions might be taken to reduce life-cycle costs to the extent possible.

An introduction to the session was provided by Michael Tortorella of Bell Laboratories, who discussed some general issues concerning warranties and life-cycle costs. Systems with different reliabilities can have substantially different production costs. In industry, given a cost model that is sufficiently precise, it is possible to offer maintenance contracts or warranties that can be profitable to the producer.

Two primary areas of focus in the field of reliability economics are risk analysis and spares management. Risk analysis involves a supplier who needs to assess the probability that a product and a warranty will be profitable, which requires estimation of system life-cycle costs. A way to think about risk analysis is that every time a supplier produces a product or warranty for sale to a customer, the supplier is placing a bet with the company's money that the product or warranty will be profitable. Reliability engineering represents an attempt to improve the odds on that bet. Spares management involves inventory investments, storage costs, transportation costs, and the consequences of outages during delays. Two approaches used are (1) stocking the spares inventory to a service continuity objective, which means stocking an inventory to ensure that, with some designated probability, a spare will be available; and (2) the preferred approach of taking into consideration the various costs associated with different stocking strategies and minimizing those costs while meeting the availability objective of the system. (For more information, see Chan and Tortorella [2001], Blischke and Murthy [1998], Murthy and Blischke [2000], and a variety of papers in Tortorella et al. [1998].)

The Blischke Paper

Wallace Blischke provided an overview of the analysis of warranties and life-cycle costs. Analysis of life-cycle costs typically is carried out from the point of view of the producer, examining the costs of a system from conception to withdrawal from the marketplace. The earlier life-cycle and associated costs can be estimated, the better it is for the decision maker, though the earlier in development these estimates are attempted, the more difficult they are to produce. Blischke stated his preference for a Bayesian approach in this effort, since that paradigm provides a basis for the use of engineering judgment and information derived from similar systems, as well as a natural method for updating predictions.

It is important not only that reliable estimates of life-cycle costs be produced, but also that reliable estimates of their uncertainty also be developed and communicated to assist in decision making. Further, an understanding of the origin of the uncertainties can help in assessing how best to improve the quality of future predictions. This is especially true for defense systems, which of course can be much more complex than consumer goods. (For example, costs for defense systems sometimes include disposal costs, which can be nontrivial.)

The Bayesian approach is initiated before initial testing with the use of all available information to form a prior distribution describing system reliability. Prototypes are then produced and tested. The data from these tests are employed using Bayes' theorem to update the prior distribution to form a posterior distribution, and the posterior distribution is used in turn to produce estimates and prediction intervals concerning parameters that govern life-cycle costs, the profitability of warranties, and related constructs.

As an example, Blischke discussed the analysis of life-cycle costs for a propulsion system in development. To achieve a required level of reliability, preliminary reliability levels are specified for the basic subsystems and components. Some of the standard tools used for this purpose are fault trees; reliability block diagrams; and failure modes, effects, and criticality analysis. One important issue is whether reliability problems are due to the design, the process, or the operations. Often, operational errors are more important than design errors. Engineering judgment based, for example, on information on components used in previous propulsion systems, can support a preliminary Bayesian assessment of system reliability (although such information will be very limited when the system involves a new tech-

nology). This analysis is followed by a detailed design analysis and then full-scale testing, leading to an operational system.

Blischke then focused on warranty concepts and costs. A warranty is a contractual agreement between the buyer and the seller that establishes buyer responsibilities and seller liability, and provides protection to both buyer and seller. Cost models are used to examine the properties of a given warranty; as the reliability of a system increases, the cost of a given warranty decreases. On the other hand, producing a highly reliable system is likely to require large up-front costs, which suggests a trade-off between the costs of fielding and those of development and production that needs to be understood and analyzed. Optimization approaches can be used in performing this analysis.

Warranty costs can be predicted empirically if enough systems are produced early. The advantage of doing so is that no modeling assumptions are needed. In the defense area, this possibility is less likely. Alternatively, one can carry out testing on prototypes or components of prototypes to obtain information on the distribution of waiting times to failure so they can be modeled.

For a simple, real example, Blischke analyzed a free-replacement, nonrenewing warranty (i.e., the replacement item is warrantied to work for the time left in the original warranty period). In this example, the supplier agreed to provide a free replacement for any failed component up to a maximum usage time. If at the end of the warranty period the purchased item was still working, the next replacement would be paid for by the buyer.

Blischke presented some mathematical details, using the following notation: w is the warranty period, $m(w)$ is expected number of replacements per item during the warranty period (the renewal function), c_s is average cost per item to seller, and c_b is average cost to buyer. The cost of offering a free-replacement warranty is analyzed as follows. First, the cost to the buyer of each new item is c_b. The expected cost to the seller, factoring in the cost of prewarranty failures, is $c_s(1 + m(w))$. For the exponential time-to-failure distribution with mean time to failure of $1/\lambda$, $m(w) = \lambda w$. In this case, it is easy to determine when $c_s(1 + m(w))$ is less than c_b. Unfortunately, for distributions other than the exponential, the renewal function can be difficult to work with analytically. However, software exists for calculating renewal functions for the gamma, Weibull, lognormal, and inverse Gaussian distributions and various combinations of these.

From the buyer's perspective (ignoring the profit of the seller), the

expected life-cycle costs in a life cycle of length L with warranty period w can also be computed. This computation involves a different renewal function that is the solution to an integral equation. This perspective may be more appropriate for DoD.

Blischke indicated that reliability-improvement warranties are popular in the defense community. These are warranties in which the seller will provide spares and field support, analyze all failures, and then make engineering changes to improve reliability for a given period of time for an extra fee. As above, to develop a model of life-cycle costs, one must model all of the various cost elements and the probabilities associated with each. One of the first models for the expected cost of an item sold under a reliability-improvement warranty is described in Balaban and Reterer (1974). This and other similar approaches are based on a comparison of expected costs. To have a model that can answer more complicated questions, however, one must have a representation of all probabilistic elements and distributional assumptions, rather than simply an analysis of expected values.

