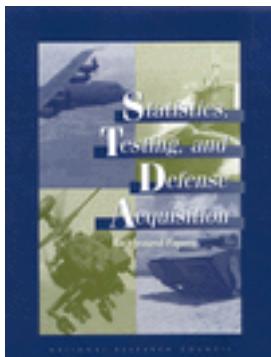


## **Statistics, Testing, and Defense Acquisition: Background Papers**



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

ISBN: 0-309-55710-0, 180 pages, 8.5 x 11, (1999)

**This free PDF was downloaded from:**

**<http://www.nap.edu/catalog/9655.html>**

Visit the [National Academies Press](#) online, the authoritative source for all books from the [National Academy of Sciences](#), the [National Academy of Engineering](#), the [Institute of Medicine](#), and the [National Research Council](#):

- Download hundreds of free books in PDF
- Read thousands of books online, free
- Sign up to be notified when new books are published
- Purchase printed books
- Purchase PDFs
- Explore with our innovative research tools

Thank you for downloading this free PDF. If you have comments, questions or just want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, [visit us online](#), or send an email to [comments@nap.edu](mailto:comments@nap.edu).

This free book plus thousands more books are available at <http://www.nap.edu>.

Copyright © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

---

# **Statistics, Testing, and Defense Acquisition**

## **Background Papers**

Panel on Statistical Methods for Testing and Evaluating Defense Systems

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

Committee on National Statistics  
Commission on Behavioral and Social Sciences and Education  
National Research Council

NATIONAL ACADEMY PRESS  
Washington, D.C.

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

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

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

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

The project that is the subject of this report is supported by Contract DASW01-94-C-0119 between the National Academy of Sciences and the Director 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 view of the organizations or agencies that provided support for this project.

International Standard Book Number 0-309-06627-1

Copyright 1999 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

## PANEL ON STATISTICAL METHODS FOR TESTING AND EVALUATING DEFENSE SYSTEMS

JOHN E. ROLPH (*Chair*), Marshall School of Business, University of Southern California  
MARION BRYSON, North Tree Management, Monterey, Marina, California  
HERMAN CHERNOFF, Department of Statistics, Harvard University  
JOHN D. CHRISTIE, Logistics Management Institute, McLean, Virginia  
LOUIS GORDON, Filoli Information Systems, Palo Alto, California  
KATHRYN B. LASKEY, Department of Systems Engineering and Center of Excellence in C<sup>3</sup>I, George Mason University  
ROBERT C. MARSHALL, Department of Economics, Pennsylvania State University  
VIJAYAN N. NAIR, Department of Statistics, University of Michigan  
ROBERT T. O'NEILL, Division of Biometrics, Food and Drug Administration, U.S. Department of Health and Human Services, Rockville, Maryland  
STEPHEN M. POLLOCK, Department of Industrial and Operations Engineering, University of Michigan  
JESSE H. POORE, Department of Computer Science, University of Tennessee  
FRANCISCO J. SAMANIEGO, Division of Statistics, University of California, Davis  
DENNIS E. SMALLWOOD, Department of Social Sciences, U.S. Military Academy, West Point, New York  
MICHAEL L. COHEN, *Study Director*  
DUANE L. STEFFEY, *Consultant*  
ANU PEMMARAZU, *Research Assistant*  
ERIC M. GAIER, *Consultant*  
CANDICE S. EVANS, *St. Project Assistant*

## COMMITTEE ON NATIONAL STATISTICS 1996-1997

NORMAN M. BRADBURN (*Chair*), National Opinion Research Center, University of Chicago  
JULIE DAVANZO, The RAND Corporation, Santa Monica, California  
WILLIAM F. EDDY, Department of Statistics, Carnegie Mellon University  
JOHN F. GEWEKE, Department of Economics, University of Minnesota, Minneapolis  
JOEL B. GREENHOUSE, Department of Statistics, Carnegie Mellon University  
ERIC A. HANUSHEK, W. Allen Wallis Institute of Political Economy, Department of Economics, University of Rochester  
RODERICK J.A. LITTLE, Department of Biostatistics, University of Michigan  
CHARLES F. MANSKI, Department of Economics, University of Wisconsin  
WILLIAM NORDHAUS, Department of Economics, Yale University  
JANET L. NORWOOD, The Urban Institute, Washington, District of Columbia  
EDWARD B. PERRIN, School of Public Health and Community Medicine, University of Washington  
PAUL ROSENBAUM, Department of Statistics, Wharton School, University of Pennsylvania  
KEITH F. RUST, Westat, Inc., Rockville, Maryland  
FRANCISCO J. SAMANIEGO, Division of Statistics, University of California, Davis  
MIRON L. STRAF, *Director*

## Contents

Preface	vii
Strategic Information Generation and Transmission: The Evolution of Institutions in DoD Operational Testing <i>Eric M. Gaier, Logistics Management Institute Robert C. Marshall, Pennsylvania State University</i>	1
On the Performance of Weibull Life Tests Based on Exponential Life Testing Designs <i>Francisco J. Samaniego and Yun Sam Chong University of California, Davis</i>	41
Application of Statistical Science to Testing and Evaluating Software Intensive Systems <i>Jesse H. Poore, University of Tennessee, Knoxville Carmen J. Trammell, CTI PET Systems, Inc</i>	124



## Preface

The Panel on Statistical Methods for Testing and Evaluating Defense Systems had a broad mandate—to examine the use of statistics in conjunction with defense testing. This involved examining methods for software testing, reliability test planning and estimation, validation of modeling and simulation, and use of modern techniques for experimental design. Given the breadth of these areas, including the great variety of applications and special issues that arise, making a contribution in each of these areas required that the Panel's work and recommendations be at a relatively general level. However, a variety of more specific research issues were either brought to the Panel's attention by members of the test and acquisition community, e.g., what was referred to as Dubin's challenge (addressed in the Panel's interim report), or were identified by members of the panel. In many of these cases the panel thought that a more in-depth analysis or a more detailed application of suggestions or recommendations made by the Panel would either be useful as input to its deliberations or could be used to help communicate more individual views of members of the Panel to the defense test community. This resulted in several research efforts. Given various criteria, especially immediate relevance to the test and acquisition community, the Panel has decided to make available three technical or background papers, each authored by a Panel member jointly with a colleague. These papers are individual contributions and are not a consensus product of the Panel; however, the Panel has drawn from these papers in preparation of its final report: *Statistics, Testing, and Defense Acquisition*. The Panel has found each of these papers to be extremely useful and they are strongly recommended to readers of the Panel's final report.

The remainder of this preface provides the reason for including each paper and a short introduction.

---

**"Strategic Information Generation and Transmission: The Evolution of Institutions in DoD Operational Testing" by Eric Gaier and Robert Marshall**

This paper examines the historical evolution of operational testing in the Department of Defense (DoD) through use of the theory of principal agent games. In this model of defense test and acquisition, test information is both strategically generated and strategically conveyed. This game theoretical model is used to better understand various incentives and hence the behavior of key participants in defense acquisition, especially the program manager for a defense system in development, DOT&E, and Congress. While the model is an oversimplification of many aspects of defense test and acquisition, it may be profitably used to indicate possible detrimental impacts of the incongruent incentives of the various participants in defense acquisition and to suggest methods for their avoidance.

**"On the Performance of Weibull Life Tests Based on Exponential Life Testing Designs" by Frank Samaniego and Yun Sam Chong**

This paper examines the consequences of using the model of exponential times to first failure during the design of a test, when the times to first failure instead follow the two-parameter Weibull distribution. When this assumption obtains, improvements to the hypothesis tests that continue to use the assumption of exponential times to first failure from use of a Weibull assumption are demonstrated. In addition, the benefits, especially the possible reduction of time on test, from making use of the Weibull assumption at the test planning stage are explored. There are two reasons to consider this paper. First, in situations when Weibull model's have been identified in the past, there are major advantages to use of the methods and tables for test design and evaluation. Further, since the Weibull model is one of several alternatives to the exponential model in a variety of reliability contexts, e.g., for testing repairable systems and systems with dependent failure rates, the gains from use of a more appropriate model can be

generalized to several other possibilities that need to be explored when the exponential model is deficient.

**"Application of Statistical Science to Testing and Evaluating Software Intensive Systems" by Jesse Poore and Carmen Trammell**

This paper examines a method for statistical testing of software. Statistical approaches to testing enable the efficient collection of empirical data that limit and measure uncertainty about the behavior of the software intensive system, and support decisions regarding the benefits of further testing, deployment, maintenance, and evolution of the software. In statistical software testing, the population is the set of all scenarios of use that are organized through an operational use model. The states of use of the system and the allowable transitions among those states are identified, represented in the form of one or more Markov chains. The methods have the advantages of being based on software architecture and are readily validated. Usage models can be represented as the solution to a linear program. Experimental design, e.g., combinatorial designs and partition testing, can be used in conjunction with this approach to achieve efficient coverage of all states. The method also provides an economic stopping criterion. This novel approach to testing software intensive systems therefore has a number of advantages over current methods used by DoD and should be considered as an alternative.

JOHN E. ROLPH, CHAIR

PANEL ON STATISTICAL METHODS FOR TESTING AND EVALUATING DEFENSE SYSTEMS



# 1

## Strategic Information Generation and Transmission: the Evolution of Institutions in DoD Operational Testing

*Eric M. Gaier, Logistics Management Institute; and Robert C. Marshall, Pennsylvania State University*

### 1. INTRODUCTION

Several important papers in the field of information and uncertainty have focused on strategic information transmission (see, for example, Milgrom, 1981; Crawford and Sobel, 1982; or Green and Stokey, 1980). The majority of this research has taken the form of principal agent games. In general, an agent observes some realization of a random variable which affects the payoff for each player. The agent then strategically signals the principal regarding the underlying realization. In the final stage, the principal takes some action which, in conjunction with the realization of the random variable, determines the payoff for each player. In equilibrium, the principal must take account of any bias in the agent's reporting strategy when determining the optimal action.

We present a model which extends the information transmission literature by allowing for a continuous choice of information quality. This is accomplished by letting the agent determine the probability with which he is able to distinguish one state from its complement. We call this stage of the game *test design*. In equilibrium, the principal must now account for the agent's selectivity in both the information generation and reporting stages. Thus, we present a model in which information is both strategically generated and strategically conveyed.

Since the preferences of the principal and the agent do not necessarily coincide, the test design and reporting process may be significantly biased in favor of the agent. The principal might choose to exercise some oversight authority in the process. The principal could do this in several ways. He might choose to extend oversight authority during the test design stage.

Alternatively, the principal might choose to extend oversight authority during the reporting stage. Our model considers each of these cases. As the main result of the paper, we show that oversight of the test design stage always improves the welfare of the principal while oversight of the test reporting stage may not. In addition, we consider the case in which the principal can extend oversight authority over both test design and test reporting.

We believe that the model describes a wide variety of interesting situations—The promotion of assistant professors in disciplines with exceptionally thin job markets, for example. Individual departments make assessments of candidates and report to the tenure committee. Although the tenure committee makes the final decision, the departments have the necessary expertise to gather the relevant data. Typically the tenure committee establishes the criteria by which individual departments judge the candidates. In the context of our model this is interpreted as oversight of the test design phase. Another interesting application is found in the operational test and evaluation procedures used by the Department of Defense. It is in this context that we develop the model.

The Department of Defense engages in two types of testing throughout the acquisition cycle. The emphasis in developmental testing is on isolating and measuring performance characteristics of individual components of a system. Developmental testing is conducted in a carefully controlled environment by highly trained technical personnel. The emphasis in operational testing, however, is on evaluating the overall capabilities and limitations of the complete system in a realistic operating environment. Operational testing is therefore conducted in a less controlled environment by trained users of the system. It is the role of this type of testing in the acquisition cycle that we investigate below.

The acquisition cycle follows a series of event based decisions called milestones.<sup>1</sup> At each milestone a set of criteria must be met in order to proceed with the next phase of acquisition. Operational testing is one of the last stages in this cycle.

When a system is ready for operational testing, the exact details of the test are prepared by the independent test agencies within each Service. Tests must be prepared in accordance with

---

<sup>1</sup> The interested reader is urged to see the interim report from the Panel on Statistical Methods for Defense Testing (National Research Council, 1995) for a complete description of the current acquisition cycle.

he Test and Evaluation Master Plan (TEMP), which spells out the critical operational issues to be addressed. The TEMP is prepared fairly early in the acquisition cycle but is continuously updated and modified. For major systems, both the TEMP and the operational test plan must receive approval from the Office of Director of Operational Test and Evaluation (DOT&E). This Congressional oversight agency was created in 1983 mainly to oversee the test design process. In this way, Congress is able to extend oversight authority into the test design portion of operational testing. Fairly regularly, resource constraints prevent the testing agencies from addressing all of the critical operational issues. In such cases, testers must determine which issues to address and which to ignore.

The independent test agencies conduct the operational tests and evaluate the results. These evaluations are conveyed directly to the Service Chief who reports the results to the relevant milestone decision authority. In the case of major systems, decision authority rests with the Undersecretary of Defense for Acquisition and Technology who is advised by the Defense Acquisition Board. If the Undersecretary approves the acquisition, a procurement request is included in the Department of Defense budget request submitted to Congress. In addition, independent evaluations of test data are conducted by DOT&E who reports directly to the Secretary of Defense and Congress. In so doing, DOT&E also exercises oversight authority in the reporting process.

The role of operational testing in the acquisition cycle has not always been characterized by the description given above. In fact, the entire procurement process has slowly evolved through a series of reform initiatives. Section 2 provides a brief description of the history of the role of operational testing in the acquisition cycle. We then introduce the model in order to gain insight into this process.

Section 3 provides an overview of the related literature. Section 4 develops the modeling framework and lists the assumptions of our model. In section 5 we introduce several games which are designed to capture the role of operational testing at various points in time. Our results are presented in sections 6 and 7. We conclude with section 8.

## 2. HISTORICAL EVOLUTION OF OT&E

The Air Force is generally considered to have been the early pioneer in operational testing. As early as May 1941, the Air Force Air Proving Ground Command was involved in the testing of new aircraft designs for possible procurement. Although operational testing in the other Services was soon initiated, the absence of strong oversight from the Department of Defense allowed each Service to develop unique regulations and procedures. Prior to 1970, for example, the Navy relied heavily on the subjective opinions of a few well-qualified officers. Little emphasis was given to the generation of verifiable data. Over the same time period, however, the Air Force had gone to great lengths to define a set of formal procedures and guidelines for the conduct of OT&E. As a result, Air Force testing generally produced objective data but lacked the flexibility to adjust to the specific requirements of individual systems.

Prior to 1971, the organization of OT&E also varied substantially across the Services. Although the Navy's test agency reported directly to the Chief of Naval Operations, the Air Force and Army test agencies were subordinate to lower levels of command. The Air Force and the Army were repeatedly criticized for allowing their testing agencies to report to organizations which were responsible for the development of new systems. Partially in response to these concerns, the Deputy Secretary of Defense directed the military services in February 1971 to designate OT&E field commands independent of the system developers and the eventual users. These agencies were instructed to report directly to the relevant Chief of Staff. Navy testing responsibility continued to reside with the Operational Testing and Evaluation Force (OPTEVFOR), while testing responsibility was assigned to the Air Force Test and Evaluation Command (AFTEC)<sup>2</sup> and the Army Operational Test and Evaluation Agency (OTEA).

Prior to 1971 the Department of Defense was not required to convey the results of operational testing to the Congress. In the absence of testing data, Congress generally deferred to DoD expertise on program funding allocations. In addition, Congress was not involved in the

---

<sup>2</sup> AFTEC has now become the Air Force Operational Test and Evaluation Command (AFOTEC).

design or implementation of operational testing. Over this time period, therefore, the Department of Defense was able to exert considerable influence over the status of individual programs.

As part of its continued effort to become more involved in the procurement process, Congress enacted Public Law 92-156 in 1971. This law requires the Department of Defense to report OT&E results to the Congress annually. Armed with these testing results, Congress began to take a more active role in determining which programs to fund and which to terminate. However, the design and conduct of operational testing continued to be the responsibility of the Department of Defense. Although Public Law 92-156 certainly reduced DoD's explicit influence over funding decisions, DoD continued to exert considerable influence over the acquisition process through its choice of operational tests. The model will show how DoD might have altered its testing strategy in light of Congressional involvement.

Over the period 1971 through 1983, Department of Defense testing procedures received strong criticism from Congress and the General Accounting Office (GAO). Many of these complaints focused on a perceived inadequacy in DoD testing. In 1983, for example, GAO determined that reliability and maintainability testing on the Army's Sergeant York Air Defense Gun had been inadequate to support the production decision (U.S. General Accounting Office, 1983). Similarly, the President's 1970 Blue Ribbon Defense Panel concluded that both developmental and operational testing of the Army M-16 rifle had been inadequate (Blue Ribbon Defense Panel, 1970). In 1979, GAO concluded that developmental testing was also inadequate in the case of the joint Air Force/Navy NAVSTAR Global Positioning Systems (GPS) (U.S. General Accounting Office, 1979a). Although such criticisms are certainly not limited to the time frame described above, the model will show in what sense testing might have been perceived as inadequate.<sup>3</sup>

As a result of allegations such as these, Congress became increasingly more concerned with the planning and conduct of DoD testing in the Department of Defense. The President's Blue Ribbon Panel also recommended the creation of a higher than Service level organization to help give direction to the operational test agencies. In 1983, Congress instructed DoD to create

---

<sup>3</sup> The Army's Aquila Remotely Piloted Vehicle (U.S. General Accounting Office, 1988a) is an example of a program which was criticized for inadequate testing outside the time period described.

the Office of Director of Operational Test and Evaluation to fill this oversight role. DOT&E is headed by a civilian who is appointed by the President and confirmed by the Congress. DOT&E is charged with two primary roles. First, DOT&E is directed to be the principle advisor to the Secretary of Defense regarding OT&E matters. Second, DOT&E is directed to report to Congress on the adequacy of operational testing and the desirability of allowing systems beyond low rate initial production.

In fulfilling these primary roles, DOT&E has assumed several responsibilities. First, DOT&E is responsible for the proscription of policies and procedures for the conduct of OT&E. Second, DOT&E provides advice to the Secretary of Defense and makes recommendations to military departments regarding OT&E in general and on specific aspects of operational testing for major systems. In this regard, operational test plans for major acquisitions require DOT&E approval. Third, DOT&E monitors and reviews the conduct of OT&E by the Services. Fourth, DOT&E is responsible for an independent analysis of the results of OT&E for each major system and must report directly to the Secretary of Defense, the Senate and House Armed Services Committees, and the Senate and House Committees on Appropriation. In each case, DOT&E is directed to analyze the adequacy of operational testing as well as the effectiveness and suitability of the tested system. Fifth, DOT&E is responsible for advising the Secretary of Defense regarding all budgetary and financial matters relating to OT&E.

It is well documented that DOT&E had only a limited impact for the first several years of its existence (U.S. General Accounting Office, 1987). The post of Director remained vacant for nearly two years while the Office continued to be underfunded and understaffed. During this time, DOT&E received criticism for failing to adequately monitor Service operational testing. In addition, the Government Accounting Office determined that DOT&E reports to the Secretary of Defense and the Congress were not composed independently as required by law. In several instances GAO found DOT&E reports which were copied verbatim from Service documents. In the first several years, DOT&E was therefore unable to fulfill one of its major responsibilities.

DOT&E was, however, largely successful in its early attempts to improve test planning and implementation. To this end, DOT&E developed a uniform set of guidelines for Service operational testing and revised Department of Defense Directive 5000.3 *Test and Evaluation*. In

1987, GAO determined that DOT&E had significantly impacted the testing process though its careful review of operational test plans (U.S. General Accounting Office, 1987). On many occasions, the Services were required to make significant revisions in operational test plans for major acquisitions in order to get DOT&E approval. GAO concluded that the adequacy of operational testing was significantly improved by DOT&E's efforts in this regard. Our model will yield considerable insight into DOT&E's decision to reform the test planning process at the expense of ignoring the reporting process.

Since the formation of DOT&E, the Department of Defense has faced renewed criticism. The Government Accounting Office and the DoD Inspector General have accused DoD officials of manipulating test results to yield the most favorable interpretation possible. The most highly publicized case involved the Navy's Airborne Self-Protection Jammer (ASPJ) (U.S. General Accounting Office, 1992). The specific allegations stemmed from the reporting of reliability growth test results which were being conducted as part of Initial Operational Test and Evaluation. After testing had begun, Navy testers changed the testing criteria to exclude certain self-diagnostic software failures as not relevant. With these failures excluded ASPJ, was reported to have passed the test criteria. However, the inclusion of this data would have resulted in a test failure. Similar allegations have been levied against other programs including the various electronic countermeasures programs of the 1980s (U.S. General Accounting Office, 1989, 1991 b, 1991 c) and the Army's Air Defense Antitank Systems (ADATS) (U.S. General Accounting Office, 1991a, 1990a). Although criticisms of the reporting process are not limited to the time period described, the model will yield considerable insight into this reporting phenomenon.<sup>4</sup>

In response to allegations such as these, DOT&E has concentrated additional efforts toward oversight of the test reporting process. DOT&E officials have begun to monitor the progress of operational testing on site. In addition, DOT&E officials currently conduct independent evaluations of operational test results. These evaluations are drawn directly from the raw test data and are not subject to DoD interpretation. DOT&E reports directly to the

---

<sup>4</sup> See any of the following GAO publications for additional criticisms of the reporting process (U.S. General Accounting Office, 1979b, 1980, 1988b).

Congress. If DoD disagrees with any of the conclusions reached by DOT&E, it may append the report to Congress with its own comments.

### 3. RELATED LITERATURE

An important avenue of research on the topic of information transmission was initiated by Milgrom (1981). As an application of more general theorems regarding the monotone likelihood ratio property (MLRP), Milgrom introduces *games of persuasion*. In a persuasion game an interested party (agent) possesses private information regarding the underlying state of nature and attempts to influence a decision maker (principal) by selectively providing data. For example, the agent might be a salesman who has information regarding the quality of his product and selectively conveys a subset of the data to a consumer. In equilibrium, the consumer accounts for the salesman's selectivity in reaching a consumption decision.

By assumption, the agent is unable (or unwilling because of infinite penalties) to communicate reports which are incorrect. Matthews and Postlewaite (1985) have described this assumption as the imposition of effective antifraud regulations. In light of these antifraud regulations, reports from the agent are limited to supersets of the truth. The salesman may, for example, claim that the product meets or exceeds some criteria if and only if the criteria is satisfied. At the discretion of the agent, however, the report may range from entirely uninformative to absolutely precise.

Milgrom shows that a Nash equilibrium always exists in which the principal resolves to ignore all reports and the agent makes only uninformative reports. However, a proposition demonstrates that every sequential equilibrium (Kreps and Wilson, 1982) of the persuasion game involves precise revelation of the truth by the agent. At the sequential equilibrium, the principal believes that any information withheld by the agent is extremely unfavorable. In the face of such extreme skepticism the agent's best response is truthful revelation.

Matthews and Postlewaite (1985) extend Milgrom's model by adding an earlier stage in which the agent chooses whether or not to become informed. They assume that the cost of

acquiring information is zero. In this context, they distinguish between mandatory disclosure and antifraud regulations. Under mandatory disclosure, an agent must disclose whether or not he has acquired information. Mandatory disclosure does not, however, require the truthful conveyance of information acquired. Truthful reporting of information is still governed by antifraud. Matthews and Postlewaite assume effective antifraud throughout the paper but consider variations of the model with mandatory disclosure and without.

Using the solution concept of sequential equilibrium, Matthews and Postlewaite examine the dependence of information acquisition upon disclosure rules. They show that the agent will acquire and fully disclose information whenever disclosure is not mandatory. When disclosure is mandatory, however, the agent may or may not acquire information. Note that in the presence of antifraud, agents who do not acquire information must report total ignorance to avoid any chance of misrepresenting the truth. In the absence of mandatory disclosure, the sequential equilibrium calls for the principal to adopt extreme skepticism toward any report of ignorance. In the face of such extreme skepticism, agents choose to acquire information and fully reveal.

The extreme skepticism on the part of the principal completely unravels any possible equilibrium claim of ignorance by the agent. Results such as these have been termed unraveling results. Avoiding this unraveling requires some type of credibility for claims of ignorance by the agent. In the context of their model, mandatory disclosure provides this credibility and impedes the unraveling.

Shavell (1994) extends the model of Matthews and Postlewaite in several important directions. Shavell allows the cost of acquiring information to be privately held by the agents. Shavell also considers cases in which the information acquired may be socially valuable. Socially valuable information increases the underlying value of the exchange between the agent and the principal. As in Matthews and Postlewaite, Shavell assumes effective antifraud and analyzes the impact of mandatory disclosure.

Shavell shows that unraveling may be impeded even in the absence of mandatory disclosure. At the sequential equilibrium, two types of agents claim ignorance. The first type have realized cost draws which exceed the expected value of acquiring information. They are truly ignorant. The second type have acquired information which was so unfavorable that they

achieve a higher payoff by claiming ignorance. In equilibrium, the principal simply assigns the appropriate probability to each type when computing his reservation value for exchanges with agents claiming ignorance. Unraveling is also impeded when the information acquired is socially valuable.

In short, the privacy of the cost draw gives credibility to the claims of ignorance by the agents. This credibility is enough to preclude the unraveling effect. Such a result is in stark contrast with Matthews and Postlewaite. This contrast highlights the critical importance of the assumption regarding the distribution of costs. When the cost distribution is not degenerate, the unraveling effect is impeded and the principal must give credibility to claims of ignorance.<sup>5</sup> However, as the cost distribution becomes degenerate the principal's skepticism completely unravels any claim of ignorance by the agent. In Matthews and Postlewaite, therefore, it is not the assumption that the costs of acquiring information are zero which drives the unraveling result. Rather it is the degeneracy of the cost distribution.

Jovanovic (1982) reaches a similar conclusion by imposing privately known costs of conveying information upon the agent. It seems clear that some private information on the part of the agent is required to avoid the unraveling effect.

Kofman and Lawarrée (1993) present a variant in which the agent takes an action which partially determines the state of nature. Although the state of nature is revealed to the principal, the action taken by the agent is not observed. In this context, the principal may employ an internal auditor to gather more accurate information regarding the agent's action. The model allows for the possibility that the internal auditor may be involved in a collusive agreement with the agent. In equilibrium, however, collusion is stymied by *bounty hunter* contracts in which the principal gives any penalty extracted from the agent directly to the auditor.

Kofman and Lawarrée also consider the case in which an external auditor may be employed. The external auditor does not have the possibility of colluding with the agent, but lacks the expertise to gather data as accurately as the internal auditor. A proposition determines the conditions under which the principal will use the internal auditor, the external auditor, or

---

<sup>5</sup> In this context, degeneracy requires only a support for the cost distribution which does not include the value of acquiring information.

both. Although they do not elaborate, Kofman and Lawarrée indicate that the model is consistent with the relationship between Congress and the Department of Defense. Perhaps DOT&E would play the role of the external auditor and the Service test agencies would play the role of the internal auditor.

Crawford and Sobel (1982) take an entirely different approach to games of information transmission. In their model, the preferences of the two parties are somewhat aligned. Crawford and Sobel completely relax the antifraud assumption to allow for a type of cheap talk communication. Although equilibrium messages will not necessarily involve full disclosure, they show that antifraud is not violated at equilibrium.

Crawford and Sobel show that all the Bayesian Nash equilibria are *partition equilibria*. In a partition equilibria, the agent introduces noise into his report by partitioning the state space and reporting only the partition in which the realization lies. The size of the individual partitions varies directly with the proximity of the parties' preferences. For identical preferences, the partitions will be infinitely small and the report will be precise. As preferences differ, the partitions grow in size and the agent attempts to pool over larger and larger realizations. If preferences are suitably different, the agent partitions the state space into a single partition which amounts to a claim of ignorance.

Crawford and Sobel show that if the preferences of the parties do not coincide, the equilibrium number of partitions is always finite. Thus information is never perfectly revealed. Such a result is also in sharp contrast with the results from Milgrom and Matthews and Postlewaite.

Green and Stokey (1980) consider a similar game from an alternate perspective. The preferences of the parties are held constant while the information structure itself is varied. Green and Stokey demonstrate that a *more informative* information structure does not necessarily imply higher welfare for the parties.<sup>6</sup> Examples are constructed in which the welfare of each party is either reduced or enhanced by improvements in the information structure. In addition, Green and Stokey identify several types of equilibria including partition equilibria. For the purpose of

---

<sup>6</sup> One information structure is said to be more informative than another if it provides higher expected utility for a decision maker regardless of the utility function. See Hirshleifer and Riley (1992) for a complete discussion.

comparative statics, focus is given to the partition equilibria. It is shown that the agent will always prefer small improvements in the information structure at a partition equilibria, while the principal may not.

#### 4. THE MODELING FRAMEWORK

The model contains three economic agents. Congress plays the role of the principal while the Department of Defense plays the role of the agent. We assume that DOT&E is a perfect agent of Congress. Thus, there are effectively two players: Congress and DoD.

There are three possible states regarding an individual program:  $A$ ,  $B$ , and  $C$ . Nature determines the state of the program according to the probabilities  $P_A$ ,  $P_B$ , and  $P_C$ , respectively. We assume that these probabilities are the pretesting beliefs of all participants and are common knowledge. In addition, we assume that the states of the world are mutually exclusive and exhaustive. That is,  $P_A + P_B + P_C = 1$  and  $P_i \geq 0$  for all  $i = A, B, C$ .

The testing of a system reveals an information partition which is a superset of the true state. Information partitions may range from very fine, as when a single state is uniquely identified, to very coarse. Let  $R_{ij}$  denote the payoff to DoD when testing reveals information partition  $(i)$  and the system is procured. For example,  $R_A$  is the payoff to DoD when a state  $A$  system is procured and  $R_{AB}$  is the payoff when an information partition  $(A,B)$  system is procured. Similarly, let  $S_{ij}$  denote the payoff to Congress when testing reveals partition  $(i)$  and the system is procured. If a system is not procured both parties are assumed to receive zero payoff.

**Assumption 1** *We make the following assumptions regarding the payoffs :*

- a.  $R_A < 0 < R_{AB} < R_B < R_{BC} < R_C$ .
- b.  $S_A < S_{AB} < S_B < S_{BC} < S_C$ .
- c.  $S_{ABC} < 0 < R_{ABC}$ .
- d.  $S_{ij} \leq R_{ij}$  for all information partitions.

e. All payoffs are common knowledge to the participants.

DoD will choose to proceed with any program that is not state A, while Congress will choose to proceed with information partitions (B,C) and C only. Clearly, the disagreement is centered on state B. Effectively, we assume that DoD would proceed with any program Congress would approve, but not the converse. These assumptions appear to be consistent with the majority of the historical disagreements between Congress and DoD. From time to time, however, Congress has approved funding for programs which DoD wished to terminate. An example of a such a program can be found in the Navy's V-22 Osprey (U.S. General Accounting Office, 1990b). In its current form, our model cannot explain programs such as this.

We give the following definition of socially valuable information:

**Definition 1** *Information is said to be socially valuable if the conditional expected value within a given information partition exceeds the actual payoff for that partition.*

If, for example,  $\frac{P_A R_A + P_B R_B}{(1 - P_C)} > R_{AB}$ , then information is said to be socially valuable. If information is not socially valuable—i.e., information does not change the way DoD behaves—then the previous statement is characterized by equality. The idea is that  $R_{AB}$ , in this case, is really a reduced form. Thus, the conditional expected value within a given information partition must always be at least as large as the actual payoff.

Intuitively, socially valuable information would be appropriate if knowing more precise information allowed DoD to adopt a better procurement strategy. For example, finer information might allow the technicians to make small changes in the design which might enhance the value of the overall program.

The total testing budget may be allocated over two types of tests. Let  $t_A$  denote resources devoted toward distinguishing state A from its complement. We term this type of testing type A testing. Similarly, let  $t_C$  denote resources devoted toward distinguishing state C from its complement. We term this type of testing type C testing. We assume that the total resources available for testing are exogenously given as  $T$ . Thus,  $t_A + t_C \leq T$  is a constraint on the test design process.

About this PDF file: This new digital representation of the original work has been recomposed from XML files created from the original paper book, not from the original typesetting files. Page breaks are true to the original; line lengths, word breaks, heading styles, and other typesetting-specific formatting, however, cannot be retained, and some typographic errors may have been accidentally inserted. Please use the print version of this publication as the authoritative version for attribution.

Let  $Z_A(t_A)$  denote the actual probability of distinguishing state A from its complement as a function of the type A testing resources. Similarly let  $Z_C(t_C)$  denote the probability of distinguishing state C from its complement.

**Assumption 2** *We make the following assumptions regarding the testing technology and test information::*

a.  $Z'_i(t_i) = \frac{\partial Z_i}{\partial t_i} > 0$  for all  $t_i \leq T$ ;  $i = A, C$ .

b.  $Z''_i(t_i) = \frac{\partial^2 Z_i}{\partial t_i^2} < 0$  for all  $t_i \leq T$ ;  $i = A, C$ .

c.  $\frac{\partial^2 Z_i}{\partial t_i \partial t_j} = 0$  for all  $t_i \leq T$ ;  $i, j = A, C$   $i \neq j$ .

d.  $Z_i(0) = 0$  for  $i = A, C$ .

e.  $Z_i(T) < 1$  for  $i = A, C$ .

f. Once the relevant player selects  $t_A$  and  $t_C$ , they are common knowledge.  $Z_{i(j)}$  is common knowledge for  $i = A, C$ .

Intuitively, we assume that additional resources increase the probability of distinguishing between a state and its complement at a decreasing rate. Furthermore, there are no learning spillovers between type A testing and type C. These assumptions will allow for the possibility of interior solutions in test resource allocation.

## 5. DECISION PROBLEMS AND GAMES

This section poses several decision problems and games which, we argue, are consistent with various time periods in the history of operational testing.

### Decision Problem 1

We begin by considering the decision problem faced by DoD in the absence of any Congressional oversight. We analyze a two stage decision problem. DoD must first determine an allocation for the total testing budget  $T$ . Then, after an information partition is revealed, DoD must decide to continue or terminate the program. Formally, we represent the decision problem as follows:

**Stage 1** DoD determines the allocation of test resources  $t_A$  and  $t_C$ .

**Stage 2** Nature reveals an information partition to DoD.

**Stage 3** DoD continues or terminates the program.

This decision problem is consistent with the procurement process prior to Public Law 92-156. Recall that this law required DoD to begin reporting operational test results to Congress. As described in section 2, DoD exercised considerable influence over the entire procurement cycle during this time period.

According to assumption 1, DoD will proceed with any program that is not state  $A$ . In light of these stage 3 preferences, the objective function for stage 1 can be expressed by equation 5.1 below:

$$\begin{aligned}\pi_1 = & (1 - Z_A(t_A))(1 - Z_C(t_C))R_{ABC} + Z_A(t_A)(1 - Z_C(t_C))(1 - P_A)R_{BC} \\ & + (1 - Z_A(t_A))Z_C(t_C)[P_C R_C + (1 - P_C)R_{AB}] \\ & + Z_A(t_A)Z_C(t_C)[P_B R_B + P_C R_C].\end{aligned}\quad (5.1)$$

Equation 5.1 can be easily interpreted. With probability  $(1 - Z_A(t_A))(1 - Z_C(t_C))$ , the completely uninformative partition is revealed. In that case, DoD would continue the program and receive benefit  $R_{ABC}$ . With probability  $Z_A(t_A)(1 - Z_C(t_C))$ , DoD can only distinguish state  $A$  from its complement. With probability  $(1 - P_A)$  then, partition  $(B, C)$  is revealed and a payoff

About this PDF file: This new digital representation of the original work has been recomposed from XML files created from the original paper book, not from the original typesetting files. Page breaks are true to the original, line lengths, word breaks, heading styles, and other typesetting-specific formatting, however, cannot be retained, and some typographic errors may have been accidentally inserted. Please use the print version of this publication as the authoritative version for attribution.

of  $R_{BC}$  is earned. However, with probability  $P_A$ , state  $A$  is revealed and a zero payoff is earned. The other entries have similar interpretations.

It is important to note that equation 5.1 would represent the social planners problem if DoD's preferences accurately reflected the society's true preferences over program quality.

