SANDIA REPORT
SAND2015-0927
Unlimited Release
Printed February 2015

A Statistical Perspective on Highly Accelerated Testing

Edward V. Thomas

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	http://www.ntis.gov/search

SAND2015-0927
Unlimited Release
Printed February 2015

A Statistical Perspective on Highly Accelerated Testing

Edward V. Thomas
Statistics and Surveillance Assessment Department
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS0829

Abstract

Highly accelerated life testing has been heavily promoted at Sandia (and elsewhere) as a means to rapidly identify product weaknesses caused by flaws in the product's design or manufacturing process. During product development, a small number of units are forced to fail at high stress. The failed units are then examined to determine the root causes of failure. The identification of the root causes of product failures exposed by highly accelerated life testing can instigate changes to the product's design and/or manufacturing process that result in a product with increased reliability. It is widely viewed that this qualitative use of highly accelerated life testing (often associated with the acronym HALT) can be useful. However, highly accelerated life testing has also been proposed as a quantitative means for "demonstrating" the reliability of a product where unreliability is associated with loss of margin via an identified and dominating failure mechanism. It is assumed that the dominant failure mechanism can be accelerated by changing the level of a stress factor that is assumed to be related to the dominant failure mode. In extreme cases, a minimal number of units (often from a pre-production lot) are subjected to a single highly accelerated stress relative to normal use. If no (or, sufficiently few) units fail at this high stress level, some might claim that a certain level of reliability has been demonstrated (relative to normal use conditions). Underlying this claim are assumptions regarding the level of knowledge associated with the relationship between the stress level and the probability of failure. The primary purpose of this document is to discuss (from a statistical perspective) the efficacy of using accelerated life testing protocols (and, in particular, "highly accelerated" protocols) to make quantitative inferences concerning the performance of a product (e.g., reliability) when in fact there is lack-of-knowledge and uncertainty concerning the assumed relationship between the stress level and performance. In addition, this document contains recommendations for conducting more informative accelerated tests.

ACKNOWLEDGMENTS

The author acknowledges helpful comments from Rene Bierbaum, Elmer Collins, Jill Glass, and Janet Sjulin during the preparation and editing of this document. The author is especially appreciative of Rene Bierbaum for providing NW context that was incorporated into the document.

CONTENTS

FIGURES

This page is intentionally blank.

1. INTRODUCTION

Today, designers and manufacturers (including Sandia) face tight deadlines and cost constraints to develop and test new products. Testing is done to improve the product during development (find and fix problems) as well as to assess/demonstrate performance. For example, in the reliability context at Sandia, the former is regarded as "improving the inherent reliability of the product", while the latter is focused upon "estimating the reliability of the product as fielded". Both are obviously important but often require different testing and sampling strategies, as well as different philosophies for interpreting the results. In response to schedule and cost constraints, engineers are under pressure to reduce both the number of units tested and the time required to complete testing. Some products must operate reliably for long periods of time over many use cycles in specific environments. Other products (one-shot devices) must operate properly once after a long period of dormant storage. In either case, it is difficult to estimate the long-term performance of such products given the constraints on testing duration. These considerations naturally lead to accelerated testing where a product is subjected to conditions (related to dormant storage and/or use) in excess of what the product normally experiences. Such conditions cause the product to fail or degrade more rapidly than when experiencing normal/use conditions. A large variety of accelerated tests can be used to assess performance (or find and fix problems). The tests are typically accelerated by: increasing the use-rate of a product, increasing the aging-rate of a product, or increasing the level of stress under which a product operates or is tested [1, pp. 468-469].

With respect to increasing the *use-rate* of a product, Meeker and Escobar [1] give an example of a toaster that may be normally used twice a day with a nominal life of 20 years. Assuming that its life is governed by the number of times that it is used, one can observe a "lifetime" of use by testing the toaster 365 times per day for 40 days. A high energy discharge capacitor is an example of a Sandia product tested in an analogous manner. Such capacitors are often used at Sandia in one-time use nuclear weapon (NW) applications. However, it will be functioned many times due to 100% production testing at the capacitor level of assembly, the next assembly level, and in some cases in sample production or post-production surveillance tests. The goal is to ensure that the capacitor still functions as required in spite of numerous previous operations. Thus, we can rapidly observe a capacitor's lifetime by functioning it repeatedly in a short period of time. Other common NW examples of use-rate aging involve sample environmental tests performed during production. These tests have the intent of emulating a weapon lifetime through application of mechanical environments and thermal cycling (all at or within normal lifetime or Stockpile to Target Sequence environmental extremes), but in a short period of time. Use-rate aging was also at the heart of the historical NW Accelerated Aging Unit (AAU) program, where a surveillance sample was temperature cycled (within the normal STS range) and then disassembled and inspected for changes. In general, the use-rate form of accelerated testing is straightforward both in its application and resulting interpretation as long as life is governed by only the number of times that the product is used without regard to the frequency per unit time of use.