In general, there are various models of life-cycle costs from either the buyer's or the seller's point of view for various kinds of warranties (e.g., pro rata, combination warranty, just rebates). To decide which defense systems should be developed to carry out a task from the perspective of minimizing life-cycle costs, one must derive estimates of life-cycle costs relatively early in the system development process, which, as mentioned above, is very difficult to do well. With respect to just the warranty component of life-cycle costs, a Bayesian approach has some real advantages. First, one collects all relevant information, including data on similar systems, similar parts, materials data, and engineering judgment, and aggregates this information into prior distributions for system reliability. All the information collected must be accompanied by an estimate of its uncertainty to elicit the spread of the prior distributions. (For details, see Blischke and Murthy, 2000; Martz and Waller, 1982.) Such reliability assessments require the use of logic models that relate the reliability of various components for which one may have some real information to the reliability of the entire system. (For specifics, see Martz and Waller, 1990; Martz et al., 1988.) These priors are then used to predict parameters of the distribution of total time to failure, which can be used to predict warranty costs. Bayesian methods are then used to update these priors based on new information on component or system reliability.

Blischke believes a comprehensive Bayesian cost prediction model can be based on an analog to a Bayesian reliability prediction model. Roughly

speaking, costs are another element besides reliability for which one acquires and updates information. This computation is very similar to those used in PREDICT. One complication is that reliability and costs are related, so a bivariate model may be needed (see, e.g., Press and Rolph, 1986). Given today's computing capabilities and the recent development of powerful new ways to carry out Bayesian computations, this approach is likely feasible. Thomas and Rao (1999) can serve as an excellent introduction to many of these ideas. Finally, to support this approach to life-cycle costs and warranties, information on systems, tests, costs, and reliabilities all must be maintained in an accessible form.

The Camm Paper

Frank Camm outlined some management hurdles that complicate the application of life-cycle cost arguments in DoD acquisition. Camm made four major points. First, the policy environment provides an important context for examining system life-cycle costs. Second, improved tools for assessment of life-cycle costs can aid DoD decision makers in their pursuit of priorities relevant to reliability as a goal in system development. However, those tools cannot change the priorities themselves. Third, as systems age the demands for reliability seem to increase, probably because of changes in the way systems are deployed. Fourth, initiating a formal maturation program provides a setting in which to conduct reliability analysis, as well as an element of acquisition planning important to the projection of future system reliability.

Most expenditures per unit time of a defense system in development are paid out during production, and the fewest are made during the initial design phase. However, the majority of expenditures for a system are postproduction, including operations and related support as well as modifications. Because these costs are viewed as being far in the future, they are to some extent ignored during acquisition.

In producing estimates of life-cycle costs for a system in development, it is usually necessary to base the analysis on several assumptions, some of which are tenuous. For example, the years for which a system will be in operation are difficult to predict. The Air Force currently has weapon systems that are expected to be in use well past their intended length of service. Further, various complicating factors that are difficult to incorporate into an analysis should be taken into account when estimating life-cycle costs. Some examples are use of the system for purposes not originally

envisioned (e.g., flying a plane faster than planned), operating an engine for more cycles than planned, flying a plane with a different profile, or using a different support plan. These factors can lead to substantial changes in the reliability of a system and hence in life-cycle costs. Also, for nonmodular systems, postproduction modifications can be very difficult to predict a priori and are often relatively costly. One needs to be aware of the sensitivity of estimates of life-cycle costs to these kind of changes and their impacts on system reliability.

Camm added that the source of failure in many systems turns out to be a single essential component that was poorly designed or improperly installed or integrated into the overall system. These are the "bad actors" (see Chapter 4 for further discussion), and if they could be identified and fixed early, performance could be substantially improved.

From the perspective of defense system development, it will often be important to choose between reducing system costs and increasing system capabilities. Alternatively, one could decide to add greater flexibility into the system so that it will be easier to make modifications when necessary. This latter approach makes it possible to learn more about the system and facilitates improvement of the system over time. The underlying question of how much reliability is enough requires a highly complex set of analyses examining a variety of difficult trade-offs between increasing design and production costs and reducing operating costs.

Some external changes will have predictable impacts on the importance of system reliability. For example, decreases in the discount rate make future money more valuable, and therefore result in a decision to build more-reliable systems. Also, as the size of a fleet grows, the payoff from system development becomes greater, and this, too, argues for greater reliability.

This perspective has other implications as well. For example, it might be useful to have much smaller fleets but still allow for the use of many components across all fleets. Doing so would increase the total number of components being built, which in turn would make it possible to learn something about the components in one weapon system that might be useful in another. This is a key aspect of system maturation. Also, if one increases the expected life of a fleet or the amount it is used, one is forced to place greater priority on system reliability. Considerations of system reliability are also dependent on deployment. When projecting the costs of a force deployed from the United States, one must consider the costs of the entire support base associated with the deployment, which is again a function of system reliability.

Questions about system reliability must be addressed within a context, and a component of that context is the incentives that exist within DoD concerning reliability. An interesting question is why DoD continues to make optimistic estimates of system reliability and costs early in development. The reason for this is that a constituency needs to be formed for systems in development, and to this end a system must be promoted. Once various groups have become committed to a system, they are more likely to support increases in costs or reductions in reliability. Developers become accustomed to viewing early estimates of final system reliability as somewhat unrealistic goals. Also, there is an incentive to wait to deliver bad news about system costs and reliability so a program is not threatened while its constituency is being developed. As a result, fixing problems becomes more costly. Early overoptimism, then, is likely not a problem of analysis, but instead one of incentives surrounding the system development process within DoD.

Another context issue lies behind the evident difficulty of making the argument that reliability improvement would be cost-effective. This difficulty is explained by at least three factors. First, there is a historical lack of emphasis of system reliability relative to system effectiveness. Second, there are very separate environments within DoD for those involved in initial design (acquisition) and in system operations (logistics), in part because of the differing time horizons of these activities. Third, Congress tends to view funding decisions within a short time frame. That is, it is difficult to argue that an extra $5 million spent this year will save $30 million over the next 10 years when those charged with distributing funds are concerned primarily with the next year or two. This problem resulted in formulation of a concept referred to as "cost as an independent variable," which asserts that cost is a relevant factor in evaluating a proposed answer to a defense need.