## Decision Problem 2

This section considers the Congressional decision problem in the absence of any DoD influence. Again we identify a multi-stage decision problem. In the first stage, DOT&E (the perfect agent of Congress) selects an allocation of test resources.<sup>7</sup> After observing an information partition, DOT&E reports to Congress, who decides whether to continue or terminate the project.

**Stage 1** *DOT&E determines the allocation of test resources  $t_A$  and  $t_C$ .*

**Stage 2** *Nature reveals an information partition to DOT&E.*

**Stage 3** *DOT&E reports the information partition to Congress.*

**Stage 4** *Congress continues or terminates the project.*

According to assumption 1, Congress will continue only those projects in information partitions ( $B, C$ ) and ( $C$ ). In light of these preferences, the stage 1 objective function for DOT&E is given by equation 5.2 below.<sup>8</sup>

$$\Pi_2 = Z_A(t_A)(1 - Z_C(t_C))(1 - P_A)S_{BC} + Z_C(t_C)P_C S_C \quad (5.2)$$

<sup>7</sup> "Congress" and "DOT&E" are identical players in this and all subsequent games. We use the different names to mimic the role of each in the actual process.

<sup>8</sup> We denote Congressional objective functions with uppercase symbols and DoD objective functions with lowercase.

It is important to note that equation 5.2 would represent the social planners problem if Congressional preferences accurately reflected the society's true preferences over program quality.

### Game 3

In this section we consider a game in which DoD determines the test resource allocation, while Congress makes the final funding decision. In this game, DoD observes the information partition and makes a report to Congress, who has not seen the information partition.

**Stage 1** DoD determines the allocation of test resources  $t_A$  and  $t_C$ .

**Stage 2** Nature reveals the information to DoD.

**Stage 3** DoD makes a report to Congress regarding the information partition.

**Stage 4** Congress continues or terminates the project.

This game is consistent with the time period between Public Law 92-156 and the formation of DOT&E. As described in section 2, DoD was required to report operational test results to Congress over this time period. However, Congress was not involved in the planning and conducting of the actual tests nor did they exercise any effective oversight of the test reporting stage.

In all of the games considered, we assume effective antifraud regulations but not mandatory disclosure. Although DoD is not forced to reveal information, any information it chooses to reveal must be correct. The absence of mandatory disclosure allows DoD to pool over information partitions. For example, they may choose to report information partition  $(B, C)$  when they observe  $B$ . DoD can always report less fine information than they observe (lack of mandatory disclosure) but cannot report finer information (antifraud).

In the final stage of the game, Congress will approve only those projects reported to be in partitions  $(B, C)$  and  $C$ . A perfect Bayesian equilibrium exists in which DoD pools over states  $B$

About this PDF file: This new digital representation of the original work has been recomposed from XML files created from the original paper book, not from the original typesetting files. Page breaks are true to the original, line lengths, word breaks, heading styles, and other typesetting-specific formatting, however, cannot be retained, and some typographic errors may have been accidentally inserted. Please use the print version of this publication as the authoritative version for attribution.

and C by reporting  $(B,C)$  for both. All other partitions are reported truthfully. Whenever DoD reports to Congress something other than  $(B,C)$ , Congress believes the report to be exactly what DoD observed. When DoD reports  $(B,C)$ , Congress Bayesian updates its prior probabilities. There are other perfect Bayesian equilibria but this one seems to best capture the salient behavior of DoD.<sup>9</sup>

At the reporting stage of this equilibrium, the stage 1 objective function for DoD is given by equation 5.3:

$$\pi_3 = Z_A(t_A)(1 - Z_C(t_C))(1 - P_A)R_{BC} + Z_C(t_C)P_C R_C + Z_A(t_A)Z_C(t_C)P_B R_B. \quad (5.3)$$

The last term is a direct result of DoD's ability to pool over  $(B,C)$  information partitions.

#### Game 4

This section considers the case in which DOT&E determines the allocation of test resources while DoD observes the actual test results. DoD then reports the test results to Congress, who may continue or terminate the program. We continue to assume effective antifraud but not mandatory disclosure.

**Stage 1** DOT&E determines the allocation of testing resources  $t_A$  and  $t_C$ .

**Stage 2** Nature reveals the information partition to DoD.

**Stage 3** DoD makes a report to Congress, regarding the information partition.

**Stage 4** Congress continues or terminates the program.

We believe that this game is consistent operational testing during the first several years after the formation of DOT&E. As described in section 2, DOT&E concentrated its early efforts

<sup>9</sup> We make no attempt to establish uniqueness of any equilibrium in any four games. We are focusing attention on equilibria that, again, best capture the salient behavior of DoD and Congress.

About this PDF file: This new digital representation of the original work has been recomposed from XML files created from the original paper book, not from the original typesetting files. Page breaks are true to the original; line lengths, word breaks, heading styles, and other typesetting-specific formatting, however, cannot be retained, and some typographic errors may have been accidentally inserted. Please use the print version of this publication as the authoritative version for attribution.

on improving test design and implementation. Standards for testing were established and DoD testing personnel were forced to comply. During this time period, however, DOT&E did not effectively oversee the reporting of test results to Congress.

Just as in game 3, a perfect Bayesian equilibrium exists in which DoD pools over states B and C by reporting  $(B, C)$  for each. All other partitions are reported truthfully. Congressional beliefs are as in game 3.

When  $(B, C)$  is reported then DoD will have observed one of three possible states— $B$ ,  $C$ , or  $(B, C)$ . If DoD observes  $(B, C)$  then the Congressional payoff is  $S_{BC}$ . If DoD observes  $B$  then, because we allow for information to be socially valuable, the Congressional payoff will be  $S_B$  even though the report is  $(B, C)$ . The same is true for  $C$ . Note that if information is not socially valuable this distinction is irrelevant. Under this assumption, the stage 1 objective function for DOT&E is given by equation 5.4:

$$II_4 = Z_A(t_A)(1 - Z_C(t_C))(1 - P_A)S_{BC} + Z_A(t_A)Z_C(t_C)P_BS_B + Z_C(t_C)P_CS_C \quad (5.4)$$

### Game 5

This section considers the case in which DoD determines the allocation of test resources, but the actual test results are observed by DOT&E. DOT&E then reports truthfully to Congress, who may continue or terminate the program.

**Stage 1** DoD determines the allocation of testing resources  $t_A$  and  $t_C$ .

**Stage 2** Nature reveals the information partition to DOT&E.

**Stage 3** DOT&E reports the information partition to Congress.

**Stage 4** Congress continues or terminates the program.

As described in section 2 Congress originally charged DOT&E with two major oversight responsibilities: test planning and test reporting. DOT&E concentrated its early efforts on improving test planning at the expense of reporting oversight. The decision to do so can be analyzed in the context of this game.

Prior to the formation of DOT&E, the status of operational testing was consistent with game 3. By concentrating its efforts in the area of test planning, DOT&E effectively shifted the operational test environment to game 4. DOT&E could have chosen to concentrate its efforts on reporting oversight. This would have shifted the testing environment to game 5. By considering game 5, we therefore gain insight into this decision.

Since DOT&E is the perfect agent of Congress, they will report truthfully the information partition revealed in stage 2. Since Congress will approve any program revealed as partition  $(B, C)$  or  $C$ , the stage 1 objective function for DoD is given by equation 5.5:

$$\pi_3 = Z_A(t_A)(1 - Z_C(t_C))(1 - P_A)R_{BC} + Z_C(t_C)P_C R_C \quad (5.5)$$

## 6. WELFARE RESULTS

This section examines the welfare of the players at the equilibria of the decision problems and games proposed in the previous section. The first stage of each game described above involves the solution of a constrained maximization problem in  $t_A$  and  $t_C$ . In all of the following analysis, we assume that the budget constraint  $t_A + t_C \leq T$  is binding. In addition, we assume that the sufficient conditions for maxima are always satisfied. In appendix B we show that this assumption requires restrictions on only three of the five problems considered.

Substituting the constraint  $t_A = T - t_C$  into the objective functions for the various games yields a series of unconstrained problems in  $t_C$ . Evaluating the welfare of the players at the relevant solution for  $t_C$  yields the following rankings:

**Proposition 1** *Under assumptions 1 and 2, Congressional welfare evaluated at the relevant solution for  $t_C$  is characterized by*

$$a. \Pi_2(t_c^*) \geq \Pi_4(t_c^*) \geq \Pi_3(t_c^*) \Pi_1(t_c^*)$$

$$b. \Pi_3(t_c^*) \geq \Pi_5(t_c^*)$$

**Proof of Proposition 1** To prove the first part of the proposition notice that for any value of  $t_C$ ,  $\Pi_4$  exceeds  $\Pi_3$  by magnitude  $Z_A(T-t_C)Z_C(t_C)P_BS_B$  (since  $S_B$  is negative by assumption). Now evaluate both  $\Pi_4$  and  $\Pi_2$  at  $t_c^*$  to yield  $\Pi_2(t_c^*) \geq \Pi_4(t_c^*)$ . Note that  $\Pi_2(t_c^*) \geq \Pi_3(t_c^*)$  completes the proof.

To formulate the Congressional objective function for game 3 we simply replace  $R_{(1)}$  with  $S_{(1)}$  in  $\pi_3$  to obtain  $\Pi_3$ . Notice that doing so yields exactly  $\Pi_4$ . Thus  $\Pi_3 = \Pi_4$ . Now since Congress selects the maximal value of  $\Pi_4$  in game 4, the welfare from game 3, in which DoD selects, cannot be higher.

By replacing  $R_{(1)}$  with  $S_{(1)}$  in  $\pi_1$  as before we obtain  $\Pi_1$ . Notice that for any  $t_C$ ,  $\Pi_1$  exceeds  $\Pi_3$  by magnitude  $(1 - Z_A(T-t_C))(1 - Z_C(t_C))S_{ABC} + (1 - Z_A(T-t_C))Z_C(t_C)(1 - P_C)S_{AB}$ . As in the first part of this proof, the maximal value of  $\Pi_1$  chosen by Congress in game 4 must exceed the maximal value of  $\Pi_3$  which in turn cannot be less than the value derived from DoD's choice in game 1.

To prove the second part of the proposition, replace  $R_{(1)}$  with  $S_{(1)}$  in  $\pi_5$  to obtain  $\Pi_5$ . Notice that  $\Pi_2 = \Pi_5$ . As above, the value of  $\Pi_2$  selected by Congress in game 2 cannot be less than the value which results from DoD's choice in game 5.

The proposition demonstrates that oversight of the test design stage (game 4) cannot decrease the welfare of the principal as compared with no oversight (game 3). However, oversight of the reporting stage (game 5) may increase or decrease the principal's welfare as compared with no oversight. Below, we examine the possibility that increased oversight may make the principal worse off.

First note that  $\pi_5$  can be expressed as a function of  $\pi_3$  by the following:

$$\pi_5 = \pi_3 - Z_A(T-t_C)Z_C(t_C)P_BR_B. \quad (6.1)$$

differentiating and evaluating the expression at the solution to game 3:

$$\Delta_{3,3} = 0 - [-Z'_A(T-t_C)Z_C(t_C) + Z'_C(t_C)Z_A(T-t_C)]P_B R_B. \quad (6.2)$$

Now if we assume that the solution to game 3 involves a relatively high value of  $t_A$  and a relatively low value of  $t_C$ , then the bracketed term above will be positive. This is a reasonable assumption given DoD's preference for  $(B,C)$  systems. When this assumption is satisfied,  $t_A^3$  will exceed  $t_C^3$ . In this case, the additional oversight by the principal has the unintended effect of reducing the type C testing. Below we show that this reduction in type C testing may lead to a reduction in the principal's welfare.

Assuming  $t_A^3 > t_C^3$  as above, we compare the principal's welfare at the solution of games 3 and 5 with the following equation:

$$\begin{aligned} \Pi_3 - \Pi_5 &= [Z_A^3(1-Z_C^3) - Z_A^5(1-Z_C^5)](1-P_A)S_{BC} \\ &+ [Z_C^3 - Z_C^5]P_C S_C + Z_A^3 Z_C^3 P_B S_B \end{aligned} \quad (6.3)$$

Specifically we are interested in the case in which  $\Pi_3 - \Pi_5 \geq 0$ . The first and third terms Of 6.3 are negative by assumption, but the second term is positive. Thus equation 6.3 will exceed 0 if  $S_C$  is suitably large. The explicit condition is given by the following inequality:

$$S_C > \frac{1}{[Z_C^3 - Z_C^5]P_C} \left( [Z_A^5(1-Z_C^3) - Z_A^3(1-Z_C^5)](1-P_A)S_{BC} - Z_A^3 Z_C^3 P_B S_B \right) \quad (6.4)$$

Thus if inequality 6.4 is satisfied, additional oversight of the reporting stage will actually reduce the principal's welfare as compared to no oversight. Inequality 6.4 is most likely to be satisfied

when information is socially valuable. In that case,  $S_C$  is large compared to  $S_{BC}$  and the condition is easier to satisfy.

If we suppose that Congressional preferences are aligned with society's preferences, then proposition 1 sheds a favorable light on the evolution of operational testing. By concentrating on the oversight of test design, DOT&E has increased social welfare. In addition, oversight of the reporting stage in conjunction with oversight of the test design stage has moved the process toward decision problem 2. This additional oversight has also improved social welfare if Congress reflects the true preferences of the society.

We obtain a similar proposition regarding DoD welfare:

**Proposition 2** Under assumptions 1 and 2, DoD welfare evaluated at the relevant solution for  $t_C$  is characterized by

$$a. \pi_1(t_C^{1*}) \geq \pi_3(t_C^{3*}) \geq \pi_5(t_C^{5*}) \geq \pi_2(t_C^{2*})$$

$$b. \pi_3(t_C^{3*}) \geq \pi_4(t_C^{4*})$$

**Proof of Proposition 2** To prove the first part of the proposition first notice that for any value of  $t_C$ ,  $\pi_1$  exceeds  $\pi_3$  by magnitude  $(1 - Z_A(T + t_C))(1 - Z_C(t_C))R_{ABC} + (1 - Z_A(T + t_C))Z_C(t_C)(1 - P_C)R_{AB}$ . As in the proof of proposition 2, the maximal value of  $\pi_1$  must therefore exceed the maximal value from  $\pi_3$ . The proof of  $\pi_3(t_C^{3*}) \geq \pi_5(t_C^{5*})$  follows from the same logic.

To formulate  $\pi_2$ , replace  $\pi_5$  with  $R_{AB}$  in  $\pi_2$ . Notice  $\pi_5 = \pi_2$  for all  $t_C$ . As in proposition 2, no other value of  $t_C$  can yield a payoff for  $\pi_5$  in excess of the value chosen by DoD in game 5. The proof of the second part of the proposition follows precisely the same logic.

If we suppose that the society's true preferences are reflected by DoD, proposition 2 sheds an unfavorable light on the evolution of operational testing. Social welfare was highest in the absence of Congressional involvement (decision problem 1). As Congressional oversight has strengthened, social welfare has progressively declined.

Propositions 1 and 2 bound social welfare during various stages in the evolution of operational testing. In all likelihood, society's true preferences are somewhere between those of

Congress and DoD. Therefore, the extent to which additional oversight has increased or decreased social welfare remains an open question.

## 7. QUALITATIVE TESTING RESULTS

In this section, we compare the equilibrium levels of testing which result from the decision problems and games posed in section 5. We continue to assume that the budget constraint is binding and the sufficient conditions for maxima are satisfied. In this section we make an additional assumption regarding the social value of information.

**Assumption 3** *Information has no social value.*

Thus in this section of the paper we assume that the payoff for any multi-state information partition is exactly equal to the conditional expected value within that information partition. So, for example,  $(1 - P_A)P_{BC} = P_B P_B + P_C P_C$  by assumption.

**Proposition 3** *Under assumptions 1, 2, and 3 the equilibrium testing induced by the decision problems and games posed in section 5 can be ranked according to*

$$t_C^{2*}, t_C^{4*} > t_C^{3*}, t_C^{5*} > t_C^{1*}. \quad (7.1)$$

The tedious but straightforward proof of proposition 3 is contained in [appendix A](#). Intuitively, the proposition orders the type C testing generated by the various models and decision problems. Below we argue that this ranking captures the major features of the evolution of operational testing in the Department of Defense.

Prior to the enactment of Public Law 92-156 in 1971, DoD exercised considerable control over the entire procurement process. This influence extended not only to test design but also into the final procurement decisions. We analyze this time period with decision problem 1.

Proposition 3 reveals the general nature of the conflict between Congress and DoD over operational testing. Namely, DoD devotes less resources toward type C testing than Congress

would like. As a first step toward resolving this conflict, Congress required DoD to report operational test results. In the context of the model, the testing process was shifted to game 3.

Game 3 represents the status of operational testing in the time between the enactment of Public Law 92-156 and the establishment of DOT&E. Over this time period, DoD received substantial criticism for what was termed inadequate testing. In the context of our model, this inadequacy might be interpreted as a lack of resources devoted to  $t_C$ . From the standpoint of DoD, the testing was completely adequate to support procurement decisions. However, Congress considered the test resource allocation to be inadequate. Proposition 3 verifies this intuition.

Section 2 describes how DOT&E concentrated its early efforts in the area of test design and implementation. DOT&E could just have easily concentrated its efforts on improving the test reporting process. However, DOT&E's limited budget probably would not have allowed them to impact both test design and test reporting. In the context of our model, this decision is simply a choice between game 4 (impact test design) and game 5 (impact test reporting). Proposition 3 reveals game 4 as the preferred option in terms of test resource allocation. As we have seen in section 6, game 5 might actually reduce the  $t_C$  testing from the game 3 level. In light of proposition 3, DOT&E's decision to impact test design appears to be a rational response to the underlying incentives.

More recently, DOT&E has taken an active oversight role in the reporting of test results. As described in section 2, DOT&E personnel are responsible for independent assessments of test data. In addition, high ranking staff members are regularly called before Congress to address the desirability of procuring new systems and the adequacy of operational testing. It is important to note that these responsibilities are in addition to DOT&E's continued oversight of test design and implementation. Therefore, DOT&E now plays a significant role in all phases of the testing process. In the context of our model, the operational testing environment is moving toward decision problem 2. We have already shown that decision problem 2 obtains the highest welfare for Congress but the lowest welfare for DoD.

## 8. CONCLUSION

We have presented a model which extends the information transmission literature to consider the question of strategic information generation. In the context of a principal agent game, strategic information generation gives the agent an added dimension in which to manipulate the process. In response, the principal might choose to extend some oversight authority. We have shown that oversight of the test design stage cannot decrease the principal's welfare while oversight of the test reporting stage may. Our analysis has shown that the model is remarkably consistent with the evolution of testing institutions in the Department of Defense.

There are many avenues in which the present model might be extended. In the context of the Department of Defense example, the next logical step might involve an endogenous total testing budget  $T$ . Despite the Congressional oversight efforts documented above, the Department of Defense continues to maintain considerable control over the total testing budget. A more complete description of the testing environment requires this feature.

## Appendix A: Proposition Proofs

This appendix contains the proof of proposition 3. We begin by considering a simple lemma.

**Lemma 1** *The following inequality is satisfied at the equilibrium of game 3 and game 5.*

$$[-Z'_A(T-t_c^*)(1-Z_c(t_c^*))+Z'_C(t_c^*)(1-Z_A(T-t_c^*))] > 0. \quad (\text{A.1})$$

**Proof of Lemma 1** Consider the first order condition from game 3:

$$\begin{aligned} 0 &= [-Z'_A(T-t_c)(1-Z_c(t_c))-Z'_C(t_c)Z_A(T-t_c)](1-P_A)R_{BC} \\ &+ Z'_C(t_c)P_C R_C \\ &+ [-Z'_A(T-t_c)Z_c(t_c)+Z'_C(t_c)Z_A(T-t_c)]P_B R_B \end{aligned} \quad (\text{A.2})$$

which can be manipulated to form the following,

$$\begin{aligned} &[-Z'_A(T-t_c)(1-Z_c(t_c))+Z'_C(t_c)(1-Z_A(T-t_c))](1-P_A)R_{BC} \\ &= [Z'_A(T-t_c)Z_c(t_c)-Z'_C(t_c)Z_A(T-t_c)]P_B R_B \\ &+ Z'_C(t_c)[(1-P_A)R_{BC}-P_C R_C]. \end{aligned} \quad (\text{A.3})$$

Thus, the lemma will hold if the right-hand side of A.3 is positive. When information has no social value, then  $(1-P_A)R_{BC}=P_B R_B + P_C R_C$ . In this case, equation A.3 reduces to the following:

$$\begin{aligned}
 & [-Z'_A(T-t_c)(1-Z_c(t_c)) + Z'_c(t_c)(1-Z_A(T-t_c))] (1-P_A) R_{BC} \\
 = & [Z'_A(T-t_c)Z_c(t_c) + Z'_c(t_c)(1-Z_A(T-t_c))] P_B R_B
 \end{aligned} \tag{A.4}$$

Since the right-hand side of A. 4 is positive, the lemma is shown to hold for game 3. Now consider the first order condition for game 5:

$$0 = [-Z'_A(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)Z_A(T-t_c)] (1-P_A) R_{BC} + Z'_c(t_c) P_C R_C. \tag{A.5}$$

When information has no social value, equation A.5 can be simplified to the following:

$$\begin{aligned}
 & [Z'_A(T-t_c)(1-Z_c(t_c)) + Z'_c(t_c)Z_A(T-t_c)] P_B R_B \\
 = & [-Z'_A(T-t_c)(1-Z_c(t_c)) + Z'_c(t_c)(1-Z_A(T-t_c))] P_C R_C.
 \end{aligned} \tag{A.6}$$

As the left-hand side of equation A.6 is positive, the lemma is shown to hold for game 5.

**Proof of Proposition 3** To show  $t_c^*$  exceeds  $\bar{t}_c^*$ , we first express the objective function from decision problem 2 in terms of the game 3 objective function:

$$\begin{aligned}
 \Pi_2 = & \pi_3 - Z_A(T-t_C)Z_C(t_C)P_B R_B \\
 + & Z_A(T-t_C)(1-Z_C(t_C))(1-P_A)[S_{BC} - R_{BC}] \\
 + & Z_C(t_C)[P_C S_C - P_C R_C].
 \end{aligned} \tag{A.7}$$

Evaluating the first order conditions from  $\Pi_2$  at the solution to  $\pi_3$  yields the following:

$$\begin{aligned}
 \Delta_{2,3} = & [Z'_A(T-t_c)Z_c(t_c) - Z'_c(t_c)Z_A(T-t_c)] P_B R_B \\
 + & [-Z'_A(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)Z_A(T-t_c)] (1-P_A)
 \end{aligned}$$

$$\begin{aligned} & \times [S_{BC} - R_{BC}] \\ + & Z'_c(t_c) [P_c S_c - P_c R_c]. \end{aligned} \quad (\text{A.8})$$

Substituting for  $[Z'_s(T-t_c)Z_c(t_c) - Z'_c(t_c)Z_s(T-t_c)]P_b R_b$  from equation A.3, we have the following:

$$\begin{aligned} \Delta_{2,3} = & [-Z'_s(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)Z_s(T-t_c)](1-P_s)R_{BC} \\ + & [-Z'_s(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)Z_s(T-t_c)](1-P_s) \\ & \times [S_{BC} - R_{BC}] \\ + & Z'_c(t_c) [P_c S_c - P_c R_c] + Z'_c(t_c)P_c R_c \end{aligned} \quad (\text{A.9})$$

which reduces to the following:

$$\begin{aligned} \Delta_{2,3} = & [-Z'_s(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)Z_s(T-t_c)](1-P_s)S_{BC} \\ + & Z'_c(t_c)P_c S_c. \end{aligned} \quad (\text{A.10})$$

Making the first term as large as possible yields the following:

$$\begin{aligned} \Delta_{2,3} \geq & [-Z'_s(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)Z_s(T-t_c)] [P_b S_b + P_c S_c] \\ + & Z'_c(t_c)P_c S_c. \end{aligned} \quad (\text{A.11})$$

Simplifying,

$$\begin{aligned} \Delta_{2,3} \geq & [-Z'_s(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)Z_s(T-t_c)]P_b S_b \\ + & [-Z'_s(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)(1-Z_s(T-t_c))]P_c S_c \end{aligned}$$

$$> 0 \quad (\text{A.12})$$

where the last inequality holds because the first term is necessarily positive and the second is positive by lemma 1.

To prove  $t_c^*$  exceeds  $t_c^{**}$ , we write the objective function from game 4 in terms of the game 3 objective function:

$$\begin{aligned} \Pi_4 &= \pi_3 - Z_A(T-t_C)(1-Z_C(t_C))(1-P_A)R_{BC} - Z_C(t_C)P_C R_C \\ &- Z_A(T-t_C)Z_C(t_C)P_B R_B + Z_A(T-t_C)(1-Z_C(t_C))(1-P_A)S_{BC} \\ &+ Z_A(T-t_C)Z_C(t_C)P_B S_B + Z_C(t_C)P_C S_C \end{aligned} \quad (\text{A.13})$$

Taking the derivative of  $\Pi_4$  and evaluating at the solution to  $\pi_3$  yields the following:

$$\begin{aligned} \Delta_{4,3} &= [-Z'_A(T-t_C)(1-Z_C(t_C)) - Z'_C(t_C)Z_A(T-t_C)](1-P_A)S_{BC} \\ &+ [-Z'_A(T-t_C)Z_C(t_C) + Z'_C(t_C)Z_A(T-t_C)]P_B S_B \\ &+ Z'_C(t_C)P_C S_C. \end{aligned} \quad (\text{A.14})$$

Simplifying and proceeding as above, we have the following:

$$\begin{aligned} \Delta_{4,3} &= -Z'_A(T-t_C)(1-P_A)S_{BC} \\ &+ [-Z'_A(T-t_C)Z_C(t_C) + Z'_C(t_C)Z_A(T-t_C)][P_B S_B - (1-P_A)S_{BC}] \\ &\geq -Z'_A(T-t_C)P_B S_B - Z'_A(T-t_C)P_C S_C + Z'_C(t_C)P_C S_C \\ &+ [Z'_A(T-t_C)Z_C(t_C) - Z'_C(t_C)Z_A(T-t_C)]P_C S_C \\ &= -Z'_A(T-t_C)P_B S_B \\ &+ [-Z'_A(T-t_C)(1-Z_C(t_C)) + Z'_C(t_C)(1-Z_A(T-t_C))]P_C S_C \\ &> 0 \end{aligned} \quad (\text{A.15})$$

where the last inequality results from the fact that the first term is necessarily positive and the second is positive by lemma 1.

To show  $t_c^{**}$  exceeds  $t_c^{*}$ , we write  $\Pi_2$  as a function of  $\pi_5$ :

$$\Pi_2 = \pi_5 + Z_A(T - t_C)(1 - Z_C(t_C))(1 - P_A)(S_{BC} - R_{BC}) + Z_C(t_C)P_C(S_C - R_C). \quad (\text{A.16})$$

Evaluating the derivative of A. 16 at the solution to game 5 and proceeding as above, we have the following:

$$\begin{aligned} \Delta_{2,5} &= [-Z'_A(T - t_C)(1 - Z_C(t_C)) - Z_C(t_C)Z'_A(T - t_C)](1 - P_A)S_{BC} \\ &\quad + Z'_C(t_C)P_C S_C \\ &= [-Z'_A(T - t_C)(1 - Z_C(t_C)) - Z'_C(t_C)Z_A(T - t_C)]P_B S_B \\ &\quad + [-Z'_A(T - t_C)(1 - Z_C(t_C)) + Z'_C(t_C)(1 - Z_A(T - t_C))]P_C S_C \\ &> 0 \end{aligned} \quad (\text{A.17})$$

where the last inequality follows from lemma 1.

To show  $t_c^{**}$  exceeds  $t_c^{*}$  we write  $\Pi_4$  as a function of  $\pi_5$ :

$$\begin{aligned} \Pi_4 &= \pi_5 + Z_A(T - t_C)(1 - Z_C(t_C))(1 - P_A)(S_{BC} - R_{BC}) \\ &\quad + Z_C(t_C)P_C(S_C - R_C) + Z_A(T - t_C)Z_C(t_C)P_B S_B. \end{aligned} \quad (\text{A.18})$$

Evaluating the derivative of equation A.18 at the solution to game 5 and proceeding as above we have the following:

$$\begin{aligned} \Delta_{4,5} &= [-Z'_A(T - t_C)(1 - Z_C(t_C)) - Z'_C(t_C)Z_A(T - t_C)](1 - P_A)S_{BC} \\ &\quad + [-Z'_A(T - t_C)Z_C(t_C) + Z'_C(t_C)Z_A(T - t_C)]P_B S_B + Z'_C(t_C)P_C S_C \end{aligned}$$

$$= [-Z'_A(T-t_c)(1-Z_c(t_c)) + Z'_c(t_c)(1-Z_A(T-t_c))]P_cS_c \\ -Z'_A(T-t_c)P_bS_b > 0 \quad (\text{A.19})$$

where the last inequality follows from lemma 1.

We begin by writing the objective function for decision problem 1 in terms of the game 5 objective function:

$$\begin{aligned} \pi_1 &= \pi_s + (1 - Z_A(T-t_C))(1 - Z_C(t_C))R_{ABC} \\ &+ (1 - Z_A(T-t_C))Z_C(t_C)(1 - P_C)R_{AB} \\ &+ Z_A(T-t_C)Z_C(t_C)P_B R_B. \end{aligned} \quad (\text{A.20})$$

The first order condition for decision problem 1 evaluated at the solution to game 5 is given by the following:

$$\begin{aligned} \Delta_{1,S} &= [Z'_A(T-t_c)(1-Z_c(t_c)) - Z'_c(t_c)(1-Z_A(T-t_c))]R_{ABC} \\ &+ [Z'_A(T-t_c)Z_c(t_c) + Z'_c(t_c)(1-Z_A(T-t_c))]R_{AB} \\ &+ [-Z'_A(T-t_c)Z_c(t_c) + Z'_c(t_c)Z_A(T-t_c)]P_b R_B \\ &= Z'_A(T-t_c)(1-Z_c(t_c))R_{ABC} \\ &+ Z'_A(T-t_c)Z_c(t_c)[(1-P_c)R_{AB} - P_b R_B] \\ &+ Z'_c(t_c)(1-Z_A(T-t_c))[(1-P_c)R_{AB} - R_{ABC}] \\ &+ Z'_c(t_c)Z_A(T-t_c)P_b R_B \\ &= Z'_A(T-t_c)(1-Z_c(t_c))R_{ABC} + Z'_A(T-t_c)Z_c(t_c)P_A R_A \\ &- Z'_c(t_c)(1-Z_A(T-t_c))P_c R_C \\ &+ Z'_c(t_c)Z_A(T-t_c)P_b R_B \end{aligned} \quad (\text{A.21})$$

where the last equality follows from the fact that information is not socially valuable. When information has no social value, the first order conditions for game 5 simplify to the following equation :

$$\begin{aligned}
 & Z'_A(T-t_c)(1-Z_c(t_c))(1-P_A)R_{BC} \\
 = & Z'_c(t_c)(1-Z_A(t_c))P_C R_c \\
 & -Z'_c(t_c)Z_A(T-t_c)P_\delta R_\delta. \tag{A.22}
 \end{aligned}$$

Combining A.22 with A.21 we obtain the following:

$$\begin{aligned}
 \Delta_{1,5} &= Z'_A(T-t_c)(1-Z_c(t_c))R_{ABC} + Z'_A(T-t_c)Z_c(t_c)P_A R_A \\
 &\quad - Z'_A(T-t_c)(1-Z_c(t_c))(1-P_A)R_{BC} \\
 &= Z'_A(T-t_c)P_A R_A \\
 &< 0 \tag{A.23}
 \end{aligned}$$

where the final inequality results from the negativity of  $R_A$ .

To show  $t_r^*$  exceeds  $t_c^*$ , we write  $\pi_1$  as a function of  $\pi_3$ :

$$\begin{aligned}
 \pi_1 &= \pi_3 + (1 - Z_A(T-t_C))(1 - Z_C(t_C))R_{ABC} \\
 &\quad + (1 - Z_A(T-t_C))Z_C(t_C)(1 - P_C)R_{AB}. \tag{A.24}
 \end{aligned}$$

Evaluating the derivative of equation A.24 at the solution to game 3 and simplifying, we have the following:

$$\begin{aligned}
 \Delta_{1,3} &= [Z'_A(T-t_c)(1-Z_c(t_c))-Z'_c(t_c)(1-Z_A(T-t_c))]R_{ABC} \\
 &\quad + [Z'_A(T-t_c)Z_c(t_c)+Z'_c(t_c)(1-Z_A(T-t_c))](1-P_C)R_{AB}
 \end{aligned}$$

$$\begin{aligned} &= Z_A'(T - t_C)P_A R_A \\ &\wedge 0. \end{aligned} \tag{A.25}$$

*This concludes the proof of proposition 3.*

## Appendix B: Second Order Conditions

This appendix details the implications of the concavity restrictions we impose on the objective functions in section 5. We begin by considering decision problem 1. It can easily be shown that the sufficient condition for interior maximization is given by  $2L_{12} - L_{11} - L_{22} > 0$ , where  $L_{ij}$  for  $i,j = 1,2$  denotes the second partial of the constrained optimization problem with respect to arguments  $i$  and  $j$ . In the context of decision problem 1, this condition is given by the following statement:

$$\begin{aligned}
 & 2Z'_A(t_A)Z'_C(t_C)[R_{ABC} - (1-P_A)R_{BC} + P_B R_B - (1-P_C)R_{AB}] \\
 - & Z''_A(t_A)(1-Z_C(t_C))[(1-P_A)R_{BC} - R_{ABC}] \\
 - & Z''_A(t_A)Z_C(t_C)[P_B R_B - (1-P_C)R_{AB}] \\
 - & Z''_C(t_C)(1-Z_A(t_A))[(1-P_C)R_{AB} - R_{ABC}] \\
 - & Z''_C(t_C)Z_A(t_A)[P_B R_B - (1-P_A)R_{BC}] - Z''_C(t_C)P_C R_C > 0. \tag{B.1}
 \end{aligned}$$

We assume that condition B.1 is always satisfied.

The sufficient condition for game 4 can be expressed by the following statement:

$$\begin{aligned}
 & 2Z'_A(t_A)Z'_C(t_C)[P_B R_B - (1-P_A)S_{BC}] \\
 - & Z''_A(t_A)[(1-Z_C(t_C))(1-P_A)S_{BC} + Z_C(t_C)P_B S_B] \\
 - & Z''_C(t_C)[P_C S_C + Z_A(t_A)P_B S_B - Z_A(t_A)(1-P_A)S_{BC}] > 0. \tag{B.2}
 \end{aligned}$$

We assume that condition B.2 is satisfied.

The sufficient condition for decision problem 2 can be expressed by the following statement:

- $2Z'_A(t_A)Z'_C(t_C)(1-P_A)S_{BC}$
- $Z''_A(t_A)(1-Z_C(t_C))(1-P_A)S_{BC}$
- $Z''_C(t_C)[P_C S_C - Z_A(t_A)(1-P_A)S_{BC}] > 0.$  (B.3)

Notice that the left-hand side of condition B.3 exceeds the left-hand side of condition B.2 everywhere. This implies that the former will be satisfied whenever the latter holds. We therefore do not need to assume concavity for decision problem 2 since it is guaranteed by condition B.2.

The sufficient condition for game 5 can be expressed as the following statement:

- $2Z'_A(t_A)Z'_C(t_C)(1-P_A)R_{BC}$
- $Z''_A(t_A)(1-Z_C(t_C))(1-P_A)R_{BC}$
- $Z''_C(t_C)[P_C R_C - Z_A(t_A)(1-P_A)R_{BC}] > 0.$  (B.4)

We assume that condition B.4 is always satisfied.

The sufficient condition for game 3 can be expressed as the following statement:

- $2Z'_A(t_A)Z'_C(t_C)[(1-P_A)R_{BC} - P_B R_B]$
- $Z''_A(t_A)[(1-Z_C(t_C))(1-P_A)R_{BC} + Z_C(t_C)P_B R_B]$
- $Z''_C(t_C)[Z_A(t_A)[P_B R_B - (1-P_A)R_{BC}] + P_C R_C] > 0$  (B.5)

Notice again that the left-hand side of condition B.5 exceeds the left-hand side of condition B.4 everywhere. Again this implies that the former will be satisfied whenever the later holds. We therefore do not need to assume concavity for game 3 since it is guaranteed by condition B.4.

