ESTIMATES OF RELIABILITY DURING THE TEST AND EVALUATION STAGE: SOME METHODOLOGICAL OBSERVATIONS

Jack Zwanziger

December 1985

P-7169

The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138

CONTENTS

TABLES

FIGURES

I. INTRODUCTION

Shortly after the Reagan administration took office, then-Deputy Secretary of Defense Carlucci issued a set of initiatives designed to reform the acquisition process. In consolidated form, they are now known as the Defense Acquisition Improvement Program initiatives. Several were designed to reduce the time required to acquire new systems, including an initiative that reduced the number of Department of Defense decision milestones major programs must undergo as they move through the acquisition process, and another that streamlined the procedure for validating mission needs and funding new concepts.

One important effect has been to elevate in importance the Defense Systems Acquisition Review Council II decision milestone that occurs at or around the start of full-scale development. Test organizations that used to make their major inputs to the Office of the Secretary of Defense decisionmaking process at Defense Systems Acquisitions Review Council III (the production decision milestone), now must make their initial inputs much earlier, generally without the benefit of as much testing as before.

Another initiative encouraged the development of policies to ensure adequate front-end funding for test hardware to compensate for the risks posed by compressing the acquisition cycle. This would be accomplished by providing more test articles earlier to accomplish more testing sooner. The same initiative called for the services to define and explain test article requirements during the planning and budgeting process. Intense competition for front-end funding has made this initiative hard to implement in practice.

Recognition of the need to improve system support and readiness also prompted several policy actions, including a requirement to establish readiness goals at Milestone I (the start of the Demonstration and Validation phase) for new systems, goals that presumably would be subjected to testing as programs moved through the acquisition process.

One of the primary objectives of operational testing is to verify that the system being tested will meet the performance levels specified in operational requirements statements. This assessment is inherently uncertain given the complexity of the weapon system, the variety of environments in which it will operate, and the statistical nature of the test process itself. As a result, the decisionmaker should be made aware of the confidence with which estimates have been made. Acquisition initiatives that compress the time available for testing may increase the risks of procuring an unreliable system by significantly increasing the uncertainty in estimating the system's operational characteristics. A compressed acquisition timetable is likely to put even greater pressure on decisionmakers given the increasing desire to field systems having better readiness and supportability characteristics.

Within the acquisition process, reliability testing is one of the methods the military has used to maintain control of the reliability characteristics of a system being developed.

Reliability testing has multiple objectives including:

1) determining compliance with contractual requirements;
2) identifying deficiencies in the system needing correction;
3) providing reliability measures for use in operational planning to estimate such critical characteristics as mission success rates and logistic support costs;
4) measuring the readiness of the system for production and subsequent operational use.

A balanced test program, which integrates Development Testing & Evaluation (DT&E) and Initial Operational Testing & Evaluation (IOT&E) requires that tests be designed to fulfill each of these roles. Developing such a test program is further complicated by the extreme budgetary and scheduling constraints often imposed by the realities of the acquisition process. When development missiles can cost $1 million each, when flying time for fighter aircraft carrying the missile can cost $3000 to $6000 per hour, when substantial additional costs must be

incurred for operating the test facilities, and when there are many users competing for those facilities, there is great pressure to keep testing at a minimum. Furthermore, within each test program there may be a tendency to emphasize reliability development growth testing, as that component of the test program which actually improves system performance, over reliability qualification testing (RQT), which is only intended to estimate the reliability of the existing system. This emphasis is understandable and possibly justified, given the limited resources available for testing. Nevertheless, RQT remains an essential component of a test program since it alone can provide an objective standard for modifying compliance with contractually required performance. In order for decisionmakers to assess the data produced by RQT, it is essential to understand some aspects of the statistical theory used and its application to reliability estimation.

Applying the statistical techniques generally used, this paper illustrates the statistical risk facing both the military purchaser and the contractor from testing less, and conversely, the gains in statistical confidence from testing more. Using the testing of air-launched missiles as an example, we analyze the confidence that can be placed in reliability estimates considering the test approach employed, the amount of testing accomplished, and the reliability theory commonly applied to evaluate test results and to forecast reliability improvement.

Two basically different kinds of reliability will be considered:

1. Binomial processes where "success" or "failure" are defined and testing is required to estimate the probability of their occurrence. It is commonly used to characterize the single shot kill probability of a missile.
2. Time-to-failure reliability where "failure" is defined and the times when failures occur are recorded. The critical quantity derived is mean time between failures (MTBF), commonly used to characterize the captive carry reliability of a missile.

These two aspects of reliability involve fundamentally different statistical models and assumptions. Section II describes reliability

theory for success-failure testing and applies it to the estimation of missile kill probabilities. Section III similarly treats time-to-failure reliability for missile captive carry tests.

II. SUCCESS-FAILURE TESTING

The theory of binomial processes is used for estimating the reliability of systems for which the only relevant outcomes are success or failure. This section describes the amount of test articles typically involved in air-launched missile tests, the impacts on the confidence level from reducing the number of test trials or increasing the reliability requirements, and concludes with some observations on the testing of these kinds of systems.

BACKGROUND

Missile test firings can occur prior to full-scale development to support engineering efforts and in some cases to contribute to source selection. Firings accomplished during operational testing normally form the basis for production decisions. IOT&E or Follow-on Operational Test and Evaluation (FOT&E) test firings normally are measured in the tens of missiles, as shown in Table 1.

The binomial distribution is used in estimating a missile's ability to destroy its target, a critical measure of operational effectiveness. Reliability requirements are usually expressed in terms of a required single shot probability of kill often broken down into the required reliabilities for the events that constitute the launch-to-intercept sequence. Missile tests are rarely simply a single event; "success" depends on an intricate series of successes occurring sequentially so that a failure at any stage will usually result in a system failure. Some tests focus on particular stages and some failures are judged irrelevant for assessing operational performance. Even these complexities could be incorporated into the simple binomial model by defining "success" appropriately, but the factor preventing the simple application of this model is the ever-present constraint on both cost and time.