With respect to increasing the *aging-rate* of a product, we are concerned about failure mechanisms that relate to changes in the product over time that are due to its background environment. For example, the dormant storage life of many types of electrochemical cells is affected by chemical reactions that consume the active materials. The reaction rates are temperature dependent. Thus, at $25°C$ a battery may have a useful storage life of 10 years while at $55°C$ it has a useful storage life of only 6 months. Hence, by increasing temperature, we can accelerate the onset of failure. Corrosion is another phenomenon leading to product failure that can be studied by accelerating the aging rate. For example, certain production and storage conditions may lead to corrosion products within energetic devices over the course of decades of storage life. It is desirable to be able to evaluate the storage life of these devices in a much shorter amount of time. In other cases, slowly-progressing chemical reactions may degrade material properties (e.g., tensile strength) over time and eventually cause the product to fail during its eventual use.

With respect to *level-of-stress* acceleration, we are concerned about failure that is accelerated by increasing the level of stress (beyond normal use conditions) in which a product operates (e.g., current, voltage, g-forces, shock). These failures are not due to age-related degradation of the product, per se, but are related to how the product is used (or transported). For example, lithium-ion rechargeable batteries experience two type of degradation (one related to aging-rate, the other related to level-of-stress). The first type (related to aging-rate) is associated with dormant storage when the battery is not in operation (i.e., not being charged or discharged). The factors that influence the degradation of the batteries in this case are temperature and state-of-charge. The second type of degradation (related to level-of-stress) is associated with active use of the battery (i.e., as it is being charged and discharged). The factors that influence the degradation of the batteries in this case include charge and discharge rates (and the number of charge/discharge cycles). In NW applications, *level-of-stress* acceleration relates to the ability of a component to survive stress (higher than use levels) without failing. For example, consider a capacitor that needs to operate for 100 cycles at low voltage over its lifetime. The level of accumulated damage at the normal operating condition may be equivalent to the accumulated damage acquired with 10 cycles at a higher voltage. Often the intent of level-of-stress testing is to understand and improve the margin of a product, with margin in this case meaning the level of a particular input or environment that the product can withstand before failing as compared to the required level.

Clearly the choice of which acceleration to use (use-rate, aging-rate, or level-of-stress) is driven by the nature of the product, the failure mechanism, the presumed model for accelerating the failure mechanism, testing capabilities, the question to be answered by the evaluation, etc. By using accelerated tests it is possible to rapidly acquire performance information to support each of these models. Nelson [2] provides a comprehensive treatise on all aspects of accelerated testing (and analysis). In addition, Nelson [3, 4] provides a relatively current bibliography of more than 100 references on statistical plans for accelerated tests.

In order to exploit accelerated testing, it is necessary to have first identified the dominating failure mechanism(s). This may be easier for materials and components and more difficult for complex assemblies and systems. It is essential to note that in order to effectively use accelerated tests to make credible performance or reliability estimates at use conditions, one must have knowledge of the acceleration mechanism(s) and an accurate model that relates performance to the level of the accelerating factor(s). For example, if a chemical reaction is the mechanism that affects performance (and limits useful life) then temperature might be used as the accelerating factor. An Arrhenius model might be useful to mathematically express the relationship between performance and temperature. Escobar and Meeker [5] describe a number of accelerating factors and models (empirical and physical).

Whether empirically or physically based, a model should be acknowledged as an imperfect approximation to reality. George Box once wrote that "all models are wrong, but some are useful." Here, the utility of a model is measured by its ability to accurately extrapolate performance from accelerated conditions to use conditions. In practice, the accuracy of the extrapolation is heavily dependent on the degree of acceleration. In general, the accuracy of the extrapolation decreases as the degree of acceleration increases. While this document relates to accelerated testing in general, <u>the emphasis is on cases where the level of acceleration is large or extreme</u>. In such cases the testing is often referred to as being <u>highly accelerated</u>.

In some cases, the performance data take the form of a continuous response variable that relates to the useful life of a product. For example, the thickness of a resistive layer grows as a consequence of a chemical reaction within a lithium-ion electrochemical cell. As this layer grows, the discharge resistance of the cell increases [6]. At some point, the resistance increases to a point where the cell's performance is unacceptably degraded. In cases where units can be measured multiple times, degradation paths can be observed and modeled as a function of the accelerating factor. Tests that provide such information are referred to as accelerated degradation tests. More commonly, accelerated life tests are conducted. Accelerated life tests, which are generally less informative than accelerated degradation tests, provide information regarding whether or not each test unit has survived a particular stress/exposure. From these tests, only the failure times for units that fail (perhaps left censored) and the running times for units that have not failed are observed.