This appendix has shown that only three of the decision problems and games considered require a concavity assumption.

## REFERENCES

- Blue Ribbon Defense Panel 1970 *Report to the President and the Secretary of Defense on the Department of Defense*. Washington, D.C.: U.S. Government Printing Office.
- Crawford, Vincent P., and Joel Sobel 1982 Strategic information transmission. *Econometrica* 50(6):1431-1451.
- Green, Jerry R., and Nancy L. Stokey 1980 A Two-Person Game of Information Transmission. Harvard Institute of Economic Research Discussion Paper Number 751.
- Hirshleifer, J., and J.G. Riley 1992 *The Analytics of Information and Uncertainty*. New York: Cambridge University Press.
- Jovanovic, B. 1982 Truthful disclosure of information. *Bell Journal of Economics* 13:36-44.
- Kofman, Fred, and Jacques Lawarrée 1993 Collusion in hierarchical agency. *Econometrica* 61(3):629-656.
- Kreps, D.M., and R. Wilson 1982 Sequential equilibria. *Econometrica* 50:863-894.
- Matthews, Steven, and Andrew Postlewaite 1985 Quality testing and disclosure. *RAND Journal of Economics* 16(3):328-340.
- Milgrom, P.R. 1981 Good news and bad news: Representation theorems and applications. *Bell Journal of Economics* 12:380-391.
- National Research Council 1995 *Statistical Methods for Testing and Evaluating Defense Systems: Interim Report*. Panel on Statistical Methods for Testing and Evaluating Defense Systems, Committee on National Statistics. Washington, D.C.: National Academy Press.

- Shavell, Steven 1994 Acquisition and disclosure of information prior to sale. *RAND Journal of Economics* 25(1):20-36.
- U.S. General Accounting Office 1979a *The NAVSTAR Global Positioning System—A Program With Many Uncertainties*. Washington, D.C.: U.S. Government Printing Office.
- 1979b *Need for More Accurate Weapon System Test Results to Be Reported to the Congress*. Washington, D.C.: U.S. Government Printing Office.
- 1980 *DoD Information Provided to the Congress on Major Weapon Systems Could Be More Complete and Useful*. Washington, D.C.: U.S. Government Printing Office.
- 1983 *The Army Should Confirm Sergeant York Air Defense Gun's Reliability and Maintainability Before Exercising Next Production Option*. Washington, D.C.: U.S. Government Printing Office.
- 1987 *Testing Oversight*. Washington, D.C.: U.S. Government Printing Office.
- 1988a *Aquila Remotely Piloted Vehicle: Its Potential Battlefield Contribution Still in Doubt*. Washington, D.C.: U.S. Government Printing Office.
- 1988b *Quality of DoD Operational Testing and Reporting*. Washington, D.C.: U.S. Government Printing Office.
- 1989 *Electronic Warfare: Reliable Equipment Needed to Test Air Force's Electronic Warfare Systems*. Washington, D.C.: U.S. Government Printing Office.
- 1990a *Army Acquisition: Air Defense Antitank System Did Not Meet Operational Test Objectives*. Washington, D.C.: U.S. Government Printing Office.
- 1990b *Naval Aviation: The V-22 Osprey—Progress and Problems*. Washington, D.C.: U.S. Government Printing Office.
- 1991a *Army Acquisition: Air Defense Antitank System's Development Goals Not Yet Achieved*. Washington, D.C.: U.S. Government Printing Office.
- 1991b *Electronic Warfare: Faulty Test Equipment Impairs Navy Readiness*. Washington, D.C.: U.S. Government Printing Office.

1991 c *Electronic Warfare: No Air Force Follow-up on Test Equipment Inadequacies*. Washington, D.C.: U.S. Government Printing Office.  
1992 *Electronic Warfare: Established Criteria Not Met for Airborne Self-Protection Jammer Production*. Washington, D.C.: U.S. Government Printing Office.

## 2

# On the Performance of Weibull Life Tests Based on Exponential Life Testing Designs

*Francisco J. Samaniego and Yun Sam Chong, University of California, Davis*

## 1. EXPONENTIAL LIFE TESTING

Applications abound in which investigators seek to make inferences about the lifetime characteristics of a "system" of interest from data on the failure times of prototypical systems placed on test. There are a good many different experimental designs that might be considered in planning a given life testing application; often, some form of data censoring (aimed at bounding the experiment's duration) or some sequential procedure (aimed at possibly resolving the test based on early failures) are part of the test plan. The analysis of life testing data is usually preceded by the setting of assumptions regarding the underlying probability distribution of system lifetimes. Among the most studied parametric life testing models are the exponential, gamma, Weibull, Pareto and lognormal families (see Lawless, 1982); nonparametric analyses under various assumptions on the distribution's hazard function or residual lifetime characteristics have also been developed (see Barlow and Proschan, 1975; Hollander and Proschan, 1984).

By far, the most comprehensive development of exact statistical procedures in life testing has occurred under the assumption of exponentiality. For virtually all other assumed models, the analysis of failure time data involves extensive use of numerical optimization methods and asymptotic approximations. The exact performance of tests and estimates developed under nonexponential assumptions has, for the most part, resisted analytical treatment, and has thus been studied mostly via simulation. The temptation to use exponential life testing methods is no doubt due, in part, to the marked lack of success in dealing with the theoretical properties of nonexponential life testing in a definitive way. The ease with which relevant distribution theory

(especially that involving ordered failure times) can be produced, and the occasional "conservatism" of the exponential assumption, have also contributed to its popularity, in spite of its notorious nonrobustness. It is important to acknowledge that the exponential assumption is very special and highly restrictive, so that its use should be discouraged except in circumstances in which there is good physical, empirical and practical support for the model. In due course, we will review the basics of exponential life testing, both to make the present paper self-contained and to set the stage for the various comparisons we wish to make with alternative analyses. First, however, we will describe the type of problem—a sort of statistical hybrid—on which the present investigation is focused.

Suppose a statistician is faced with an application in which two hypotheses concerning the mean life  $\mu$  of a new system are to be tested. He wishes to resolve the test of  $H_0: \mu = \mu_0$  against the alternative  $H_1: \mu = \mu_1$ , where  $\mu_1 < \mu_0$  are fixed and known, with certain predetermined probabilities  $\alpha$  and  $\beta$  for type I and type II errors (also often called the producer's and consumer's risks). Having no pressing reason to doubt exponentiality in the application at hand, the statistician determines (using the Department of Defense's *Handbook H108*, for example; U.S. Department of Defense, 1960) that these goals can be accomplished with an experimental design calling for some specific number of observed failures (say  $r$ ), rejecting  $H_0$  in favor of  $H_1$  if the total time on test  $T$  at the time of the  $r$ th failure is less than the threshold  $T_0$ . Among the advantages afforded by an exponential life test plan is the fact the the resources required to perform the test (that is, the number of systems that must be placed on test and the maximum amount of testing time needed to resolve the test) may be calculated in advance. The fact that the duration of the test, in real time, can be controlled and made suitably small by placing  $n > r$  systems on test while still resolving the test upon the  $r$ th failure is also an important advantage.

Consider, now, the analysis stage of this life testing experiment. Suppose that when the data have been collected, their characteristics suggest that they are definitely not exponential. It then falls upon the statistician to analyze the available data under some alternative model or, perhaps, nonparametrically. Let us suppose, as will be tacitly assumed in the sequel, that the two-parameter Weibull distribution is taken as an appropriate underlying model for the

experiment in question. It is then incumbent upon the statistician to test the means  $\mu_0$  vs  $\mu_1$  under the Weibull assumption. The goal of this paper is to examine the consequences of this paradigm shift. We will study the resultant error probabilities associated with the Weibull test, and will explore the potential that exists for resource savings (smaller sample sizes, less testing time) when the Weibull model is entertained during the design stage rather than only at the analysis stage of the experiment. Our study has enabled us to identify the circumstances under which rather substantial resource savings are possible. For a study which examines similar questions in the contrast of interval estimation, see Woods (1996).

We now turn to a brief description of the mechanics of exponential life testing. Let us first suppose that a sample  $X_1, \dots, X_r$  of system lifetimes is available for observation, and that these data are independent and identically distributed according to the exponential distribution  $Exp(\theta)$  with density function

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0. \quad (1.1)$$

For short, we will write  $X_1, \dots, X_r \stackrel{iid}{\sim} Exp(\theta)$ . The statistic  $T_{rr} = \sum_{i=1}^r X_i$ , which may be described as the "total time on test" for the  $r$  systems taken together, is a sufficient statistic for  $\theta$  and is distributed according to the gamma distribution  $\Gamma(r, \theta)$  with density function

$$f(t) = \frac{1}{\Gamma(r)\theta^r} t^{r-1} e^{-t/\theta}, \quad t > 0. \quad (1.2)$$

We use the subscript  $r : r$  to reflect the fact that the experimental design calls for sampling  $r$  failure times out of a random sample of size  $r$ . The standard estimate of  $\theta$  based on  $Tr:r$  is the sample mean

$$\hat{\theta}_{rr} = T_{rr} / r, \quad (1.3)$$

which is both the maximum likelihood estimate and the minimum variance unbiased estimate of  $\theta$ . For any fixed  $\alpha \in (0, 1)$ , the best test of size  $\alpha$  for testing  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta = \theta_1 < \theta_0$  is the test which rejects  $H_0$  if and only if  $\hat{\theta}_{rr} < c$ , where  $c$  is determined by the equation

$$P(\hat{\theta}_{rr} < c | \theta = \theta_0) = \alpha. \quad (1.4)$$

Since, given  $\theta = \theta_0$ ,  $2T/\theta_0$  is distributed as  $\chi^2_{2r}$  variable, it is clear that the threshold for rejection is given by  $c = (\theta_0/2r)\chi^2_{2r,1-\alpha}$ , where  $\chi^2_{2r}$  is such that  $P(X > \chi^2_{2r}) = \alpha$  when  $X \sim \chi^2_{2r}$ . The test which rejects  $H_0$  when

$$\hat{\theta}_{rr} < \frac{\theta_0}{2r} \chi^2_{2r,1-\alpha} \quad (1.5)$$

is, in fact, uniformly most powerful for testing against  $H_1 : \theta < \theta_0$ , and, in particular, maximizes the power (or minimizes the "consumer's risk"  $\beta$ ) at the alternative  $\theta = \theta_1$ . If we assume that the levels of  $\alpha$  and  $\beta$  are fixed and determined in advance, then it remains to find the sample size  $r$  for which these levels obtain. Since  $r$  must satisfy the equation

$$P(\hat{\theta}_{rr} \geq \frac{\theta_0}{2r} \chi^2_{2r,1-\alpha} | \theta = \theta_1) = \beta, \quad (1.6)$$

it follows that the required sample size is the smallest integer  $r = r_0$  for which

$$\chi^2_{2r,1-\alpha} \geq \frac{\theta_1}{\theta_0} \chi^2_{2r,\beta}. \quad (1.7)$$

The fact that the required sample size  $r_0$  is completely determined by the values of  $\alpha$ ,  $\beta$  and the "discrimination ratio"  $\theta_1/\theta_0$  is a special feature of exponential life testing that facilitates the automated application of this methodology. Once a sample size  $r = r_0$  is obtained through (1.7), the rejection threshold  $c$  in (1.4) may be represented as

$$c = \frac{\theta_0}{2r_0} \chi^2_{2r_0,1-\alpha}. \quad (1.8)$$

The constant  $c/\theta_0$ , which is independent of model parameters, will appear in several of our tabulations as the multiplier which, together with the  $\theta_0$  of interest, determines the rejection threshold of the desired test.

Execution of the exponential life test above is perhaps most easily described in terms of the "total time on test" (TTT) function. If  $X_{(1)} < \dots < X_{(r)}$  are the ordered failure times in our sample of size  $r$ , then the TTT function may be written, for  $t \in [X_{(j)}, X_{(j+1)}]$  as

$$T_{rr}(t) = \sum_{i=1}^j X_{(i)} + (r-j)t, \quad (1.9)$$

where  $X_{(0)} = 0$  and  $j = 0, 1, \dots, r - 1$ . The TTT function keeps track of the total amount of test time logged by working systems up to a fixed time  $t$ . Clearly

$$T_{r,r}(X_{(r)}) = \sum_1^r X_{(j)}. \quad (1.10)$$

The TTT function is itself a useful tool in reliability modeling. Plots involving a rescaled version of this function will be discussed in the next section.

Returning to the problem of testing  $H_0: \theta = \theta_0$  vs  $H_1: \theta = \theta_1$ , we note that the test may be resolved as follows: if the  $r$ th failure occurs before the total time on test exceeds the threshold  $r_0c$ , that is, if  $T_{r,r}(X_{(r)}) < r_0c$ , then  $H_0$  is rejected in favor of  $H_1$ ; otherwise,  $H_0$  is accepted. In the latter case, the experiment is completed at time  $t_0$ , where

$$T_{r,r}(t_0) = r_0c,$$

while in the former case, the experiment is terminated at time  $t = X_{(r)} < t_0$ . Thus the threshold  $r_0c$ , with  $c$  given in (1.8), represents the maximum total test time that could be required to resolve the test, that is, to be able to accept or reject  $H_0$  on the basis of the data. Together,  $r_0$  and  $c$  describe the total resources that must be committed to guarantee successful completion of the life test.

Extension of the above discussion to type II censored data is immediate. If  $|X_1, \dots, X_n \sim \text{Exp}(\theta)|$ , and if the experiment is terminated upon the occurrence of the  $r$ th failure, then the statistic

$$T_{r,n} = \sum_1^r X_{(j)} + (n-r)X_{(r)}, \quad (1.11)$$

is sufficient for  $\theta$ . Moreover, since  $T_{r,n}$  has precisely the same distribution as  $T_{r,r}$ , that is, since

$$T_{r,n} \sim \Gamma(r, \theta),$$

the best test of  $H_0: \theta = \theta_0$  vs  $H_1: \theta = \theta_1$  has the same form as before, that is, rejects  $H_0$  if

$$\hat{\theta}_{r,n} < c, \quad (1.12)$$

where

$$\hat{\theta}_{r,n} = T_{r,n} / r \quad (1.13)$$

is the MLE (and UMVUE) of  $\theta$ .

Similarly, the sample size required to resolve this test, given set values for  $\alpha$ ,  $\beta$  and  $\theta_1/\theta_0$ , is  $r_0$  derived via (1.7), and the maximum total testing time needed is again the constant  $r_0c$ , where  $c$  is given in (1.8). The number  $n$  of systems on test influences test performance only with regard to the test's duration. Let us expand the definition of the total time *on test* function to accommodate the case of type II censoring; for  $t \in [X_{j\cdot}, X_{(j+1)}], j = 0, \dots, r - 1$ , define

$$T_{r,n}(t) = \sum_{i=1}^j X_{(i)} + (n-j)t. \quad (1.14)$$

Then, under type II censoring, the experiment is terminated at time  $t = X_{(r)}$  if  $T_{r,n}(X_{(r)}) \leq r_0c$  or, otherwise, at time  $t = t_0$ , where  $t_0$  satisfies the equation

$$T_{r,n}(t_0) = r_0c.$$

It is easy to see that the random time  $\min(X_{(r)}, t_0)$  at which the experiment is terminated is bounded above by the factor  $r_0c/(n-r_0)$ . Thus, the waiting time until the test is completed can be made suitably small for any fixed  $r_0$  by choosing the sample size  $n$  sufficiently large. This strategy of course is based on a tacit assumption of the correctness of the exponential model in the application of interest; when exponentiality fails, this practice can yield highly misleading results.

There are a host of other experimental designs for exponential life testing, including type I censoring (that is, censoring at a fixed time  $t$ ), random record designs (that is, observing only record breaking failure times) and sequential designs. The type of study which will be pursued in this paper can be carried out analogously for other designs, but we have chosen to focus exclusively on complete and type II censored data. This choice is motivated by the fact that these two designs are frequently encountered in practice and also by our belief that the general lessons learned from analyzing these particular designs will hold more broadly. For example, the distribution theory developed in Samaniego and Whittaker (1986) shows that inverse sampling from an exponential distribution until the occurrence of the  $r$ th record value (that is, successive minimum) yields a test statistic (again, the total time on test) that has properties identical to those of the designs mentioned above. In particular, the resources required to resolve testing problems for predetermined values of  $\alpha$ ,  $\beta$ ,  $\theta_0$  are again given by the pair  $(r_0, c)$  of (1.7) and (1.8).

Instead of pursuing greater breadth in the designs considered, we will direct our efforts at examining two particular designs (complete samples and type II censoring) in depth.

As a guide for military applications of exponential life testing, DoD *Handbook H108* provides tabled values of the required sample size  $r_0$  and the constant  $c/\theta_0$  through which the total test time required by a particular application can be computed. An excerpt from Table 2B-5 of that *Handbook*, showing the five tabled values given corresponding to error probabilities  $\alpha = .1$  and  $\beta = 1$ , appears in [Table 1](#). If, for instance, one wishes to test  $H_0: \theta = 1,000$  hrs vs  $H_1: \theta = 500$  hrs, and one sets  $\alpha = .1 = \beta$ , then [Table 1](#) indicates that 15 or more systems should be put on test, and that a total test time required to ensure resolution of the test is  $15(.687)(1,000) = 10,305$  hrs.

Before proceeding with our study of alternatives to exponential life tests, we briefly review what is known about their lack of robustness. Of special interest to us is the behavior of exponential life tests when the underlying distribution is a nonexponential Weibull, since it is then that the procedures we investigate in the sequel stand to provide improved performance. We thus restrict ourselves to this particular circumstance and describe the findings of Zelen and Dannemiller (1961), who studied the performance of exponential life tests for Weibull data in exhaustive detail. In that paper, four specific life testing designs were studied: complete samples, type II censored samples, truncated type II censored samples, and samples obtained sequentially. We quote from Zelen and Dannemiller's discussion section:

None of the four life testing procedures studied in this paper is robust with respect to Weibull alternatives. In particular, the censored life test and the truncated nonreplacement test are strikingly non-robust. It is obvious from the graphs of the O.C. curves that lots having low mean failure time have a high probability of acceptance when the failure times follow a Weibull distribution with shape parameter  $p > 1$ . This tendency is increased as  $p$  increases.... We have tried to show that dogmatic use of life testing procedures without a careful verification of the assumption that failure times follow the exponential distribution may result in a high probability of accepting "poor quality" equipment.

In the case of complete and type II censored samples, the operating characteristics plotted by Zelen and Dannemiller indicate the extent to which the risk of a high probability of acceptance of a hypothesized mean of 1,000 occurs at mean values less than 1,000.

The performance of the exponential test of  $H_0: \theta = 1,000$  vs  $H_1: \theta = 500$  at the nominal values  $\alpha = .1 = \beta$  is shown there to deteriorate as the Weibull shape parameter increases from 1 to 3. It is interesting to note that at  $\theta = 500$  and  $\theta = 1,000$ , the probabilities  $\alpha$  and  $\beta$  of error actually decrease in the complete sample setting; this is a manifestation of the conservative nature of these tests. Since Weibull distributions with shape parameter greater than 1 are lighter tailed than the exponential, these distributions are more tightly concentrated about the mean, rendering it easier to distinguish between two candidate mean values on the basis of a Weibull sample. For complete samples, the nonrobustness of which Zelen and Dannemiller write becomes evident as the mean value at which the probability of accepting  $H_0$  is being computed moves toward the null value 1,000 from the alternative value of 500. At  $\theta = 750$ , for example, the probability of accepting  $H_0: \theta = 1,000$  goes from .615 under exponentiality to .837 under a Weibull distribution with shape parameter equal to 3. In spite of this type of inflation, it is clear that exponential life tests carried out with complete samples offer reasonable performance in that even under rather severe departures of the Weibull type, they deliver error probabilities at selected key parameter values  $\theta_0$  and  $\theta_1$  that are smaller than those set at the planning stage. The question that will interest us as we proceed is: since the achieved  $\alpha$  and  $\beta$  levels are both lower than planned for or required, what savings might be possible with a test that is calibrated to achieve the nominal values of  $\alpha$  and  $\beta$  when the data are Weibull?

The case of censored samples is markedly different from the above. In an example involving  $n = 28$  systems on test with censoring at the 14th failure, Zelen and Dannemiller note that the probability of acceptance of  $H_0: \theta = 1,000$  is exceedingly high for all potential mean values between the alternative 500 and the null 1,000. Remarkably, the probability of accepting the null hypothesis of mean 1,000 when the true mean is 500 is .985 when the sample is drawn

from a Weibull distribution with shape parameter 3. Even when the shape parameter is 1.5, this probability is unduly high (.463).

The lessons to be learned from the phenomena documented above include (1) exponential life testing based on complete samples works fairly well in a Weibull environment, but there should be opportunities for saving resources when that environment is recognized in advance; and (2) exponential life testing based on censored samples works very poorly in a Weibull environment, and alternative procedures should be considered when the exponential assumption is suspect. The sequel is largely devoted to the study of ways of addressing these two issues.

Before proceeding, let us make special mention of the scope of this paper, and its attendant limitations. We have begun by discussing exponential life testing based on complete or type II censored samples. In sections 3 and 4, we will develop a comparable analysis under the assumption that the underlying distribution of the observable failure time data is, instead, a nonexponential Weibull. Both analyses assume that it is a random sample of items simultaneously and independently placed on test. Because of the memoryless property of the exponential distribution, exponential life testing methods can be validly applied (assuming the model is appropriate) to data on time between failures of repairable systems by treating time between failures as independent exponential observations. Such an extension will not generally be valid under Weibull assumptions. In the latter case, the alternative analysis developed in this paper would be applicable only when each repair following an observed failure could reasonably be considered "perfect" in the sense of restoring the item to its condition when new. When such an assumption cannot be justified, the appropriate reanalysis of data should be based on a more elaborate modeling of the failure process, perhaps as a nonhomogeneous Poisson process. Nonparametric alternatives in this setting have been developed by Nelson (1995) and by Lawless and Nadeau (1995) and have been shown to work very well in a variety of applications (without the restrictive NHPP and independence assumptions). Such analyses lie beyond the scope of the present paper.

Other issues not covered in the present report include the treatment of systems with multiple failure modes and the treatment of accelerated life testing data. Parallel developments

in those areas, where Weibull alternatives to exponentiality assumptions are developed, would certainly be worthwhile.

## 2. WEIBULL CONSIDERATIONS

The Weibull distribution is arguably the most popular parametric alternative to the exponential distribution in reliability applications. Like the gamma model, it contains the exponential distribution as a special case, so that the adoption of a Weibull assumption represents a broadening from the exponential model rather than a rejection of it. Often, statistical extreme value theory forms the basis for the applicability of the Weibull model; when system failure can be attributed to the failure of the weakest of its many components, the Weibull model will tend to describe failure data quite well. The parametrization we will employ for the Weibull is as follows:  $X$  has a Weibull distribution with parameters  $A > 0$ ,  $B > 0$  (henceforth denoted as  $X \sim W(A, B)$ ) if  $X$  has distribution function

$$F(x) = P(X \leq x) = 1 - e^{-\frac{x^A}{B}}, \quad x > 0 \quad (2.1)$$

and density function

$$f(x) = \frac{A}{B} x^{A-1} e^{-\frac{x^A}{B}}, \quad x > 0, \quad (2.2)$$

where  $A$  is the "shape" parameter and  $B^{1/A}$  the scale parameter of the distribution. The mean and variance of  $X \sim W(A, B)$  can be written as:

$$\mu = \Gamma\left(\frac{A+1}{A}\right) B^{1/A} \quad (2.3)$$

and

$$\sigma^2 = B^{2/A} \left[ \Gamma\left(\frac{A+2}{A}\right) - \Gamma^2\left(\frac{A+1}{A}\right) \right]. \quad (2.4)$$

The coefficient of variation  $cv = \sigma/\mu$  is independent of the parameter  $B$  and may be written as

$$cv = \frac{\left[ \Gamma\left(\frac{A+2}{A}\right) - \Gamma^2 \left(\frac{A+1}{A}\right) \right]^{1/2}}{\Gamma\left(\frac{A+1}{A}\right)}. \quad (2.5)$$

It is apparent from (2.2) that the  $W(1,B)$  distribution is simply the exponential distribution  $Exp(B)$ . A more interesting and valuable connection between the Weibull and exponential models is the fact that if  $X \sim W(A,B)$ , then  $X^A \sim Exp(B)$ .

There is a rather substantial literature on modeling and inference involving the Weibull distribution. A keyword search of the *Current Index to Statistics*, volumes 1-19 (American Statistical Association, 1975 to 1995), shows that there were 647 articles published in statistics journals between 1975 and 1993 on Weibull-related topics. Much of this literature deals with estimation issues, with goodness of fit questions, with separate families tests (for example, testing Gamma vs Weibull) or with robustness issues. Good overviews on estimation and testing procedures may be found in the recent books by Lawless (1982), Sinha (1987) and Bain and Engelhardt (1991). Other references with extensive discussion of inference for the Weibull distribution include Mann, Shafer and Singpurwalla (1972), Sinha and Kale (1979) and Nelson (1982 and 1990).

Of particular interest to us are testing procedures which seek to distinguish between two mutually exclusive collections of Weibull models. In the sequel, we will examine and compare various approaches to testing competing hypotheses about a Weibull mean. The literature on this latter problem is rather sparse. When the shape parameter is assumed known, the test of interest can be executed easily after transforming the data into exponential variables. With the scale parameter known, Bain and Weeks (1965) developed tests and confidence intervals for the unknown shape parameter. For the general problem, when both  $A$  and  $B$  are unknown, there is rather limited guidance on how to proceed. Thoman, Bain and Antle (1969) have developed MLE-based confidence intervals for each parameter when the other parameter is unspecified. However, it is known that large sample methods based on the asymptotic behavior of maximum likelihood estimates behave rather poorly for small and moderate samples (see Lawless, 1975). Likelihood ratio tests for

$$H_0 : B = B_0 \text{ vs } H_1 : B \neq B_0,$$

$$H_0 : A = A_0 \text{ vs } H_1 : A \neq A_0, \text{ and}$$
$$H_0 : \xi(p) = \xi_0 \text{ vs } H_1 : \xi(p) \neq \xi_0,$$

where  $\xi(p)$  is the  $p$ th quantile of the underlying probability distribution, are discussed in Lawless (1982) and recommended as preferable to tests based on the large sample distributions of MLEs. Lawless (1982:195-197) discusses Weibull life test plans briefly, stating that "life test plans under the Weibull model have not been thoroughly investigated it is almost always impossible to determine exact small-sample properties or to make effective comparisons of plans, except by simulation.... Therefore, little formal discussion of the merits of different plans has taken place Further development of test plans under a Weibull model would be useful." It is our hope that the discussion of Weibull life testing in this paper will contribute to a better understanding of the possible advantages and risks these methods involve.

As we have described the problem of interest in the introductory section, the statistician, after collecting data under an exponential life test plan, takes the opportunity to reconsider his distributional assumptions. "Physics of failure" considerations might, in certain cases, point to an alternative model. In the case that Weibull alternatives to the exponential are considered sufficiently broad, one can carry out a formal test of the hypothesis  $H_0 : A = 1$  (that is,  $X$  is exponential) vs  $H_1 : A \neq 1$  (that is,  $X$  is nonexponential Weibull). Such a test is outlined in Thoman et al. (1969). More generally, there is a variety of existing goodness of fit tests through which an alternative model, Weibull or otherwise, might be identified. Attractive and usually quite effective alternatives to formal or analytical procedures are two widely used graphical methods: total time on test (TTT) plots and plots of transformed failure times on suitably chosen probability paper. We discuss these two methods below as possible tools in determining the viability of exponential assumptions against Weibull alternatives.

Total time on test plots were introduced by Barlow, Bartholomew, Bremner and Brunk (1972), and their properties have been studied further by Barlow and Campo (1975), Barlow (1979), Chandra and Singpurwalla (1981) and Neath and Samaniego (1992). Barlow, Toland and Freeman (1988) employ TTT plots in a large scale accelerated life testing experiment as a guide to appropriate modeling. Such plots are widely used as goodness of fit indicators for the

exponential distribution. In what follows, we will restrict attention to the total time on test function in (1.14) since the complete sample version in (1.9) is subsumed by (1.14) when  $n = r$ . We note that the TTT function in (1.14) has domain  $[0, X(r)]$  and range  $\left[0, \sum_{i=1}^r x_i^{(n)} + (n-r)x_r^{(n)}\right]$ . To render TTT plots both manageable and comparable, a rescaled version of the function is generally used. For an arbitrary positive random variable  $X$  with distribution  $F$  and finite mean  $\mu$ , the total time on test transform  $\tau$  is defined as

$$\tau(x) = \int_0^{\tau^{-1}(x)} \bar{F}(t) dt, \quad 0 \leq x \leq 1, \quad (2.6)$$

where  $F^{-1}(x) = \inf\{u | F(u) \geq x\}$ , and the survival function  $\bar{F}(t) = 1 - F(t)$ . As is well known,  $\tau(1) = \mu$ . The empirical counterpart  $\tau_n$  of  $\tau$  is obtained by replacing  $F$  in (2.6) by the empirical cdf  $F_n$ . If  $n$  items are placed on test, and the ordered lifetimes  $\{X_{(1)}, X_{(2)}, \dots\}$  are observed, then  $\tau_n$  may be evaluated at  $x = j/n$  as

$$\tau_n\left(\frac{j}{n}\right) = \frac{1}{n} \sum_{i=1}^j (n-i+1)(X_{(i)} - X_{(i-1)}). \quad (2.7)$$

Thus,  $n\tau_n(r/n)$  is precisely the cumulative survival time of all tested items at the time of the  $r$ th failure. The transform  $\tau_n$  is continuous and is linear for  $x \in \left[\frac{j}{n}, \frac{j+1}{n}\right]$  for each  $j$ . For complete samples, the function

$$\tau_r^*(x) = \frac{\tau_r(x)}{\tau_r(1)}$$

is plotted for  $x \in [0,1]$ , while for type II censored data, the function

$$\tau_{rn}^*\left(x \cdot \frac{n}{r}\right) = \frac{\tau_n(x)}{\tau_n(r/n)}$$

is plotted for  $x \in [0,1]$ . In both cases, the plots lie in the unit square. It is easy to verify that the TTT transform  $\tau$  of the exponential distribution is linear, and that the rescaled transform of the exponential is the diagonal line in the unit square.

The failure rate of a distribution is defined as  $f(t)/F(t)$ . The failure rate of the Weibull distribution  $W(A, B)$  is given by

$$r(t) = \frac{A}{B} t^{A-1}, \quad t > 0, \quad (2.8)$$

which is decreasing for  $A < 1$  and increasing for  $A > 1$ . It is known that the rescaled TTT transform of a distribution with increasing failure rate (IFR) is concave and that for a distribution with decreasing failure rate (DFR) is convex. Thus, the TTT plot based on a sample from a nonexponential Weibull might be expected to exhibit some nonlinearity—concavity when  $A > 1$  and convexity when  $A < 1$ . Plotting a scaled TTT transform for data collected according to an exponential life testing plan is an excellent way to detect possible departures from exponentiality. In Figures 1 to 6, we display the TTT plots from six consecutive simulated Weibull experiments, each featuring complete samples of size 20 from eight Weibull distributions with varying shape parameters. These figures give a feeling for the variability in TTT plots for a fixed value of  $A$ , and for the general character of the plots as  $A$  varies from 0.3 to 3.0.

From the TTT plots above, it should be evident that detecting departures from an exponentiality assumption is not an exact science. For example, five of the six simulated TTT plots of data drawn from the  $W(3.0, 1.0)$  distribution show convincing IFR behavior, while one, from the second simulation, is much less definitive. While a TTT plot may not be conclusive, it will often be quite suggestive of possible nonexponentiability; as such, it seems reasonable to suggest that such graphical investigations be a standard part of the analysis of life test data. For formal tests for exponentiality based on the TTT transform, see Barlow and Proschan (1969) and Klefsjö (1980).

It is, of course, true that a TTT plot does not point directly toward a Weibull alternative when it casts doubt upon the exponential assumption. Detecting IFRness or DFRness is a good start, but the checking for Weibullness requires more. Nelson (1990) and others advocate plotting life testing data on Weibull probability paper. A strong linear trend in such plots is indicative of an underlying Weibull distribution. These plots are based on the following considerations: The Weibull survival function is

$$\bar{F}(x) = e^{-\frac{x^A}{B}}, \quad t > 0,$$

and thus

$$\ln(-\ln \bar{F}(x)) = -\ln B + A \ln x. \quad (2.10)$$

Equation (2.10) is the basis for the expected linearity in a Weibull plot. The plot itself is simply a scatter diagram consisting of the points  $(\ln X_{(i)}, \ln(-\ln(1 - \frac{i-1}{r})))$   $i = 1, \dots, r$ . The parameters  $-B$  and  $A$  are generally estimated by the intercept and slope of the least squares line fitted to these points. The figures that follow show Weibull plots for eight simulated samples of size 20 from Weibull distributions with varying shape parameters. These plots (Figures 7 to 14) appear as scatter diagrams of the points  $(X_{(i)}, -\ln(1 - \frac{i-1}{r}))$  on log-log paper.

Through use of the graphical methods described above, or otherwise, assume that the statistician, after gathering data according to an exponential life test plan, determines that the data are more appropriately modeled as a nonexponential Weibull. It will then be necessary to proceed with an analysis appropriate for these broadened assumptions. The next two sections are dedicated to an examination of various ways of carrying out a Weibull life test.

### 3. WEIBULL LIFE TESTING—PART I

The classical theory of hypothesis testing yields its strongest results in problems in which the null and alternative hypotheses are simple, that is, specify the underlying probability model completely. In that circumstance, it is possible to construct tests that will minimize the consumer's risk  $\beta$  among tests with producer's risk less than or equal to some fixed level  $\alpha$ . When one or both of the hypotheses of interest are composite rather than simple, optimal testing procedures exist only in rather special circumstances. The problem of interest here is of this latter type, and no "optimal" tests have been devised for solving this problem. Specifically, we are interested in tests which compare two prespecified values of the population mean based on an available sample (be it complete or censored) drawn from a Weibull distribution. When the basic observable lifetime is distributed according to  $W(A, B)$ , then the null hypothesis  $H_0 : \mu = \theta_0$  actually represents the complex composite hypothesis that the parameter pair  $(A, B)$  satisfies the equation

$$\Gamma\left(\frac{A+1}{A}\right)B^{1/A} = \theta_0. \quad (3.1)$$

Thus, testing  $H_0 : \mu = \theta_0$  vs  $H_1 : \mu = \theta_1$  forces one to consider whether the parameter pairs  $(A, B)$  consistent with  $H_0$  provide an adequate explanation of the data by comparison to the explanation provided by  $(A, B)$  pairs satisfying  $H_1$ . We will eventually examine three specific ways of testing  $\theta_0$  vs  $\theta_1$  in the context above. We first treat a simpler problem for which an exact and optimal solution is available. Our purpose is to construct a "gold standard" against which solutions to the original problem can be compared.