One approach to increasing defense system reliability through a change in context is the introduction of warranties for defense systems. Two possibilities have been suggested. The first is compliance with specifications on delivery. This is basically an acceptance inspection to ensure that specifications are met. Once the specifications are met, the warranty is over. The second possibility is performance warranties, which promise a certain level of performance from a system over a period of time. This approach is being promoted by many as an answer to the reliability shortcomings of defense systems, but it raises some difficult questions.

First, this increase in reliability would come at a price since warranties

are not free. Paying for these warranties raises some of the contextual problems identified above. Assuming funding can be arranged, there are real analytic and implementation problems. One key problem is identifying a reasonable set of incentives or penalties for keeping or not keeping the promises made. Also, it is difficult to enforce these warranties because of the many factors that are not under the control of the developer and producer of the weapon system. These factors include the behavior of the user, the environment of use, the operational support system, and the specific counterforces used against the system. It is difficult to prove that a system has not met performance requirements.

Discussion of the Blischke and Camm Papers

Allen Beckett agreed that reliability estimation is extremely challenging. How does one estimate the reliability of an engine that has been overhauled? What are its failure modes? How does one develop a spares budget for a system in development? Beckett acknowledged that collecting relevant data is necessary and that there are promising modeling approaches. But the key is to understand the operational issues so the models will represent all of the complexities.

The related issue of surveillance testing was raised by Rob Easterling. The objectives of such testing are to find and fix reliability problems, and then update the estimate of system reliability. For defense systems, surveillance testing poses some difficult problems. First, in the case of complicated defense systems, it is unlikely that a large number of replications for surveillance testing will be available. Second, there are a multitude of environments and missions with potentially different reliabilities and failure modes for a given system. The fundamental complication, however, is that it is difficult to quantify the goals of surveillance testing. For example, what is the tolerable probability of failing to detect a fault leading to a reduction in reliability of more than 10 percent in 2 years of testing? It may be that surveillance test plans cannot be driven solely by statistical arguments.

However, statistical constructs should be part of the decision process regarding surveillance testing. Like a warranty, surveillance testing should be considered a type of insurance policy, whereby some amount of protection is being purchased for the price of additional testing. The general point is that, regardless of how a surveillance test plan and the associated decision rule based on the test results are derived—whether through eco-

nomic analysis, classical hypothesis testing, Bayes methods, or the like—a test has (statistical) operating characteristics. These operating characteristics are the probabilities of making various decisions based on underlying properties of the system under test. (The operating characteristics of a simple hypothesis test are the probability of rejecting the null hypothesis when it is true and the probability of failure to reject the null hypothesis when it is false.) Generalizing the notion of operating characteristics would provide the correct basis for decision rules; for example, a decision rule based on a specific test design would ideally have a high probability of passing a system that met the requirement and a high probability of failing a system that did not meet the requirement. Estimates of the operating characteristics of a test should be communicated to decision makers and recognized in the decision process. Finally, a decision rule could be enriched through use of information from such sources as simulations and developmental test results (a relevant paper is Fries and Easterling, 2002).

4

Further Discussion and Next Steps

This chapter begins by providing some general remarks concerning the collaboration of statisticians and DoD in carrying out research on reliability measurement. This is followed by discussion of the physics of failure, the need to develop and utilize reliability models that reflect the physical attributes of the materials used and the stresses to which these materials are subjected, and the need for procedures for identifying important modes and sources of failure in components and systems. The final section begins with a summary of some of the discussion at the workshop on appropriate ways to think about and quantify the uncertainties that accompany the statistical modeling of a complex system, and then presents highlights from the closing panel discussion and a summary of the general discussion that followed.

COLLABORATION BETWEEN STATISTICIANS AND THE DEFENSE COMMUNITY

The successful early collaboration between the statistics and defense communities was epitomized by the achievements of the Statistical Research Working Groups during World War II. Thereafter, academic research in experimental design, reliability estimation, and other areas at the interface of statistics and engineering was strongly supported by DoD. However, the level of collaboration has fallen off of late, possibly because the research was not fully targeted to DoD's most pressing needs.

Nevertheless, academic and industrial research on reliability methods has continued, and substantial progress has been made in the last 20–25 years. Areas of recent progress include: (1) methods for combining information across test environments, including methods for incorporating subjective assessments; (2) fatigue modeling; (3) statistical methods for software engineering; (4) nonparametric or distribution-free methods, specifically for reliability growth, but also more generally (an important example being for variance estimation); (5) alternative methods for modeling reliability growth; (6) treatment of censored or missing data; (7) use of accelerated testing methods; and (8) greater use of physics-of-failure models and procedures that are helpful in identifying the primary sources of failure.

Chapter 3 presented an argument expressed by several workshop speakers: that DoD needs not only to upgrade the "tried and true" reliability methods that could be disseminated in a handbook, but also to stay abreast of methods on which current research is being carried out or for which the full extent of the applicability of recent methods to defense systems is still unclear. Application of these methods may still require greater resources, but many of them are likely to provide important, substantial advantages over current methods. The issue is how the test service agencies and other members of the test and evaluation community can gain easier access to contemporary reliability methods. As discussed at the workshop, one important way to address this issue would be to identify properties of a reference book that could be made available to help provide this linkage between the defense and statistical communities. The primary means suggested for accomplishing this was updating or redesigning the RAM Primer.

PHYSICS-OF-FAILURE MODELS AND METHODS FOR SEPARATELY MODELING FAILURES FROM DIFFERENT SOURCES

While no session was devoted specifically to either greater use of physics-of-failure models (i.e., models that directly represent the physical basis for failure) in modeling reliability for defense systems or methods for separately modeling failures due to distinct sources, these two related topics arose repeatedly during various workshop sessions. Several speakers supported the greater use of physics-of-failure models whenever possible to acquire a better understanding of the sources and effects of component and

system failure. Of course, models that make no use of direct understanding of specific failure modes can still be useful in certain broad contexts, but their validity is, generally speaking, more questionable.