This pressure expresses itself as the need to maximize the information derived from an extremely limited number of tests. The probability of success of each stage can be estimated directly from the

Table 1

NUMBERS OF TEST FIRINGS

System	Test Phase	Missile Firing Attempts		
		Air Force	Navy	Total
A	IOT&E/OPEVAL	NA	NA	31
B	FOT&E Phase I/OT-VA	3	3	6
C	FOT&E	34	--	34
D	IOT&E Phase I/OPEVAL	NA	NA	16
	IOT&E Phase II/OPEVAL	8	11	19
E	IOT&E	12	--	12

data, once "success," "failure," "no test" and the precise conditions for the test are defined. Extending these estimates to an estimate of the system's operational reliability involves a careful analysis of how this estimate will be used and the individual characteristics of the system being tested. The difficulty in validating such extensions and the extremely limited number of trials result in highly uncertain estimates. Some of these difficulties will be discussed in greater detail at the conclusion of this section after considering the basic theory of binomial processes and some of its implications.

The theory of binomial processes assumes the following:

1. that the probability of "success" (or "failure") is identical for each test;
2. that the tests are statistically independent.

Given the reliability and the number of test articles, it is easy to calculate the probability of observing any number of failures. More typically, one can estimate the probability that the system reliability exceeds a given threshold reliability from observing the number of failures occuring during n tests.

SENSITIVITY OF CONFIDENCE LEVELS TO AMOUNT OF TESTING

Although a test program of any length will produce an estimate of reliability, as the number of trials decreases, the confidence of that estimate, and therefore its usefulness as a basis for decisionmaking, diminishes. Figures 1-4 illustrate some of the relationships involved in assessing the need for more testing.

Figures 1a-1c show the relationship between the probability that the reliability is no lower than threshold and the number of trials for several different sets of demonstrated reliabilities. The rate of increase in confidence with the number of trials decreases steadily, an "elbow" appearing in the 10-20 trial region. If the demonstrated reliability is 25 percent greater than threshold, for example .75 rather than .6 on Fig. 1b, then 80 percent confidence is reached by 13 trials. On the other hand, if the demonstrated reliability is only .69, or 15

percent above threshold, a much greater number of trials is required to reach the same level of confidence, around 30 in this case.

Figure 2 illustrates more graphically how the rate of increase in the confidence level falls off as the number of trials increases. If the demonstrated reliability is likely to be only slightly (less than 10 percent) above threshold then there is a consistent increase in confidence with testing but the confidence attained is relatively low.

TESTING REQUIRED TO DEMONSTRATE HIGHER RELIABILITY

System reliability must improve if more stringent readiness and sustainability goals are to be met. This requirement has implications for the amount of testing needed to demonstrate with confidence the achievement of higher levels of reliability.

Figure 3 illustrates several critical aspects of the relationships involved if binomial assumptions are met. Firstly, it is clear that past thirty or so trials there is relatively little reduction of the demonstrated reliability required for a given threshold reliability. For example, given a threshold reliability of .7, increasing the number of trials from 25 to 50 only reduces the demonstrated reliability required to achieve 80 percent confidence from approximately .8 to .78. Secondly, as the threshold reliability is increased, the stringency of the demonstrated reliability required for acceptance increases dramatically, a relationship partially obscured by the smoothed curves shown in Figure 3. In fact, for a threshold reliability of .8 and for fewer than 20 trials, two or more failures would result in rejection if 80 percent confidence is required. Past 20 trials or so, an increase of .1 in threshold reliability roughly translates into an increase of .1 in the demonstrated reliability required to achieve 80 percent confidence.

Figure 4 illustrates the same relation but focuses on the 80 percent lower confidence limit resulting from a given observed reliability. Once again, as the number of trials increases, the 80 percent success threshold increases only slightly for any given demonstrated reliability.

The clear conclusion is that in testing systems based upon binomial processes, increasing the threshold reliability will increase dramatically the difficulty in achieving the same confidence as for