Let us, then, assume that a random sample  $X_1, \dots, X_r$  is available from what was originally thought to be an exponential distribution, and that the sample size  $r$  was determined on the basis of an exponential life test plan for testing  $H_0 : \mu = \theta_0$  vs  $H_1 : \mu = \theta_1$ , where  $\theta_0 > \theta_1$ , at fixed predetermined values of the error probabilities  $\alpha$  and  $\beta$ . Assume further that, once the data was collected, the assumption

$$X_1, \dots, X_r \stackrel{iid}{\sim} W(A, B) \quad (3.2)$$

was adopted. Finally, let us suppose that the shape parameter  $A$  is known precisely. Then we may transform the data to

$$X_1^A, \dots, X_r^A \stackrel{iid}{\sim} Exp(B), \quad (3.3)$$

and test  $H_0 : \mu = \theta_0$  vs  $H_1 : \mu = \theta_1$  on that basis. We note, however, that the discrimination ratio  $\theta_1/\theta_0$  in the original problem is affected by the transformation in (3.3). Specifically, the original hypotheses may now be rewritten as

$$H_0 : B = B_0 = \left( \frac{\theta_0}{\Gamma\left(\frac{A+1}{A}\right)} \right)^A \text{ vs } H_1 : B = B_1 = \left( \frac{\theta_1}{\Gamma\left(\frac{A+1}{A}\right)} \right)^A, \quad (3.4)$$

to be tested on the basis of data from  $Exp(B)$ . The discrimination ratio in this latter problem is

$$\frac{B_1}{B_0} = \left( \frac{\theta_1}{\theta_0} \right)^A, \quad (3.5)$$

which is a decreasing function of  $A$ . This change is significant in that the performance characteristics of exponential life tests depend very strongly on this parameter. The

discrimination ratio is a measure of the distance between the null and alternative hypotheses. If that ratio is reduced sharply, the new testing problem can be resolved with much greater power. In particular, given the same sample size, the  $\alpha, \beta$  values that can be realized in the transformed environment will be much smaller than the nominal levels with respect to which the test was planned. Alternatively, the same nominal  $\alpha, \beta$  values could have been achieved with a substantially smaller sample size. The flip side of these outcomes must also be mentioned. In the DFR case, we see that the discrimination ratio increases under a power transformation. It will then be generally true that error probabilities are larger than their nominal values when the sample size is held constant, and it may take a substantially larger sample size to achieve the nominal levels of  $\alpha$  and  $\beta$ . Thus, life testing in a Weibull environment is not necessarily advantageous to the tester. Fortunately, in many engineering applications of the Weibull distribution, the shape parameter  $A$  turns out to be substantially larger than 1, and the opportunity exists for the execution of tests with smaller error probabilities or tests requiring less in the way of resources for their implementation. When infant mortality (and an initial decreasing failure rate) is present, the institution of a bum-in phase, in which key defects are detected early and removed, will tend to result in burned-in systems whose lifetimes are well modelled by an IFR distribution.

One of the main goals of the present study is to characterize the magnitude of the gains or losses attributable to a shift from the exponential to the Weibull paradigm. In order to do this, it is necessary to have tables like our **Table 1** (that is, Table 2B-5 in DoD *Handbook H108*) in much more extensive form. As mentioned in the discussion following **Table 1**, an exponential life test for testing  $H_0 : \theta = 1,000$  hrs vs  $H_1 : \theta = 500$  hrs with  $\alpha = \beta = .1$  requires that at least 15 systems be placed on test, and requires a total time on test of up to 10,305 hrs. Now, suppose it is determined that the Weibull distribution  $W(2,B)$  is the appropriate model under which the data from this experiment should be analyzed. In this particular Weibull environment, after squaring each observed failure time, we are testing the hypotheses  $H_0 : B = 1,273,231$  hrs<sup>2</sup> vs  $H_1 : B = 318,208$  hrs<sup>2</sup> based on data from the exponential distribution  $Exp(B)$ . Since the discrimination ratio in this new problem is 1/4, we'd now like to determine, for comparison purposes, the test resources that would be required to carry out the latter test at  $\alpha = \beta = .1$ . Two

difficulties arise in trying to do this. First, the discrimination ratio 1/4 does not appear in [Table 1](#), so that the values of the required sample size  $r_0$  and critical threshold  $c/\theta_0$  can only be roughly determined by interpolation. Second, the critical threshold in the new problem relates to a function other than total test time—in general, it is a threshold which provides a bound for the statistic  $\sum_{i=1}^r \frac{x_i}{\theta}$  rather than for the total time on test  $\sum_{i=1}^r x_i$ . We will now consider each of these matters carefully.

In order to be able to examine the impact of an arbitrary power transformation from  $W(A,B)$  to  $Exp(B)$ , we need to have a table comparable to [Table 1](#), but containing values of  $r_0$  and  $c/\theta_0$  for any value of the discrimination ratio between 0 and 1. The computations involved are conceptually simple—we need to find, for fixed values of  $\alpha$ ,  $\beta$  and the ratio  $\theta_1/\theta_0$ , the smallest integer  $r_0$  satisfying inequality (1.7) and the associated value of  $c$  in (1.8). In developing our tabulations, we have utilized the Peizer-Pratt approximation (see Alfers and Dinges, 1984) for  $X^2$  tail probabilities for degrees of freedom  $\theta$ , and an approximation based on the Central Limit Theorem for degrees of freedom  $>100$ . Normal tail probabilities were approximated using an "error function" which is a special case of the incomplete gamma function (see Press et al., 1992:220). Whenever the inverse of the normal or  $X^2$  distribution was needed, we employed numerical approximations for the quantiles of interest based on Newton-Raphson iterations. For computations involving the gamma function, we used table of  $\Gamma(x)$  for  $x = 1.0(0.01)2.0$  found in Abramowitz and Stegun (1964), with linear interpolation as needed. We note that stable, highly accurate algorithms for the functions above are available in various popular software packages (NAG, IMSL, S+, netlib). In the first four columns of each page in Tables [2](#), [3](#), [4](#) and [5](#), we have recorded the discrimination ratio, the required number  $r_0$  of systems on test, the critical threshold  $c/\theta_0$ , where  $c$  is computed via equation (1.8), and the realized value of  $\beta$  when the exponential life test is executed at the indicated nominal significance level  $\alpha$ .

Our expansion of DoD *Handbook H108*'s Table 2B-5 is restricted to four typical choices of  $\alpha$ ,  $\beta$ :  $\alpha = \beta = .01$  ([Table 2](#)),  $\alpha = \beta = .05$  ([Table 3](#)),  $\alpha = \beta = .10$  ([Table 4](#)) and  $\alpha = \beta = .25$  ([Table 5](#)). In each of these four tables, we record, for different values of the shape parameter  $A$

of the underlying Weibull parameter (for  $A = .1 (.1)3$ ), the values of four measures of the impact of carrying out a Weibull life test: SSR (for "sample size ratio"), TTR (for "total time on test ratio") BR (for the ratio of error probabilities  $\beta$  of the Weibull test and the planned exponential test at the same fixed values of  $\alpha$  and  $r$ ) and  $r/n$  (for the censoring fraction at which the Weibull analysis approximately achieves the nominal error probabilities in the exponential test plan). We now turn to a description of the reasoning and computations involved in producing these four measures.

As the discrimination ratio changes from  $\theta_1/\theta_0$  to  $(\theta_1\theta_0)^A$  in the course of shifting from exponential to Weibull assumptions, so do the sample size requirements for any fixed  $\alpha$  and  $\beta$ . In general, the smallest integer  $r_0$  satisfying (1.7) is an increasing function of the discrimination ratio, so that, when  $A > 1$ , the sample size needed to resolve the Weibull test will be smaller than that called for in the original test plan (and, conversely, will be larger when  $A < 1$ ). The ratio of these two sample sizes is recorded as SSR. If, for example, for  $\alpha = \beta = .1$ , and the original discrimination ratio is .5, then  $r_0 = 15$  for the exponential test plan. If  $A = 2$ , then the new discrimination ratio is .25, and the new required sample size is  $r_0 = 4$ . In the column labeled SSR under  $A = 2.0$ , and in the row for  $\theta_1/\theta_0 = .5$ , one finds the tabulated value SSR = .267 which, of course, is equal to 4/15.

Determining the maximal total time on test that might be required in the Weibull environment is a little trickier. Indeed, it cannot be determined exactly, though useful (indeed, sharp) upper and lower bounds can be obtained. In the case of complete samples from  $W(A,B)$ , we know that the Weibull life test will stop as soon as

$$\sum_{i=1}^r X_i^A = rc, \quad (3.6)$$

where  $c$  may be obtained from Table 4 through the use of the value of  $c/\theta_0$  corresponding to the discrimination ratio  $(\theta_1\theta_0)^A$  and the formula

$$c = \frac{c}{\theta_0} \left( \frac{\theta_0}{\Gamma(\frac{1}{A})} \right)^A. \quad (3.7)$$

The question then arises: what are the possible values of the total time on test  $\sum_{i=1}^r X_i$  under the constraint in (3.6)? That question is answered by the following result:

Lemma 1: Let  $X$  be the set of all vectors  $\mathbf{x} = (x_1, \dots, x_r)$  of nonnegative real numbers such that

$$\sum_{i=1}^r x_i^A = K. \quad (3.8)$$

Then, if  $A > 1$ ,

$$K^{1/A} \leq \sum_{i=1}^r x_i \leq (r^{A-1} K)^{1/A} \quad \forall \mathbf{x} \in X, \quad (3.9)$$

and if  $A < 1$ ,

$$\left(\frac{K}{r^{1-A}}\right)^{1/A} \leq \sum_{i=1}^r x_i \leq K^{1/A} \quad \forall \mathbf{x} \in X. \quad (3.10)$$

Moreover, these bounds are sharp.

Proof: The upper bound in (3.9) and the lower bound in (3.10) may be obtained quite readily by Lagrangian optimization. The other two bounds may be obtained by variational and/or geometric arguments. We eschew these approaches in favor of a simple argument based on majorization ideas (see Marshall and Olkin, 1979). If  $\mathbf{x}$ ,  $\mathbf{y}$  are vectors in  $\mathbb{R}^n$ , then  $\mathbf{x}$  is *majorized* by  $\mathbf{y}$  ( $\mathbf{x} \prec \mathbf{y}$ ) if  $\sum_{i=1}^j x_{(i)} \leq \sum_{i=1}^j y_{(i)}$  and the ordered vectors, with  $x_{(1)} \leq \dots \leq x_{(n)}$  and  $y_{(1)} \leq \dots \leq y_{(n)}$  satisfy the inequalities

$$\sum_{i=j}^n x_{(i)} \leq \sum_{i=j}^n y_{(i)}, \quad j = 1, \dots, n. \quad (3.11)$$

A real valued function  $\phi$  is *Schur convex* if  $\phi(\mathbf{x}) \leq \phi(\mathbf{y})$  whenever  $\mathbf{x} \prec \mathbf{y}$ , and *Schur concave* if  $\phi(\mathbf{x}) \geq \phi(\mathbf{y})$  whenever  $\mathbf{x} \prec \mathbf{y}$ . It is clear from these definitions that a Schur convex function is maximized, among vectors  $\mathbf{x} \in \mathbb{R}^n$  with nonnegative components and a fixed sum  $S$ , by the vector  $\mathbf{x}$  that majorizes all the rest, namely

$$\mathbf{x}_M = (0, 0, \dots, 0, S), \quad (3.12)$$

and is minimized, within this same class of vectors, by the vector  $\mathbf{x}$  majorized by all the rest, namely

$$\mathbf{x}_m = (S/n, S/n, \dots, S/n). \quad (3.13)$$

Similarly, among the class of vectors of interest, a Schur concave function will be maximized by  $x_m$  and minimized by  $x_M$ . It remains to show that these ideas provide a solution to the problem posed by the Lemma.

It is well known (see Hardy, Littlewood and Pólya, 1929) that if  $g(x)$  is a real valued convex function, then the function

$$\phi(\mathbf{x}) = \sum_i g(x_i) \quad (3.14)$$

is Schur convex, and if  $g$  is concave,  $\Phi$  in (3.14) is Schur concave. Now, since the function

$$g(x) = x^p \quad (3.15)$$

is convex for  $x \in [0, \infty)$  when  $p > 1$  and is concave for  $x \in [0, \infty)$  when  $0 < p < 1$ , we have that

$$\phi(\mathbf{x}) = \sum_i x_i^p \quad (3.16)$$

is Schur convex on  $[0, \infty)^r$  for  $p > 1$  and is Schur concave on  $[0, \infty)^r$  for  $0 < p < 1$ . Now, consider the functions of  $\mathbf{x}$  given in (3.8) and (3.9). By defining  $y_i = x_i^A$ , we may rewrite the problem at hand as: minimize and maximize

$$\phi^*(\mathbf{x}) = \sum_i y_i^{1/A} \quad (3.17)$$

among  $\mathbf{y} \in \left\{ \mathbf{y} \in [0, \infty)^r \mid \sum_i y_i = K \right\}$  Since  $\Phi^*$  in (3.17) is Schur concave when  $A > 1$ , we have that

$$K^{1/A} = \phi(0, \dots, 0, K) \leq \phi^*(\mathbf{y}) \leq \phi^*\left(\frac{K}{r}, \dots, \frac{K}{r}\right) = \left(r^{A-1} K\right)^{1/A}. \quad (3.18)$$

Transforming back from  $\mathbf{y}$  to  $\mathbf{x}$  yields (3.9). It follows, similarly, by the Schur convexity of  $\square^*$  when  $A < 1$ , that, in this case,

$$\left(\frac{K}{r^{1/A}}\right)^{1/A} = \phi^*\left(\frac{K}{r}, \dots, \frac{K}{r}\right) \leq \phi^*(\mathbf{y}) \leq \phi^*(0, \dots, 0, K) = K^{1/A}. \quad (3.19)$$

Replacing  $y_i$  by  $x_i^A \forall i$  in (3.19) yields (3.10).

The lemma above allows us to bound the total time on test required in a Weibull environment to achieve the nominal  $\alpha$  and  $\beta$  levels. We will use these bounds differently, depending on whether  $A > 1$  or  $A < 1$ . If  $A > 1$ , then the total time on test required to resolve the Weibull test based on the transformed data will generally be smaller than the maximal TTT called for in the exponential life test plan. In this case, the upper bound in (3.9), with  $K = r^*c^*$ , where  $r^*$  and  $c^*$  correspond to the sample size and critical threshold for the transformed discrimination ratio, serves as an indicator of the potential savings in TTT in the new environment. It is, of course, an upper bound; the true savings may be substantially greater! The total time on test ratio (TTTR) tabulated for each combination of discrimination ratio and shape parameter, is the ratio of the maximal (that is, upper bound) TTT in the Weibull environment to the planned-for TTT in the original exponential environment. For example, for  $\alpha = \beta = .1$  and discrimination ratio is .5, the TTT required to guarantee resolution of the exponential life test is, as we have seen, 10,305 hrs. If one then assumes a  $W(2,B)$  environment, the new discrimination ratio is .25, requiring, to achieve the same  $\alpha$  and  $\beta$  levels, that  $r = 4$  observations be placed on test and that a maximum value for  $\chi^2_{\alpha}$  of  $K = 4(.436)(1,273,239) = 2,220,529$  hrs<sup>2</sup>. Thus, the upper bound on the total time on test is  $\sum_{i=1}^{r_1} \chi^2_{2r_i,1-\alpha} = 2,980.3$  hrs. From this, we find that TTTR = 2,980.3/10,305 = .289, as recorded in Table 4 in the TTTR column under  $A = 2.0$  and across from  $\theta_1/\theta_0 = .5$ . It is possible to give a closed form expression for TTTR as a function of  $A$ ,  $\alpha$ , and the sample sizes  $r_0$  and  $r_1$  in the exponential and Weibull environments:

$$TTTR = \frac{\left( (2r_1)^{A-1} \chi^2_{2r_1,1-\alpha} \right)^{1/A}}{\chi^2_{2r_0,1-\alpha} \cdot \Gamma\left(\frac{A+1}{A}\right)}$$

We will define TTTR differently when  $A < 1$ . In these cases, the TTT required to resolve the test in the Weibull environment will tend to be larger than that required in the original exponential test plan. We are thus interested in determining a bound which the TTT will exceed with certainty. In doing so, we employ the lower bound in (3.10). For  $A < 1$ , our tabled values of TTTR represent the ratios of the smallest possible value of the required total test time in the

Weibull environment to the required TTT in the exponential test plan. As an example, consider the computation of TTTR in testing the hypotheses  $H_0 : \mu = 1,000$  vs  $H_1 : \mu = 250$  at  $\alpha = \beta = .1$ . Assume that the true distribution is  $W(1/2, B)$ . Since the discrimination ratio is .25, the exponential life test plan calls for setting 4 systems on test, and the TTT required to ensure the resolution of the test is  $4(.436)(1,000) = 1,744$  hrs. To achieve  $\alpha = \beta = .1$  in the  $W(1/2, B)$  environment, given that the new discrimination ratio is .5, one needs to place 15 systems on test and require a TTT commensurate with the equation

$$\sum_{i=1}^{15} X_i^{1/2} = 15(.607) \left( \frac{1,000}{\Gamma(3)} \right)^{1/2} = 230.4 = K.$$

From (3.10), it follows that

$$TTT = \sum_{i=1}^{15} X_i \geq \frac{K^2}{r} = 3,539.8.$$

Thus, a lower bound on the ratio of the total time on test in the Weibull vs exponential environments is given by  $TTTR = 3,539.8/1,744 = 2.03$ . This latter value, or more precisely, the value 2.027, is recorded in **Table 4** in the TTTR column under  $A = 0.5$  and across from  $\theta_1/\theta_0 = .25$ . As an aside, we note that the upper bound provided in (3.10) indicates that the total time on test might be as much as 30.4 times as large as that required by an exponential life test plan; thus, while 1,744 hrs of testing are required by the original plan, the TTT in the Weibull environment with  $A = 1/2$  will fall between 3,540 and 53,018 hrs.

The BR column in Tables 2 to 5 is more or less self-explanatory. The exponential test plan corresponding to a fixed discrimination ratio stipulates a certain sample size  $r$  as necessary to achieve fixed, nominal  $\alpha$  and  $\beta$  values. If the data is actually drawn from  $W(A, B)$  with  $A$  known, and if the same sample size  $r$  is used in executing the Weibull test at significance level  $\alpha$ , then a new level of  $\alpha$  is attained—all it  $\beta$ . We define BR as the ratio  $\beta/\alpha$ . An asterisk in the BR column means  $BR < .0005$ .

We now turn to the subject of censoring and, in particular, to the interpretation of the column labeled  $r/n$  in Tables 2 to 5. While type II censoring does not alter the power function of an exponential life test, but only serves to accelerate the completion of the test, its impact in

Weibull life testing is quite different. In particular, the behavior of the total time on test statistic in Weibull life testing is strongly influenced by the number of systems on test; indeed, we demonstrate below that the upper bound on TTT in Weibull environment with  $A > 1$ , tends to infinity as  $n$  grows in  $r$ -out-of- $n$  life tests. Thus, while one can increase  $n$  with impunity in exponential life tests, one must be very careful in using censored life test designs when the underlying distribution is Weibull. We will motivate below a guideline for identifying what might be considered a reasonable upper bound for the amount of censoring one should entertain in a particular application. The tabled ratio  $r/n$  identifies this bound. If  $r/n = 4$ , for example, then the number  $n$  of systems on test should not exceed  $n = r/4 = 2.5r$ . We will return to our  $r/n$  computation momentarily. First, we extend Lemma 1 to a result which applies to type II censored samples and provides the bounds utilized in that computation.

Lemma 2: Let  $X$  be the set of all vectors  $\mathbf{x} = (x_1, \dots, x_r)$  of nonnegative real numbers such that

$$\sum_{i=1}^r w_i x_i^A = K, \quad (3.20)$$

where  $w_1, \dots, w_r$  are positive weights. Then for  $A > 1$ ,

$$\sum_i w_i x_i \leq \left[ \left( \sum_i w_i \right)^{A-1} K \right]^{1/A} \quad \forall \mathbf{x} \in X, \quad (3.21)$$

and for  $A < 1$ ,

$$\sum_i w_i x_i \geq \left[ \frac{K}{\left( \sum_i w_i \right)^{1-A}} \right]^{1/A} \quad \forall \mathbf{x} \in X. \quad (3.22)$$

Proof: We show, by Lagrangian methods, that the bounds in (3.21) and (3.22) are the values associated with the unique critical points of the function  $\sum_i w_i x_i$  in each case. Then, an easy geometric argument involving hyperplanes above or below the surface in (3.20) will demonstrate that these values are extrema. Let

$$f(\mathbf{x}, \lambda) = \sum_i w_i x_i - \lambda \left( \sum_i w_i x_i^A - K \right). \quad (3.23)$$

To find critical points of the function  $\sum_i w_i x_i^A$  under the constraint (3.20), we solve the equations

$$\frac{\partial}{\partial x_i} f(x, \lambda) = w_i - \lambda w_i A x_i^{A-1} = 0, \quad i = 1, \dots, r$$

and

$$\frac{\partial}{\partial \lambda} f(x, \lambda) = \sum_i w_i x_i^A - K = 0.$$

It is evident from the above that there is a single critical point of  $f$ , namely

$$x_i = \left( \frac{1}{\lambda A} \right)^{\frac{1}{A-1}}, \quad i = 1, \dots, r, \quad (3.24)$$

where  $\lambda$  is chosen so that

$$\sum_i w_i \left( \frac{1}{\lambda A} \right)^{\frac{A}{A-1}} = K. \quad (3.25)$$

This results in the values

$$x_i = \left( \frac{K}{\sum_i w_i} \right)^{1/A}, \quad i = 1, \dots, r. \quad (3.26)$$

The value of the function  $\sum_i w_i x_i^A$  at this critical point is

$$\sum_i w_i \cdot \left( \frac{K}{\sum_i w_i} \right)^{A/A} = \left( \sum_i w_i \right)^{A/A} K^A,$$

which is equivalent to the upper bound in (3.21) when  $A > 1$  and to the lower bound in (3.22) when  $A < 1$ .

We can now introduce the  $r/n$  column in Tables 2 to 5. The TTTR computation in these tables is based on the assumption that complete (that is, uncensored) samples of size  $r$  are available. If, however, the life test plan calls for type II censoring, and is terminated, at the latest, when the  $r$ th failure occurs among the  $n$  systems on test, then the TTT statistic becomes

$\sum_{i=1}^{r-1} X_i + (n-r+1)X_r$ , which is of the form  $\sum_i w_i X_i$  with unequal, positive weights. From Lemma 2, we see that, when  $A > 1$ , the maximum total time on test under the exponential test plan constraint  $\sum_i X_{i0}^A + (n-r+1)X_{r0}^A = K$  is given by

$$TTT_{rn}^* = \left( n^{A-1} K \right)^{1/A}. \quad (3.27)$$

A worrisome feature of this value is that it is unbounded in  $n$ , when  $A > 1$ , as assumed above. For the complete sample design based on  $r$  observed failures,

$$TTT_{rn}^* = \left( r^{A-1} K \right)^{1/A}, \quad (3.28)$$

the constraint on  $\sum_i X_i^A$  being the same as in the censored case, given that the transformed problem is based on an underlying exponential distribution. Now the ratio we have called TTTR measures (conservatively) the extent to which TTT could have been reduced if an appropriate Weibull life test had been conducted. Suppose we take particular values of the discrimination ratio,  $\alpha$ ,  $\beta$  and  $A > 1$  as fixed and given. We could then ask the question: what sample size  $n$  would yield a censored sampling plan in the Weibull environment that has (at worst) an equivalent TTT requirement as that of the original exponential test plan? The answer is the following: the value of  $n$  yielding a maximum TTT no larger than the required TTT in the exponential test plan is the solution to the equation

$$\frac{\left( r^{A-1} K \right)^{1/A}}{\left( n^{A-1} K \right)^{1/A}} = TTTR, \quad (3.29)$$

that is,

$$n = r (TTTR)^{\frac{A-1}{A}} \quad (3.30)$$

or, as we will record it,

$$\frac{r}{n} = (TTTR)^{\frac{A-1}{A}}. \quad (3.31)$$

The formula for  $r/n$  in (3.31) is derived in the same way for  $A < 1$ , and will be applied in Tables 2 to 5 for arbitrary  $A$ . However, since TTTR has different interpretations for  $A > 1$  and  $A < 1$ , so

too does the fraction  $r/n$ . As stated above, when  $A > 1$ ,  $r/n$  represents the censoring ratio that yields a test in the Weibull environment for which the required total time on test is no greater than the TTT specified in the exponential test plan. For  $A < 1$ ,  $r/n$  represents the censoring ratio for which the minimum possible TTT in the Weibull environment is approximately equal to the TTT specified in the exponential test plan. The actual TTT experienced in executing the Weibull life test can, of course, be much larger than that minimum. Thus, caution must be exercised in interpreting an  $r/n$  ratio when  $A < 1$ . Further, it is possible, when  $A < 1$ , for  $r/n$  to exceed 1. Such an outcome simply points to the fact that the minimum possible TTT in the Weibull environment will be less than or equal to the required TTT in the exponential test plan when one reduces the sample size from  $r$  to the value of  $n$  for which  $r/n$  is the tabulated value.

As an example of the computation of  $r/n$ , suppose  $\theta_1/\theta_0 = .5$ ,  $A = 2.0$  and  $\alpha = \beta = 0.1$ . From [Table 4](#), we find that TTTR = .289, so that  $r/n = (.289)^2 = .0835$ , which is recorded as  $r/n$ -.084 in [Table 4](#). From this, we deduce that a test plan which places 48 systems on test and resolves the test upon the 4th failure would have a total test time no larger than 10,305 hrs, the test time associated with the exponential test plan based on 15 observed failures.

Given the definitions of SSR, TTTR, BR and  $r/n$  in the preceding paragraphs, we now present the tables in which these measures appear.

While Tables 2 to 5 largely speak for themselves, a few comments on them seem warranted at this point. In general, these tables confirm that there are potential resource savings available when one recognizes an IFR Weibull environment and carries out a Weibull life test instead of an exponential one. Similarly, more resources are required in carrying on Weibull life tests when the DFR Weibull analysis is carried out instead of an exponential life test. Since IFR Weibull distributions arise with some frequency in life testing applications, the magnitude of the measure SSR, TTTR, BR and  $r/n$  when the shape parameter is large is of special interest. Contour plots showing level curves of each of these measures are especially revealing.

As an example, a rough sketch of the level curves of TTTR as a function of the discrimination ratio  $\theta_1/\theta_0$  and the value of the shape parameter  $A$  is shown in [Figure 15](#) for the case  $\alpha = \beta = .1$ . These plots are only approximate, of course, since the discreteness of sample

size selection causes the computed values of TTTR (and the other measures as well) to be a rather choppy function of the discrimination ratio for each fixed value of  $A$ .

From [Figure 15](#), one may infer that the most substantial savings in TTT are made in situations in which both the discrimination ratio and the Weibull shape parameter are high. In applications, the costs associated with life tests when the discrimination ratio is high (say, greater than .7) are generally prohibitive; thus, even though the resource savings afforded by a Weibull life test might be substantial, the cost of the alternative analysis is still likely to be prohibitive. If the discrimination ratio is .9, for example, an exponential life test plan for, say,  $H_0 : \theta = 1,000$  vs  $H_1 : \theta = 900$  at  $\alpha = \beta = .1$  would require at least  $r = 593$  systems on test and a total test time of 561,571 hrs. If  $A = 2.5$ , say, the Weibull test with the same error probabilities could be accomplished with  $r = 96$  systems on test and a total test time no greater than 102,206 hrs. While these savings are striking, the experiment may still be too costly to perform. It appears that the kind of problems in which recognizing a Weibull environment and performing a Weibull life test will be both feasible and economically attractive will be those in which  $.3 \leq \theta_0/\theta_1 \leq .7$  and  $A \geq 1.5$ .

We will return to our discussion of Tables 2 to 5 in the concluding section. It is perhaps worth noting here that the measure BR shows quite dramatically the power of Weibull life tests when the shape parameter  $A$  is reasonably large; for fixed values of  $A$ , BR appears to vary inversely with the discrimination ratio. We also note that, for fixed  $A > 1$ , the amount of censoring that can be accommodated per the  $r/n$  computation is an increasing function of the total time on test ratio, which in turn tends to increase as a function of the discrimination ratio.

#### 4. WEIBULL LIFE TESTING—PART 2

In section 3, we studied the performance characteristics of Weibull life tests under the simplifying assumption that the Weibull shape parameter  $A$  was known. The assumption is not totally whimsical, since engineering experience with a particular type of application might make such an assumption quite reasonable. The exponential assumption is, after all, nothing more than

the assumption that the Weibull shape parameter is known to be equal to one. One might consider the results of section 3 to apply to the situation in which the statistician guesses (or estimates) the value of the Weibull shape parameter, and happens to guess it correctly. It is, of course, necessary to move beyond this first step, and to engage seriously the question of how to execute a Weibull analysis in the general, two-parameter problem. This section is devoted to examining three specific possibilities in that regard.

Suppose  $x_1, \dots, x_n \sim W(A, B)$ , where  $A > 0, B > 0$  are unknown, and we wish to test hypotheses  $H_0 : \mu = \mu_0$  vs  $H_1 : \mu \neq \mu_1$ . The approach of section 3 immediately suggests a possible approach to this testing problem: estimate  $A$  from data, and carry out a Weibull test, as in the preceding section, with the estimate  $\hat{A}$  taken as the known value of  $A$ . The performance of the resulting test procedure is, naturally, dependent upon the quality of the estimate  $\hat{A}$ . This "plug-in" method has a history. In an estimation framework, Gong and Samaniego (1981) described the large sample behavior of the solutions of a reduced system of likelihood equations when certain (nuisance) parameters were replaced by  $\sqrt{n}$ -consistent estimators. In the context of testing composite hypotheses, Neyman (1959) gave conditions under which tests utilizing  $\sqrt{n}$ -consistent estimators of the nuisance parameters have a certain type of asymptotical optimality. More specifically, Neyman showed that, under fairly standard regularity conditions, test statistics which were uncorrelated with the logarithmic derivative of the likelihood with respect to the nuisance parameters (under a null hypothesis specifying a fixed value of a given parameter) were "locally asymptotically most powerful" in testing that null hypothesis against its compliment. Such tests were named  $C(\alpha)$  tests by Neyman, in deference to the similarity of his regularity conditions to those posited by Cramér (1946) in his work on the large sample theory of maximum likelihood estimators.

The essence of a  $C(\alpha)$  test is the substitution of one or more unknown parameters by  $\sqrt{n}$ -consistent estimators, and the testing of hypotheses concerning a lower dimensional parameter space. We will examine two tests based on such an approach. The first of these is based on the fact that the coefficient of variation of the Weibull distribution, given in (2.5), depends only on

the shape parameter of the distribution and is independent of the second parameter  $B$ . A  $\sqrt{n}$ -consistent estimator of the parameter  $cv$  is readily available; a natural estimate is the sample  $cv$ , or

$$\hat{cv} = \bar{s} / \bar{x}, \quad (4.1)$$

where  $\bar{x}$  and  $s$  are the sample mean and standard deviation, respectively. Now the relationship between  $cv$  and  $A$  in (2.5) is not easily inverted; we will therefore deal with that inversion numerically. In spite of that slight complication, the inversion which expresses  $A$  as a function of  $cv$  will, by virtue of the continuity and differentiability of the functional relationship, provides us with a  $\sqrt{n}$ -consistent estimator of  $A$ . Table 6 represents a numerical compilation from which one can obtain an estimated shape parameter from an estimated  $cv$ . We have relied upon this table for obtaining  $\hat{A}$  when  $\hat{cv} \leq 1$ , since interpolation in these cases provides acceptable accuracy. When  $\hat{cv} > 1$ , the bisection method was used to calculate  $\hat{A}$ , pivoting on the expression in (2.5) until sufficient accuracy in  $\hat{A}$  was attained.

The testing method for which results are recorded under the "cv" column in Tables 7 to 9 is the  $C(\alpha)$ -type test with  $A$  estimated by the appropriate function  $\hat{A}$  of the sample coefficient of variation. Once  $A$  is set equal to  $\hat{A}$ , the test is carried out as in section 3, leading to rejection of the hypothesis  $H_0 : \mu = \mu_0$  in favor of  $H_1 : \mu < \mu_0$  if the statistic  $\sum x_i^2$  is sufficiently small.

The second testing procedure we study here is the likelihood ratio test of the hypotheses  $H_0 : \mu = \mu_0$  vs  $H_1 : \mu < \mu_0$  based on a Weibull sample. Execution of such a test requires the maximization of the Weibull likelihood  $L$ , given by

$$L(x|A, B) = \left( \frac{A}{B} \right)^r \left( \prod_i x_i \right)^{A-1} e^{-\frac{\sum x_i^2}{B}}, \quad (4.2)$$

over  $\Omega = \{(A, B) | A > 0, B > 0\}$  and over  $\Omega_0 = \{(A, B) | A > 0, B > 0, \Gamma(\frac{A+1}{A}) B^{1/A} = \mu_0\}$ . The likelihood ratio statistic we compute is

$$\Lambda = \frac{\max_{(A,B) \in \Omega_0} \mathcal{L}(x|A,B)}{\max_{(A,B) \in \Omega} \mathcal{L}(x|A,B)}. \quad (4.3)$$

Standard theory implies that, for sufficiently large  $r$ , the statistic  $-2\ln\Lambda$  is approximately distributed as a  $\chi^2_1$  random variable. To compute  $\Lambda$  in (4.3), we employed two-dimensional Newton-Raphson iterations in the unrestricted maximization required by the denominator of  $\Lambda$ , and carried out the (essentially) one-dimensional search required by the numerator of  $\Lambda$  using the "golden-section" search algorithm as described by Luenberger (1989:199). Given that we were interested in a one-sided test, we defined the rejection region of the test based on the likelihood ratio as follows: if the unrestricted maximization of  $L$  in (4.2) results in MLEs  $\hat{A}$  and  $\hat{B}$  for which  $\hat{\mu} = \Gamma\left(\frac{1}{\hat{A}}\right)\hat{B}^{1/\hat{A}} \geq \mu_0$ , accept  $H_0$ . Otherwise, compute the likelihood ratio statistic  $\Lambda$ , and reject  $H_0$  if

$$-2\ln\Lambda > \chi^2_{1,2\alpha}.$$

Since  $-2\ln\Lambda$  is expected to be large under departures from  $\mu = \mu^0$  in either of two directions, we doubled the nominal tall probability of the  $\chi^2$  distribution and reject only when the data is indicative of a mean value smaller than  $\mu_0$ . For sufficiently large  $r$ , this procedure should have a significance level close to  $\alpha$ . In Tables 7 to 9, the performance of this test is recorded in the columns labeled "*er*".

We might mention, at this point, the fact that an additional possibility exists for constructing a  $C(\alpha)$ -type test: one could estimate the shape parameter  $A$  by the maximum likelihood estimate  $\hat{A}$ , and carry out the test in section 3 with  $A$  replaced by that estimate. We have confirmed, via simulation, that the performance of that test, in small and moderate samples, is essentially indistinguishable from the likelihood ratio procedure described above. We have therefore excluded the MLE-driven  $C(\alpha)$ -type test from the simulation results reported here.

Finally, we will investigate the performance of a third approach to tests involving Weibull means. As discussed in section 2, it is possible to examine the Weibull assumption through appropriate plots on Weibull probability paper. Formal estimation and hypothesis tests may be developed from these fitting procedures. Chernoff and Lieberman (1956) gave

conditions under which certain plots were optimal for estimating particular parameters. Nair (1984) has studied the large sample behavior of estimators of model parameters derived from probability plots and, in particular, showed that, under suitable regularity conditions, estimates obtained via ordinary least squares are  $\sqrt{n}$ -consistent and asymptotically normal. This latter work suggests that one might test hypotheses concerning Weibull means by first estimating the shape parameter from the least squares fit to data plotted on Weibull probability paper, and then carrying out the appropriate exponential test based on the transformed data  $X_1^A, \dots, X_r^A$  as if  $A = \hat{A}$  is the true shape parameter. The hypothesis  $H_0 : \mu = \mu_0$  is rejected when  $\sum x_i^A$  is smaller than an appropriate threshold. The performance of the test based on an least squares estimator of the Weibull shape parameter from a Weibull probability plot is recorded in Tables 7 to 9 in the column labeled " . . . . "

We are now in a position to describe the contents of Tables 7 to 9. Each table summarizes a simulation study for a fixed value of the discrimination ratio<sup>1</sup> Table 7 ( $\theta_1/\theta_0 = .25$ ), Table 8 ( $\theta_1/\theta_0 = .50$ ), and Table 9 ( $\theta_1/\theta_0 = .75$ ). In all three studies, simulations were carried out for assumed Weibull samples with five possible sample sizes, and with Weibull shape parameters ranging from 0.1 to 3.0 in increments of 0.1. The median sample size among the five used for each table was set to be equal to the sample size required by an exponential test plan with the set value of  $\theta_1/\theta_0$  for that table and the error probabilities set to  $\alpha = \beta = .1$ . For example, when  $\theta_1/\theta_0 = .25$ , the exponential test plan calls for a minimum of  $r = 4$  systems on test ( $r = 15$  when  $\theta_1/\theta_0 = .5$  and  $r = 81$  when  $\theta_1/\theta_0 = .75$ ). For a particular sample size, the error probabilities  $\alpha$  and  $\beta$  realized in 100 repetitions of each of four tests are recorded. The first column, labeled "kno," records the  $\alpha$  and  $\beta$  achieved by the test in section 3, where the true value of the shape parameter  $A$  of the underlying Weibull distribution is taken as known and is used in testing for mean life. In the column labeled "cv," the error probabilities  $\beta$  and  $\beta$  are given for a test which treats the estimate  $\hat{A}$  based on the sample coefficient of variation as if it were the true value of the shape parameter  $A$ . In the column labeled " lr,"  $\alpha$  and  $\beta$  are recorded for the version of the likelihood ratio test discussed above. Finally, in the column labeled " lse,"  $\alpha$  and  $\beta$  are given for the test in

which the least squares estimator  $\hat{A}$  from a Weibull probability plot is used as if it were the true value of the shape parameter. The first and third tests were carried out so as to achieve a significance level of 0.1. The rejection region of the second and fourth tests were chosen to yield significance level 0.1 under the assumption that  $A = \hat{A}$ . The results of our simulations appear below.