To make progress in the development and implementation of physics-of-failure models will require the interaction of statisticians and other scientists. Some systems or components will not benefit from this type of approach since the physics underlying the phenomenon, for one reason or another, is not mature enough. For many different types of defense systems, however, even partial knowledge of the underlying mechanism for various failure modes can be extremely helpful to assist in the statistical estimation of reliability. Some areas addressed by the workshop in which these points were made were (1) fatigue modeling, to help apply the proper acceleration function; (2) early assessments of system reliability for PREDICT; (3) help in categorizing failure modes into types A and B in the Integrated Reliability Growth Strategy (IRGS); (4) improved understanding of whether combining information from developmental and operational test is reasonable; and (5) Meeker's research linking developmental, operational, and field use by employing various acceleration models.

A related issue is separate modeling of the failure-time distributions for failures from different sources. This topic arose at the workshop in several somewhat unrelated contexts. First, IRGS considers separately fault modes that are characteristic of a mature component and those that are characteristics of a component still capable of further development, referred to as type A and B failure modes. Second, Frank Camm mentioned that failures in the field are overrepresented by poorly produced components, referred to as "bad actors," possibly resulting from a poorly controlled manufacturing process. Third, Bill Meeker noted that some failures are unpredictable (possibly because of changes in the manufacturing process) and therefore in need of separate modeling, which he referred to as "special-cause failures" as distinguished from "common-cause" failures.

Two related notions—the separation of a system's components into those that are mature and immature and the separation of the results of an industrial process into those systems that are indicative of the proper and improper functioning of the process—arose in these and other parts of the workshop. This separation of failures due to components or processes that are or are not functioning as intended is clearly worth greater investigation for its applicability to reliability analysis for defense systems.

There are problems involved in making this idea operational, that is, in designating which failures are due to mature or immature components, and

which production processes can be considered stable and which prototypes may have been poorly produced. However, an approach for estimating system or fleet reliability that would be consistent with this line of reasoning would be to analyze the pattern of failures from each source alone and separately model the impact on reliability. The expectation would be that decisions at the boundary would not make that much difference in such analyses.

A number of advantages could potentially result from this general approach. For example, better decisions could be made in separating out faulty designs from faulty processes. Also, some design failures might be attributed to a single component and easily remedied.

CONCLUDING PANEL SESSION AND GENERAL DISCUSSION

Models now exist or are being developed for representing how weapon systems work, such as missile intercept models and models of earth penetration. These models have input variables, for example, impact velocities. If one has a multivariate distribution for the inputs, one can run simulations and estimate a number of characteristics concerning model performance, such as system reliability. A partial means of understanding how useful these estimates are and how they should be compared or combined with real data is model validation.

Consider a computer model that is intended to simulate a real-world phenomenon, such as the functioning of a defense system. The validity of a computer model necessarily focuses on the differences between the model's predictions, y^*, and the corresponding observations of the phenomenon, y, that is, the prediction errors. To learn efficiently about the prediction errors, one designs an experiment by carefully choosing a collection of inputs, x, and then running the model, observing the system, and computing the prediction errors at those inputs. The prediction errors are then analyzed, often through development of a model of those errors as a function of the inputs. A candidate starting point for a prediction error model is that the prediction errors are normally distributed with parameters that are dependent on x; that is, $y = y^* + e_x$, where $e_x \sim N(m_x, \sigma_x)$. The objective of the prediction error model is to characterize the e_x's, which can be difficult to do since (1) x is typically high-dimensional; (2) the model may be numerically challenging and hence may take a long time to converge; (3) observing what actually happens can be extremely expensive, and as a result the number of separate experimental runs may be highly limited; and (4)

some of the input variables may be very difficult to control. A great deal of progress has been made on the simple linear version of this problem, but this is not the case for highly complicated nonlinear versions. Further, if one is required to make a prediction in a region of the x space where no tests have been run, one needs to extrapolate from the prediction error model, which necessitates a rather complete understanding of the underlying physics. This response-surface modeling could be equivalent in difficulty to building the computer model in the first place.

In the concluding panel session, Rob Easterling described some approaches that might address this problem: (1) leaving some x's out of the model, (2) using simplified variable and parameter spaces, (3) using simplified prediction error models, and (4) using fractional two-level factorial experiments. Easterling said that it can be shown that

$$\mathrm{Var}_x(y) = \mathrm{Var}_x(y^* + m_x) + E_x\left(\sigma_x^2\right).$$

where the subscript x under the variance and expectation operators indicates averaging over x as it varies according to its multivariate distribution. Since m_x is likely to vary much less than y^*, this equation can be typically modified as follows:

$$\mathrm{Var}_x(y) \approx \mathrm{Var}_x(y^*) + E_x\left(\sigma_x^2\right).$$

There is a large body of research exploring how much variance there is in y^* given the random variation in the inputs, using such methods as Latin Hypercube sampling. However, that term is a good estimate of the variance of y only if the second term, the variance of the prediction errors, is negligible. This is currently a relatively unexplored area of research.

A second panelist, Steve Pollock, stated that the DoD community needs to determine how best to apply the various methods described at the workshop. Doing so would entail directly implementing some methods when applicable and otherwise tailoring them, if necessary, to DoD's specific needs. Pollock added that additional workshops (structured similarly to this one) should be organized at regular intervals to help keep DoD abreast of recent advances in reliability methods.

A third panelist, Marion Williams, recommended that great care be exercised in using developmental test results in combination with opera-

tional test results. He suggested that the failure modes are so distinct that this linkage is unlikely to be useful for many systems. He added that he thought there was a place for Bayesian models in the evaluation of defense systems in development. A key issue for him is the justification of test sizes for operational testing. Williams' comments elicited a discussion of the place of Bayes methods in operational evaluation. David Olwell said that the key issue is that priors need to be selected objectively. Francisco Samaniego added that a sensitivity analysis using an appropriately broad collection of possible priors would be especially important in DoD applications since assessment of the influence of prior assumptions should be part of the subsequent decision-making process concerning a system's suitability.

A fourth panelist, Jack Ferguson (substituting for Hans Mark), returned to the theme of reliability management. He is convinced that testing and analysis must be moved upstream so that the system design is improved with respect to its operational performance as early as possible. These are the types of systems that generally work well in the field. Further, field data need to be used more often to update estimates of the costs of spares, maintenance, and so on.