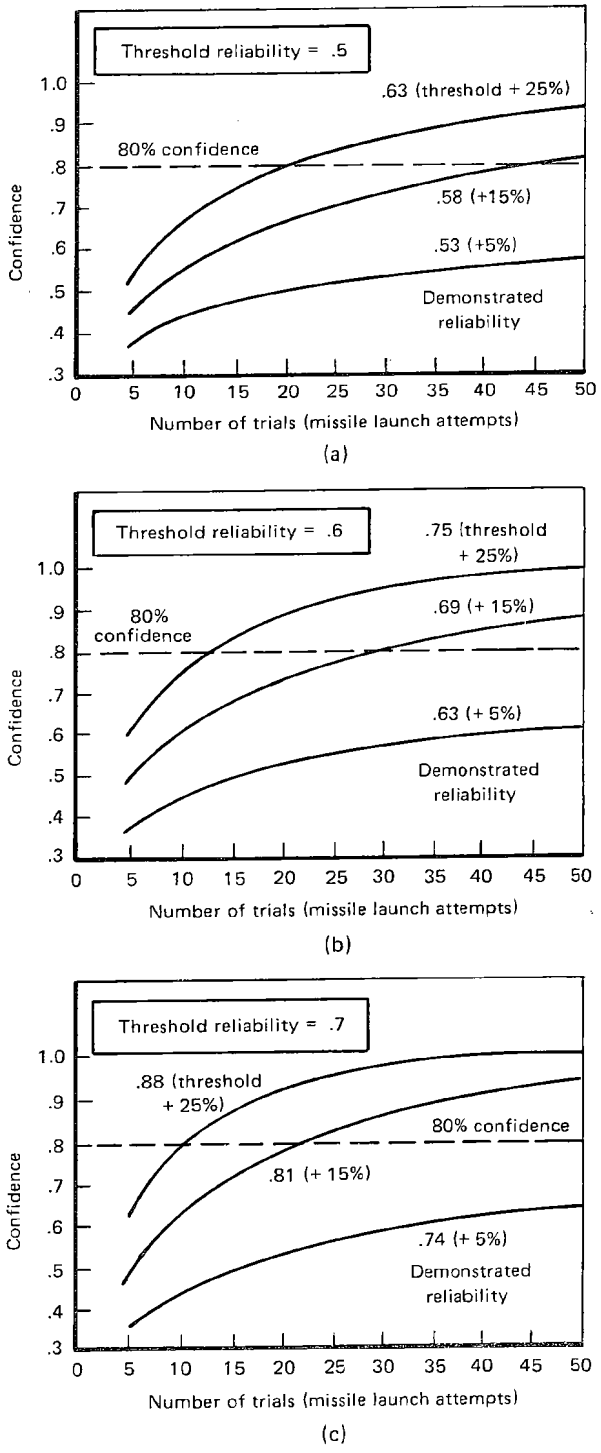

Fig. 1 — Confidence as a function of the number of tests

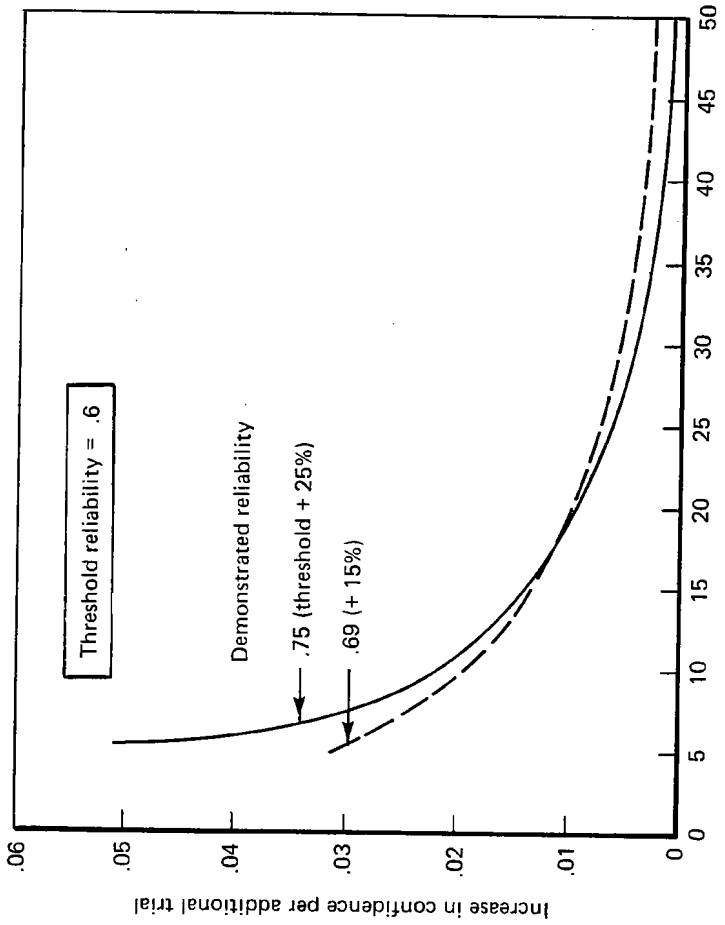

Fig. 2 — The effect of increased testing

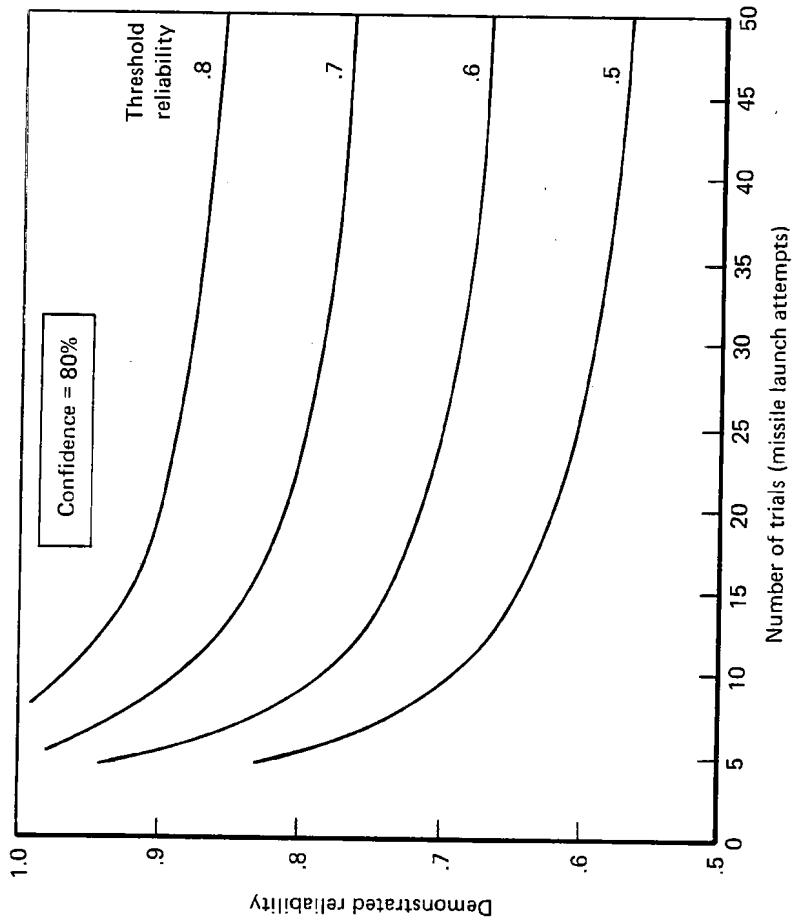

Fig. 3 – Influence of increased reliability

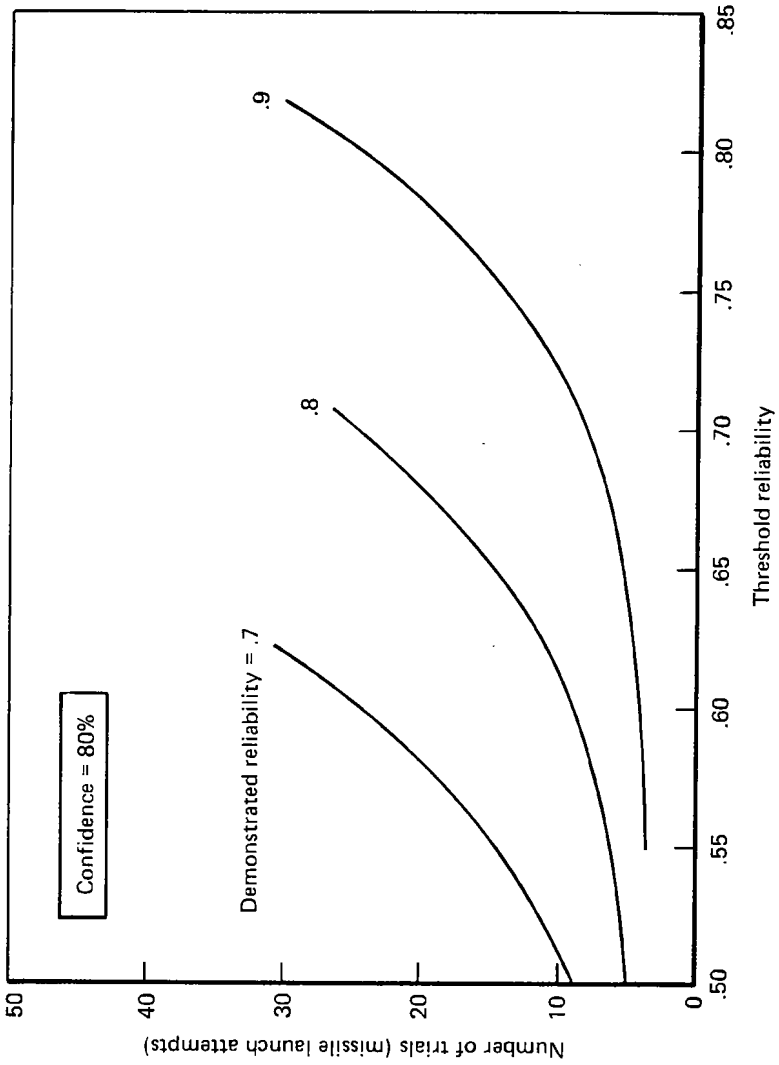

Fig. 4 — Influence of increasing test reliability

lower reliabilities. Since the corresponding increases in the demonstrated reliability required allow for very small failure probabilities, increasing the number of tests will only slightly decrease this allowed failure probability.

RECONSIDERING THE ASSUMPTIONS

The preceding analysis has been based on the simple theory of binomial processes described above. However, applying this theory to the testing of actual systems is far from straightforward. To illustrate some of the complexities involved, let us reexamine the two basic assumptions of the theory.

First, counting the number of successes is difficult since in most cases "success" is a composite event. For example, a successful missile firing consists of a successful launch, successful guidance, successful fuzing, and warhead detonation within a lethal miss distance. Each event has its own probability of success. If these probabilities are considered to be independent, an assumption commonly made, then the overall probability of success is the product of the corresponding probabilities for each event so that one can calculate the observed probability of success.

Test results can be "success," "failure," and "no test," where a "no test" can occur if the event considered did not have an opportunity to be tested. For example, an unsuccessful launch may not permit a testing of fuzing abilities. Alternatively, a failure due to some factor external to the missile, such as an aircraft radar failure or a target failure, can prevent a fair test of the missile and create a "no test." Since "no tests" are ignored in calculating the reliability of each stage, each stage is likely to have a different number of trials so that the number of trials to be used in the calculations is left ambiguous. Lloyd and Lipow argue that the appropriate choice in estimating the system reliability is the minimum number of trials for any of the observed stages.[1] This choice does provide an unambiguous and conservative method of estimating system reliability. The validity of this approach depends on the assumption that the reliability of each stage is independent of the other. This assumption must obviously be considered carefully for each system.

A second difficulty involves the assumption that the probability of a success on each trial is constant. Tests are usually conducted in a wide variety of different environments and conditions. This variation is essential in assessing the operation of a system under a variety of possible circumstances, but it may result in probabilities of success that vary significantly from one test condition to another. Therefore, in estimating reliability, to use the theory properly, one would have to segregate trials into categories where one could plausibly assume that reliability is essentially the same for all the trials in that category and then calculate its reliability. Following this procedure, however, may result in only a few trials in each group; the resulting wide confidence interval for system reliability would reduce its usefulness to the decisionmaker.

OBSERVATIONS

A basic problem facing test planners, then, involves an assessment of the differences in system reliability which different test environments are likely to produce. The extent of these differences will greatly influence the tradeoffs they must make between the number of test articles and the uncertainty in the estimates of reliability. Given the number of test missiles commonly used in reliability tests, if there is no significant variation in reliability over the different test environments, then significant increases in confidence would require a large increase in test size. If large variations are anticipated, then current test programs will produce highly uncertain estimates.