In this document, the primary focus is on the use of accelerated life tests to improve, predict, and demonstrate reliability. Section 2 contains a summary of HALT (highly accelerated life test) which is a *specific class* of highly accelerated testing methods for finding and fixing problems during design, development, and preproduction. By doing so, HALT can improve reliability. Section 3 discusses the quantification of reliability, particularly in the NW context. First, this section briefly illustrates how reliability might be characterized by using both parametric and nonparametric approaches with data acquired at use conditions. Next, various approaches for using accelerated tests to characterize reliability are described. Section 4 compares the various approaches for accelerated testing as a means to characterize reliability and discusses the attributes and potential pitfalls of each. Section 5 summarizes and contains recommendations for conducting informative accelerated tests.

2. HALT

HALT (as described by Hobbs [7] and McLean [8]) is a *specific class* of highly accelerated testing methods for finding and fixing problems during design and preproduction. McLean [8, p. xix] views HALT as a "process for the ruggedization of preproduction products." According to McLean [8, p.2], "HALT constitutes both singular and multifaceted stresses that, when applied to a product, uncover defects. These defects are analyzed and driven to the root cause, and corrective action is implemented. Product robustness is a result of adhering to the HALT process." HALT is broadly viewed as a prescriptive process for testing pre-production units under extreme levels of one or more stress factors (e.g., extreme temperatures, thermal cycles, and vibration). The implication is that by following HALT during product development, the portion of the unreliability of a product *related to margin* can be reduced by making the product more robust to exposures to extreme environments. However, HALT is generally viewed as being *qualitative* in nature and is not intended to make quantitative inferences (e.g., regarding reliability). Also, note that HALT is not intended to address issues in production (e.g., process shifts) that generate defects which can also adversely influence reliability. HASS (highly accelerated stress screen), which is not discussed further here, is the process recommended in [7] and [8] for ensuring that units with such defects are not delivered to customers.

It is widely believed that HALT can be used to improve reliability. McLean [8, pp. 28-30] gives a number of examples where HALT was found to be beneficial. However, it is somewhat surprising that there aren't more examples in the open literature. McLean [8, p. 135] offers the following in this regard – "Since 1990, some users of these techniques have been willing to share their product and business improvements, but this number remains relatively small and lacking in detail. The majority have not had the desire or time to share these techniques because of the competitive advantages that these techniques can realize, and they don't want their competitors to know why they're pulling away from the pack. Also, time is a luxury that many can ill-afford to invest in publishing their findings."

There are caveats associated with using HALT for its intended purpose (see e.g., [2, pp. 37-39]). For example, Nelson [2, p. 38] describes a case in which hundreds of thousands of a type of television transformer had been manufactured and gone into service. An engineer conducted a HALT experiment that revealed a new failure mode. A re-design corrected the failure mode. However, no transformer from the old design ever failed from the new mode. In this case, the re-design was unnecessary! Thus, when there is a failure during a HALT experiment, it is necessary to find and carefully study the failure's root cause and assess whether the failure mode could occur in actual use conditions. Subject-matter knowledge (often combined with fundamental modeling of the particular failure mode observed) is essential for making such a determination.

HALT is generally regarded by the statistical community as a non-statistical, engineering approach [2, 3, 9] that cannot properly be used to demonstrate or estimate reliability, or otherwise make quantitative inferences at use conditions. Nelson [2, p. 194] states “An entirely different purpose of accelerated testing is to force the product to fail to *discover* failure modes that would occur in actual use. Then the product or process is improved to reduce those failure modes. Such testing is used during product development or debugging of the production process, and includes HALT, HASS, and environmental stress screening. This bibliography does not include references to such techniques as they are nonstatistical, and do not yield estimates of product life.”

In fact, while engineers are able to use HALT to find and fix potential problems without a statistical basis, there is not perfect clarity among its proponents regarding whether HALT can or should be used to make quantitative inferences, such as reliability assessments. Regarding HALT and HASS, Hobbs [7; p.1-2] states “The HALT and HASS methods are designed to improve the reliability of the products, not to determine what the reliability is.” On the other hand, McLean [8, p. 132] states that “Presently, HALT does allow one to calculate a reliability number (see the last section in this chapter)”. The section referred to [8, p.145] is rather vague and relates to issued (or to be filed) patents on this topic. [A literature search revealed three possibly relevant issued US patents; 7,120,566, 7,149,673, and 7,260,509.] Without providing further details, McLean states “Results indicate an excellent correlation between the results from HALT and the field. This estimator will be available to practitioners on a Web site in the future.”