Finally, we summarize discussions concerning the RAM Primer's current value in disseminating state-of-the-art reliability methods to the DoD test and evaluation community and the possible form of an updated version. These discussions occurred primarily in the concluding panel session of the workshop.

The DoD test and evaluation community currently has limited access to expert statistical advice (see, e.g., National Research Council, 1998). It is typical (and has been for decades) in all four services for both operational test planning and operational test evaluation to be carried out by individuals with relatively limited statistical training. For this reason, the RAM Primer served an important function for many years in communicating omnibus, easily applied techniques for test planning and test evaluation with respect to measurement of system reliability, availability, and maintainability. The chapters of the RAM Primer cover basic definitions, reliability measures, test planning, reliability models and estimation, hypothesis testing and confidence intervals for reliability performance, data analysis, and reliability growth estimation. Also included are tables and charts for assistance in applying the methods described.

Steve Pollock pointed out a number of areas in which the RAM Primer is currently deficient. These include the lack of discussion of physics-of-

failure models, nonparametric and robust methods, variance estimation (e.g., jackknife, bootstrap), stress testing, accelerated testing, decision-analytic issues, repair and replacement policies, methods for dependent components, current methods in experimental design, and Bayesian approaches. (Since the application of physics-of-failure and Bayesian models is highly specific to the system at hand, it is not clear that any omnibus approaches to the use of these models could be represented in an updated RAM Primer. However, it might be helpful to suggest the utility of these models and provide a casebook of successful and unsuccessful applications of these models for estimating or evaluating the reliability of defense systems.) The view expressed was that the RAM Primer does not currently serve any set of potential users very well.

Many participants at the workshop believe that, 20 years after its last revision, the RAM Primer is substantially out of date. Jim Streilein observed that in many respects, it was already limited in its utility in 1982. One indication of its obsolescence is that it contains a large section on statistical tables and graphs that provide critical values for tests, whereas today a modest amount of embedded software would provide better information. Other documents may also be obsolete; recall that Paul Ellner called for the updating of Military Handbook 189, on reliability growth modeling. Streilein was concerned more broadly about the training of tomorrow's reliability analysts.

To address this problem, a number of speakers strongly argued that the RAM Primer should be fully updated, possibly in a substantially different format. The suggestion was that a small planning group be charged with responsibility for deciding the goals and form of a new RAM Primer. The possibilities include (1) a primer, (2) a self-contained introductory text, (3) a set of standards, (4) a handbook, (5) a set of casebooks, and (6) a state-of-the-art reference book that well-educated and well-trained professionals could use to remind themselves of various methods. The form selected depends to some extent on whether the users are likely to be inexperienced junior analysts or experienced analysts. Perhaps several of these forms could be developed simultaneously. One interesting suggestion was for the RAM Primer to be a web-based document with embedded software or be linked to interactive software to carry out the variety of calculations necessitated by modern methods.[1] Taking this approach would provide greatly ex-

[1]An example of a handbook constructed in ths manner, the National Institute of Standards and Technology/SEMATECH Engineering Statistics Handbook, can be found at http://www.itl.nist.gov/div898/handbook/toolaids/sw/index.htm.

panded capabilities to the user with only modest demands for understanding the underlying theory.

Chapter 9 of the RAM Primer describes methods for modeling reliability growth. Current practice in DoD acquisition (i.e., absent the methods described in Chapter 2) makes it likely that a complex defense system will enter into the latter stages of development with system reliability substantially lower than that expected upon maturation. The expectation is then that system reliability will be improved through a series of steps of test, analyze, and fix. Typically for the latter stages of development, a limited amount of time is allocated for testing. To formally quantify system reliability at some point in time, one could rely solely on the results from the last testing carried out. However, the limited amount of testing implies less precise estimation than would be possible if more of the pattern of system reliability were somehow used. A key challenge here is that reliability under operational conditions differs from that under laboratory conditions, and appropriate linkage of the two is essential for the best estimation.

Several participants argued for greater use of reliability growth models that are consistent with the maturation process of test, analyze, and fix. Since Chapter 9 of the RAM Primer is focused on applying the power law process to all reliability growth problems, one objective of a revised RAM Primer would be either to augment the presentation there with a description of alternatives or to focus the presentation on these newer models. In this way, defense reliability growth modeling would (more often) take into account explicitly the process of fixing the faults found in testing.

SUMMARY

Ernest Seglie, the fifth panelist, provided a thorough summary of the workshop. First, the RAM Primer is antiquated, and a more useful tool needs to be developed. Second, the reason many systems are deficient with respect to their reliability when tested or fielded is in the management of the system development process. In particular, it must be understood that a change in taking a system to a new environment, using a new manufacturing process, or employing new users will cause a break in the reliability growth curve. Analysis should be able to illuminate how these various changes will affect system reliability. One related problem is that almost all of the operational testing for a system is clustered immediately before the decision point on proceeding to full-rate production. This result is a tre-

mendous amount of risk for that decision. This is another aspect of the management problem.

Third, the problem with reliability growth modeling is not only the weakness of specific approaches, but also the inability to assess the uncertainty of predictions, an area in which statisticians need to make progress. Fourth, with respect to life-cycle costs, the current inability to forecast a system's costs of ownership correctly is bankrupting the military. It is crucial that DoD improve its ability to forecast system life-cycle costs.

Fifth, with respect to the development of models for combining information, the linkage required in relating developmental test performance to operational test performance is not the only requirement. It is also important to link test performance to performance in the field.

Finally, Seglie agreed that it is important to get the science underlying reliability assessment right. Therefore, it is important to understand and use models of the physics of failure. A related need is understanding of the impact of "bad actors" on models and estimates. To make progress in this regard, it is important for statisticians and the relevant scientists to work together closely.

The concern was expressed at the workshop that much of the methodological progress described by the participants is not represented in the reliability assessments of defense systems. This is unfortunate since many of these methods offer substantial benefits relative to those used in the 1970s and 1980s. The newer methods are often more efficient, which is important as data collection becomes more expensive. They also make better use of datasets with either subjective elements, censoring, or missing data, again providing more reliable estimates for the same amount of data. Finally, they offer greater flexibility in handling alternative distributional forms, and as a result, the estimates derived are often more trustworthy.