The wide range of reliability estimates can be reduced by increasing the number of tests, since significant increases in confidence occur with an increase in the amount of testing when the initial number of tests in each environment is small. Another approach to reducing the confidence interval would be to combine the reliability estimates in each category to produce some average reliability. While this may result in an estimate similar to that of the "naive" approach of lumping together all the test results, it would make explicit the assumed relative importance of each environment to the "average" operational experience of the missile. It would also enable one to

estimate these reliabilities with anticipated changes in operating environment.

Yet another approach for dealing with these variations in reliability estimates is to develop a model of the relationship between test conditions and success probabilities; this model would be compared to the observed reliability of each group of trials and if verified could provide some useful information on the dependence of reliability on the operating environment. Models are already used today prior to test firing events to insure that upcoming tests have a reasonable probability of success. Results from such models may also provide more confidence when aggregating test results from different parts of the envelope for statistical analysis--if the models suggest each firing event has a comparable probability of success.

Overall, despite the complexities involved in its application, the binomial distribution is a "robust" one in the sense that it is unlikely to be seriously misleading. It involves simple assumptions, and with some necessary manipulation, it can be applied to these systems. However, before attempting to use the theory of the binomial process, one must develop a detailed definition of what outcome would constitute a success, verify that the stages to be observed are independent and assess the likely influence of variable test environments on reliability. All these exercises would in any case be a useful part of planning a test program.

III. TIME-TO-FAILURE TESTING

This major test category focuses on the frequency of system failure as the measure of weapon system performance. It is applicable to aircraft radars, for example, which are expected to fail and where logistic support functions are largely determined by their failure rates. The systems which will be considered here are air-to-air missiles which can fail as a result of being captive carried by aircraft in flight. This section will discuss reliability testing for such missile systems, investigate some implications of increasing test times and of requiring higher reliabilities, concluding with an assessment of the applicability of the theory to the systems being tested.

BACKGROUND

Some missile systems are intended to survive a substantial number of flights while being captive carried by jet aircraft. For such weapon systems, primarily air-to-air missiles and, to a lesser extent, the training versions of air-to-ground missiles, a critical measure of operational suitability is the captive carry failure rate. These failures are caused both by the stresses imposed on the missile from carriage on an aircraft in flight, as well as from handling on the ground during transport and maintenance. The rate at which these failures occur has a critical impact on both the cost of logistic support and on operational effectiveness. To control such impacts, design requirements are established, usually expressed as a minimum acceptable mean time between failures (MTBF).