Meanwhile, practicing engineers may be tempted to use results acquired from HALT (and other types of “highly accelerated” life tests) as a basis for making quantitative inferences. For example, suppose that a HALT experiment revealed no new failures. With certain strong assumptions (e.g., regarding the relationship between the level(s) of the accelerating factors and the probability of product failure), engineers may be inclined to make quantitative inferences about performance (e.g., reliability) at use conditions based on exposure of a limited number of units to extremely elevated levels of stress. The purpose of the following sections is to explain how highly accelerated tests (including HALT) might be used to attempt to make such inferences and the pitfalls of doing so.

3. QUANTIFICATION OF RELIABILITY

The performance of a product might be characterized in many different ways such as accuracy, versatility, speed, availability, durability, and reliability. Here we will focus on the quantitative characterization of reliability. Bazovsky [10] attributes the following definition of reliability to Knight et al. [11] "Reliability is the probability of a device performing its purpose adequately for the period of time intended under the operating conditions encountered." For many commercial "continuously operating" products, reliability is generally given as a probability distribution relating to the period of time associated with failure-free operation. In such cases, the period of failure-free operation is often measured by time-to-failure or cycles-to-failure. For many military and aerospace products, reliability relates to the ability of an item to perform over the duration of a particular mission given a specified set of use conditions (environment and operation).

Specific to the NW context, nuclear weapon reliability is defined as the probability that a unit selected from the stockpile will achieve the specified yield at the target, given proper inputs and assuming operation anywhere across the entire range of normal STS environments, and throughout the life of the weapon. In the NW context, reliability is quantified by considering relevant system and component test data [12]. In doing so, there are rather stringent criteria in place to judge the usability of test data and other evidence in making a reliability assessment:

- The information must be representative of stockpile performance – this means that the tested hardware must be sufficiently representative of that in the stockpile and that the tester and test procedure must faithfully report performance as it would occur in a real operation.
- One must be able to infer from the data whether or not yield at target would have been achieved.
- One must understand well the scope of the evaluation in identifying the full range of defectiveness that might be present in a sample. As a simple example, tests conducted at hot temperature are incapable of detecting defects only manifested at cold temperature. In general this means that multiple complementary tests must be done in adequate quantity in order to ensure that all defects are findable. No one evaluation type is sufficient to do this. Thus, it is important to realize the limitations and benefits of each part of the test program as they relate to understanding performance in the large environmental/operational/lifetime state space over which weapons are assessed.

Design robustness to various stressors is of course important. Ensuring margin during design and development is a critical foundation upon which long-term reliability is built. However although it is a necessary part of the overall test program, it is not in and of itself sufficient.

While the reliability of a product can be legitimately assessed directly only through fielded experience, it is important that the producer have some notion of a product's capability (in terms of reliability or margin) prior to being fielded. In general in the commercial world, if there is sufficient time and/or a sufficient number of units available to be tested, reliability can be characterized via testing at use conditions. However, due to tight deadlines, there is often a need for accelerated testing to provide a quicker assessment of a product's capability. In the case of nuclear weapons, there is clearly a desire to have some understanding of performance over the lifetime without requiring test durations of decades. Note that defects due to uncontrolled variation during later production may keep the product's reliability from achieving the capability that was observed during development and early production. On the other hand, improvements in the production process (that result in fewer defects) may allow the actual reliability to exceed the capability that was observed during development and early production.

During development and early production, there are various approaches for characterizing reliability. In most of the approaches discussed here, reliability is inferred directly from failure data (i.e., a model of reliability as a function of either time or stress is developed based upon the point at which the failures occurred). The time-to-failure analogy will be used here for illustration. The subsections that follow describe some of these approaches where reliability is described with reference to a requirement (e.g., in terms of a minimum time-to-failure). The first two approaches do not involve accelerated tests (testing is performed at use conditions) but serve to illustrate the use of parametric and nonparametric methods. The third approach involves accelerated testing and is fully parameterized in terms of a model for probability of failure as a function of the stress factor(s). The parameters are estimated via an experiment. The fourth approach is a variant of the third approach where it is assumed that some of the model parameters are known, allowing for a nonparametric "pseudo demonstration" of reliability. The fifth approach involves the use of degradation data to characterize reliability. In this approach, the results of the testing will be expressed as some performance measure as a function of time. By itself, this is not reliability. To make an inference about reliability, one must then compare the degradation data to a requirement that has a well-understood relationship to performance. Knowledge of such a performance requirement will be assumed here, but finding it may be difficult and is outside of the scope of this discussion. Lastly, a brief description of some Bayesian approaches is provided.