Reinstitution of more active collaboration between the statistical and defense acquisition communities, along with leading to better statistical methods, could also increase the number of academics interested in the most pressing problems faced by the defense acquisition community. In addition, such collaboration could increase the chances of attracting highly trained statisticians to careers in defense testing and acquisition. Many participants in the workshop were strongly in favor of greater interaction between the two communities.

References

Arcones, M.A., P.H. Kvam, and F.J. Samaniego
2002 Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *Journal of the American Statistical Association* 97: 170-182.

Ascher, H., and H. Fiengold
1984 *Repairable Systems Reliability*. New York: Marcel Dekker.

Balaban, H., and B. Reterer
1974 The use of warranties for defense avionics procurement. Pp. 363-368 in *Proceedings of the Annual Reliability and Maintainability Symposium.* New York: Institute of Electrical and Electronics Engineers.

Benton, A.W., and L.H. Crow
1989 Integrated reliability growth testing. P. 160 in *Proceedings of the Annual Reliability and Maintainability Symposium.* New York: Institute of Electrical and Electronics Engineers.

Blischke, W.R., and D.N.P. Murthy
1998 *Reliability: Modeling, Prediction, and Optimization*. New York: John Wiley & Sons.

Castillo, E., and A.S. Hadi
1995 Modeling lifetime data with application to fatigue models. *Journal of the American Statistical Association* 90 (431): 1041-1054.

Chan, C.K., and M. Tortorella
2001 Spares-inventory sizing for end-to-end service availability. Pp. 98-102 in *Proceedings of the Annual Reliability and Maintainability Symposium*. New York: Institute of Electrical and Electronics Engineers.

Cohen, D.M., S.R. Dalal, J. Pareluis, and G.C. Patton
1996 The Combinatorial Design Approach to Automatic Test Generation. *Proceedings of the Seventh International Symposium of Software Reliability Engineering (ISSRE),* White Plains, NY.

Collas, G.
1991 Dynamic reliability prediction: How to adjust modeling and reliability growth. *Proceedings of the IEEE Annual Reliability and Maintainability Symposium*. New York: Institute of Electrical and Electronics Engineers.

Crow, L.H.
1984 Methods for assessing reliability growth potential. *Proceedings of the Annual Reliability and Maintainability Symposium*. New York: Institute of Electrical and Electronics Engineers.
1998 Using reliability growth models in the management and structure of a failure review board. K. Garquhar and A. Mosleh, eds. *Proceedings of Workshop on Reliability Growth Modeling: Objectives, Expectations, and Approaches*. College Park, Maryland: Center for Reliability Engineering, University of Maryland.

Dalal, S.R., M. Hamada, and T.J.Wang
1999 How to improve performance of software systems: A methodology and a case study for tuning performance. *Annals of Software Engineering* 8: 53-84.

Defense Science Board
2000 *Report of the Defense Science Board Task Force on Test and Evaluation Capabilities*. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics.

Duane, J.T.
1964 Learning curve approach to reliability monitoring. *IEEE Transactins on Aerospace & Electronic Systems* 2 (April): 563-566.

Durham, S.D., and W.J. Padgett
1991 A probabilistic stress-strength model and its application to fatigue failure in gun barrels. *Journal of Statistical Planning and Inference* 29: 67-74.
1997 A cumulative damage model for system failure with application to carbon fibers and composites. *Technometrics* 39: 34-44.

Ericson, W.A.
1969 A note on the posterior mean of a population mean. *Journal of the Royal Statistical Society, Series B (Methodological)* 31 (2): 332-334.
1970 On the posterior mean and variance of a population mean. *Journal of the American Statistical Association* 65: 649-652.

Folks, J.L., and R.S. Chhikara
1989 The inverse Gaussian distribution and its statistical application: A review. *Journal of the Royal Statistical Society* Series B (Methodological) 40: 263-275.

Fries, A., and R. Easterling
2002 Sample size methodologies. *Handbook of Statistics*. North-Holland.

Fries, A., and A. Sen
1997 A survey of discrete reliability growth models. *IEEE Transactions on Reliability* 42: 582-604.

Gaver, D.P., P.A. Jacobs, and A. Fries
1997 Prediction of changeover performance: Operational test (OT) rates from developmental test (DT) rates via meta-analysis. Pp. 149-153 in *Proceedings of the Section on Government Statistics and Section on Social Statistics*. American Statistical Association.

Gaver, D.P., P.A. Jacobs, K.D. Glazebrook, and E.A. Seglie
2000 *Probability Models for Sequential-Stage System Reliability Growth via Failure Mode Removal*. NPS-OR-00-006. Monterey, California: Naval Postgraduate School.

Hartigan, J.
1969 Linear Bayesian methods. *Journal of the Royal Statistical Society*, Series B (Methodological) 31: 446-454.

Jewell, W.
1984 A general framework for learning curve reliability growth models. *Operations Research* 32.

Kalbfleisch, J.D., and J.F. Lawless
1988 Estimation of reliability in field-performance studies. *Technometrics* 30(4): 365-378.

Kalbfleisch, J.D., J.F. Lawless, and J.A. Robinson
1991 Methods for the analysis and prediction of warranty claims. *Technometrics* 33: 273-285.

Martz, H.F., and R.A. Waller
1982 *Bayesian Reliability Analysis*. New York: John Wiley & Sons.
1990 Bayesian reliability analysis of complex series/parallel systems of binomial subsystems and components. *Technometrics* 32: 407-416.

Martz, H.F., R.A. Waller, and E.T. Fickas
1988 Bayesian reliability analysis of series systems of binomial subsystems and components. *Technometrics* 30: 143-159.

Meeker, W.Q., L.A. Escobar, and H. Wu
2002 Unpublished manuscript.

Mosemann, L.
1994 Predictability. *Crosstalk: The Journal of Defense Software Engineering* 7(8). Software Technology Support Center, U.S. Department of Defense.