There has been a clear trend for requiring substantial increases in the mean time between failures (with the guidance head as the element which is most likely to fail during captive carry). This approach assumes that flight time is the factor which determines the number of failures. If a significant number of failures are the result of either slow deterioration during storage or improper handling or maintenance, then the failure rates estimated from tests which ignore these sources of failures will be substantially below those actually experienced in the field.

Operational testing has as one of its objectives the measurement of the mean time between failures and its comparison with stated MTBF requirements. These requirements are usually expressed as a threshold value at a given confidence level or as point estimates, and often specify which MIL-STD-781 tests are to be used.[1] In contrast to the relatively straightforward binomial theory which applies to the probability of kill, time to failure testing uses a more involved statistical theory based on the exponential distribution. Despite its approximate nature, the convenient properties of this distribution have encouraged its widespread use. This theory applies when failure rates are constant over time. Systems are believed to have a roughly constant failure rate between their initial "burn-in" and final "wear out" phases; this middle portion of a missile's life cycle is believed to characterize most of its operationally useful life when failures are most likely to be purely random events. Suggestions have been made, however, that these assumptions are incorrect and that therefore this approximation may result in misleading MTBF estimates. These objections and some alternative approaches will be considered in the conclusion of this chapter.

Captive carry testing can begin as early as during combined DT&E and IOT&E testing. Demonstration of satisfactory captive carry reliability is often a prerequisite for unconditional DSARC approval to proceed with production, although other factors often override this consideration. Captive carry testing usually occurs after the production decision as well, during FOT&E, to validate the correction of deficiencies identified during earlier testing and to qualify final production configurations. It is not unusual for additional captive carry testing to be scheduled if previous testing indicates unsatisfactory reliability. Additional captive carry testing may also be scheduled at any time to qualify missiles produced by a second source. Test programs for air-to-air missiles usually amount to several thousand hours of captive flight spread across multiple test articles (see Table 2).

---

[1] *MIL-STD-781C*, "Reliability Design Qualification and Production Acceptance Tests: Exponential Distribution," establishes the standard

Table 2

TYPICAL TOTAL CAPTIVE CARRY TEST TIMES

System	Test Phase	Articles	Captive Carry Hours		
			Air Force	Navy	Total
A	IOT&E/OPEVAL	56			4664
B	FOT&E Phase I/OT-VA	44	2279	2216	4495
C	FOT&E	47	1440		1440
D	IOT&E Phase I/OPEVAL	23	546	498	1044
	IOT&E Phase II/OPEVAL	40	1060	1225	2285
E	IOT&E	--	292	--	292

test and associated statistical techniques for systems assumed to have exponential failure rates.

The need to balance test time and costs severely constrains test design. One of the clear implications of these constraints is to question the importance of any statistical analysis of test results. The statistical model almost invariably used in analyzing such time-to-failure reliability is based on the assumption that failures are the result of a Poisson process. One implication of this assumption is that the time between failures is exponentially distributed.

An exponential distribution of time between failures is, of course, a particularly convenient assumption for acceptance testing under severe constraints on test time and the number of test articles since exponentiality implies that test articles have "no memory." Testing 10 test articles 100 hours each is equivalent to testing one article for 1000 hours, so that test time per article and the number of articles can freely be traded off for each other as long as total test time is kept constant. This conclusion will be further considered in assessing the plausibility of the exponential failure distribution for these weapon systems.

SENSITIVITY TO THE AMOUNT OF TESTING

Several related probabilities were calculated in the course of analyzing both the confidence level currently achieved in testing for time-to-failure reliability, and the sensitivity of this confidence level to total test time. (As noted above, for exponential distributions it is the total test time that enters into the probability distribution, not the time each article was tested.) The risks to both the buyer and the contractor, were calculated as a function of different combinations of test time, actual MTBF and demonstrated MTBF.

First, to illustrate the risk to the buyer--that the actual MTBF is less than threshold MTBF--we consider Fig. 5 which uses data drawn from an initial phase of System A's IOT&E test program. During the initial phase 3613 captive carry hours were accumulated and a demonstrated MTBF of 251 hours was observed. After the initial phase, the probability that the true MTBF was greater than the threshold 80 percent lower confidence limit reliability of 200 hours was about .77, hence the equipment marginally failed to meet the customary 80 percent confidence level standard.

Additional testing was ordered of articles incorporating design changes. These corrective actions were successful and the missile satisfied the requirements in a subsequent 882-hour test. Note the gradual and decreasing rate at which increased test time reduces risk. Increasing test time from 800 to 1200 hours reduces the probability of a below-threshold MTBF from .48 to .41, or 13 percent, with a rate of decrease of 15.5 percent/100 hours. Going from a 3600- to a 4000-hour test program, results in a decrease from .23 to .21, a 9 percent decrease, with the rate of decrease of .5 percent/100 hours. Therefore, the marginal benefit of increased testing, in terms of decreased risk to the buyer, decreases substantially as the test hours are accumulated.

Vendors have the opposite problem; the weapon system may be rejected even though its true MTBF is greater than the threshold MTBF. Given a test program with a prespecified total test time, the contractor can only reduce the risk facing him by increasing the design reliability of his equipment. The contractor, of course, has only a limited ability to control with precision the reliability of the system being developed. The means at his disposal involve decisions such as the use of high reliability parts, derating components, using more screening, and more fundamental design changes. Of course, these methods of increasing reliability can increase a contractor's costs--although they may reduce overall lifecycle costs to the military.

Figure 6 illustrates this risk/reliability/test time relationship for a system with a required MTBF of at least 200 hours as the 80 percent lower confidence limit. As test time increases, the risk to the contractor falls sharply. For example, a contractor willing to accept a 20 percent risk of rejection has to design to an MTBF of approximately 475 hours if the test plan is 1200 hours or can "afford" an MTBF of approximately 300 hours if a test plan calls for a total time of 4400 hours. These relationships may have to be considered by the buyer if the price quoted by the vendor reflects his assessment of the degree of risk and/or the cost of reducing it.