The nature of the data used to characterize reliability may depend on the testing approach that is used. In some cases in which units are tested to failure, failure times are observed directly. In other instances, the failure times are censored. That is, we might only know that a unit failed prior to some specified time (left-censored observation). This is very common for NW components, since they reside in dormant storage and thus will not manifest failure until explicitly tested even though the failure may actually have occurred much earlier or could have even been observed "at birth". In other cases, we might only know that a unit continues to operate beyond the last running time. Such observations are referred to as being right-censored. In other cases, failure times are interval-censored. That is, the failure time is known only to have occurred within a specified time interval. When all of the failure times are observed directly, the resulting data are said to be complete. When at least some of the failure times are censored, the resulting data are said to be incomplete. In general, complete data are more informative than incomplete data. However, more time and effort may be needed to acquire complete data. Methods for analyzing a data set that includes incomplete data differ from methods for analyzing complete data (see e.g., [1] and [13]).

3.1. Parametric Approach for Characterizing Reliability via Life Testing at Use Conditions

Here, test data obtained at use conditions are used to estimate the parameters of a probability distribution. It is assumed that some failures can be observed in a reasonable short time during normal use, such as in the case of incandescent light bulbs. While this is grossly unlike NW applications of interest at Sandia (where products are fielded for decades), it is useful for illustrating the parametric approach without acceleration. The test data might be complete or incomplete. For example, consider a case where complete test data (testing units to failure) are used to fit an exponential model, where $F(t) = 1 - exp(-\lambda \cdot t)$ represents the proportion of units that fail prior to time t. Here, λ (hazard rate) is the model parameter. The test data are used to obtain an estimate of λ, given by $\hat{\lambda}$. The fitted model, $\hat{F}(t) = 1 - exp\left(-\hat{\lambda} \cdot t\right)$, is then evaluated with respect to the requirement (t_*) in order to predict reliability. That is, the proportion of units that have a failure time exceeding t_* is estimated to be $\hat{R} = 1 - \hat{F}(t_*) = exp(-\hat{\lambda} \cdot t_*)$. It is assumed that the units tested are selected at random from the population of interest. It is also assumed that the model form is accurate.

Figure 1 illustrates a case where n=30 units are tested to failure (i.e. actual failure times are observed) and analyzed assuming an exponential distribution. The failure times are indicated by black *'s. The fitted model, $(1 - \hat{F}(t)$, is indicated by the thick blue curve. The average failure time in this case is $\bar{X} = 2{,}648$ hours. The requirement in this case is $t_* = 100$ hours. Here the estimated hazard rate is $\hat{\lambda} = \frac{1}{\bar{X}} = 4.052 \cdot 10^{-4}$, so that $\hat{R} = exp\left(-\hat{\lambda} \cdot t_*\right) = 0.960$. By assuming an exponential distribution, one can construct a $100 \cdot (1 - \alpha)\%$ upper confidence bound for λ. That bound is given by the inequality, $\lambda < \hat{\lambda} \cdot \frac{\chi^2_{1-\alpha,\ 2 \cdot n}}{2 \cdot n}$, where $\chi^2_{1-\alpha,\ 2 \cdot n}$ is the $(1 - \alpha)$ percentile of a chi-squared distribution with $2 \cdot n$ degrees of freedom. Thus, a 95% upper confidence bound for the hazard rate is $\lambda_{ub} = 5.34 \cdot 10^{-4}$. It follows that a 95% lower confidence bound for R is $R_{lb} = exp(-\lambda_{ub} \cdot t_*) = 0.948$. Note that the failure times used in this example were simulated from an exponential distribution with $\lambda = 0.0005$. Thus, $R = exp(-0.05) = 0.951$. Figure 2 illustrates the use of a probability plot and the Anderson-Darling statistic to judge the adequacy of the exponential model assumption. Here, as expected, there is no evidence to reject that assumption (p-value = 0.279).