Tables 7 to 9 have some rather striking features. It will be clear from these tables that when the underlying Weibull distribution is strongly DFR, that is, when  $A$  is quite near zero, testing hypotheses concerning the mean value is an extremely difficult proposition. Even tests which substitute the correct true value of the shape parameter into the density have very low power at the alternative hypothesis, even for relatively large sample sizes. The exact, best test of  $H_0: \mu_1$  has suitably small  $\alpha$  and  $\beta$  values if  $A$  is not too small, and requires a somewhat larger value  $A$  to ensure such behavior when the sample sizes are small. For sample sizes equal to the required sample size for exponential life tests with  $\alpha = \beta = .1$ , the test with known  $A$  performs very well, with  $\alpha$  and  $\beta$  hovering at or below 0.1 for all  $A > 1$ . Fortunately, in practice, the recurrence of  $A$  near zero is not at all common.

The most surprising and encouraging aspect of Tables 7 to 9 is the fact that the three procedures for testing means in the general two parameter problem each performs nearly as well as the best test when  $A$  is known. The appropriate ground for comparison purposes is the collection of tests in which the true parameter  $A$  exceeds 1.0. This is the domain of primary interest in applications, and is the domain in which the "gold standard", that is, the test based on known  $A$ , achieves acceptable error probabilities. Inspection of Tables 7 to 9 reveals that all three general tests perform well in these settings. The values  $A_0$  for which  $\alpha \leq .15, \beta \leq .15$  for  $A \geq A_0$  are roughly estimated in Table 10 as a function of the ordered sample sizes  $r_{(1)} < \dots < r_{(5)}$ . (So the  $r_{(1)}$  through  $r_{(5)}$  of Table 10 represent 2,3,4,8, and 22 for Table 7; 4,7,15,29, and 87 for Table 8; and 21, 36, 81, 164, and 498 for Table 9.)

As an example of the surprising competitiveness of a two parameter Weibull life test, consider testing  $H_0: \mu = 1,000$  vs  $H_1: \mu = 500$  at  $\alpha = .1$ . Suppose 15 systems are placed on test, as prescribed by an exponential life test plan with  $\alpha = \beta = .1$ . If the data happens to be

governed by a Weibull distribution with shape parameter  $A = 1.2$ , and the fact that  $A = 1.2$  is somehow revealed to the experimenter, the best test of  $H_0$  vs  $H_1$  of size  $\alpha = 0.1$  can be executed after appropriately transforming the data to  $X_1^{*2}, \dots, X_{15}^{*2}$ . Table 8 shows this test as having error probabilities  $\alpha = .11$  and  $\beta = .02$  in our simulation. We expect  $\beta$  to be less than 0.1 here since a shape parameter of 1.2 has served to decrease the effective discrimination ratio, so that 15 observations is more than actually required to achieve  $\alpha = \beta = .1$  in the life test based on transformed data. For the tests which did not benefit from knowledge of the true  $A$ , we find that the  $cv$ -based test had  $\alpha = .11, \beta = .02$ , the likelihood ratio test had  $\alpha = .13, \beta = .02$ , and the least-squares-based test had  $\alpha = .09, \beta = .03$ . The performance of all three procedures are clearly indistinguishable from that of the best test in this instance. A general perusal of Tables 7 to 9 shows that this example is not an isolated instance of this type of performance.

We interpret the excellent performance of all four tests we have studied (including the  $C(\alpha)$ -type test based on the maximum likelihood estimate of  $A$ ) as constituting compelling evidence that Weibull life testing is both feasible and efficient. The lack of sensitivity of  $C(\alpha)$ -type tests to the precision of the estimated shape parameter, and the ability of such tests to achieve error probabilities comparable to those of the best possible test when  $A$  is known, provides strong support for using such tests in practice. Among the four tests we've studied, we would favor the MLE-based  $C(\alpha)$ -type test, since it has exhibited competitive small sample behavior, and is, of course, defensible asymptotically as well.

As we have seen, Weibull life testing does not enjoy the immunity from the effects of censoring that characterizes exponential life testing. It is thus important to extend the investigation above to the censored data case. In what follows, we examine the performance of a particular procedure for testing two hypothesized Weibull means in the general two parameter problem under a type II censoring design. More specifically, we have selected for study the censored data version of the "*lse*" test based on the least squares fitting of transformed censored data with a straight line of the form (2.10). The execution of this test, that is, the development of a Weibull probability plot under censoring, involves no increased complexity. The fundamental question of interest will be whether the estimated shape parameter  $\hat{A}$  obtained from such a plot

has sufficient accuracy and precision to provide reasonable performance in the associated  $C(\alpha)$ -type test in samples of small or moderate size. Our simulations will address this question. Before describing our findings in this regard, we pause to discuss briefly the other two procedures we've studied in the complete sample case.

*lse*

An extension of the *cv* test has not been pursued for lack of a reliable estimate of the Weibull coefficient of variation from censored data. The problem does not appear to have been treated in the literature; various ad hoc estimates with which we experimented proved unsatisfactory. We had greater success in extending the *er* test to censored data. Conceptually, the latter problem can be dealt with adequately. Programs have been written to obtain the required likelihood ratio statistic from censored data. However, as of this writing, we have not satisfactorily resolved the attendant numerical issues. For these and other reasons, we have not to date completed a simulation study on censored data *cv* or *er* tests comparable to the study on which we report below. Since our primary purpose in examining the censored data case is to determine whether tests exist which provide satisfactory (that is, nearly optimal) performance in cases of practical interest (that is, moderate sample sizes, shape parameter moderately large), the simulation we have done will suffice and provides an affirmative answer to this question.

In [Table 11](#), we record the realized error probabilities for two tests, the first being the optimal test (*kno*) when the Weibull shape parameter  $A$  is fixed and known and the second being the *lse* test in which  $A$  has been estimated from the Weibull probability plot. The six sections of this table provide  $\alpha$  and  $\beta$  for censored Weibull samples with  $A = .1(1)3.0$  and for censoring fraction  $r/n \geq .9, .8, .7, .6, .5$  and  $.4$  in succession. For each lower bound on the censoring fraction, we have estimated  $\alpha$  and  $\beta$  for five different possible sample sizes  $n$ . The jagged lines drawn across the table represent, roughly, the lower boundary for the shape parameter  $A$  for which the *lse* test is, for practical purposes, comparable to the optimal test.

Several conclusions may be reliably drawn from [Table 11](#). The performance of the *lsetest* based on censored data improves as  $A$  increases, as  $r/n$  increases and as  $n$  increases (while holding the other two of these parameters fixed). Our simulation gives strong support to

the claim that under moderate censoring (say  $r/n \geq .7$  in a Weibull environment with sufficiently large shape parameter (say  $A > 1.5$ ), and large enough  $n$  (say  $n > 20$ ), Weibull life testing using the *lse* procedure provides excellent performance, with error probabilities  $\alpha$  and  $\beta$  quite close to the best possible values.

## 5. DISCUSSION

In the fall of 1992, the Committee on National Statistics of the National Research Council hosted a workshop, co-sponsored by the Department of Defense, on statistical issues in defense analysis and testing (see Samaniego, 1993; and Rolph and Steffey, 1994). Much discussion during that workshop was centered around the intriguing question "how much testing is enough?" The question was considered more than just interesting. Efficient use of the resources available for testing in the DoD acquisitions programs is always of interest, but is especially pressing in the face of declining budget allocations for operational testing and evaluation. At least part of the motivation for the present study is drawn from the workshop's (and subsequent) discussion of resource-related issues. A second source of motivation for this study is the apparent overuse of exponential life testing methodology in both military and civilian applications. It seems that a careful study on the cost-saving potential of alternative treatments of life testing data might have rather broad utility.

This paper begins with a review of exponential life testing methods, and discusses basic properties of the Weibull distribution and how that distribution might be identified as an appropriate model for life testing data. Of particular interest to us has been the mechanics of Weibull life testing and the statistical performance, and cost, of this alternative approach. The results of section 3 show quite graphically that Weibull life tests can, in certain circumstances, provide substantially greater statistical power than exponential life tests based on the same sample size, and can offer substantial savings in both sample size and testing time when the goal is to match the statistical power of a planned exponential life test. While Tables 2 to 5 ostensibly offer guidance only for the very special case in which the Weibull shape parameter  $A$  is known,

and while these tables were constructed, primarily, to compare exponential and Weibull life testing for complete samples, they are more widely applicable. We will elaborate on this shortly. Yet even in the narrowest domain of applicability (that is, complete samples,  $A$  known), these tables provide important insights. First, they show quite emphatically that exponential life testing can be especially misleading when the underlying distribution is a DFR Weibull; it is clear that a much larger sample size and much greater testing time are needed to achieve any given nominal error probabilities than what an exponential life test plan would prescribe. The good news carried by Tables 2 to 5 is that, when the underlying distribution is an IFR Weibull, considerable savings are possible. Our results confirm and expand upon some of the findings in Anderson's (1994) thesis. From Tables 2 to 5, one can see that a 20-30% reduction in needed test resources is typical when the shape parameter  $A = 1.2$ , a 60-70% reduction is typical when  $A = 2.0$  and a 70-90% reduction is typical when  $A = 3.0$ . Potential cost savings of such magnitudes should certainly provide a strong incentive for life testing practitioners to try to detect an IFR Weibull environment when it is present, and to utilize Weibull life testing methods when Weibull modeling is deemed appropriate.

The results reported in Tables 7 to 9 of section 4 carry important practical implications. There tabulations show, in general, that the three approaches we've considered for real-life Weibull life testing (that is, when both Weibull parameters are unknown) perform nearly as well, in IFR Weibull environments, as the optimal test with the shape parameter assumed known. This suggests that the savings available in the idealized setting of section 3 can also be realized in real, practical life testing scenarios. Especially interesting to us is the fact that the general tests, particularly the *cvt*- and *lse*-based tests, are competitive with the ideal test even for small sample sizes. This is clear by inspection of the columns of achieved  $\alpha$  and  $\beta$  levels of the four tests for sample size  $r \leq 4$  in Table 7, for sample size  $r \leq 15$  in Table 8 and for sample size  $r \leq 81$  in Table 9. The general tests perform adequately when the Weibull shape parameter  $A > 1$ , and perform exceptionally well when  $A > 2$ . In the testing problems we have examined, it is clear that the estimation of the shape parameter prior to executing a formal test has only a second-order, and quite modest, effect on test performance. Thus, even with relatively unstable estimators of  $A$ , like those obtained from Weibull probability plots based on small or moderate samples, one can

still test hypotheses concerning Weibull means quite reliably, provided the censoring fraction is reasonably small.

Let us now consider how one might use the tabulations in sections 3 and 4 in a practical problem. For now, we'll restrict our attention to the case of complete samples, that is,  $r = n$ . In many engineering applications in which Weibull models are routinely entertained, the investigators involved have a good feel for the range of Weibull shape parameters that tend to arise. Some sort of IFR behavior is a common occurrence, and it is possible, even likely, that reasonable bounds can be placed a priori on the anticipated value of the shape parameter  $A$ . All that's really needed to employ Tables 2 to 5 with profit is a reliable lower bound on  $A$ . Suppose, for example, that one wishes to test  $H_0 : \mu = 1,000$  vs  $H_1 : \mu = 500$  at  $\alpha = \beta = .05$ . From Table 3, we see that an exponential life test plan requires that at least  $r = 23$  systems be placed on test, and that a total testing time of 15,709 hrs be planned for. Assume that the experiment is judged to be well modeled by a Weibull distribution, and assume that the experimentors can assert with some confidence that the shape parameter  $A$  is bounded below by, say, 1.5. Now note that a Weibull life test for known  $A = 1.5$  can achieve  $\alpha = \beta = .05$  with a sample size of  $r = 23(.478) = 11$  and a maximum testing time of  $15,709(.527) = 8,279$  hrs. If  $A > 1.5$ , further savings would be possible. For example, if  $A$  is known to be 2.5, the test can be accomplished with a sample of size  $r = 5$  and a total test time no greater than 3,880 hrs.

It should be emphasized that Weibull life tests having characteristics such as those above require knowledge of the value of the shape parameter. Recall, however, that under precisely the circumstances with which we are dealing, the general Weibull tests of section 4 may be employed with confidence. Our recommendation would therefore be: carry out a Weibull (cv-or *lse*-based) test with a sample of  $r = 11$  systems, terminating the test on the basis of the value of the statistic  $\sum x^2$ . Provided that the underlying model is indeed Weibull with shape parameter  $> 1.5$ , this procedure should secure savings of 50% or more in sample size and total testing time while maintaining error probabilities in the neighborhood of the nominal levels used in the design of the experiment.

It is natural to ask how the above generalizes to censored-data designs. We can give at least a partial answer to this question. We should first note that in the Weibull test of section 3, based on a data transformation depending on the known true value of the Weibull shape parameter  $A$ , the statistic

$$T = \sum_{i=1}^r X_{(i)}^A + (n-r)X_{(r)}^A \quad (5.1)$$

has the distribution  $\Gamma(r, B)$  irrespective of the number of systems  $n$  on test, so that the realized values of  $\alpha$  and  $\beta$ , and the reported value of SSR, hold for these tests under type II censoring. As we have noted previously, however, the maximal total time on test associated with a fixed value of  $T$  in (5.1) does depend on  $n$  and grows without bound as  $n \rightarrow \infty$ . Thus, in order to identify a value  $T_0$  such that the test is sure to be resolved with TTT no greater than  $T_0$ , one would need to bound the value of  $n$ . It should be recognized that this upper bound on TTT may be considerably larger than the realized TTT in actual Weibull life testing, especially since that upper bound is achieved only when all observed failure times are identical and satisfy a constraint of the form of (3.20). In any case, the influence of type II censoring on the tables in section 3 is exclusively through the total time on test ratio. The measure TTTR is increasing as a function of  $n$ , and reaches the value 1 (corresponding to the circumstance in which the TTT in exponential and Weibull tests have the same maximal value) when  $n$  is equal to  $r$  times the reciprocal of the recorded value of  $r/n$ .

Our investigations regarding Weibull life testing in the censored data case, while not as comprehensive as we'd like, support the general conclusion that it is possible to test competing Weibull means reliably in the presence of censoring. In applications of that type, our simulations indicate that extreme censoring can be dangerous, that is, can lead to quite inflated error probabilities unless the shape parameter  $A$  is very large. In standard applications of the Weibull model, where the shape parameter  $A \in (1.5, 3)$ , Weibull life testing based on moderate sample sizes ( $20 \leq n \leq 30$ ) and modest censoring ( $r/n \geq .7$ ) provides excellent performance. Our conclusions are based on our study of the performance of the *lse* procedure. We conjecture that an appropriately implemented *cv* or *er* procedure will have similar performance characteristics.

One might rightly seek an intuitive explanation of the phenomena that have been observed in this paper. Why would one expect to be able to test hypotheses concerning means more efficiently under an IFR Weibull model than under an exponential one? The best explanation may be in terms of the variance of observed lifetimes under the two models. For a model with a given mean, an IFR Weibull has smaller variance than an exponential, making it possible to detect departures from a hypothesized value of the mean more economically when IFR Weibull assumptions hold. This fact suggests that the resource-saving potential of life testing in an IFR Weibull environment is likely to arise as well for certain gamma and lognormal models as well.

There are a number of unresolved issues that must be left to future studies. Studies of alternative life testing designs, including type I censoring and sequential designs, would be of interest. Even within the framework of the present study, extension of our results and findings to various cases in which  $\alpha \neq \beta$  would be useful, as would expansion of our tables to larger values of the shape parameter, including, at a minimum,  $\alpha \in (3.0, 4.0)$ . Also, similar studies for other models, especially for the gamma family, would provide useful guides to alternative life testing methodology. In the context of the hybrid statistical problem discussed in the introductory section—where an alternative model is selected after data has been gathered according to an exponential life test plan—it would be important, in practice, to be able to carry out an appropriate analysis for the model identified as most suitable, be it the Weibull or some other failure-time model deemed to be applicable to the life testing experiment of interest.

TABLE 1 Life Test Sampling Plans for  $\alpha = .1 = \beta$ .

$\theta_1\theta_2$	$r_0$	$c/\theta_0$
2/3	41	.806
1/2	15	.687
1/3	6	.525
1/5	3	.367
1/10	2	.266









































TABLE 6 Shape Parameter A Corresponding to the Coefficient of Variation cv

cv	A	cv	A	cv	A
428.8314	0.10	0.7238	1.40	0.3994	2.70
47.0366	0.15	0.7006	1.45	0.3929	2.75
15.8430	0.20	0.6790	1.50	0.3866	2.80
8.3066	0.25	0.6588	1.55	0.3805	2.85
5.4077	0.30	0.6399	1.60	0.3747	2.90
3.9721	0.35	0.6222	1.65	0.3690	2.95
3.1409	0.40	0.6055	1.70	0.3634	3.00
2.6064	0.45	0.5897	1.75	0.3581	3.05
2.2361	0.50	0.5749	1.80	0.3529	3.10
1.9650	0.55	0.5608	1.85	0.3479	3.15
1.7581	0.60	0.5474	1.90	0.3430	3.20
1.5948	0.65	0.5348	1.95	0.3383	3.25
1.4624	0.70	0.5227	2.00	0.3336	3.30
1.3529	0.75	0.5112	2.05	0.3292	3.35
1.2605	0.80	0.5003	2.10	0.3248	3.40
1.1815	0.85	0.4898	2.15	0.3206	3.45
1.1130	0.90	0.4798	2.20	0.3165	3.50
1.0530	0.95	0.4703	2.25	0.3124	3.55
1.0000	1.00	0.4611	2.30	0.3085	3.60
0.9527	1.05	0.4523	2.35	0.3047	3.65
0.9102	1.10	0.4438	2.40	0.3010	3.70
0.8718	1.15	0.4341	2.45	0.2974	3.75
0.8369	1.20	0.4279	2.50	0.2938	3.80

*(Table continued from previous page)*

<i>cv</i>	A	<i>cv</i>	A	<i>cv</i>	A
0.8050	1.25	0.4204	2.55	0.2904	3.85
0.7757	1.30	0.4131	2.60	0.2870	3.90
0.7487	1.35	0.4062	2.65	0.2838	3.95







TABLE 10 Bounds on Shape Parameter Values Achieving Error Probabilities < .15.

$\theta_1 \theta_0$	.25	.50	.75
$r_{(1)}$	2.1	1.8	1.9
$r_{(2)}$	1.6	1.3	1.4
$r_{(3)}$	1.1	1.0	0.9
$r_{(4)}$	0.8	0.7	0.7
$r_{(5)}$	0.8	0.5	0.6



		0.70				0.60					
		n = 124	n = 41	n = 21	n = 10	n = 5	n = 145	n = 48	n = 25	n = 11	n = 6
		r	r	r	r	r	r	r	r	r	r
0.1 $\alpha$	0.06	0.40	0.08	0.31	0.06	0.28	0.11	0.36	0.05	0.38	0.09
	B	0.67	0.60	0.84	0.66	0.86	0.58	0.86	0.65	0.89	0.63
0.2 $\alpha$	0.14	0.28	0.09	0.30	0.10	0.29	0.10	0.29	0.07	0.35	0.06
	B	0.51	0.60	0.68	0.60	0.68	0.58	0.83	0.63	0.89	0.64
0.3 $\alpha$	0.14	0.24	0.07	0.22	0.12	0.24	0.08	0.33	0.08	0.25	0.04
	B	0.33	0.45	0.55	0.53	0.70	0.66	0.83	0.67	0.86	0.63
0.4 $\alpha$	0.12	0.21	0.08	0.16	0.15	0.28	0.11	0.21	0.09	0.27	0.07
	B	0.06	0.29	0.45	0.58	0.67	0.65	0.78	0.69	0.75	0.56
0.5 $\alpha$	0.11	0.24	0.09	0.18	0.09	0.17	0.12	0.20	0.08	0.26	0.11
	B	0.01	0.10	0.28	0.40	0.48	0.45	0.68	0.58	0.77	0.53
0.6 $\alpha$	0.11	0.17	0.06	0.20	0.12	0.19	0.16	0.18	0.11	0.21	0.09
	B	0.00	0.10	0.17	0.32	0.32	0.33	0.59	0.57	0.74	0.73
0.7 $\alpha$	0.16	0.17	0.09	0.15	0.12	0.22	0.09	0.14	0.12	0.23	0.08
	B	0.00	0.08	0.12	0.24	0.29	0.37	0.63	0.58	0.63	0.59
0.8 $\alpha$	0.06	0.09	0.12	0.13	0.09	0.15	0.13	0.27	0.11	0.18	0.16
	B	0.00	0.01	0.02	0.14	0.19	0.34	0.55	0.52	0.50	0.55
0.9 $\alpha$	0.14	0.15	0.06	0.08	0.07	0.13	0.16	0.17	0.08	0.14	0.11
	B	0.00	0.00	0.04	0.14	0.13	0.28	0.41	0.42	0.70	0.57
1.0 $\alpha$	0.11	0.11	0.10	0.10	0.09	0.09	0.04	0.14	0.07	0.17	0.05
	B	0.00	0.00	0.01	0.07	0.04	0.20	0.35	0.41	0.50	0.52
1.1 $\alpha$	0.14	0.17	0.07	0.19	0.08	0.11	0.13	0.20	0.09	0.14	0.06
	B	0.00	0.00	0.00	0.03	0.07	0.20	0.29	0.40	0.54	0.43
1.2 $\alpha$	0.14	0.11	0.11	0.13	0.17	0.22	0.14	0.15	0.12	0.15	0.08
	B	0.00	0.00	0.00	0.03	0.01	0.14	0.21	0.37	0.41	0.45
1.3 $\alpha$	0.10	0.08	0.06	0.10	0.07	0.06	0.07	0.15	0.15	0.21	0.13
	B	0.00	0.00	0.00	0.03	0.00	0.13	0.16	0.31	0.38	0.33
1.4 $\alpha$	0.09	0.10	0.16	0.11	0.09	0.08	0.11	0.15	0.10	0.15	0.06
	B	0.00	0.00	0.00	0.00	0.00	0.06	0.07	0.21	0.33	0.32
1.5 $\alpha$	0.07	0.10	0.14	0.15	0.12	0.16	0.15	0.17	0.10	0.16	0.09
	B	0.00	0.00	0.00	0.01	0.00	0.04	0.05	0.18	0.24	0.36
1.6 $\alpha$	0.13	0.09	0.06	0.07	0.06	0.06	0.05	0.13	0.07	0.11	0.08
	B	0.00	0.00	0.00	0.01	0.00	0.05	0.10	0.20	0.26	0.37
1.7 $\alpha$	0.11	0.08	0.14	0.17	0.06	0.10	0.14	0.16	0.08	0.08	0.12
	B	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.13	0.18	0.20
1.8 $\alpha$	0.13	0.12	0.08	0.08	0.10	0.12	0.11	0.11	0.10	0.15	0.12
	B	0.00	0.00	0.00	0.00	0.01	0.04	0.12	0.13	0.21	0.27
1.9 $\alpha$	0.11	0.10	0.10	0.11	0.07	0.09	0.09	0.10	0.11	0.18	0.13
	B	0.00	0.00	0.00	0.01	0.00	0.00	0.05	0.13	0.22	0.27
2.0 $\alpha$	0.11	0.11	0.19	0.16	0.13	0.12	0.09	0.17	0.09	0.15	0.10
	B	0.00	0.00	0.00	0.00	0.00	0.01	0.08	0.16	0.00	0.00
2.1 $\alpha$	0.11	0.07	0.15	0.12	0.07	0.08	0.09	0.16	0.09	0.13	0.06
	B	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.17	0.00	0.00
2.2 $\alpha$	0.11	0.09	0.09	0.10	0.13	0.17	0.10	0.12	0.11	0.18	0.15
	B	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.15	0.00	0.00
2.3 $\alpha$	0.11	0.10	0.10	0.10	0.07	0.07	0.10	0.12	0.12	0.11	0.15
	B	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.09	0.00	0.00
2.4 $\alpha$	0.13	0.11	0.11	0.14	0.08	0.06	0.11	0.12	0.07	0.12	0.11
	B	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.13	0.00	0.00
2.5 $\alpha$	0.10	0.10	0.07	0.06	0.06	0.07	0.08	0.12	0.06	0.09	0.09
	B	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.08	0.00	0.00
2.6 $\alpha$	0.09	0.10	0.11	0.11	0.12	0.13	0.16	0.15	0.14	0.17	0.12
	B	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.08	0.00	0.00
2.7 $\alpha$	0.09	0.09	0.08	0.08	0.06	0.09	0.14	0.09	0.11	0.12	0.17
	B	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.11	0.00	0.00
2.8 $\alpha$	0.07	0.06	0.09	0.07	0.11	0.11	0.09	0.12	0.10	0.12	0.21
	B	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.06	0.00	0.00
2.9 $\alpha$	0.11	0.13	0.10	0.13	0.13	0.12	0.11	0.10	0.06	0.09	0.17
	B	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.00	0.00
3.0 $\alpha$	0.11	0.11	0.13	0.12	0.12	0.11	0.08	0.09	0.14	0.09	0.11
	B	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03

	0.50					0.40				
	n = 174	n = 58	n = 30	n = 14	n = 8	n = 217	n = 72	n = 37	n = 17	n = 10
	r = 87	r = 29	r = 15	r = 7	r = 4	r = 87	r = 29	r = 15	r = 7	r = 4
	kno lse									
0.1 $\alpha$	0.14 0.36	0.09 0.33	0.04 0.34	0.08 0.36	0.12 0.43	0.10 0.41	0.07 0.41	0.09 0.44	0.10 0.37	0.11 0.48
B	0.70 0.54	0.84 0.60	0.86 0.57	0.87 0.53	0.93 0.53	0.74 0.61	0.90 0.63	0.89 0.64	0.87 0.58	0.88 0.52
0.2 $\alpha$	0.10 0.28	0.11 0.28	0.05 0.41	0.13 0.35	0.10 0.36	0.10 0.40	0.09 0.34	0.08 0.35	0.07 0.36	0.10 0.54
B	0.49 0.52	0.78 0.68	0.77 0.70	0.83 0.58	0.83 0.56	0.53 0.61	0.76 0.65	0.72 0.51	0.82 0.56	0.85 0.57
0.3 $\alpha$	0.10 0.22	0.09 0.39	0.11 0.40	0.12 0.37	0.07 0.31	0.12 0.29	0.13 0.36	0.06 0.35	0.10 0.40	0.13 0.44
B	0.23 0.46	0.54 0.52	0.74 0.65	0.76 0.60	0.91 0.63	0.27 0.44	0.61 0.53	0.70 0.59	0.76 0.58	0.83 0.59
0.4 $\alpha$	0.10 0.29	0.09 0.20	0.05 0.22	0.08 0.31	0.03 0.34	0.10 0.29	0.11 0.24	0.13 0.36	0.07 0.38	0.18 0.39
B	0.14 0.31	0.45 0.44	0.70 0.58	0.79 0.56	0.77 0.56	0.07 0.38	0.40 0.44	0.60 0.53	0.67 0.58	0.77 0.58
0.5 $\alpha$	0.16 0.26	0.07 0.25	0.15 0.27	0.08 0.25	0.08 0.33	0.11 0.35	0.06 0.21	0.12 0.30	0.10 0.40	0.11 0.30
B	0.03 0.24	0.33 0.33	0.43 0.52	0.61 0.60	0.72 0.57	0.02 0.28	0.28 0.47	0.52 0.56	0.72 0.53	0.71 0.47
0.6 $\alpha$	0.15 0.29	0.11 0.20	0.16 0.25	0.11 0.30	0.09 0.35	0.05 0.23	0.09 0.27	0.11 0.27	0.07 0.34	0.07 0.31
B	0.00 0.14	0.17 0.32	0.44 0.53	0.64 0.56	0.74 0.57	0.00 0.11	0.15 0.39	0.37 0.44	0.67 0.53	0.67 0.48
0.7 $\alpha$	0.08 0.18	0.07 0.15	0.11 0.25	0.14 0.28	0.09 0.21	0.10 0.22	0.17 0.34	0.11 0.23	0.13 0.28	0.10 0.33
B	0.00 0.12	0.07 0.34	0.33 0.47	0.54 0.53	0.77 0.59	0.00 0.12	0.09 0.36	0.30 0.41	0.51 0.44	0.67 0.55
0.8 $\alpha$	0.07 0.18	0.03 0.25	0.10 0.20	0.11 0.25	0.10 0.24	0.11 0.20	0.09 0.22	0.07 0.24	0.08 0.30	0.15 0.28
B	0.00 0.08	0.05 0.27	0.21 0.30	0.53 0.44	0.64 0.52	0.00 0.07	0.02 0.22	0.17 0.33	0.49 0.44	0.64 0.48
0.9 $\alpha$	0.10 0.25	0.12 0.20	0.12 0.21	0.06 0.19	0.09 0.23	0.10 0.30	0.04 0.29	0.05 0.25	0.10 0.25	0.08 0.28
B	0.00 0.01	0.04 0.22	0.10 0.29	0.34 0.42	0.62 0.45	0.00 0.04	0.01 0.19	0.14 0.26	0.49 0.54	0.65 0.50
1.0 $\alpha$	0.11 0.20	0.12 0.21	0.09 0.19	0.12 0.24	0.14 0.25	0.12 0.30	0.10 0.21	0.09 0.17	0.13 0.23	0.09 0.37
B	0.00 0.00	0.00 0.05	0.09 0.24	0.37 0.51	0.56 0.45	0.00 0.03	0.00 0.11	0.05 0.23	0.36 0.47	0.48 0.42
1.1 $\alpha$	0.10 0.24	0.09 0.15	0.12 0.21	0.07 0.17	0.07 0.24	0.09 0.25	0.06 0.29	0.12 0.21	0.04 0.26	0.09 0.28
B	0.00 0.01	0.00 0.11	0.05 0.23	0.28 0.45	0.48 0.48	0.00 0.03	0.00 0.16	0.09 0.28	0.22 0.38	0.45 0.42
1.2 $\alpha$	0.10 0.16	0.14 0.24	0.04 0.12	0.09 0.26	0.10 0.24	0.11 0.23	0.06 0.29	0.08 0.14	0.08 0.19	0.11 0.34
B	0.00 0.01	0.00 0.07	0.01 0.16	0.21 0.28	0.51 0.42	0.00 0.00	0.00 0.15	0.05 0.18	0.14 0.39	0.45 0.34
1.3 $\alpha$	0.08 0.21	0.08 0.21	0.08 0.17	0.10 0.25	0.12 0.25	0.09 0.25	0.09 0.21	0.09 0.20	0.10 0.34	0.14 0.27
B	0.00 0.00	0.00 0.06	0.03 0.14	0.13 0.31	0.35 0.41	0.00 0.00	0.00 0.09	0.05 0.18	0.13 0.34	0.36 0.43
1.4 $\alpha$	0.10 0.15	0.06 0.15	0.06 0.21	0.11 0.22	0.14 0.25	0.08 0.21	0.09 0.16	0.07 0.15	0.07 0.23	0.14 0.27
B	0.00 0.00	0.00 0.02	0.00 0.12	0.08 0.26	0.27 0.33	0.00 0.00	0.00 0.03	0.00 0.11	0.15 0.26	0.26 0.40
1.5 $\alpha$	0.04 0.13	0.06 0.15	0.11 0.26	0.09 0.22	0.09 0.23	0.10 0.22	0.12 0.22	0.05 0.24	0.11 0.24	0.05 0.19
B	0.00 0.00	0.00 0.03	0.01 0.06	0.07 0.27	0.35 0.40	0.00 0.01	0.00 0.04	0.00 0.14	0.05 0.27	0.23 0.25
1.6 $\alpha$	0.12 0.11	0.05 0.17	0.15 0.22	0.09 0.17	0.10 0.18	0.13 0.23	0.09 0.22	0.12 0.18	0.09 0.13	0.11 0.28
B	0.00 0.00	0.00 0.02	0.00 0.10	0.02 0.17	0.22 0.29	0.00 0.00	0.00 0.04	0.00 0.08	0.08 0.20	0.23 0.35
1.7 $\alpha$	0.14 0.18	0.09 0.14	0.10 0.18	0.11 0.19	0.09 0.21	0.11 0.20	0.12 0.16	0.12 0.21	0.11 0.24	0.08 0.29
B	0.00 0.00	0.00 0.01	0.00 0.06	0.03 0.27	0.20 0.33	0.00 0.00	0.00 0.02	0.00 0.07	0.03 0.22	0.16 0.31
1.8 $\alpha$	0.07 0.10	0.08 0.10	0.12 0.19	0.09 0.18	0.08 0.20	0.04 0.13	0.12 0.21	0.12 0.18	0.12 0.22	0.10 0.28
B	0.00 0.00	0.00 0.00	0.00 0.06	0.01 0.14	0.16 0.35	0.00 0.00	0.00 0.01	0.00 0.11	0.01 0.24	0.19 0.27
1.9 $\alpha$	0.08 0.13	0.10 0.15	0.08 0.14	0.11 0.17	0.13 0.32	0.20 0.30	0.14 0.24	0.12 0.15	0.08 0.16	0.06 0.18
B	0.00 0.00	0.00 0.03	0.00 0.03	0.00 0.15	0.08 0.31	0.00 0.00	0.00 0.00	0.00 0.05	0.04 0.17	0.15 0.22
2.0 $\alpha$	0.08 0.09	0.14 0.15	0.03 0.10	0.13 0.19	0.04 0.16	0.08 0.17	0.13 0.13	0.10 0.17	0.07 0.21	0.06 0.19
B	0.00 0.00	0.00 0.00	0.00 0.03	0.01 0.10	0.06 0.29	0.00 0.00	0.00 0.01	0.00 0.05	0.00 0.16	0.08 0.21
2.1 $\alpha$	0.10 0.11	0.11 0.17	0.12 0.23	0.18 0.21	0.08 0.23	0.17 0.15	0.08 0.18	0.11 0.15	0.17 0.23	0.14 0.24
B	0.00 0.00	0.00 0.00	0.00 0.01	0.00 0.14	0.08 0.23	0.00 0.00	0.00 0.00	0.00 0.01	0.01 0.18	0.06 0.28
2.2 $\alpha$	0.09 0.13	0.06 0.12	0.12 0.17	0.07 0.16	0.10 0.27	0.08 0.19	0.12 0.18	0.11 0.20	0.11 0.15	0.10 0.23
B	0.00 0.00	0.00 0.00	0.00 0.02	0.00 0.07	0.05 0.20	0.00 0.00	0.00 0.00	0.00 0.04	0.00 0.12	0.06 0.21
2.3 $\alpha$	0.15 0.21	0.11 0.19	0.13 0.18	0.07 0.12	0.05 0.09	0.06 0.16	0.11 0.21	0.15 0.18	0.16 0.24	0.09 0.31
B	0.00 0.00	0.00 0.00	0.00 0.02	0.00 0.05	0.04 0.24	0.00 0.00	0.00 0.00	0.00 0.01	0.00 0.14	0.04 0.25
2.4 $\alpha$	0.06 0.14	0.09 0.13	0.10 0.10	0.08 0.09	0.10 0.20	0.11 0.16	0.06 0.09	0.14 0.19	0.11 0.25	0.13 0.24
B	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.07	0.00 0.16	0.00 0.00	0.00 0.00	0.00 0.01	0.00 0.08	0.04 0.19
2.5 $\alpha$	0.07 0.16	0.13 0.20	0.10 0.10	0.09 0.17	0.16 0.23	0.06 0.08	0.11 0.15	0.08 0.25	0.11 0.20	0.11 0.23
B	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.02	0.00 0.22	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.07	0.00 0.21
2.6 $\alpha$	0.08 0.11	0.08 0.15	0.09 0.10	0.12 0.15	0.15 0.20	0.11 0.16	0.08 0.18	0.09 0.23	0.17 0.23	0.15 0.31
B	0.00 0.00	0.00 0.00	0.00 0.01	0.00 0.06	0.00 0.14	0.00 0.00	0.00 0.00	0.00 0.03	0.00 0.07	0.01 0.14
2.7 $\alpha$	0.11 0.15	0.11 0.09	0.05 0.19	0.05 0.12	0.08 0.20	0.07 0.17	0.13 0.19	0.11 0.19	0.07 0.16	0.12 0.21
B	0.00 0.00	0.00 0.00	0.00 0.06	0.00 0.17	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.04	0.02 0.19	
2.8 $\alpha$	0.12 0.11	0.08 0.18	0.08 0.13	0.08 0.14	0.12 0.18	0.13 0.20	0.10 0.14	0.08 0.19	0.06 0.20	0.10 0.25
B	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.13	0.00 0.00	0.00 0.00	0.00 0.02	0.00 0.09	0.00 0.10	
2.9 $\alpha$	0.10 0.13	0.10 0.19	0.07 0.15	0.09 0.15	0.09 0.22	0.07 0.16	0.09 0.16	0.09 0.13	0.07 0.17	0.12 0.17
B	0.00 0.00	0.00 0.00	0.00 0.04	0.00 0.12	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.03	0.00 0.11	
3.0 $\alpha$	0.05 0.11	0.12 0.17	0.10 0.17	0.03 0.11	0.14 0.22	0.09 0.14	0.09 0.14	0.07 0.15	0.10 0.23	0.10 0.21
B	0.00 0.00	0.00 0.00	0.00 0.01	0.00 0.05	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.03	0.00 0.13	

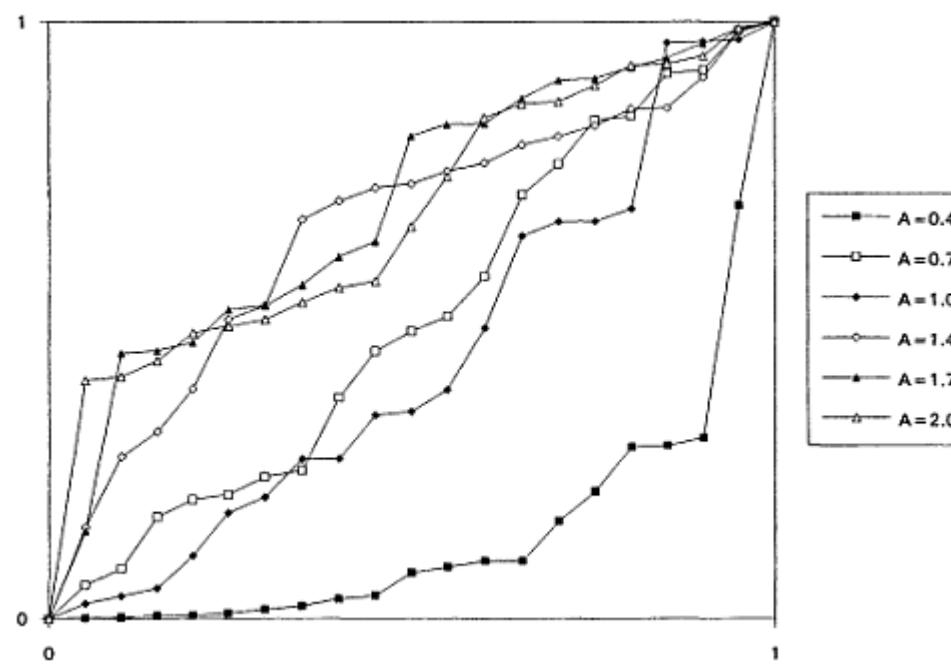


Figure 1  
First simulation.