As test time increases, a demonstrated reliability exceeding the threshold translates into greater confidence that the true MTBF exceeds the threshold. Figure 7 illustrates this for systems with required

reliabilities of 200 hours when demonstrated MTBFs are 240 and 360 hours respectively. The y coordinate is the confidence level of an MTBF of 200 hours. Confidence increases with total test time for both observed MTBFs, but for 360 hours, approaches 1.0 very rapidly, flattening out after about 2400 hours. For the 240 hours observed MTBF on the other hand, the increase is steadier and quite slow, so that even with over 4000 hours of test time, confidence of exceeding threshold is still less than .75.

TESTING REQUIRED TO DEMONSTRATE HIGHER RELIABILITY

Captive carry reliability requirements have been increasing steadily, and this has a substantial impact on the test time required to achieve a given confidence level. To consider some specific systems, a hypothetical comparison of two programs serves to illustrate the effect on test time of increasing the required reliability. Figure 8 is a plot of confidence of having exceeded their respective thresholds as a function of test time. One system has a threshold MTBF of 240 hours at the 80 percent lower confidence limit; the other's comparable contractual threshold is 733 hours. For a demonstrated MTBF of 1.25 times threshold, the test time required to achieve comparable confidence for the latter is roughly triple that for the former. So the test time required triples as the reliability to be demonstrated triples. Testing increasingly reliable systems requires the accumulation of more test hours to achieve the same degree of confidence, a result to be considered when trying to compress the acquisition cycle.

RECONSIDERING THE ASSUMPTIONS

The use of the exponential model, with its highly restrictive assumptions, must obviously be justified. A theoretical result often cited as the basis for its use is the theorem proved by Drenick that a system of many parts all failing randomly which is repaired and placed back in service immediately, exhibits a failure rate which is approximately constant.[4] In contrast to mechanical systems, where "wearout" is expected, predominantly electronic systems are expected to exhibit constant failure rates as a result of randomly occurring stresses. This assumes that failure rates are purely usage driven, and are independent of other aspects such as storage and handling.

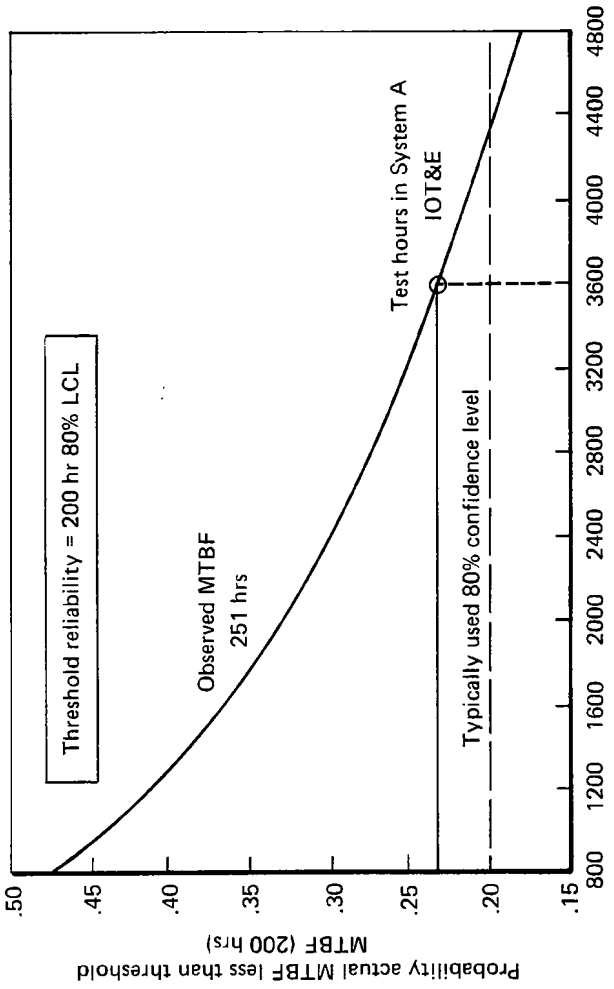

Fig. 5 — Sensitivity of buyer's risk to test time

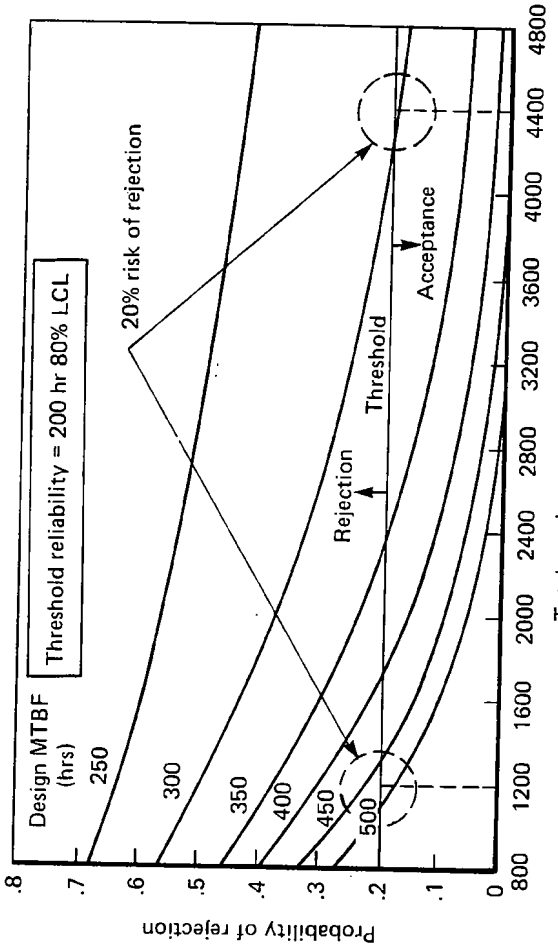

Fig. 6 – Sensitivity of producer's risk of rejection to test time and design reliability