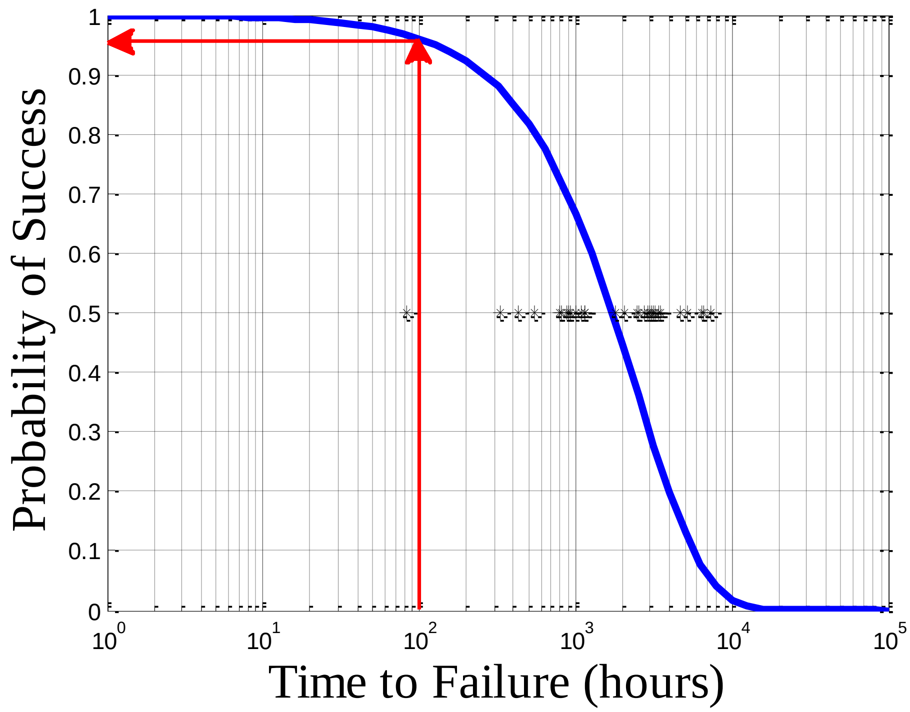

Figure 1. Predicting reliability from fitted exponential model.

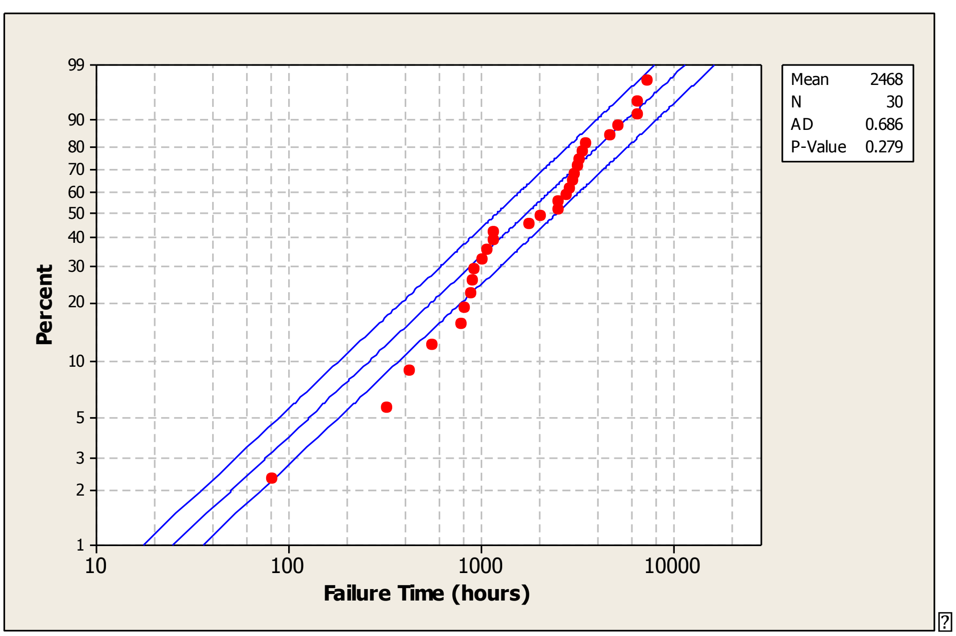

Figure 2. Probability Plot.

There are similar approaches for estimating reliability (and establishing lower bounds) for other distributional models, many with multiple parameters (see e.g., [1]). The Weibull model, where $F(t) = 1 - exp\left[-\left(t/_{\alpha}\right)^{\beta}\right]$ represents the proportion of units that fail prior to time *t*, is often used to model failure-time data. Both of the model parameters (α and β) can be treated as unknown and estimated by the failure-time data. Alternatively, one might assume that β is known (e.g., see page 70 in [14]). The Weibull model, which is a flexible generalization of the exponential model, will be referred to later in this document.

Sandia examples which may be associated with time-dependent reliability (albeit at a much longer time scale: decades rather than hours) are related to corrosion (e.g., moisture inadvertently sealed inside of integrated circuits, degradation of energetics due to contamination), stress voiding of integrated circuits (where voids gradually grow in size until an open circuit condition occurs in a via, resulting in failure of the component), and electrical component performance changes due to exposure to low-dose radiation. When functional test failures have been observed in these cases, the failure times have been left-censored and thus the parametric model was not used due to large uncertainty in failure times. Accelerated tests may be useful to investigate such degradation processes although it has also been observed that sometimes increasing the level of a stress factor actually serves to prevent rather than induce these defects more quickly (a notable example being stress voiding).