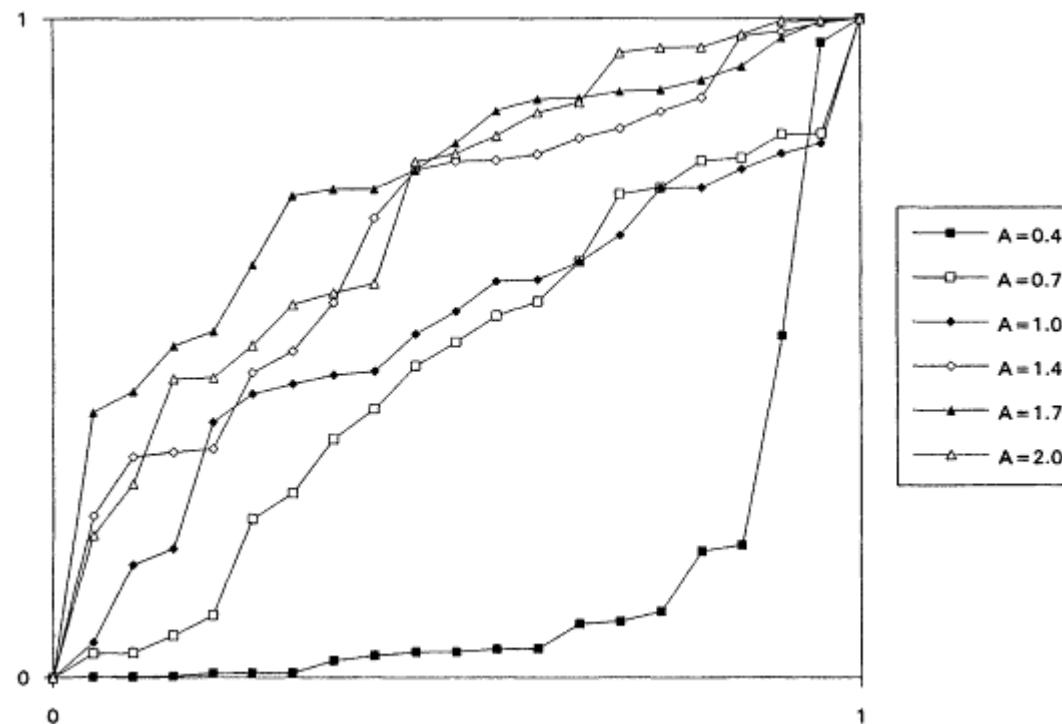


Figure 2  
Second simulation.

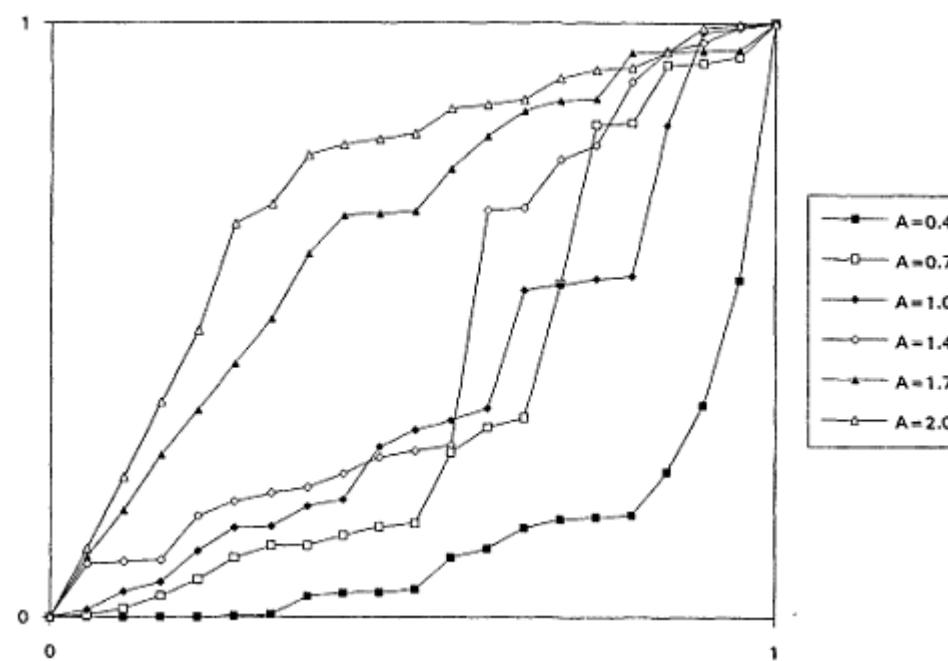


Figure 3  
Third simulation.

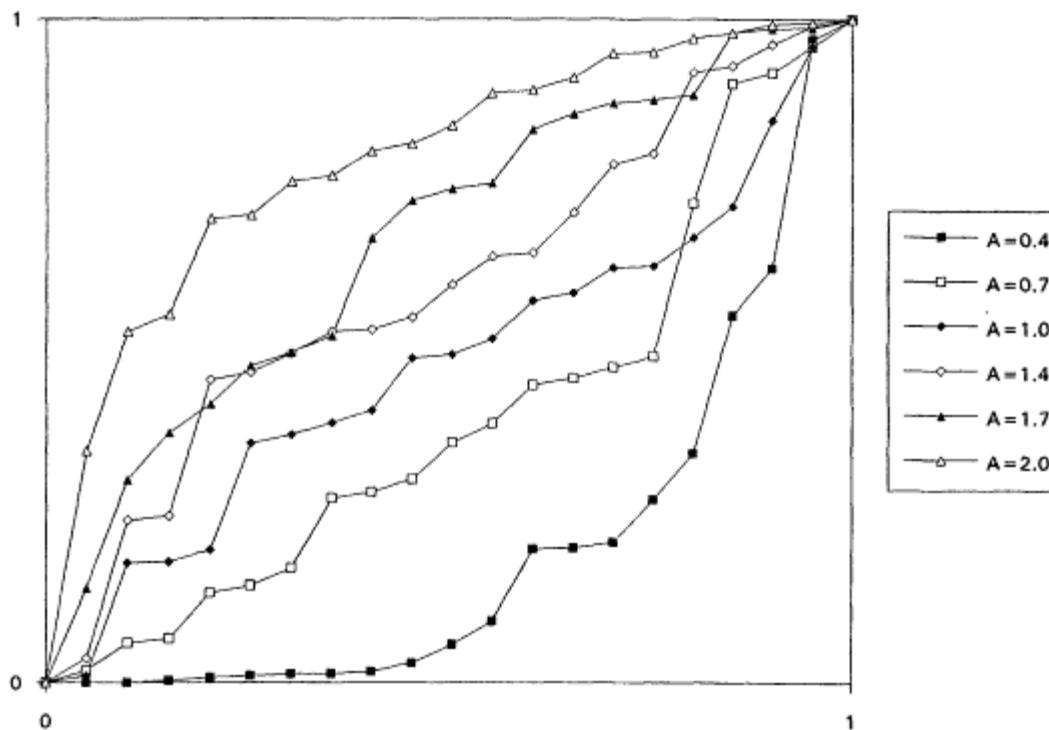


Figure 4  
Fourth simulation.

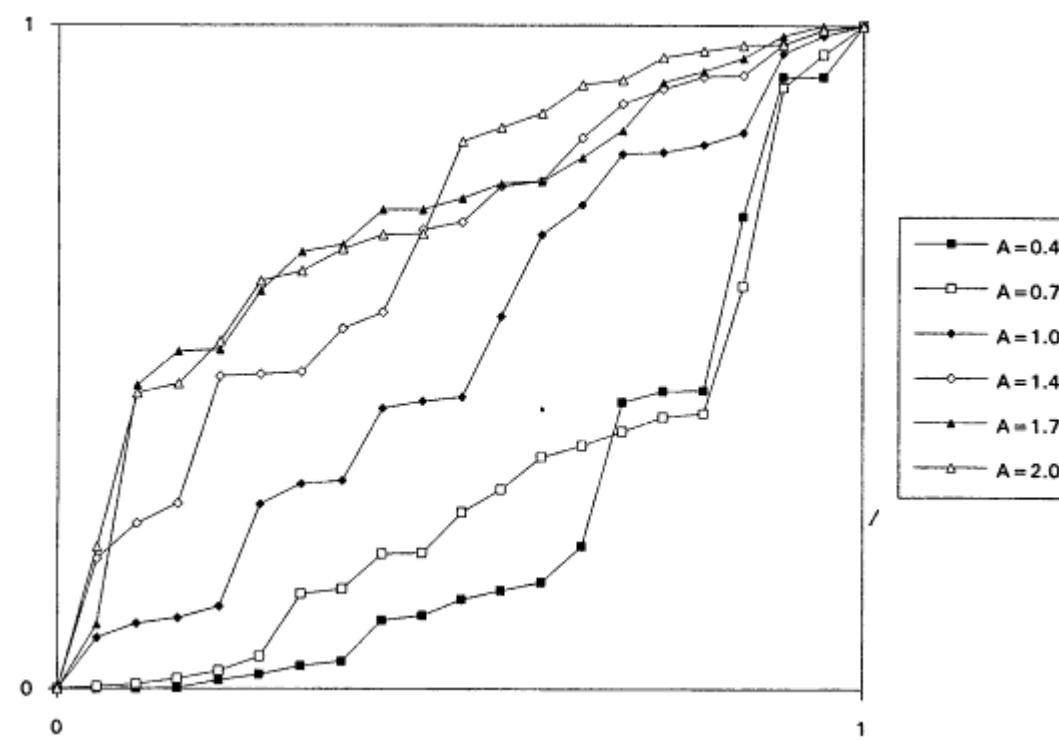


Figure 5  
Fifth simulation.

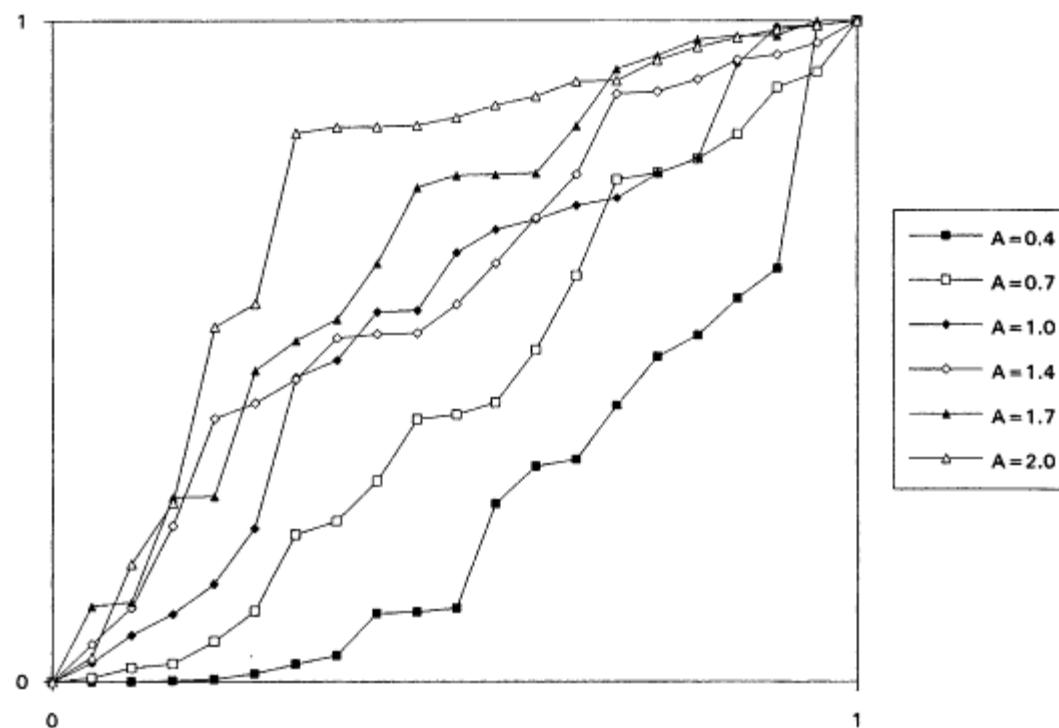


FIGURE 6  
Sixth simulation.

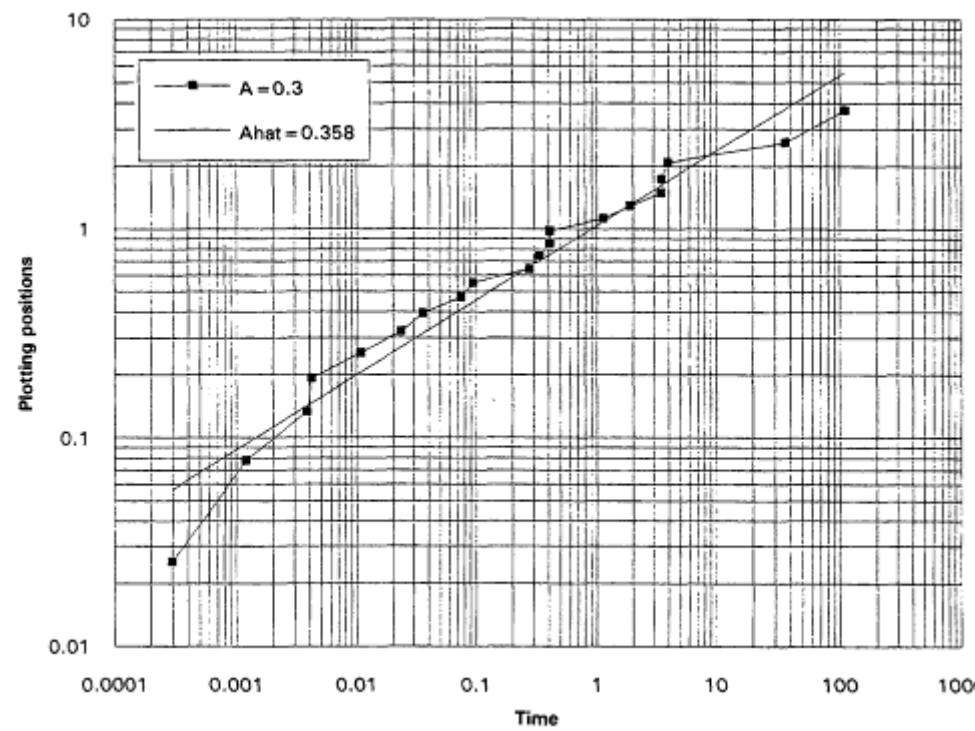


Figure 7  
First plot.

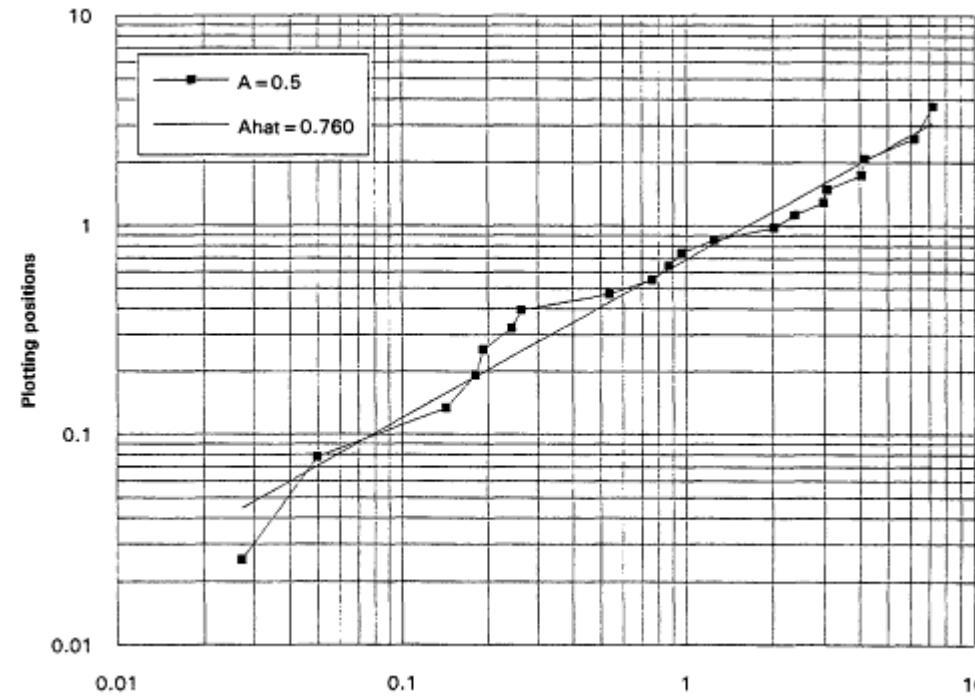


Figure 8  
Second plot.

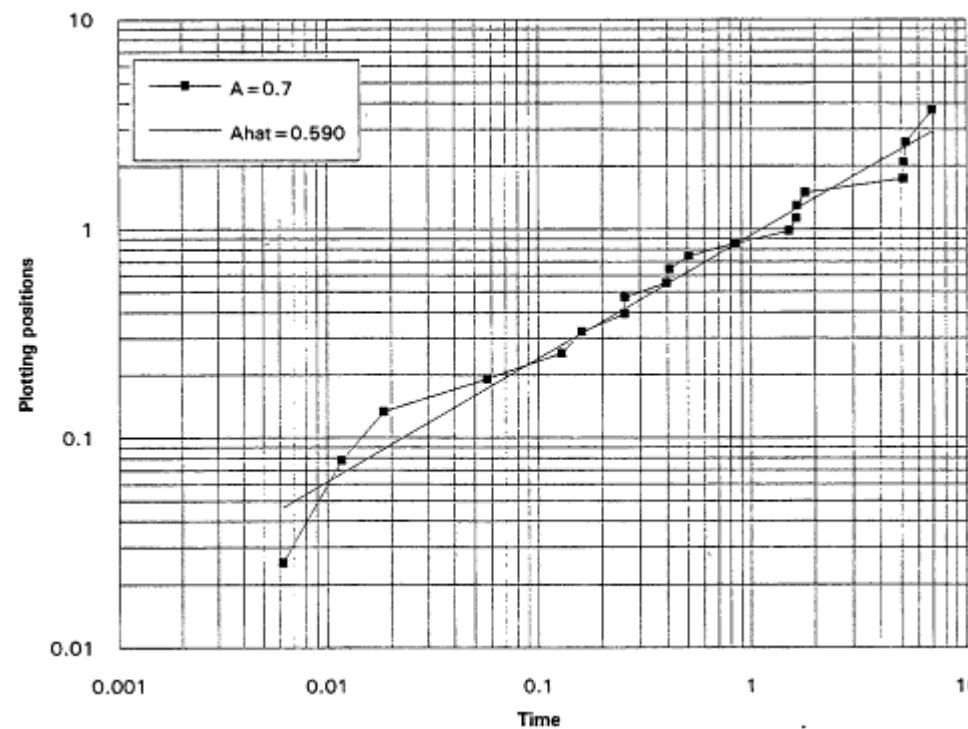


Figure 9  
Third plot.

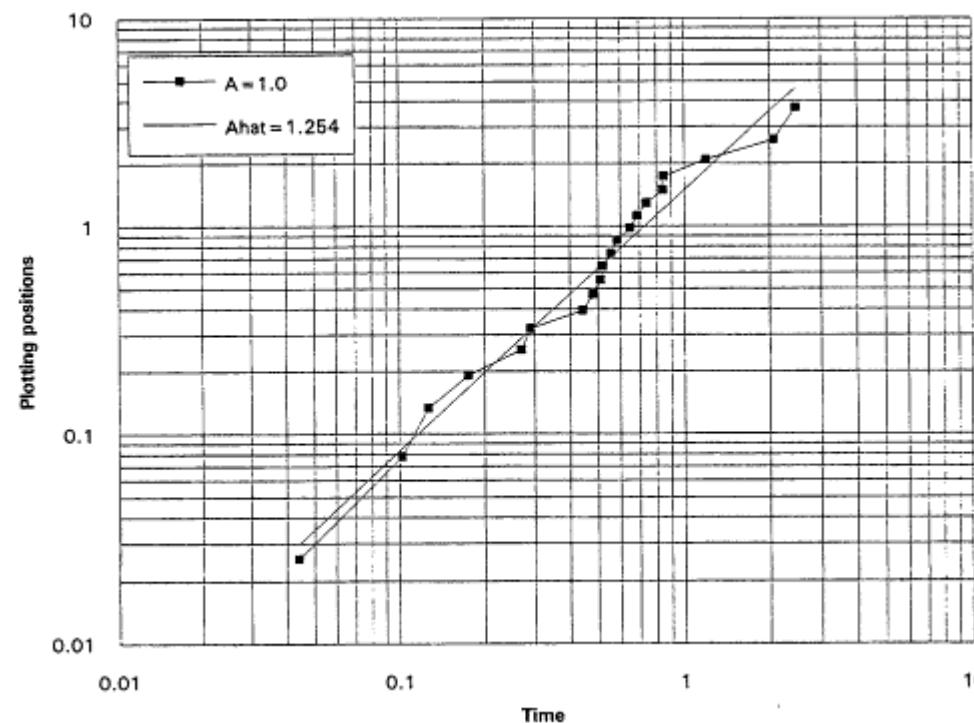


Figure 10  
Fourth plot.

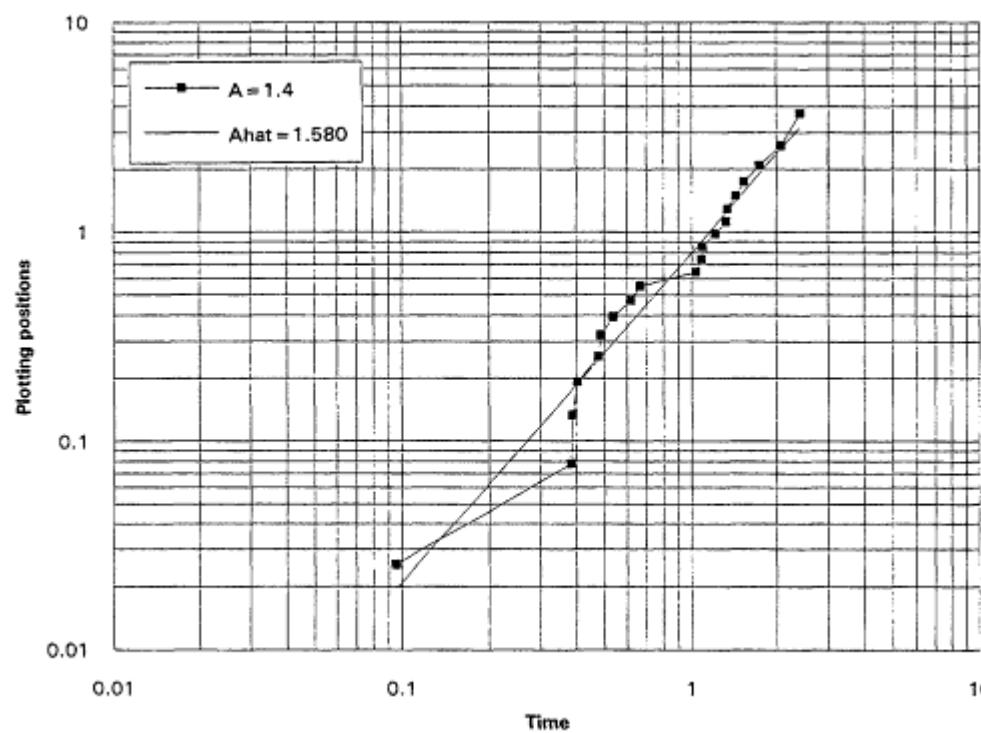


Figure 11  
Fifth plot.

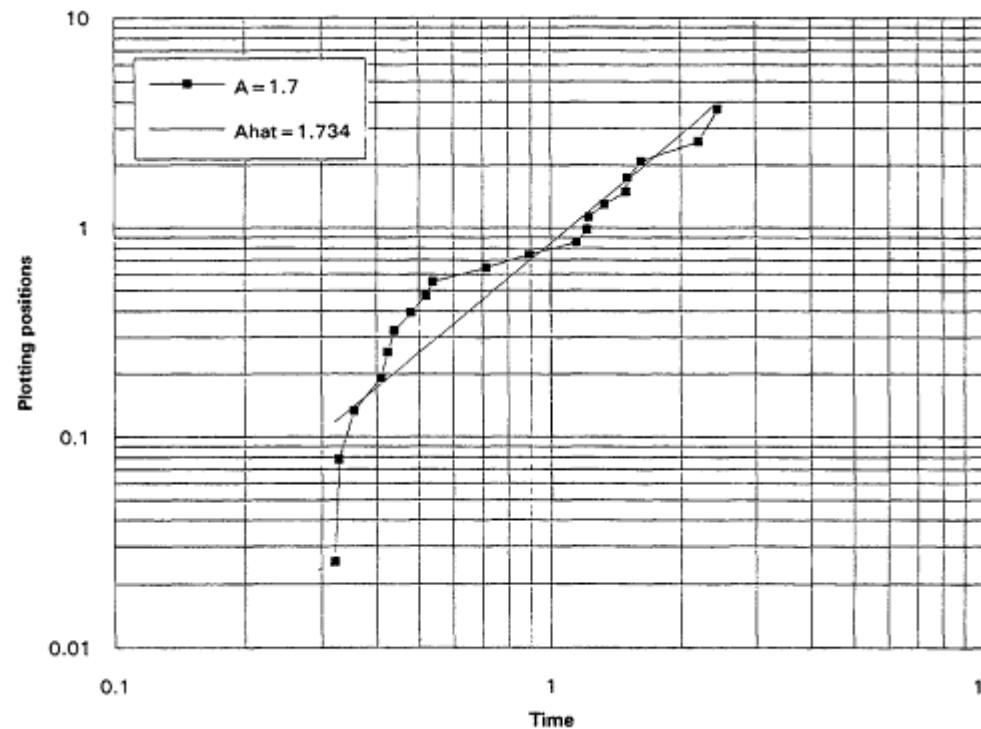


Figure 12  
Sixth plot.

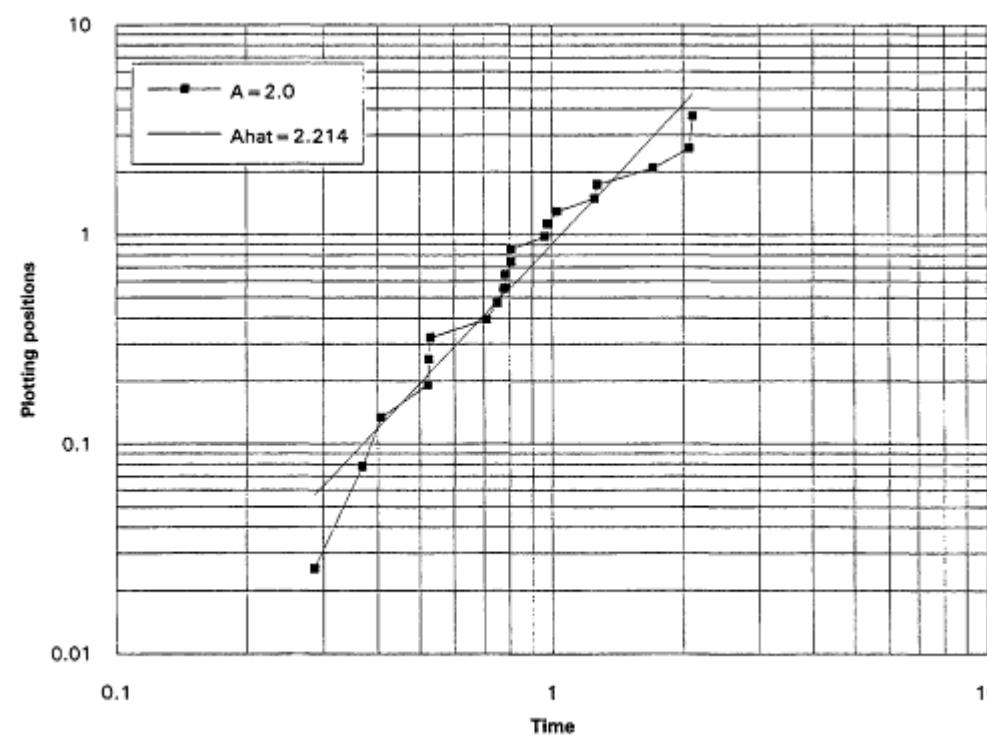


Figure 13  
Seventh plot.

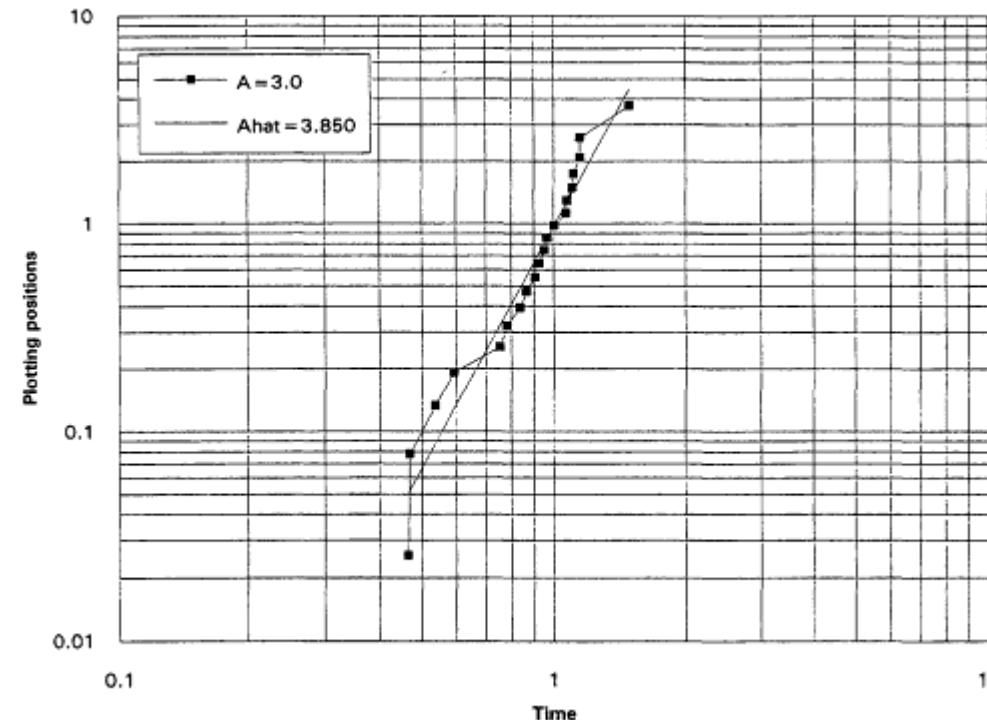


Figure 14  
Eighth plot.

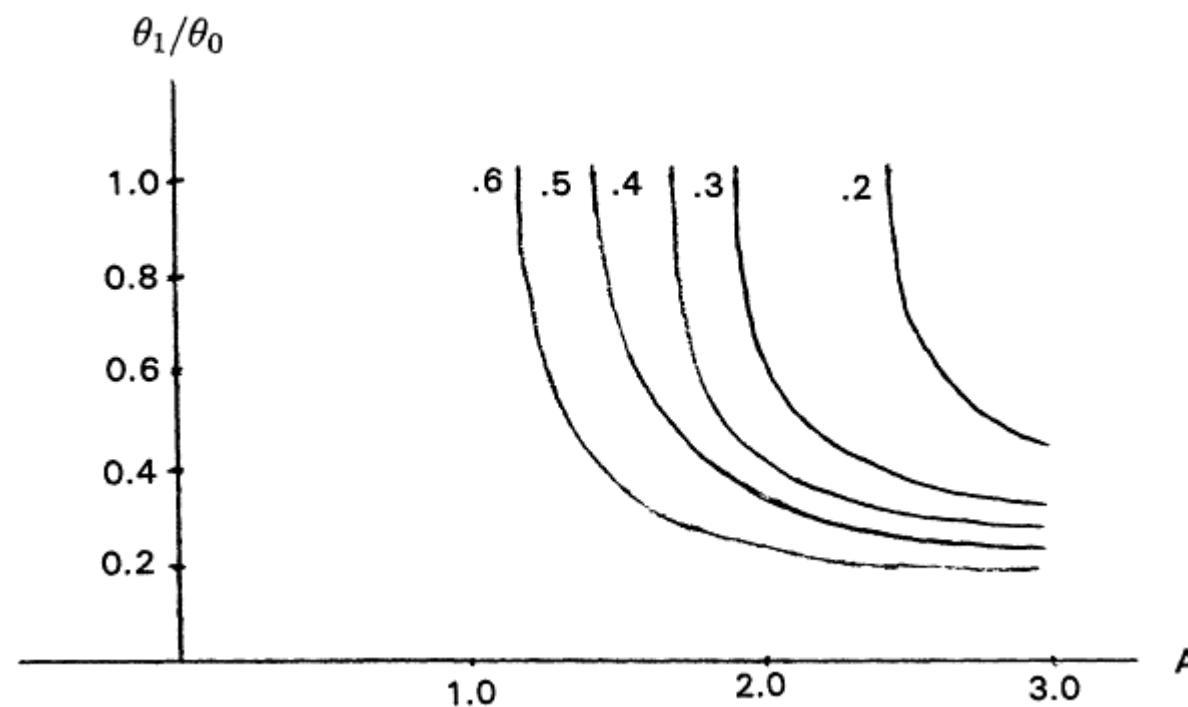


Figure 15  
Level curves for TTTR when  $\alpha = \beta = .1$ .