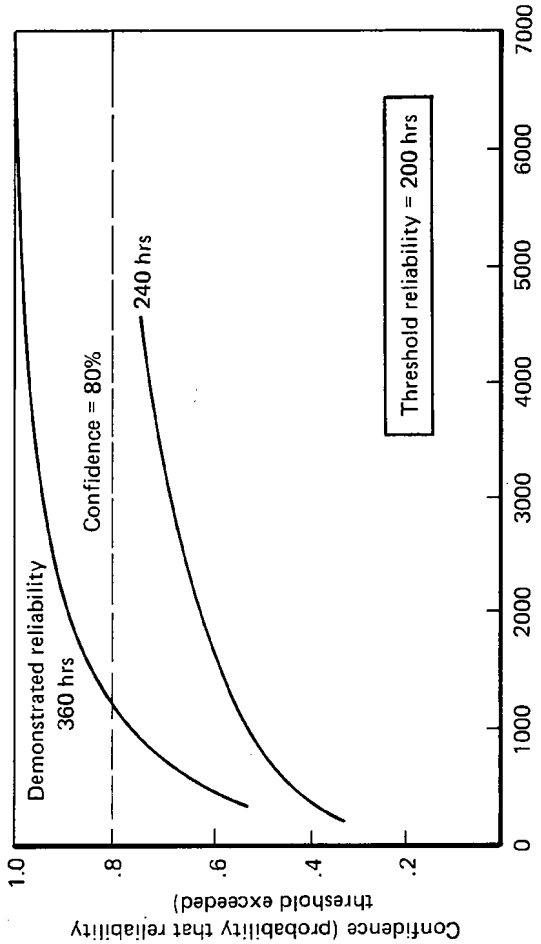

Fig. 7 — Sensitivity of confidence to amount of testing

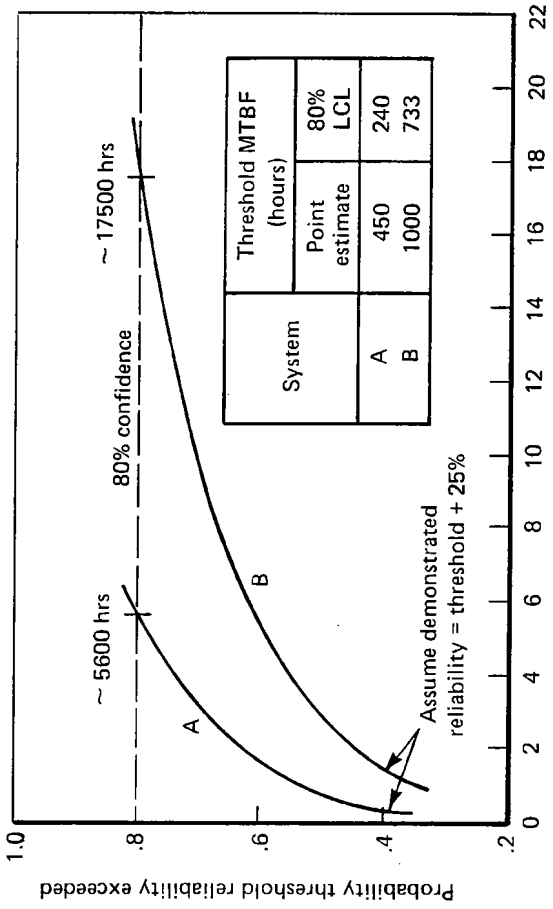

Fig. 8 — Increased testing to demonstrate higher levels of reliability

Furthermore, the exponential distribution is especially convenient for testing systems with severe constraints on the number of test articles provided. Therefore there is a great reluctance to abandon this useful distribution.

Criticism of its use has focused on the following aspects:

1. The weak theoretical basis
2. The data recording system
3. Variability in the test environment
4. Weak empirical evidence.

To consider these in turn:

Theoretical Basis

The theorem cited does predict an exponential-like distribution but only in the long run; it does not address shorter intermediate-run results.[4] Moreover, the result assumes that the entire system, not only the part repaired or replaced, is as good as new after repair, an unrealistic assumption.

A recent paper shows that if the system has not yet reached its long-run equilibrium, then the test results produced by MIL-STD-781C could be seriously misleading. Since average time to failure for individual components tends to be quite large, it is quite likely that weapon systems being tested are still composed of relatively "new" components and so have not yet reached equilibrium. If components tend to have increasing failure rates with age, the MTBF calculated assuming that the system has reached equilibrium would substantially overestimate the system's reliability.[5]

An alternative approach based on the assumption that after repair the system is as "bad as old" leads to substantially different results with non-constant failure rates.[6] This paper by Ascher also shows how an analysis based on an assumption of exponentiality can be misleading, when even non-exponential distributions can be made to appear to produce a reasonably constant failure rate. Overall, these arguments imply that there is little theoretical basis for *assuming* a constant failure rate.

Data Recording System

In order to calculate the exponential failure rate from failure data, one must record the actual time of failure and use these times in Eq. (13). It is rarely possible to record exactly when failures occur, since the status of test articles is usually monitored at intervals. As a result, failures are usually recorded as having occurred during a certain time interval. The test documents we reviewed described three methods for estimating the failure rate from time interval data, the "simplest," maximum likelihood, and least squares methods, each with different degrees of plausibility and applicability. Unfortunately, each can yield significantly different results. Their application to the test results from one program yielded MTBF estimates ranging from 40 to 57 hours depending on the method used. Such large variations lend support to the proposition that rather than assume an exponential distribution which can never be calculated from actual results, one should translate reliability statements into "binomial" language and talk about the probability of failure over some operational time interval.[7]

Variability in the Test Environment

A typical operational test involves subjecting the test article to a series of different operating environments. Even if one can assume exponentiality in a given environment, the failure rate is likely to vary substantially with variation in the test environment. Correspondingly, the time between failures may depend critically on the sequence of different environments during test cycles. The extent to which this effect would distort the MTBF estimate is unknown although a calculation could be performed assuming a different exponential constant failure rate in each environment and cycling the system through a MIL-STD-781 test. The results of this calculation would provide some indication as to how significant this effect is likely to be. If the MTBF is several test cycles long, this effect should not be too serious a source of error. This effect, however, could be a more serious source of error for testing on aircraft which may be operating in very different environments during different tests. Of course in translating

to field data where there may be consistent differences in mission profiles, such variations in failure rates may result in large differences in observed reliability.

Weak Empirical Evidence

The final criticism to be considered is based on two observations. The first is that empirically both tests of the failure distributions of specific pieces of equipment and analysis of field data do not point to constant failure rates over time.[8] Furthermore, the standard methods used to verify an exponential distribution, the Kolmogorov-Smirnoff or Bartlett tests, are highly imprecise for the small number of data points which are produced by a typical test program. These tests would usually be satisfied for a large number of possible distributions.