Murthy, D.N.P., and W.R. Blischke
2000 Strategic warranty management: A life-cycle approach. *IEEE Transactions on Engineering Management* 47 (1): 40-54.

National Research Council
1998 *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements*. Panel on Statistical Methods for Testing and Evaluating Defense Systems. Michael L. Cohen, John B. Rolph, and Duane L. Steffey, eds. Committee on National Statistics. Washington, D.C.: National Academy Press.

Press, S.J., and J.E. Rolph
1986 Empirical Bayes estimation of the mean in a multivariate normal distribution. *Communications in Statistics-Theory and Methods* 15: 2201-2228.

Samaniego, F.J., and D.M. Reneau
1994 Towards a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association* 89: 947-959.

Samaniego, F.J., D. Steffey, and H. Tran
2001 *Towards a Theory for Combining Information from "Related" Experiments*. Technical Report. Department of Statistics, University of California, Davis.

Samaniego, F.J., and E. Vestrup
1999 On improving upon standard estimates via linear empirical Bayes methods. *Statistics and Probability Letters* 44: 309-318.

Saunders, S.C.
2001 *Studies of Reversible First-Order Reactions During Hydrolysis in Polymer Coatings and the Effect of Changing Temperature and Humidity*. Research Note, Scientific Consulting Services.

Scholz, F.
1986 Software reliability modeling and analysis. Pp. 25-31 in *IEEE Transactions on Software Engineering,* Vol. SE-12, No. 2. Institute of Electrical and Electronics Engineers.

Sen, A.
1998 Analysis of repairable systems—Past, present, and future. Pp. 317-336 in *Frontiers in Reliability Analysis*, A.P. Basu, S.P. Mukherjee, and S.K. Basu, eds. World Scientific Publishing Co.

Sen, A., and G.K. Bhatttacharya
1993 A piecewise exponential model for reliability growth and associated inference. Pp. 331-335 In *Advances in Reliability,* A.P. Basu, ed. Elsevier Science Publishers.

Shaked, M., and J.G. Shanthikumar
1994 *Stochastic Orders and Their Applications.* Boston: Academic Press.

Thomas, M.U., and S.S. Rao
1999 Warranty economic decision models: A summary and some suggested directions for future research. *Operations Research* 47: 807-820.

Tortorella, M., F.J. Samaniego, and M. Zuo (eds.)
1998 *IIE Transactions on Quality and Reliability Engineering* (Special Issue on Reliability Economics) 30 (12): Kluwer.

U.S. Department of Defense
1981 *Military Handbook 189. Reliability Growth Management.* MIL-HDBK-189, Philadephia: Naval Publication and Form Center.
1982 *Test and Evaluation of System Reliability, Availability, and Maintainability: A Primer.* Report No. DoD 3235.1-H, third edition. Washington, DC: U.S. Department of Defense.

Ushakov, I.
1994 *Handbook of Reliability Engineering*. New York: John Wiley & Sons.

Walton, G.H., J.H. Poore, and C.J. Trammel
1995 Statistical testing of software based on a usage model. *Software Practice & Experience* 25(1): 97-108.

Whittaker, J.A., and M.G. Thomason
1994 A Markov chain model for statistical software testing. *IEEE Transactions on Software Engineering* 30(10): 812-824.

Wu, H. and W.Q. Meeker
2002 Early detection of reliability problems using information from warranty data bases. *Technometrics* 44: 120-133.

Appendix

WORKSHOP AGENDA AND PARTICIPANTS

Committee on National Statistics
Workshop on Reliability Issues for DoD Systems

National Academy of Sciences
2101 Constitution Avenue, N.W.
Washington, D.C.
Lecture Room

AGENDA

Friday, June 9

8:00 a.m. Continental Breakfast

8:30 a.m. Opening Remarks

Nancy Spruill, Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, DoD
Francisco Samaniego, University of California, Davis

8:45 a.m. Reliability Test Designs to Direct Defense System Development

Development and management of comprehensive programs, from system concept through fielding, for the design, test, and assessment of reliability performance.

Ronald Glaser (Chair), Lawrence Livermore National Laboratory
Jane Booker, Los Alamos National Laboratory, on "PREDICT: A New Approach to System Performance Prediction"
Lawrence Crow, General Dynamics, on "An Integrated Reliability Growth Strategy at General Dynamics Advanced Technology Systems"
Walter Hollis (Discussant), Deputy Under Secretary of the Army
Arthur Fries (Discussant), Institute for Defense Analyses

10:30 a.m. Break

10:45 a.m. Recent Developments in Reliability Growth Modeling

Patricia Jacobs (Chair), Naval Postgraduate School
Ananda Sen, Oakland University, on "Recent Developments in Reliability Growth Modeling: Duane Curve and Beyond"
Donald Gaver, Naval Postgraduate School, on "Reliability Growth in Stage-wise-Functioning Systems by Failure Mode Removal"
Paul Ellner (Discussant), Army Materiel System Analysis Activity

12:00 noon Lunch

1:00 p.m. Use of Statistical Modeling to Join Developmental and Operational Testing Data

Pooling of data from diverse information sources (e.g., laboratory and developmental testing, operational testing, field and training exercise data, simulations).

Asit Basu (Chair), University of Missouri-Columbia
Duane Steffey, San Diego State University, on "Combining Information for Reliability Assessment"
Francisco Samaniego, University of California, Davis, on

"Nonparametric Alternatives to Exponential Life Testing"
Ernest Seglie (Discussant), Science Advisor to the Director, Operational Test and Evaluation
Fred Myers (Discussant), Office of the Under Secretary of Defense (Acquisition, Technology and Logistics)

2:15 p.m. Estimating Reliability from Field Performance Data

Philip Boland (Chair), University College, Dublin
William Meeker, Iowa State University, on "Extracting Information from Field-Failure and Warranty Data Bases: An Important Opportunity"
Fritz Scholz, The Boeing Company, on "Assessing Reliability Growth from Field Performance Data"
James Crouch (Discussant), Wright-Patterson Air Force Base