## REFERENCES

- Abramowitz, M. and Stegun, I.A. (eds.) 1964 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Wiley and Sons.
- Alfers, D., and Dinges, H. 1984 A normal approximation for beta and gamma tail probabilities. *Zeit. Wahr.* 65:399-419.
- American Statistical Association 1975 *Current Index to Statistics*: Volumes 1-19. Alexandria, Va.: American Statistical -93 Association.
- Anderson, T.P. 1994 Current Issues Concerning Reliability Estimation in Operations Test and Evaluation Unpublished Master's Thesis, Naval Postgraduate School, Monterey, Calif.
- Bain, L.J., and M. Engelhardt 1991 *Statistical Analysis of Reliability and Life Testing Models, Theory and Methods*. Second edition. New York: Dekker.
- Bain, L.J., and D.L. Weeks 1965 Tolerance limits for the generalized gamma distribution. *Journal of the American Statistical Association* 60:1142-1152.
- Barlow, R.E. 1979 Geometry of the total time on test transformation. *Naval Research Logistics Quarterly* 26:393-402.
- Barlow, R.E., D. Bartholomew, J. Bremner, and H. Brunk 1972 *Statistical Inference Under Order Restrictions*. New York: John Wiley and Sons.

- Barlow, R.E., and R. Campo 1975 Total time on test processes and applications to failure data analysis. Pp. 451-481 in R.E. Barlow, R. Fussell, and N.D. Singpurwalla, eds., *Reliability and Fault Tree Analysis*. Philadelphia, Pa.: SIAM.
- Barlow, R.E., and F. Proschan 1969 A note on tests for monotone failure rate based on incomplete data. *Annals of Mathematical Statistics* 40:595-600.
- 1975 *Statistical Theory of Reliability and Life Testing*. New York: Holt, Reinhart and Winston.
- Barlow, R.E., R.H. Toland, and T. Freeman 1988 A Bayesian analysis of the stress-rupture life of Kevlar/Eposy spherical pressure vessels. In C. Clarotti and D. Lindley, eds., *Accelerated Life Testing and Experts' Opinions in Reliability*. Lindley, Amsterdam: North-Holland.
- Chandra, M., and N.D. Singpurwalla 1981 Relationships between some notions which are common to reliability theory and economics. *Math. Operational Research* 6:113-121.
- Chernoff, H., and G. Lieberman 1956 The use of generalized probability paper for continuous distributions. *Annals of Mathematical Statistics* 27:806-818.
- Cramér, H. 1946 *Mathematical Methods of Statistics*. Princeton: Princeton University Press.
- Gong, G., and F.J. Samaniego 1981 Pseudo maximum likelihood estimation: Theory and methods. *Annals of Statistics* 9:861-869.
- Hardy, G., J. Littlewood, and G. Pólya 1929 Some simple inequalities satisfied by convex functions. *Messenger Mathematics* 58:145-152.

- Hollander, M., and F. Proschan 1984 Nonparametric concepts and methods in reliability. Pp. 613-655 in P.R. Krishniah, and P.K. Sen, eds., *Handbook of Statistics, Volume 4: Nonparametric Methods*. Amsterdam: North-Holland.
- Klefsjö, B. 1980 Some tests against aging based on the total time on test transform. *Statist. Res. Report* No. 1979-4, University of Umeå (Umeå, Sweden).
- Lawless, J.F. 1975 Construction of tolerance bounds for the extreme value and Weibull distributions. *Technometrics* 17:255-261.
- 1982 *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons.
- Lawless, J.F., and C. Nadeau 1995 Some simple robust methods for the analysis of recurrent events. *Technometrics* 37:158-168.
- Luenberger, D. 1989 *Linear and Nonlinear Programming*. Second edition. Reading, Mass.: Addison Wesley.
- Mann, N.R., R.E. Schafer, and N.D. Singpurwalla 1974 *Methods for Statistical Analysis of Reliability and Lifetime Data*. New York: John Wiley and Sons.
- Marshall, A., and I. Olkin 1979 *Inequalities: The Theory of Majorization and Its Applications*. New York: Academic Press.
- Nair, V. 1984 On the behavior of some estimators from probability plots. *Journal of the American Statistical Association* 79:823-831.

- Neath, A., and F. Samaniego 1992 On the total time on test transforms of an IFRA distribution. *Statistics and Probability Letters* 14:289-291.
- Nelson, W. 1982 *Applied Data Analysis*. New York: John Wiley and Sons.
- 1990 *Accelerated Testing: Statistical Models, Test Plans and Data Analyses*. New York: John Wiley and Sons.
- 1995 Confidence limits for recurrence data—Applied to cost a number of product repairs. *Technometrics* 37:147-157.
- Neyman, J. 1959 Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, ed., *Probability and Statistics, the Harold Cramér Volume*. New York: John Wiley and Sons.
- Press, W., S. Teukolsky, W. Vetterling, and B. Flannery 1992 *Numerical Recipes in C: The Art of Scientific Computing*. Second edition. Cambridge, U.K.: Cambridge University Press.
- Rolph, John E. and Duane L. Steffey, eds. 1994 *Statistical Issues in Defense Analysis and Testing: Summary of a Workshop*. Committee on National Statistics and Committee on Applied and Theoretical Statistics, National Research Council. Washington, D.C.: National Academy Press.
- Samaniego, F.J. 1993 On the Needs of the DoD Testing Community and the Expertise in the Statistical Research Community: A Look at the Interface. Technical Report #286, Division of Statistics, University of California, Davis, Calif.
- Samaniego, F.J., and L.R. Whittaker 1986 On estimating population characteristics from record-breaking observations: I. parametric results. *Naval Research Logistics Quarterly* 33:531-543.

- Sinha, S.K. 1987 *Reliability and Life Testing*. New York: John Wiley and Sons.
- Sinha, S.K., and B.K. Kale 1979 *Life Testing and Reliability Estimation*. New Delhi: Wiley Eastern Limited. Thoman, D.R., L.J. Bain, and C.E. Antle 1969 Inferences on the parameters of the Weibull distribution. *Technometrics* 11:445-460.
- U.S. Department of Defense 1960 *Handbook H108: Sampling Procedures and Tables for Life and Reliability Testing (Based on the Exponential Distribution)*. Washington, D.C.: U.S. Department of Defense.
- Woods, W.M. 1996 Using wearout information to reduce reliability demonstration test time, Proceedings of the First Annual U.S. Army Conference on Applied Statistics, Army Research Laboratory, Adelphi, MD, Publication QRL-SR-43.
- Zelen, M., and M. Dannemiller 1961 The robustness of life testing procedures derived from the exponential distribution. *Technometrics* 3:29-49.

# 3

## Application of Statistical Science to Testing and Evaluating Software Intensive Systems

*Jesse H. Poore, University of Tennessee; and Carmen J. Trammell, CTI PET Systems, Inc.*

### 1. INTRODUCTION

Any large, complex, expensive process with myriad ways to do most activities, as is the case with software development, can have its cost-benefit profile dramatically improved by the use of statistical science. Statistics provides a structure for collecting data and transforming it into information that can improve decision making under uncertainty. The term "statistical testing" as typically used in the software engineering literature has the narrow reference to randomly generated test cases. The term should be understood, however, as the comprehensive application of statistical science, including operations research methods, to solving the problems posed by industrial software testing. Statistical testing enables efficient collection of empirical data that will remove uncertainty about the behavior of the software intensive system and support economic decisions regarding further testing, deployment, maintenance and evolution. The operational *usage model* is a formalism that enables the application of many statistical principles to software testing and forms the basis for efficient testing in support of decision making.

The software testing problem is complex because of the astronomical number of scenarios of use and states of use. The domain of testing is large and complex beyond human intuition. Because the software testing problem is so complex, statistical principles should be used to guide testing strategy. In general, the concept of "testing in quality" is costly and ineffectual; software quality is achieved in the requirements, architecture, specification, design and coding activities. Although not within the scope of this essay, verification or reading techniques (Basili et al., 1996) are of critical importance to achieving quality software and may efficiently and effectively displace some testing. The problem of doing just enough testing to

remove uncertainty regarding critical performance issues, and to support a decision that the system is of requisite quality for its mission, environment or market, is a problem amenable to solution by statistical science. The question is not whether to test, but when to test, what to test and how much to test.

Testing can be justified at many different stages in the life cycle of a software intensive system. There is, for example, testing at various stages in development, testing of reusable components, testing associated with product enhancements or repairs, testing of a product ported from one hardware system to another, and "customer testing" (i.e., field experience). Service in the field is the very best "testing" information, because it is real use, often extensive, and free except for the cost of data collection. The usage model can be the framework, the common denominator, for combining test and usage experience across different software engineering methods and life cycle phases so that maximum use can be made of all available testing and field-use information.

A statistical principle of fundamental importance is that a population to be studied must first be characterized, and that characterization must include the infrequent and exceptional as well as the common and typical. It must be possible to represent all questions of interest and all decisions to be made in terms of this characterization. All experimental design methods require such a characterization and representation, in one form or another, at a suitable level of abstraction. When applied to software testing, the population is the set of all possible scenarios of use with each accurately represented as to frequency of occurrence.

One such method of characterization and representation is the operational usage model. The states-of-use of the system and the allowable transitions among those states are identified, and the probability of making each allowable transition is determined. These models are then represented in the form of one or more highly structured Markov chains (a type of statistical model, see e.g. Kemeny and Snell, 1960), and the result is called a usage model (Whittaker and Poore, 1993; Whittaker and Thomason, 1994).

From a statistical point of view, all of the topics in this paper follow sound problem solving principles and are direct applications of well-established theory and methodology. From a software testing point of view, the applications of statistical science discussed below are not in

widespread use, nor is the full process presented here in routine use. Many methods and segments of the process are used in isolated pockets of industry, on both experimental and routine bases. This paper is a composite of what is in hand and within reasonable reach in the application of statistical science to software testing.

Statistical testing based on usage models can be applied to large and complex systems because the modeling can be done at various levels of abstraction and because the models effectively allow analysis and simulation of *use* of the application rather than the application itself.

Many of the methods that follow are well within the capability of most test organizations, with a modest amount of training and tool support. Some of the ideas are more advanced and would require the services of a statistician the first few times they are used, or until packaged in specialized tool support. Some of the advanced methods would require a resident analyst. However, the methods lend themselves to small and simple beginnings with big payoff, and to systematic advancement in small steps with continued good return on the investment.

## **2. FAILURES IN THE FIELD**

Failures in the field, and the cost (social as well as monetary) of failures in the field, are the motivation behind statistical testing. The collection, classification and analysis of field failure reports on software products has been standard practice for decades for many organizations and is now routine for most software systems and software intensive systems regardless of the maturity of the organization. Field data is analyzed for a variety of reasons, among them the ability to budget support for the next release, to compare with past performance, to compare with competitive systems, and to improve the development process.

Field failure data is unassailable as evidence of need for process improvement. The operational front line is the source of the most compelling statistics. The opportunities to compel process changes move upstream from the field, through system testing, code development, specification writing, and into stating requirements. Historically, the further one moves

upstream, the more difficult it has been to effect a statistically based impact on the software development process that is designed to reduce failures in the field. Progress has been made, however, in applying statistical science to prevention of field failures: given an operational usage model, it is possible to have a statistically reasoned and economically beneficial impact on all aspects of the software life cycle.

### **3. UNDERSTANDING THE SOFTWARE INTENSIVE SYSTEM AND ITS USE**

A usage model characterizes the population of usage scenarios for a software intensive system. Usage models are constructed from specifications, user guides, or even existing systems. The "user" might be a human, a hardware device, another software system, or some combination. More than one model might be constructed for a single system if there is more than one environment of interest.

For example, a cruise control system for tracks has both human and hardware users since it exchanges information with both. The usage model would be based on the states of use of the system—system off, system on and accelerating, system on and coasting, etc.—and the allowable transitions among the states. The model could be constructed without regard to whether the supplier will be Delco or Bosch. It will be irrelevant that one uses a processor made by Motorola and the other by Siemens and that they have very different internal states, that one is programmed in C and the other in Ada. It is conceivable that the system would be tested in two environments of use—operation as a single vehicle and operation in a convoy with vehicle-to-vehicle communication.

#### **First Principles**

When a population is too large for exhaustive study, as is usually the case for all possible uses of a software system, a statistically correct sample must be drawn as a basis for inferences

about the population. [Figure 1](#) shows the parallel between a classical statistical design and statistical software testing. Under a statistical protocol, the environment of use can be modeled, and statistically valid statements can be made about a number of matters, including the expected operational performance of the software based on its test performance.

### Statistical Testing Process

Statistical testing can be initiated at any point in the life cycle of a system, and all of the work products developed along the way become valuable assets that may be used throughout the life of the system. The statistical testing process involves the six steps depicted in [Figure 2](#).

### Building Usage Models

An operational usage model is a formal statistical representation of all possible uses of a system. A usage model may be represented in the familiar form of a state transition graph, where the nodes represent states of system use and the arcs represent possible transitions between states (see [Figure 3](#)). If the graph has any loops or cycles (as is usually the case), then there is an infinite number of finite sequences through the model, thus an infinite population of usage scenarios. In such graphical form, usage models are easily understood by customers and users, who may participate in model development and validation. As a statistical formalism, a usage model lends itself to statistical analysis that yields quantitative information about system properties.

The basic task in model building (Walton, Poore, and Trammell, 1995) is to identify the states-of-use of the system and the possible transitions among states-of-use. Every possible scenario of use, at the chosen level of abstraction, must be represented by the model. Thus, every possible scenario of use is represented in the analysis, traceable on the model, and potentially generated from the model as a test case.

There are both informal and formal methods of discovering the states and transitions. Informal methods such as those associated with "use cases" in object-oriented methods may be used. A formal process has been developed (Prowell, 1996; Prowell and Poore, 1998) that drives the discovery of usage states and transitions. The process is based on the systematic enumeration of sequences of inputs and leads to a complete, consistent, correct and traceable usage specification.

Usage models are finite state, discrete parameter, time homogeneous, recurrent Markov chains. Inherent in this type of model is the property that the states have no memory; some transitions in an application naturally do not depend on history, whereas others must be made independent of history by state-splitting, making the states sufficiently detailed to reflect the relevant history. This leads to growth in the number of states, which must be managed. A usage model is developed in two phases—a structural phase and a statistical phase. The structural phase concerns possible use, and the statistical phase, expected use. The structure of a model is defined by a set of states and an associated set of directed arcs that define state transitions. When represented as a stochastic matrix, the 0 entries represent the absence of arcs (impossible transitions), the 1 represents the certain transitions, and all other cells have transition probabilities of  $0 < x < 1$  (see Table 1). This is the structure of the usage model.

The statistical phase is the determination of the transition probabilities, i. e., the  $x$ 's in the structure. There are two basic approaches to this phase, one based on direct assignment of probabilities and the other on deriving the values by analytical methods.

Models should be designed in a standard form consisting of connected sub-models with a single-entry and single-exit. States and arcs can be expanded like macros. Sub-models of canonical form can be collapsed to states or arcs. This permits model validation, specification analysis, test planning, and test case generation to occur on various levels of abstraction. The structure of the usage models should be reviewed with the specification writers, real or prospective users, the developers, and the testers. Users and specification writers are essential to represent the application domain and the workflow of the application. Developers get an early opportunity to see how the system will be used, and look ahead to implementation strategies that take account of use and workflow. Testers, who are often the model builders, get an early

opportunity to define and automate the test environment.

### **Software Architecture**

The architecture of the software intensive system is an important source of information in building usage models. If the model reflects the architecture of the system, then it will be easier to evolve the usage model as the system evolves. The architecture can be used to directly identify how models should be constructed and how testing should proceed.

A product line based on a common set of objects used through a graphical user interface might have a model for each object as well as a model for the user interface. Each object could be certified independently and the object interactions as permitted by the user interface would be certified with the interface. A new feature might be added later by developing a new object and a modification to the interface; this would require a new model for the new object and an updating of the model for the interface. Importance sampling might be used to emphasize testing of the changed aspects of the interface.

Protocols and other standards established by the architecture can also be factors in usage model development. For example, a usage model for the SCSI protocol has been developed and used in constructing models of several systems that use the SCSI protocol.

### **Assigning Transition Probabilities**

Transition probabilities among states in a usage model come from historical or projected usage data for the application. Because transition probabilities represent classes of users, environments of use, or special usage situations, there may be several sets of probabilities for a single model structure. Moreover, as the system progresses through the life cycle the probability set may change several times, based on maturation of system use and availability of more information.

When extensive field data for similar or predecessor systems exists, a probability value may be known for every arc of the model (i.e., for every nonzero cell of the stochastic matrix of transition probabilities, as in column 4 of [Table 2](#)). For new systems, one might stipulate expected practice based on user interviews, user guides, and training programs. This is a reasonable starting point, but should be open to revision as new information becomes available.

When complete information about system usage is not available, it is advisable to take an analytical approach to generating the transition probabilities, as will be presented in section 5. In order to establish defensible plans, it is important that the model builder not overstate what is known about usage or guess at values.

In the absence of compelling information to the contrary, the mathematically neutral position is to assign uniform probabilities to transitions in the usage model. [Table 2](#) column 3 represents a model based on [Figure 3](#) with uniform transition probabilities across the exit arcs of each state.

#### 4. MODEL VALIDATION WITH CUSTOMER AND USER

A usage model is a readily understandable representation of the system specification that may be reviewed with the customer and users. The following statistics are assured to be available by the mathematical structure of the models and are routinely calculated.

- *Long run probability.* This is the long-run occupancy rate of each state, or the usage profile as a percentage of time spent in each state. These are additive, and sums over certain states might be easier to check for reasonableness than the individual values (for example, Model in [Figure 3](#)).
- *Probability of occurrence in a single sequence.* This is the probability of occurrence of each state in a random use of the software.
- *Expected number of occurrences in a single sequence.* This is the expected number of times each state will appear in a single random use or test case.

- *Expected number of transitions until the first occurrence.* For each state, this is the expected number of randomly generated transitions (events of use) before the state will first occur, given that the sequence begins with Invocation. This will show the impracticality of visiting some states in random testing without partitioning and stratification.
- *Expected sequence length.* This is the expected number of state transitions in a random use of the system and may be considered the average length of a use case or test case. (Using this value and transitions until first occurrence, one may estimate the number of test cases until first occurrence.)

These statistics should be reviewed for reasonableness in terms of what is known or believed about the application domain and the environment of use. Given the model, these statistics are derived without further assumptions, and if they do not correspond with reality then the model must be changed. These and other statistics describe the behavior that can be expected in the "long ran," i.e., in ongoing field use of the software. It may be impractical for enough testing to be done for all aspects of the process to exhibit long run effects; exceptions can be addressed through special testing situations (as discussed below).

### Operational Profiles

Operational profiles (Lyu, 1995) describe field use. Testing based on an operational profile ensures that the most frequently used features will be tested most thoroughly. When testing schedules and budgets are tightly constrained, profile-based testing yields the highest practical reliability; if failures are seen they would be the high frequency failures and consequent engineering changes would be those yielding the greatest increase in reliability. (Note that critical but infrequently used features must receive special attention.)

One approach to statistical testing is to estimate the operational profiles first and then create random test cases based on them. The usage model approach is to first build a model of system use (describe the stochastic process) based on many decisions as to states of use,

allowable transitions and the probability of those transitions, and then calculate the operational profile as the long run behavior of the stochastic process so described.

### **Operational Realism**

A usage model can be designed to simulate any operational condition of interest, such as normal use, nonroutine use, hazardous use, or malicious use. Analytical results are studied during model validation, and surprises are not uncommon. Parts of systems thought to be unimportant might get surprisingly heavy use while parts that consume a large amount of the development budget might see little use. Since a usage model is based on the software specification rather than the code, the model can be constructed early in the life cycle to inform the development process as well as for testing and certification of the code.

### **Source Entropy**

Entropy is defined for a probability distribution or stochastic source (Ash, 1965) as the quantification of uncertainty. The greater the entropy, the more uncertain the outcome or behavior. As new information is incorporated into the source, the behavior of the source generally becomes more predictable, and less uncertain. One interpretation of entropy is the minimum average number of "yes or no" questions required to determine the result of one outcome or observation of the random event or process (Ash, 1965).

Each state of a usage model has a probability distribution across its exit arcs to describe the transitions to other states, which appears as a row of the transition matrix. State entropy gives a measure of the uncertainty in the transition from that state.

Source entropy is by definition the probability-weighted average of the state entropies. Source entropy is an important reference value because the greater the source entropy, the greater the number of sequences (test cases) that it would be necessary to generate from the usage model,

on average, to obtain a sample that is representative of usage as defined by the model.

### Specification Complexity

Some systems are untestable in any meaningful sense. Some systems have such a large number of significant paths of use and such high cost of testing per path, that there is not sufficient time and budget to perform an adequate amount of testing by any criteria, even with the leverage of statistical sampling. This situation can be recognized early enough through usage modeling to be substantially mitigated.

A usage model represents the capability of the system in an environment of use. All usage steps are probability weighted. Any model with a loop or cycle (other than the one from Termination to Invocation) has an infinite number of paths; however, only a finite number have a probability of occurring that is large enough to consider them. The *complexity* of a model can be viewed as the number of statistically typical paths (to be thought of as "paths worth considering"). Note that this concept of complexity has nothing to do with the technical challenge posed by the requirements, nor with the intricacies of the ultimate software implementation. It is simply a measure of how many ways the system might be used (how broadly the probability mass is spread over sequences) and, therefore, a measure of the size of the testing problem.

Complexity analysis can be used to assess the extent to which modification of the specification (and usage model) would reduce the size of the testing problem. By excluding states and arcs from the model, such what-if calculations can be made. For example, mode-less display systems that allow the user to switch from any task to any other task are far more expensive to test than modal displays that restrict tasks to categories. It is possible, also, to compare the differences in complexity associated with different environments of use (represented by different sets of transition probabilities, as in Tables 3 and 4). Complexity analysis can be used to assess the impact on testing of changes in the requirements and system implementation. Because the usage model is based on the specification, the model can be developed, validated,

and analyzed before code is written. An analysis of the complexity of the model might lead to simplification of the specification in various ways, before code development begins.

When the system cannot be changed to reduce complexity and the test budgets cannot be made adequate, usage models can help to focus the budgets on the most important states, arcs and paths. Certain usage states might be critical to achieve (or to avoid) and the number of pathways by which one might achieve (or avoid) these states could be very important. In a slightly more complex situation, there may be two or more states among which passage should be quick and easy (or virtually impossible). *Trajectory entropy* provides a measure of the uncertainty in selecting a path from a set of paths. A variation on the techniques of Ekroot and Cover (1993) produces the measure of specification complexity (Walton, 1995). Trajectory entropy is the sum of the uncertainty of the first step in the path plus the conditional uncertainty of the rest of the path, given the first step. This value is the ratio of the source entropy to the stationary probability of the invocation state and is used as an index of specification complexity, the minimum average number of yes-no questions one would have to ask to identify the path taken. When 2 is raised to this power, an estimate of the number of paths worth considering is obtained. Many well-posed questions involving states, arcs and paths can be expressed in a mathematical model with a closed form solution.

### Simulation

All the analyses mentioned above have closed form solutions and are applicable to all usage models because of their mathematical structure. However, questions do arise for which the analytical expressions or solutions might not exist or might be difficult to formulate. In some of these cases an effective bound on the solution is available, in other cases not.

For example, the expected number of sequences before each state first appears in random testing is computed analytically as shown in Table 3 and 4. However, the expected number of sequences before all the states in an arbitrary subset of model states appear at least once in random testing is a harder question and the answer depends upon the arrangement of the states in

the graph, as well as the probabilities. While this solution can be bounded analytically, a simulation is easily constructed to get an estimate of the answer to the question.

Moreover, many questions that depend strictly on the structure of the graph will not have a general solution. In these cases simulations can usually be constructed to approximate an answer. Since there can be a great deal of variation among the realizations of a model, the simulations might have to run for millions of sequences, but the cost of such simulations is generally not prohibitive. Not all aspects of the sequences will reach asymptotics at the same rate, so some simulations may resolve quickly while others might require analysis of a great many sequences.

It is an extremely valuable aspect of the usage model that simulations are readily available when analytical solutions do not exist or are very difficult to work out.

### **Model Revision and Revalidation**

As mentioned above, these statistics and analyses flow from the usage model without further assumptions. If the structure of the model represents the capability of the system and if the probabilities represent the environment of use, then the conclusions are inescapable. If they do not agree with what is known or believed about the application, then the model must be changed.

Even small models embody a great deal of variation. Consequently, it is not always obvious how to change a model in order to change its statistics. Moreover, small changes in the probabilities can have large and unanticipated side effects. An alternative to the cycle of setting probabilities, analyzing statistics and revising probabilities is to analytically generate models with stochastic matrices guaranteed to have certain statistics, as described in the next section.

## 5. REPRESENTING USAGE MODELS WITH CONSTRAINTS

An alternative to the direct assignment of transition probabilities discussed in section 3 is generation of transition probabilities with the aid of mathematical programming (specifically, convex programming) (Walton, 1995). Usage models can be represented by a system of constraints and the matrix of transition probabilities can be generated as the solution to an optimization problem. In general, three forms of constraints are used to define a model: structural, usage, and test management constraints.

### Structural Constraints

Structural constraints are so named because they define model structure: the states themselves and both possible and impossible transitions among the usage states.

Structural constraints are of four types:

- $P_{i,j} = 0$  defines an *impossible transition* between usage state  $i$  and usage state  $j$ .
- $P_{i,j} = 1$  defines a *certain transition* between usage state  $i$  and usage state  $j$ .
- $0 < P_{i,j} < 1$  defines *probabilistic transition* between usage state  $i$  and usage state  $j$ .
- Each row of  $P$  must sum to one.

### Usage Constraints

If one has no information about the expected usage of the system, one should generate uniform probabilities for the possible transitions from each state. As new information arises, it is recorded in the form of constraints:

- About this PDF file: This new digital representation of the original work has been recomposed from XML files created from the original paper book, not from the original typesetting files. Page breaks are true to the original; line lengths, word breaks, heading styles, and other typesetting-specific formatting, however, cannot be retained, and some typographic errors may have been accidentally inserted. Please use the print version of this publication as the authoritative version for attribution.
- $P_{i,j} = c$  may be used for *known usage probabilities*, i.e., probability values that are exactly known on the basis of historical experience or designed controls.
  - $a \leq P_{i,j} \leq b$  defines *estimated usage probabilities* as a range of values. Defining an estimate as being within a range allows information to be given without being overstated.
  - $P_{i,j} = P_{k,m}$  defines *equivalent usage probabilities*, values that should be the same whether or not one knows what the value should be.
  - $P_{i,j} = d P_{k,m}$  defines *proportional usage probabilities*, where one value is a multiple of another.

Probability values can be related to each other by a function to represent what is known about the relationship, without overstating the data and knowledge. More complex constraints may be expressed as:

- $P_{i,j} = f(P_{k,m})$ , where one value is a function of another.
- $a \leq f(P) \leq b$ , where the value of a function of the matrix  $P$  is bounded, for example, to constrain the average test case length to a certain range.

Most usage models can be defined with very simple constraints.

### Test Management Constraints

Finally, constraints may be used to represent test management controls. Management constraints are of the same forms as usage constraints. A limitation on revisiting previously tested functionality, for example, may be represented in the form of a known usage probability in the section above—a constant that limits the percentage of test cases entering a certain section of the model.

### Objective Functions

Mathematical programming is a technique for determining the values of a finite set of decision variables which optimize an objective function subject to a specified set of mathematical constraints. The general problem of optimizing any function subject to a set of unrestricted constraints can be analytically or computationally intractable. The problem is tractable when it is restricted to convex programming, the minimization of a convex objective function subject to a finite set of convex constraints.

When mathematical programming is used to generate transition probabilities, the solution generated is optimized for some objective function while satisfying all structural, usage and management constraints. Theoretically, one could construct a system of constraints for which there is no solution. In practice, if one does not overstate data and knowledge, this is unlikely.

Analysis of a usage model invariably leads to modification of the transition probabilities in order to incorporate new information, or to change focus at different phases of the process. With complex usage models, individual changes in transition probabilities may result in unintended, poorly understood and unwanted side effects. Better control and understanding is maintained if models are amended through revised or additional constraints and regenerated relative to an optimization objective, rather than by estimation of individual transition probabilities.

Objective functions can be formulated, for example, to minimize cost of testing or to maximize value of testing. Also, entropy measures can be used in objective functions in order to minimize or maximize the uncertainty or variability in the model and, consequently, in the sequences randomly generated from the model.

There are, in general, many sets of transition probabilities that collectively satisfy a system of constraints. Even when the usage profile (stationary distribution) is fully prescribed, many sets of transition probabilities are possible for the usage model which have the same usage profile. Consequently, the certification strategy must be based on a carefully reasoned choice among them that produces the desired overall effects. Mathematical programming can be used to make that choice.

## 6. THE BENEFITS OF USAGE MODELING

As early as possible in the life cycle, one or more usage models is developed (both the structure and the transition probabilities), and the model is validated. To the best ability of the model developers, with the information available to them, the model represents the operational capability of the system at the desired level of abstraction, and the statistics agree with what is known or believed about the intended environment of use. The following is a summary of the many beneficial uses of the model in planning, managing and conducting testing.

### Testing Scripts

A script is associated with each arc of the usage model. This script constitutes the instructions for testing the transition from one state of use to another as represented by the arc. Scripts should be developed by experienced testers and should be validated. The scripts are a significant factor in assuring experimental control during testing.

In the case of testing performed by humans, the script can tell the tester what to do, what inputs to give the system, and what to look for in deciding that the transition was made correctly or not. Testing can be a tedious activity which degenerates in effectiveness unless specific measures are taken to keep the testers focused on what to do and what to look for. Furthermore, testing effectiveness can vary greatly from one person to another unless steps are taken to assure uniformly effective testing. Every test is a traversal of a series of arcs through the model; if the scripts are granular and are followed, they will assure uniform testing.

In the case of automated testing, the scripts will be commands to testing software or equipment and in some cases will contain the information needed to verify correct performance. Lines of code have been used as scripts in such a way that the test case literally becomes a

program to be compiled and executed by the automated test facility.

### Recording Testing Experience

The usage model provides a method of recording testing experience that can be used in assessing test sufficiency and other aspects of the software development process.

A *testing chain* is a representation of testing experience. A testing chain is started by using just the structure of states and arcs (no transition probabilities) for the usage model. As test sequences are executed, each arc successfully traversed (no failure) is marked and the relative frequencies across the exit arcs of each state are calculated. Given enough random sequences, these relative frequencies will converge to the probabilities of the usage model. The measure of similarity between the weights on the usage model (expected traffic) and the weights on the testing chain (tested traffic) is discussed later as a stopping criterion for testing.

Two types of failures are possible. The first type does not impair or distort the functioning of the system, and the transition to the next state can be made. For example, a spelling error might appear in a message on the screen, or a window might be in the wrong location. In such cases, a new state is created to represent the failure and two new arcs are created, one from the departure state to the failure state and one from the failure state to the destination state, and each of the two new arcs gets a mark. Any time in the future that the same failure appears from the same departure state, these two arcs will each be marked again.

A second type of failure is one in which it makes no sense to continue the test case. For example, if the system crashes and it is impossible to continue, or if the failure renders further steps meaningless as in the case of a destroyed file. In such cases, a new state is created to represent the failure and two new arcs are created, one from the departure state to the failure state and one from the failure state to the termination state, and each of the two new arcs gets a mark. Any time in the future that the same failure appears from the same departure state, these two arcs will each be marked again.

Several testing chains can be maintained. One testing chain could be maintained from the

beginning of all testing and another might be maintained for each version of the system, with a new testing chain started each time the code is changed. The cumulative data may be used for process analysis and the data on each version for product analysis. The testing chain can represent all testing experience, special cases as well as random testing, or it can represent just random testing. It is possible to instrument code to maintain a "testing chain" based on actual field experience as well.

### Support for Experimental Design

Design of statistical experiments is being used increasingly in testing software intensive systems (Nair et al., 1998). Although the use of experimental design in software testing is not widespread, the variety of applicable techniques has great potential to transform the testing field.

Designed experiments tell in advance how much testing and what kind of testing will be required to achieve desired results. Indeed, with most of these methods it is possible to influence product design decisions in order to make such testing feasible and more economical. Some characterization of the population under study is necessary for any application of experimental design. The usage model can be of value in all cases.

- *Combinatorial design.* A class of statistical experimental design methods known as combinatorial designs is used to generate test sets that cover the n-way combinations of inputs (Cohen, 1997). For certain types of applications, including data entry screens, this approach has been used to minimize the amount of testing required to satisfy use-coverage goals. Combinatorial design deals with test factors, levels within factors, and treatments (combinations of factor levels) but leaves other issues unaddressed; for example, one must choose among many different test cases that cover all pairs of factor levels. Given a usage model, the treatments will appear as visitation of states of use in specific sequences and the likelihood of these sequences arising in use may be taken into account. Both combinatorial design and operational profiles may be used to plan testing.

- *Partition testing.* Partitioning is a standard statistical technique for increasing the efficiency of random sampling. It is applicable to increasing the efficiency of random testing as well. Partitions can be identified and defined in terms of the usage model. For example, based on [Figure 3](#) test cases might be partitioned into those that include system reconfiguration (state 3) and those that do not; those that include visiting states in Mode-1 and those that do not. The reliability model of Miller et al. (1992) can be used since the probability mass of each block of the partition can be calculated from the model, as can the probability mass of test cases run in each block.
- *Rare events and accelerated rate testing.* Some testing must address infrequent but highly critical situations in order to remove uncertainty or estimate reliability that takes rare events into account. Experimental design has been used to determine the most efficient approach to testing combinations of factors associated with rare events and reliability models have been developed for these situations (Alam et al., 1997; Ehrlich et al., 1997). Usage models can be built from many different perspectives, including process flow. Critical states, transitions, and sub-paths that would have low likelihood of arising in field use (or in a random sample) can be identified from the usage model. The probability of reaching any given state or transition can be calculated directly from the model, as can the traversal of any sub-path. For example, in [Figure 3](#) one might look for unlikely system configurations (state 3) in combination with high rate data (state 10).
- *Sequential testing.* In some cases each test is so expensive to run or to evaluate that it is important to decide based on the outcome of each test whether or not additional testing is justified. The degree to which the variety and extent of testing is representative of the variety and extent of use expected in the field can be calculated directly from the usage model and the testing record.
- *Economic testing criteria.* Different forms or modes of failures in the field can result in different operational economic loss. Usage models together with mathematical programming methods can be used to design testing to minimize the potential of economic loss from field failure (Sherer, 1996).
- *Economic stopping criteria.* Mathematically optimal rules have been developed for

supporting decisions to stop testing, based on the known cost of continued testing versus the expected cost of failure in the field (Dalal and Mallows, 1988). Quantitative analysis of the usage model can assist in assessing the cost of continued testing and the risk of failure in the field.

### Guiding Special Test Situations

Application of statistical science includes creating special, non-random, test cases. Such testing can remove uncertainty about how the system will perform in specific circumstances of interest, aid in understanding the sources of variation in the population, and contribute to effectiveness and control over all testing. In all instances, however, the usage model is the road map for planning where testing should go and recording where testing has been. A few of the many special situations that can be represented in terms of the usage model are as follows.