The consequences of specifying incorrectly the reliability distribution can be extremely significant. A simulation study has been performed to test departures from an exponential to a Weibull distribution where the failure rate would increase with time.[9] This showed that systems with Weibull distributions could have MTBFs significantly below the required threshold MTBF and still have a high probability of acceptance if the analysis of the test results assumes that failure times are exponentially distributed. Tests where some test articles survive to the end of the test period ("censored" tests) are particularly liable to such errors since little information is available regarding the statistical distribution of the longer times to failure. For such "censored" tests, systems that have failure rates increasing with time could easily be accepted even if over their lifetime the MTBF will be lower than threshold MTBF.[2]For example, with the combination of these two factors, the inability to distinguish an exponential distribution of failure times from a wide variety of other distributions, and the resulting potential for significant mis-estimation of the MTBF are cause for concern as this would affect support costs and the system's operational suitability.

---

[2]28 test articles of which 14 fail and using the AGREE acceptance criteria, a Weibull distribution with shape parameter p = 3 has an 80 percent probability of acceptance at an MTBF ~ 400 hours, whereas an exponential distribution requires an MTBF of ~ 900 hours to be accepted 80 percent of the time.[10] The long-term consequences of accepting

OBSERVATIONS

The assumption that the failure rate is constant, while tempting, is almost certain to be incorrect. The issue then is, how seriously does a departure from exponentiality distort the estimates of significant parameters such as MTBF. It would appear that the commonly used tests such as the Bartlett or the Kolmogorov-Smirnoff are unable to specify the distribution being observed very precisely. Thus the true uncertainty in the estimate of MTBF is due not only to sample randomness but also to the lack of knowledge as to the form of the actual distribution of failure times.

One method of increasing the ability to identify the distribution of failure times is to test some test articles to several times the threshold MTBF. This would provide much more information regarding the "tail" of the failure distribution and possibly reveal different failure modes. Cost constraints may require that the number of test articles be reduced correspondingly, but if total test time is kept roughly constant, this will not substantially affect the confidence level of the result. Then a maximum likelihood Weibull distribution could be estimated from the data and tested to determine whether the departure from exponentiality is significant.

A more radical approach would involve a focus on the actual operational reliability required. Operational scenarios would be defined together with the required probability of a mission success. A threshold probability of success would be specified for each mission. The purpose of testing would then be the verification of these probabilities of success. The major disadvantage of this approach is that at least 10-20 tests of each scenario would be required to have any confidence in the resulting reliability estimate. That many tests of each scenario are unlikely to be feasible. A non-statistical advantage of this approach would be that it would require a close examination of the required reliability level and a specification of the anticipated operating environment.

---

this system would be very large, since over a long period of time it will fail roughly twice as often as expected, based on the initial test.

IV. CONCLUSIONS

Attempting to estimate the reliability of a complex weapon system is inherently difficult. Designing a test program will always involve a series of difficult tradeoffs because a system's performance depends on a wide variety of factors in its operating environment, it is often not fully developed at the point of testing, and both test time and test articles are often severely constrained. The statistical methods used to estimate reliability reflect these tensions. Two statistical models, the binomial and the exponential, have been commonly used to estimate the reliability of these weapon systems.

The reliability estimates which result from the use of the binomial model are unlikely to be too misleading, especially if, where necessary, provisions are made to estimate separately system reliability in different test environments and then to combine them explicitly. Some measure of the confidence of these estimates should be included in the test reports sent to decisionmakers as the statistical uncertainty of these estimates may be a factor in their considerations. Reducing the number of test articles would reduce the confidence in these reliability estimates, especially as the reliability required from the systems is increased. If kill probability varies significantly in different test environments, it will be necessary to ensure at least 10-20 tests of the system should be conducted in each environment or else reliability estimates will be highly uncertain.

The exponential distribution, while convenient analytically, may produce seriously misleading reliability estimates if it is applied to a system whose failure rate is not constant with operating time. Once again practical constraints on test time and articles severely limit the test planner, but some steps may reduce the potential for mis-estimating a system's MTBF. Firstly, explicit provision should be made for measuring the degree of variability of the MTBF with operating environment. Secondly, at least some test articles could be tested to several times threshold MTBF. Then some simple but non-exponential distributions, such as the Weibull, could also be considered during the

course of the analysis. Alternatively, all testing could be reduced to the success-failure situation, where explicit operating scenarios would be designated, and success, failure, and no-test defined so that the analysis would involve no assumption regarding the distribution of failure times. Both of these approaches are likely to require at least some increase in the size of test programs.

REFERENCES

1. Lloyd, David K. and Myron Lipow, "Reliability Management Methods and Mathematics," *Prentice Hall*, 1962, pp. 226-229.

2. "Reliability Design Qualification and Production Acceptance Tests: Exponential Distribution," *MIL-STD-781C*, Par. 4.5.2:1.

3. Kapur and Lamberson, "Reliability in Engineering Design," John Wiley & Sons, 1977, pp. 285-287.

4. Drenick, R. F., "The Failure Law of Complex Equipment," *Jour. SIAM,* Vol. 8, No. 4, pp. 680-690.

5. Blumenthal, S., J. A. Greenwood and L. H. Herbach, "Series Systems and Reliability Demonstration Tests," Operations Research, Vol. S2, No. 3, pp. 641-648.

6. Ascher, H. E., "Evaluation of Repairable System Reliability Using the Bad-As-Old Concept," *IEEE Transaction on Reliability*, Vol. R-17, No. 2, June 1968, pp. 103-110.

7. Elliott, T., "Common Errors in Application of MIL-STD-781, *Proc. of Annual Reliability and Maintainability Symposium*, 1984, pp. 139-140.

8. Balaban, H. S., and R. A. Kowalski, "Field Data: The Final Measure," *Proc. of Annual Reliability and Maintainability Symposium*, 1984, pp. 123-125.

9. Zelen, M., and M. C. Dannemiller, "The Robustness of Life Testing Procedures," *Technometrics*, Vol. 3, No. 1, 1961, pp. 29-49.

RAND/P-7169 ESTIMATES OF RELIABILITY DURING THE TEST AND EVALUATION STAGE: SOME METHODOLOGICAL OBSERVATIONS Jack Zwanziger