4:00 p.m. Break

4:15 p.m. Modeling Fatigue

Henry Block (Chair), University of Pittsburgh
Samuel Saunders, Washington State University, on "Backward and Forward in the Prediction of Service-Lives"
William Padgett, University of South Carolina, on "Some Cumulative Damage Approaches to Modeling Materials or System Failure"
Theodore Nicholas (Discussant), United States Air Force Research Laboratory

5:30 p.m. Adjourn

Saturday, June 10

8:00 a.m. Continental Breakfast

8:30 a.m. Reliability of Software-Intensive Systems

Simon Wilson (Chair), Trinity College, Dublin

Sidhartha Dalal, Telcordia Technologies, on "The State of Software Reliability Research"
Jesse Poore, University of Tennessee, on "An Approach to Software Testing"
Margaret Myers (Discussant), Director, Information Technology, Acquisition and Defense
Jack Ferguson (Discussant), Software Intensive Systems

10:00 a.m. Break

10:15 a.m. Reliability Economics—Life-Cycle Costs

Michael Tortorella (Chair), Bell Laboratories
Wallace Blischke, University of Southern California, on "Life-Cycle Costing, Reliability, and Warranty Presentation for NAS Workshop on Reliability"
Frank Camm, The RAND Corporation-Washington, on "A Life Cycle Context for Reliability Assessment of DoD Systems"
Allen Beckett (Discussant), Deputy Under Secretary of Defense for Logistics

12:00 noon Lunch and Panel Discussion on "Implications for Future DoD Practices"

Francisco Samaniego (Moderator), University of California, Davis
Stephen Pollock, University of Michigan
Ernest Seglie, Director, Operational Test and Evaluation
Marion Williams, Air Force Operational Test and Evaluation Command
Robert Easterling, Sandia National Laboratories
Hans Mark, Director, Defense Research and Engineering

2:00 p.m. Adjourn

PARTICIPANTS

Presenters

Allen Beckett, Principal Assistant, Deputy Under Secretary of Defense for Logistics and Materiel Readiness Logistics
Wallace Blischke, University of Southern California
Henry Block, University of Pittsburgh
Philip Boland, University College, Dublin
Jane Booker, Los Alamos National Laboratory
Frank Camm, The RAND Corporation
James Crouch, Wright-Patterson Air Force Base
Larry Crow, General Dynamics Advanced Technology Systems
Siddhartha Dalal, Telcordia Technologies
Robert Easterling, Sandia National Laboratories
Paul Ellner, U.S. Army Materiel Systems Analysis Activity
Jack Ferguson, Software Intensive Systems, Office of the Under Secretary of Defense (Acquisition, Technology, and Logstics)
Arthur Fries, Institute for Defense Analyses
Donald Gaver, Naval Postgraduate School
Ronald Glaser, Lawrence Livermore National Laboratory
Walter Hollis, Office of the Deputy Under Secretary of the Army (Operations Research)
Patricia Jacobs, Naval Postgraduate School
William Meeker, Iowa State University
Fred Myers, Office of the Under Secretary of Defense (Acquisition, Technology and Logistics)
Margaret Myers, Director, Information Technology, Acquisition and Investments
Theodore Nicholas, U.S. Air Force Research Laboratory
William Padgett, University of South Carolina
Jesse Poore, University of Tennessee
Frank Samaniego, University of California, Davis
Sam Saunders, Washington State University
Friedrich Scholz, Boeing
Ernest Seglie, Office of the Director, Operational Test and Evaluation
Ananda Sen, Oakland University
Nancy Spruill, Office of the Under Secretary of Defense (Acquisition, Technology, and Logistics)

Duane Steffey, San Diego State University
Michael Tortorella, Bell Laboratories
Marion Williams, Air Force Operational Test & Evaluation Command
Simon Wilson, Trinity College, Dublin

Invited Guests

Anthony Adessa, BMDO/TEP
Asit Basu, University of Missouri-Columbia
Michael Bell, Office of the Director, Operational Test and Evaluation
Barry Bodt, U.S. Army Research Laboratory
Paul Bricker, Maintenance Policy Program & Resources
Mike Bridgman, Logistics Management Institute
David S. Chu, The RAND Corporation
Daniel Cork, Carnegie Mellon University
Kathleen Diegart, Sandia National Laboratories
William Eddy, Carnegie Mellon University
Robert Ernst, Naval Air Systems
Joseph Ferrara, Deputy Director, Acquisition Management
Lee H. Frame, Office of the Director, Operational Test and Evaluation
Donald Henry, Office of Deputy Assistant Army Secretary (Research and Technology)
Paul Hoffman, Naval Air Systems Command
Charles Horton, Office of the Director, Operational Test and Evaluation
Shane Knighton, Air Force Operational Test and Evaluation Command
Paul Kvam, Georgia Institute of Technology
David Lee, Logistics Management Institute
Thomas Louis, The RAND Corporation
Milton Margolis, Logistics Management Institute
Maj. Kyle McKown, U.S. Air Force
Andy Monje, Naval Air Systems Command
Robert Nemetz, Office of the Under Secretary of Defense (Acquisition, Technology, and Logistics)
David Oliver, Office of the Secretary of Defense
David Olwell, Naval Post Graduate School
Russell S. Pimm, CSC/Nichols Research
Stephen Pollock, University of Michigan
Pete Poppe, Marine Corps Operational Test and Evaluation Activity
Shane Reese, Los Alamos National Laboratory

Philip Rodgers, Office of the Under Secretary of Defense (Acquisition, Technology, and Logistics)
Richard Sabo, Naval Air Systems Command
Michael Saboe, U.S. Army Tank Automotive, Research, Development & Engineering Center
Dennis Smallwood, U.S. Military Academy
James Streilein, U.S. Army Evaluation Center
David Thomen, Office of the Director, Operational Test and Evaluation
Ron Wagner, Georgia Tech Research Center
Steven Whitehead, Commander, Operational Test and Evaluation Force
Daniel Willard, Office of the Deputy Under Secretary of the Army (Operations Research)
Alyson Wilson, Los Alamos National Laboratory
Paul Zimmerman, Naval Air Systems Command

Staff of the Committee on National Statistics

Jamie Casey, Research Assistant
Michael Cohen, Study Director
Agnes Gaskin, Senior Project Assistant
Andrew White, Director

Index