3.2 Nonparametric Approach for Characterizing Reliability via Life Testing at Use Conditions

The example in the previous sub-section illustrates a parametric approach for estimating reliability, since it involves a model with parameters. Note that there are ways to characterize the failure time distribution without assuming a parametric model (hence a nonparametric approach). While such methods naturally require fewer assumptions, they are generally useful only for interpolation. For instance, consider the empirical cumulative distribution function of the failure times, denoted by $\hat{E}(t)$. To illustrate, consider the example in the previous section. Figure 3 compares the complementary empirical cumulative distribution function, $1 - \hat{E}(t)$, with $1 - \hat{F}(t)$ based on the exponential model. In this particular case, the estimate of reliability given by $1 - \hat{E}(t_*)$ is a special case of the Kaplan-Meier estimator [15] since the failure time data are complete. Note that the general form of the Kaplan-Meier estimator can be used with certain types of incomplete data as well. Other related approaches are in Chapter 3 in [13].

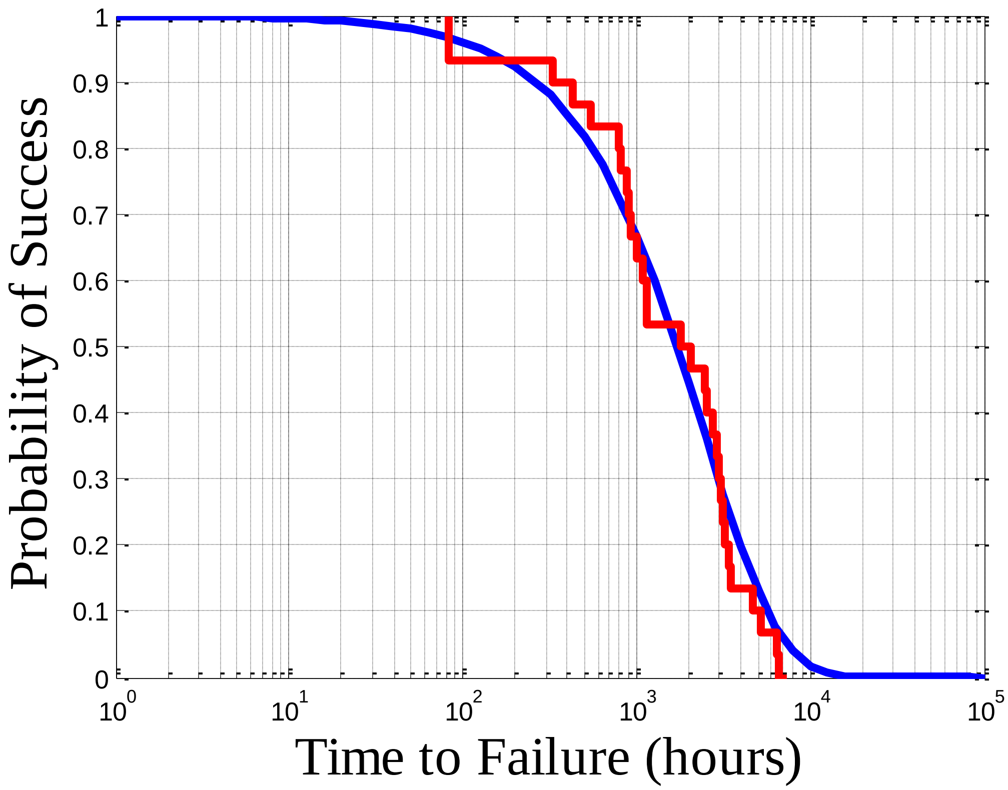

Figure 3. Probability of success: $1 - \hat{F}(t)$ (blue) and $1 - \hat{E}(t)$ (red).