- *Model coverage tests.* Using just the structure of the model, a graph-theoretic algorithm generates the minimal sequence of test events (least cost sequence) to cover all arcs (and therefore all states) (Gibbons, 1985). If it is practical to conduct this test, it is a good first step in that it will confirm that the testers know how to conduct testing and evaluate the results for every state of use and every possible transition. For large models, even this compelling testing strategy may not be affordable!
- *Mandatory tests.* Any specific test sequences that are required on contractual, policy, moral, or ethical grounds can be mapped onto the model and rim.
- *(Non-random) regression tests.* Existing regression test suites can be mapped to the model. This is an effective way to discover the redundancy in the test suite and assess its omissions. One can calculate the probability mass accounted for by the test suite. Of course, one may use the model to create or enhance a regression test set.
- *Most likely use.* The most likely test scenarios can be generated in rank order to some number of scenarios, or to some cumulative probability mass.

Some balance must be reached between the amount of test time and money that will be spent in special testing and the amount that will be reserved for testing based on random sampling. Only random testing supports inferences about expected operational performance.

### **Generating Random Samples of Test Cases**

Random test cases can be automatically generated from the usage model, constituting a random sample of uses as the basis for statistical estimation about the population. Each test case is a "random walk" through the stochastic matrix, from the initial state to the terminal state. The script associated with each arc of the model is generated at each step of the random walk. Test cases are generated as scripts for human testers or as input sequences for automated testing. One may generate as large a set of test cases as the budget and schedule will bear and establish bounds on test outcomes before incurring the cost of performing the tests.

A random sample of test cases is still a random sample when used multiple times. Thus, it is legitimate to rerun the test set after code changes (regression testing) and to use the results in statistical analysis, provided the code was not changed to specifically execute correctly on the test set. It is not uncommon to see situations where the code always works on the test set, but does not work in the field; developers in some organizations literally learn what the testers are testing. Bias in evaluation must also be avoided. Testers may expect correct results because they have always been correct in the past; testers may learn the test set as well. If testing and the random test sets are independent of the developers and maintenance workers, reuse of the random test sets is a valid statistical testing strategy that can facilitate automated testing and substantial reductions in the time and cost of testing.

### **Importance Sampling**

As was mentioned above, it is generally the case that many sets of transition probabilities exist that satisfy all known constraints on usage. In other words, there are many usage models (same structure, different transition probabilities) that are consistent with what is known about the environment of use.

Objective functions are used to choose the model that satisfies all constraints and is optimal relative to some criterion. By a combination of additional management constraints and objective functions the resulting model can emphasize aspects of the system or of the testing process that are important to testers. The following are among the controls that are possible:

- costs can be associated with each arc, and one can minimize cost;
- value can be associated with each arc, and one can maximize value;
- probabilities associated with exiting arcs that control critical flow can be manipulated;
- certain long run effects can be regulated by constraints;
- some entropy measures can be maximized to increase uncertainty and increase variability in the sequences;
- some entropy measures can be minimized to reduce variability.

One must always be wary of constructing an overly complex model that might be ill-conditioned relative to the numerical methods used in calculating the solution. Too many constraints that are functions of long run behavior are not advised. (Source entropy of a Markov chain is not a convex function. It becomes convex if the stationary distribution or operational profile is fixed.) A statistical analyst must be involved in this kind of modeling.

Recent theoretical developments (Gutjahr, 1997) hold promise for dynamic revision of probabilities as testing progresses in order to optimize sampling relative to an importance objective.

## Test Automation

Usage models have led to increased test automation in almost every situation in which they have been used. Test automation is attractive because it vastly increases the number of tests that can be run and greatly reduces the unit cost of testing. Test automation is more cost effectively done when planned as a companion to the system development, but can also be cost effective for existing systems for which long-term evolution is anticipated.

Test automation depends upon three things: (1) generation of test cases in quantity in a form suitable for automated test runners, (2) an oracle or means of confirming that the system executes the test case correctly, and (3) a test runner that can initiate testing and report results.

The usage model is an excellent means of controlled generation of test cases in any desired quantity. Control is achieved by setting probabilities in order to implement importance sampling. Test cases are produced by walking the graph with a random number generator.

The oracle is the means by which one confirms that each step of the test case does what it is supposed to do (sufficient correctness) and nothing more (complete correctness). This is generally the difficult issue. Some systems have natural and easy oracles much like double inversion of a matrix or squaring a square root; for example, a disk drive control unit might be tested by writing a file to disk and then reading it back. Sometimes a predecessor system can be used because the behavior of the new system is to be identical to the behavior of the old system. Sometimes the testing will be manual the first time as the correct behavior is recorded and then automated on subsequent runs of the test suite.

There are several test runners on the market to which scripts can be sent and which will return information for constructing the testing chain.

## Testing

Testing is expensive; industry data indicates that about half the software budget is spent on testing. Testing costs are best attacked in the development process, by clarifying and simplifying requirements, providing for testability and test automation, and verifying code against specifications. When high quality software reaches the test organization, there are two

goals: (1) provide the development organization with the most useful information possible as quickly as possible in order to shorten the overall development cycle, and (2) certify the system as quickly and inexpensively as possible. Just "more testing" will certainly add cost, but will not necessarily add new information or significantly improve reliability estimates.

- *Resource and schedule estimation.* Calculations on a usage model provide data for effort, schedule, and cost projections for such goals as coveting all states and transitions in the model or demonstrating a target reliability. Estimating the time and cost required to conduct the test associated with each arc of the usage model can lead to estimates for sequences; average sequence lengths can be used to estimate the time and cost of executing test sets.
- *Reliability analysis, with failures.* The testing chain provides the basis for a data-driven estimation of reliability. In the presence of failures, reliability can be assessed without additional mathematical assumptions (in contrast to reliability growth models). The failure states of the testing chain are made absorbing and the reliability of the system is defined to be the probability of going from the invocation state to the termination state without being absorbed in a failure state. The failure states in a testing chain can be ranked with respect to their impact on reliability, which is used to help decide the order in which to work on code corrections.
- *Reliability analysis, no failures.* In the absence of failures, reliability models based on the binomial are sometimes used (Parnas, 1990). Alternatively, the reliability models of Miller et al. (1992), which are based on partitioning the sample space, can be used to take advantage of the structure of the model in order to improve the confidence in the reliability estimate.
- *Test sufficiency analysis.* A stopping criterion can be calculated directly from the statistical properties of the usage model and testing chain. The log likelihood ratio (Kullback, 1958) (Kullback discriminant) can be calculated for these Markov chains and provides evidence for or against the hypothesis that the two stochastic processes are equivalent. The Kullback discriminant is a measure of the difference between expected field usage (usage model) and actual experience in testing (testing chain) that can be monitored during testing because the testing chain is changing with each test event (transition). A graph of the discriminant will have a terrace-like appearance of declines and plateaus as failure-free testing proceeds; when a failure

About this PDF file: This new digital representation of the original work has been recomposed from XML files created from the original paper book, not from the original typesetting files. Page breaks are true to the original; line lengths, word breaks, heading styles, and other typesetting-specific formatting, however, cannot be retained, and some typographic errors may have been accidentally inserted. Please use the print version of this publication as the authoritative version for attribution.

occurs it will step up. This is an information theoretic comparison of the usage and testing chains to assess the degree to which the testing experience has become representative of expected field use. As the testing chain converges to the usage model, it becomes less likely that new information will be gained by further testing generated from the usage model.

### **Simulation**

Testing can be simulated by generating the test cases and marking the testing chain as if the test cases has been run. Simulated testing is an inexpensive way to answer important questions. For example:

- which specific states, arcs, paths will be covered in the course of executing a random test set, and which should be covered by special test cases?
- how many additional random sequences would have to be run without a new failure in order to reach the reliability goal?
- if a failure occurs in a given transition, what will be the impact on the estimated reliability?
- how much will the measured reliability, discriminant, state coverage, arc coverage, etc., improve if a certain number of additional sequences are run?

By analysis and simulation, one can be assured that a test plan has the potential to answer the questions and address the issues of interest to test management. Similarly, one has the information to know at any stage if the failures encountered have rendered the testing goals (quality demonstration) unattainable.

## **7. PRODUCT AND PROCESS IMPROVEMENT**

## Certification

The certification process involves ongoing evaluation of the merits of continued testing. Stopping criteria are based on reliability, confidence, and uncertainty remaining. Decisions to continue testing are based on an assessment that the goals of testing can still be realized within the schedule and budget remaining.

In most cases users of statistical testing methods release a version of the software in which no failures have been observed in testing. Reliability estimates such as those in Miller et al. (1992) are recommended in this case.

Software is sometimes released with known faults. If the test data includes failures, then reliability and confidence may be calculated from the testing chain. The reliability measure computed in this manner reflects all aspects of the sequences tested, including the probability weighting defined by the usage model.

Certification is always relative to a protocol, and the protocol includes the entire testing process and all work products. An independent audit of testing must be possible to confirm correctness of reports. An independent repetition of the protocol should produce the same conclusions, to within acceptable statistical variation.

## Incremental Development

The Cleanroom software engineering process (Poore and Trammell, 1996) uses the testing approach described in this paper. Cleanroom produces software in a stream of increments to be tested. An increment may be "accepted," indicating that the development process is working well, by different (less stringent) criteria than will be used to certify the final product. If increment certification goals are not met, review of experience may show that changes are needed in the process itself, for example, better verification, changes to the usage model, improved record keeping, more frequent analysis of test data, or rethinking of the entire increment plan. If certification goals are met, the process moves ahead with the next increment.

or system acceptance.

The historical testing chain and related statistics will reflect consequences of all failures seen and fixed from the very beginning of testing through all versions of the system to the one released. The historical chain may be used to review the development and testing processes across increments. The historical testing chain and the collection of testing chains from version to version can be used to assess reliability growth.

### **Reliability Growth Modeling**

Many reliability growth models exist in the literature (Lyu, 1995). In each case the model developer sets out a list of basic beliefs about the phenomenon of improving software reliability in a failure-repair cycle. A mathematical model is devised that embodies the basic beliefs as its assumptions. The model is applied to the stream of field failure data, and statistical techniques are used to predict the future failure profile based on the trend of past data. If the software maintenance organization and the user base become sufficiently stable and repetitive, however good or bad, a suitable model for which the assumptions are materially met can be identified by statistical methods. Reliability growth can be measured to support post-deployment decisions.

Reliability growth can also be measured during system test and used to support deployment decisions. When the predicted level of failures is within acceptable intensity, duration and cost, the product can be released. Moreover, one can compare actual field experience with predicted field experience to assess the suitability of the reliability growth model being used.

Reliability growth modeling is being applied successfully to products that permit software changes in the field, and for which there is a degree of tolerance for failures, such as in central office telephone switches, computer operating systems and the like. Its use is problematic once the code is embedded in a system and inaccessible for change; in these cases the goal of testing is to demonstrate that the version of the software that is to be embedded in the system is

of sufficiently high reliability to render a recall sufficiently unlikely.

### **Combining Testing Information**

There are many situations in the life cycle of a software intensive system where it would be beneficial to use existing testing and field use information to identify and minimize the additional testing that is needed to support a decision regarding the system. These situations can be generally classified as development, reengineering, maintenance, reuse and porting. Effective configuration control over the software and correct association of the testing records with code or system versions is required to use such information in a statistically valid way. Furthermore, a common basis for describing the testing done and for interpreting the data is necessary. Usage models have the potential to be the common denominator for test planning and evaluation of results in all testing situations in the system life cycle. The specific statistical models for combining information have not been worked out for every situation, but some progress has been made and work continues in this effort to unify testing. The theoretical path seems clear in all cases.

### **Development**

The incremental development cycle of the Cleanroom software engineering process concludes each cycle with statistical testing to support the decision to move forward with development of the next increment. In the case of the final increment the decision is to deploy or accept the system. The testing protocol for incremental development is based on a comprehensive usage model in which one section is opened to generate test cases in the first increment; a second section is opened additionally for the second increment, etc. In each increment some testing is done in the previous section, but most is in the new section. The statistics associated with each increment of testing have the same usage model as a basis for

combining test information.

An evolutionary procurement of a defense system is based on the concept of a series of fielded systems with past performance, new requirements and new technology coming together for each successive version of the system. Previous testing records and field data are available after the first version. The usage model for the fielded version would be a subset of the model for the new version and the starting point for testing the new development. The field experience from various environments of use could be expressed in terms of the usage model and, together with planned revisions and changes, form the basis for testing.

### **Reuse**

Many systems involve reuse of existing systems or system components, with or without reengineering. Object oriented reuse ranges from pattern instantiation, to framework integration, to class-subclass hierarchy extensions with polymorphic methods. If a component is to be reused without change, then the usage model originally used to certify the component can be used to assess the testing necessary for the new use. One would have the original usage model and testing records (testing chains and statistics), plus field use summarized as an estimate of sequences actually run. A set of transition probabilities would describe the new use. It is straightforward to compare the new use against the records of previous testing and use to determine whether or not further testing is required for the new use.

### **Reengineering**

Reengineering typically involves changing the technology from which the system is made, for example, one or more of the hardware processor, memory units, power supplies or even the programming language and data structures, but generally preserves the way the system is used (otherwise, it would be new development or maintenance rather than reengineering).

Usage specifications usually survive, with varying degrees of change. The original usage model might change in structure. Usage states and arcs can be associated with underlying changes in the technology. Usage models might be used to assess the extent of change and to guide testing; the greater the change, the harder the testing problem.

### Maintenance

Maintenance is usually associated with small change to an operational system. Thus, the developmental testing and field use records are available. Field experience indicates that good understanding of both the usage model (states and arcs) and the architecture and implementation of the system are required in order to map maintenance changes to relevant parts of the usage model. Testing after maintenance must address both known impact areas and the possibility of unanticipated impacts, and be planned using records of prior model-based testing and field use.

### Porting

Porting is the process of moving a system to new or additional "platforms," usually meaning different operating systems for the same hardware, new hardware running the same operating system, or the hardware and operating system of a different vendor. Given a good software architecture and design that anticipated the porting, the changes to the system will be minimal but the services provided by the hardware and operating systems can be significantly different. Given multiple platforms on which the system must run, what is the optimal amount of testing to be done on each platform in order to support a decision to deploy each? Generating test cases and recording test results based on a common usage model and a common set of statistics makes this a tractable problem.

A common framework for planning and recording all testing and field use in the life cycle of a system can lead to substantial cost savings in testing and much better information to support

About this PDF file: This new digital representation of the original work has been recomposed from XML files created from the original paper book, not from the original typesetting files. Page breaks are true to the original; line lengths, word breaks, heading styles, and other typesetting-specific formatting, however, cannot be retained, and some typographic errors may have been accidentally inserted. Please use the print version of this publication as the authoritative version for attribution.

decisions.

## CONCLUSION

Most usage modeling experience to date is with embedded real-time systems (Oshana, 1997), application program interfaces (APIs) and graphical user interfaces (GUIs). Models as small as 20 states and 100 arcs have proven very useful. Typical models are on the order of 500 states and 2,000 arcs; large models of more than 2,000 states and 20,000 arcs are in use. Even the largest models developed to date are small in comparison to similar mathematical models used in industrial operations research, and are manageable with available tool support. Large models must be accepted as appropriate to many software systems and to the testing problem they pose. The size is not to be lamented, because the larger and more complex the testing problem, the greater the need for the assistance that modeling and simulation afford.

Since 1992, the IBM Storage Systems Division has developed and applied Markov chain usage models for certification of tape drives, tape controllers, tape libraries, disk drives and disk controllers. Some products are tested based on up to five different usage models, including models of customer use, a data communication protocol model, a model keyed to the injection of hardware and media failures, and a stress model. Many of these models are reused from product to product because only the technology of the product changes and not the architecture of the product, the way it is used or the standards to which it is built. Transition probabilities have been determined by using instrumentation measurements collected during internal use combined with external customer command traces originally collected for performance analysis of previous products. The test facility is highly automated and employs compiler writing technology to automatically compile executable test cases from abstract arc labels, which permits testing a large number of scripts. Stopping criteria are based on both high reliability estimates and substantial agreement between testing experience and expected field experience. Use of this technology has significantly reduced the testing effort and improved field reliability.

Special tools are available for integration into existing testing environments that support

those activities of testing based on usage models that have settled into "standard practice," namely:

- systematic enumeration of input sequences to drive the discovery of behavior and identification of states and allowable transitions,
- usage model editing,
- generating uniform probabilities when specific values are not given,
- generating a standard set of statistics for model validation,
- generating the minimum cost state and arc coverage scripts,
- generating test cases with scripts,
- automatic construction of testing chains based on sequences generated and feedback of testing results,
- sorting failures according to the impact of repair on expected reliability,
- monitoring reliability, the discriminant and coverage measures as testing proceeds.

General tools must be used for aspects in which there has been insufficient experience to identify and justify specialized tools; for example, constraint-based generation of transition probabilities and related optimization problems must be solved using general purpose convex programming solvers. General purpose graph editors must be used for graphical model editing. Also, simulations must be programmed in a general purpose language.

Some activities of statistical testing are computationally intensive, with run-time for analyses a function of the number of states or the number of arcs in a usage model. While the computations would seem routine to an operations research analyst, they might seem prohibitive to some software engineers. Ironically, software engineering environments tend to be computationally starved. Computation times do not become an issue in work flow until the models reach about 2,000 states. In general, all the time and effort spent in modeling, analysis, and simulation will be recouped with dividends in the avoidance of uninformative testing.

Usage models facilitate the application of many aspects of statistical science to software testing, because (1) they offer a formalism in which to represent the statistical issues of interest

and (2) they have the ability to generate test cases and record the results of testing. They facilitate test planning as a part of the product or system requirements definition, which in turn makes it possible to optimize the amount of testing done at each stage of system development. Usage models set the stage for getting the best information from testing for the least cost. The economics of software failures in the field and the cost of testing complex systems make the use of statistical science imperative to software testing.

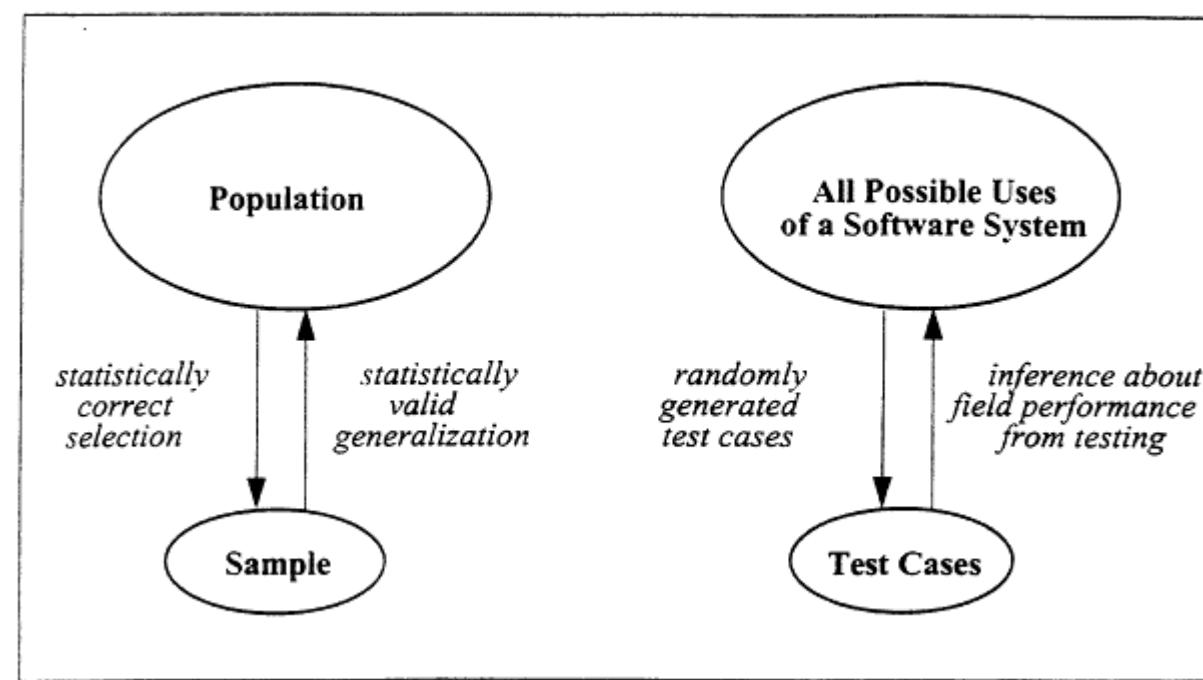


Figure 1  
Parallel between statistical inference and software testing.

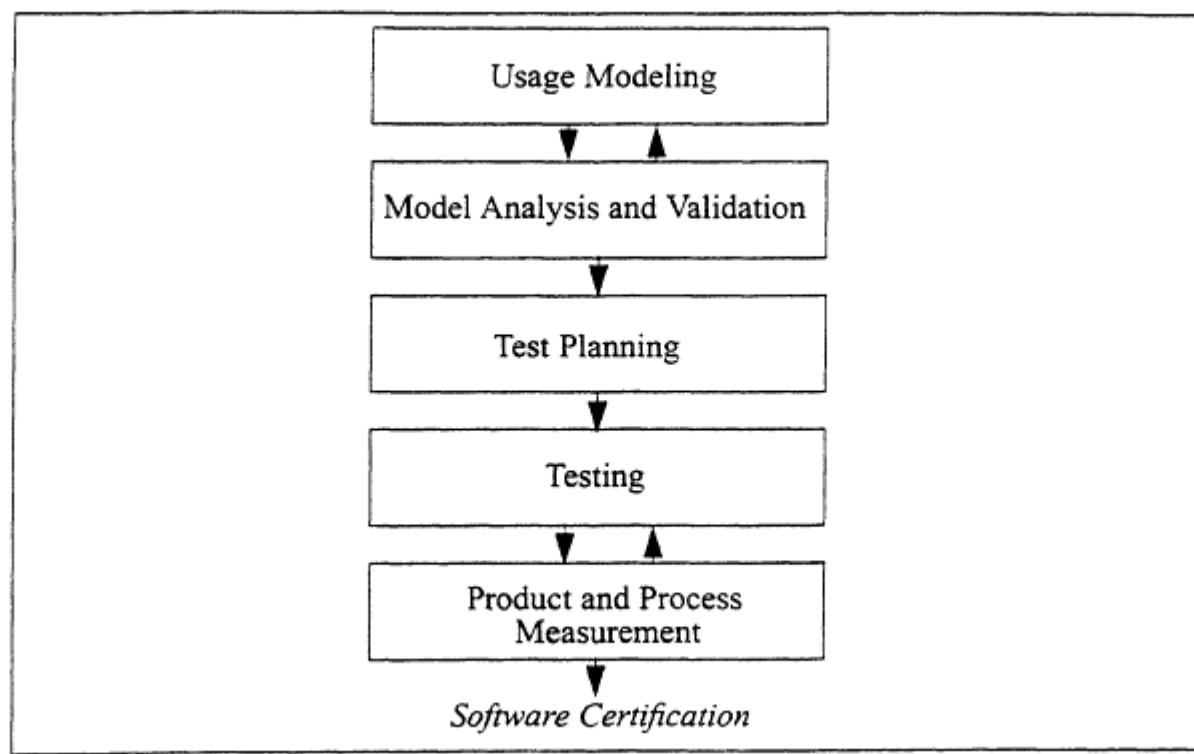


Figure 2  
The statistical testing process.

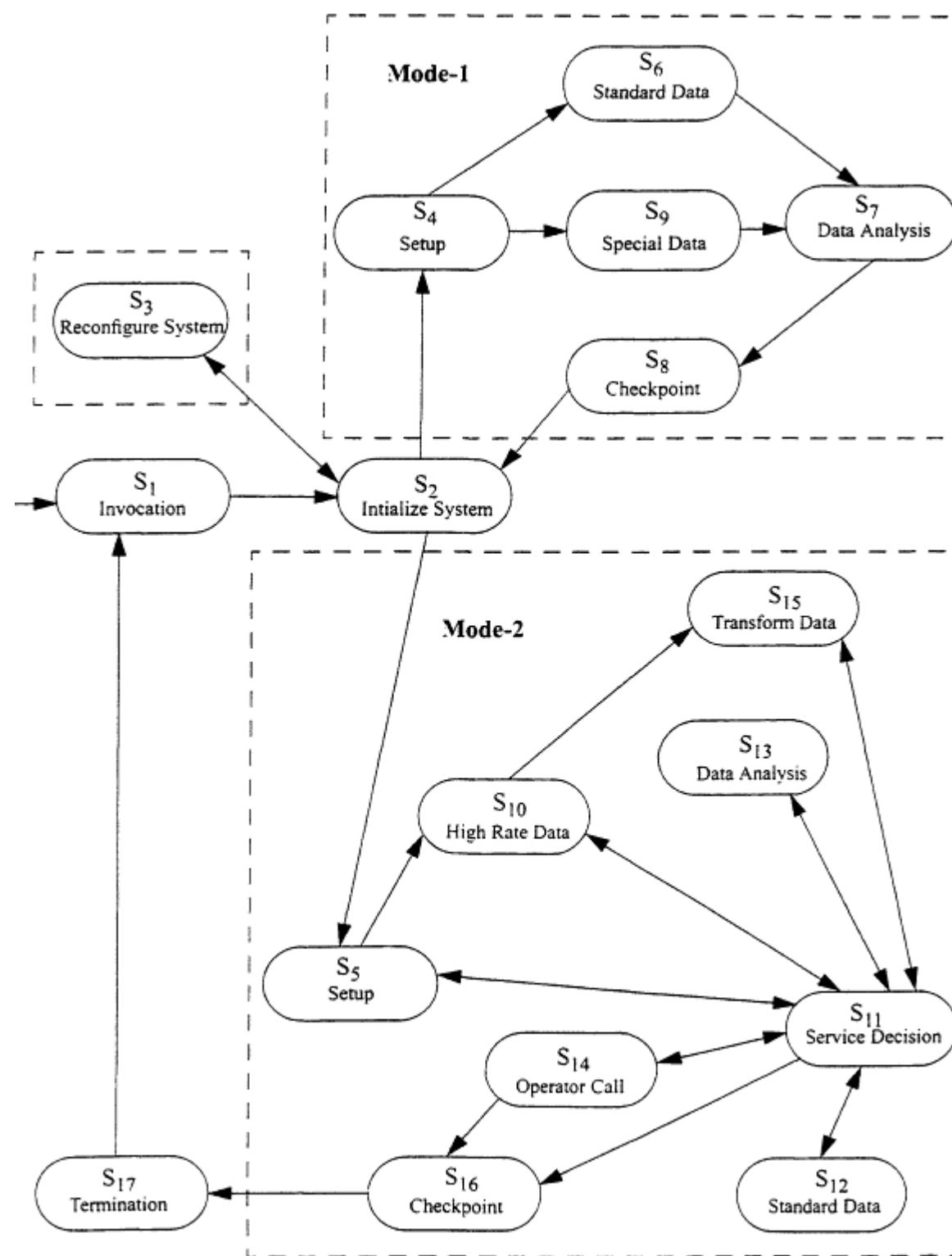


Figure 3  
Example Usage model structure, directed graph format.

TABLE 1 Example of a Usage Model Structure, Transition Matrix Format

To State

From State	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
To State	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	x	x	x	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	x	0	0	x	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	x	x	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	x	0	0	0	x	0	0
11	0	0	0	0	x	0	0	0	0	x	0	x	x	x	x	x	0
12	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	x	0	0	0	0	x	0
15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLE 2 Example Usage Models, One Structure, Two Matrices of Transition Probabilities

From-State	To-State	Uniform Probabilities	Specific Environment
(1) Invocation	(2) Initialize System	1	1
(2) Initialize System	(3) Reconfigure System	1/3	1/12
(2) Initialize System	(4) Mode-1 Setup	1/3	8/12
(2) Initialize System	(5) Mode-2 Setup	1/3	3/12
(3) Reconfigure System	(2) Initialize System	1	1
(4) Mode-1 Setup	(6) Mode-1 Standard Data	1/2	3/4
(4) Mode-1 Setup	(9) Mode-1 Special Data	1/2	1/4
(5) Mode-2 Setup	(10) Mode-2 High Rate Data	1/2	1/12
(5) Mode-2 Setup	(11) Mode-2 Service Decision	1/2	11/12
(6) Mode-1 Standard Data	(7) Mode-1 Data Analysis	1	1
(7) Mode-1 Data Analysis	(8) Mode-1 Checkpoint	1	1
(8) Mode-1 Checkpoint	(2) Initialize System	1	1
(9) Mode-1 Special Data	(7) Mode-1 Data Analysis	1	1
(10) Mode-2 High Rate Data	(11) Mode-2 Service Decision	1/2	3/4
(10) Mode-2 High Rate Data	(15) Mode-2 Transform Data	1/2	1/4
(11) Mode-2 Service Decision	(5) Mode-2 Setup	1/7	2/16
(11) Mode-2 Service Decision	(10) Mode-2 High Rate Data	1/7	1/16
(11) Mode-2 Service Decision	(12) Mode-2 Standard Data	1/7	4/16
(11) Mode-2 Service Decision	(13) Mode-2 Data Analysis	1/7	4/16
(11) Mode-2 Service Decision	(14) Mode-2 Operator Call	1/7	1/16
(11) Mode-2 Service Decision	(15) Mode-2 Transform Data	1/7	3/16

TABLE 2 Example Usage Models, One Structure, Two Matrices of Transition Probabilities

(11) Mode-2 Service Decision	(16) Mode-2 Checkpoint	1/7	1/16
(12) Mode-2 Standard Data	(11) Mode-2 Service Decision	1	1
(13) Mode-2 Data Analysis	(11) Mode-2 Service Decision	1	1
(14) Mode-2 Operator Call	(11) Mode-2 Service Decision	1/2	1/2
(14) Mode-2 Operator Call	(16) Mode-2 Checkpoint	1/2	2/2
(15) Mode-2 Transform Data	(11) Mode-2 Service Decision	1	1
(16) Mode-2 Checkpoint	(17) Termination	1	1
(17) Termination	(1) Invocation	1	1

TABLE 3 Usage Statistics for the Model with Uniform Probabilities on the Exit Arcs

State Identification Number	Long Run Probabilities	Prob. of Occurrence in a Single Sequence	Expected Number of Occurrences in a Single Sequence	Expected Number of Transitions Until Occurrence	Expected Number of Sequences Until Occurrence
1	0.0449	1.0000	1.00	22.25	1
2	0.1348	1.0000	3.00	1.00	1
3	0.0449	0.5000	1.00	22.25	1
4	0.0449	0.5000	1.00	19.25	1
5	0.0749	1.0000	1.67	9.00	1
6	0.0225	0.3333	0.50	42.50	2
7	0.0449	0.5000	1.00	21.25	1
8	0.0449	0.5000	1.00	22.25	1
9	0.0225	0.3333	0.50	42.50	2
10	0.0674	0.7500	1.50	16.67	1
11	0.2097	1.0000	4.67	10.75	1
12	0.0300	0.4000	0.67	43.12	2
13	0.0300	0.4000	0.67	43.12	2
14	0.0300	0.5000	0.67	36.75	2
15	0.0637	0.6538	1.42	21.53	1
16	0.0449	1.0000	1.00	20.25	1
17	0.0449	1.0000	1.00	21.25	1

Number of arcs is 28.

Expected sequence length is approximately 23 states.

Expected number of sequences to cover the least likely state is approximately 3.

Expected number of sequences to cover the least likely arc is approximately 7.

The log base 2 source entropy is approximately 1.0197.

The specification complexity index is approximately 22.71 (or 6,861,000 sequences).

TABLE 4 Usage Statistics of Model for Specific Environment

State Identification Number	Long Run Probabilities	Prob. of Occurrence in a Single Sequence	Expected Number of Occurrences in a Single Sequence	Expected Number of Transitions Until Occurrence	Expected Number of Sequences Until Occurrence
1	0.0250	1.0000	1.00	40.08	1
2	0.0998	1.0000	4.00	1.00	1
3	0.0083	0.2499	0.33	120.28	3
4	0.0665	0.7273	2.67	12.03	1
5	0.0582	1.0000	2.33	16.00	1
6	0.0499	0.6667	2.00	18.04	1
7	0.0665	0.7273	2.67	14.03	1
8	0.0665	0.7273	2.67	15.03	1
9	0.0166	0.4000	0.67	58.11	2
10	0.0215	0.4843	0.86	58.52	2
11	0.2662	1.0000	10.67	17.10	1
12	0.0665	0.7273	2.67	31.13	1
13	0.0665	0.7273	2.67	31.13	1
14	0.0166	0.5000	0.67	66.67	2
15	0.0553	0.6935	2.22	33.82	1
16	0.0250	1.0000	1.00	38.08	1
17	0.0250	1.0000	1.00	39.08	1

Number of arcs is 28.

Expected sequence length is approximately 40 states.

Expected number of sequences to cover the least likely state is approximately 4.

Expected number of sequences to cover the least likely arc is approximately 16.

The log base 2 source entropy is approximately 0.9169.

The specification complexity index is approximately 36.68 (or  $1.1 \times 10^{11}$  sequences).

## REFERENCES

- Alam, M.S., et al. 1997 Assessing software reliability performance under highly critical but infrequent event occurrences. In review.
- Ash, R. 1965 *Information Theory*. New York: Wiley.
- Basili, V., et al. 1996 The empirical investigation of perspective-based reading. *Empirical Software Engineering* 1:133-164.
- Cohen, D.M. 1997 The AETG system: An approach to testing based on combinatorial design. *IEEE Transactions on Software Engineering* 23 (7):437-444.
- Dalal, S., and C. Mallows 1988 When should one stop testing software? *Journal of the American Statistical Association* 83(403).
- Ehrlich, W., et al. 1997 Application of accelerated testing methods for software reliability assessment. In review.
- Ekroot, L., and T.M. Cover 1993 The entropy of Markov trajectories. *IEEE Transactions on Information Theory* 39(4):1418-1421.
- Gibbons, A.M. 1985 *Algorithmic Graph Theory*. Cambridge, U.K.: Cambridge University Press.
- Gill, P.E., W. Murray, and M.H. Wright 1981 *Practical Optimization*. New York: Academic Press.
- Gutjahr, W.J. 1997 Importance sampling of test cases in Markovian software usage models. *Probability in the Engineering and Informational Sciences* 11:19-36.

- Kemeny, J.G., and J.L. Snell 1960 *Finite Markov Chains*. D. Van Nostrand Company, Inc.
- Kullback, S. 1958 *Information Theory and Statistics*. New York: John Wiley and Sons.
- Lyu, M.R. 1995 *Handbook of Software Reliability Engineering*. New York: McGraw-Hill.
- Miller, K., et al. 1992 Estimating the probability of failure when testing reveals no failures. *IEEE Transactions on Software Engineering*. January.
- Nair, V.N., D. James, W. Ehrlich, and J. Zevallos 1998 A statistical assessment of some software testing strategies and applications of experimental design techniques. *Statistica Sinica* 8:165-184.
- Oshana, R. 1997 Software testing with statistical usage based models. *Embedded Systems Programming* 10(1):40-55.
- Parnas, D. 1990 An evaluation of safety critical software. *CACM* 23(6):636-648.
- Poore, J.H., and C.J. Trammell 1996 *Cleanroom Software Engineering: A Reader*. Oxford, United Kingdom: Blackwell Publishers.
- Prowell, S.J. 1996 Sequence-Based Software Specification. Ph.D. Dissertation, University of Tennessee, Knoxville, Tenn.
- Prowell, S.J., and J.H. Poore 1998 Sequence-based specification of deterministic systems. *Software—Practice & Experience* 28(3):329-44.
- Sherer, S.A. 1996 Statistical software testing using economic exposure assessments. *Software Engineering Journal* September:293-297.

- Walton, G.H., J.H. Poore, and C.J. Trammell 1995 Statistical testing of software based on a usage model. *Software —Practice & Experience* 25(1):97-108.
- Walton, G.H. 1995 Generating Transition Probabilities for Markov Chain Usage Models. Ph.D. Dissertation, University of Tennessee, Knoxville, Tenn.
- Whittaker, J.A., and J.H. Poore 1993 Markov analysis of software specifications. *ACM Transactions on Software Engineering and Methodology* 2(1):93-106.
- 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.