Note that it is not necessary to parameterize (or even characterize $F(t)$ over a range of t) in order to establish a lower bound for reliability. Again, suppose that the lifetime requirement is given by t_*. Also assume that units are tested to the requirement. As a result of the testing, one might only know whether (or not) a unit failed prior to the requirement. In this case suppose that a number of units (n) are tested (at use conditions) to the minimum time-to-failure requirement. Depending on the number of units tested and the number of failures observed (m), one can claim (with a certain level of confidence) that some level of reliability has been demonstrated. For example, suppose that seventy units have been tested (n=70) with no failures observed prior to the requirement (m=0). Then, one can claim (with caveats discounting changes in performance during later production as well as assumptions regarding test efficacy) that the reliability is at least 0.99 with 50% confidence. This claim can be made since if the reliability was less than 0.99, the probability that we would have observed at least one failure is at least 0.50. In general, with n units tested with m failures observed, we can claim with $C \cdot 100$ % confidence that the reliability is at least R if $1 - \sum_{k=0}^{m} \binom{n}{k} \cdot R^{n-k} \cdot (1-R)^k \geq C$. This confidence statement is based only on the binomial distribution of successes/failures. The only assumptions are that the units tested are selected at random (therefore representative) from the population of interest and that tests performed on a sample would detect a defect if one were present. Nothing is assumed about the probability distribution of failure times. Therefore, such a nonparametric approach is often considered the most robust approach for demonstrating reliability of fielded product since its accuracy is not impacted by incorrect assumptions regarding the underlying distribution (including population homogeneity, which is seldom observed) and it does not depend upon time-of-failure information (seldom available for NW components because they reside in dormant storage). This approach is often used by Sandia reliability engineers to provide a lower bound for a component's reliability and is referred to here as the "nonparametric-binomial approach." However, disregarding information about the distribution (if available and credible) can result in a conservative bound. To illustrate, based on the data in Figure 1, suppose that we had observed only whether units passed or failed the 100 hour requirement. If so, we would have n=30 and m=1. Without presuming a distributional form, we would only be able to claim with 95% confidence that the reliability exceeds 0.851. Clearly, this bound is not as precise as the one developed earlier which used complete data and a presumption of the model form in its development. The added precision is a direct consequence of the complete data and the added assumption. One must decide if the risk of making this assumption (possibly overstating reliability) is outweighed by the benefit (increased precision). Note also that testing to a fixed requirement offers no ability to assess performance at a more demanding requirement (or stressful condition).

3.3 Parametric Approach for Characterizing Reliability via Accelerated Life Testing

Engineers are driven to perform accelerated tests in order to minimize cost and save time due to constraints associated with budget and schedule. When accelerated testing is used to assess reliability, additional knowledge/assumptions are required. First, subject-matter experts need to identify the dominant failure modes/mechanisms that can occur within normal use conditions. Next, a relevant process for accelerating the identified failure modes (manifested in terms of a controllable stress factor) needs to be established. Finally, an accurate model providing a quantitative relationship between the onset of failure and stress level needs to be developed. This model must be valid across the range of conditions where it is developed and applied (use conditions through accelerated conditions) in order to have utility in predicting reliability at use conditions from data acquired at accelerated conditions. In general the wider this range of conditions is, the less likely it is that the model will be sufficiently accurate.

When accelerated testing is used in conjunction with a parametric probability model, the parameters of the probability model are presumed to be functions of one or more stress factors. For example, consider the exponential probability model used in conjunction with an Arrhenius relationship between the hazard rate and temperature. That is, λ in the exponential model depends on absolute temperature (T) via the relation $\lambda(T) = \alpha \cdot exp\left(-\frac{\beta}{T}\right)$. Effectively the model is that of an age-rate defect and increased temperature is being used to emulate increased age. Temperature is the stress factor while α and β are model parameters to be estimated. The Arrhenius relation can be useful for modeling the effects of thermally activated mechanisms associated with various materials (e.g., growth of layers/films on an exposed surface). Such a relationship might also be a reasonable model for corrosion growth in NW components. Changes in electrical properties (e.g., resistance) or mechanical properties (e.g., tensile strength) giving rise to failure might be well approximated by an Arrhenius relationship (see Section 3.5).

Other commonly used stress models include variations of the inverse power relationship (law). A generic representation of this relationship is $\tau(S) = {}^{\alpha}/_{S^{\gamma}}$, where $\tau(S)$ is the lifetime of a product as a function of the stress level (S). Typically, these relationships are empirically based and are often associated with mechanical stresses (level-of-stress acceleration). The parameters (α and γ) depend on the product and need to be estimated experimentally. Specific examples of this generic form are the Coffin-Manson relationship and Palmgren's equation. The Coffin-Manson relationship models fatigue failure of metal subjected to thermal cycling. In this relationship, the number of cycles-to-failure (N) is given by $N = {}^{A}/_{(\Delta T)^B}$, where ΔT is the temperature range of the thermal cycle and A and B are model parameters to be estimated. Palmgren's equation is used in applications where life is assessed in terms of the level of mechanical load. At Sandia, the inverse power relationship combined with a Weibull model has been used to model the lifetime of a capacitor as a function of applied voltage [16]. In terms of damage (or degradation), the model might be expressed as $\delta(S, C) = {}^{C}/_{S^{\gamma}}$, where δ represents the level of damage and C represents the number of cycles (or amount of time).

These models, applicable to level-of-stress acceleration, are particularly relevant for investigating margin to mechanical stress (e.g., shock/vibration) that might occur during the transportation or use of NW components.

Escobar and Meeker [5] and Nelson [2] describe a number of other accelerating factors and models (both empirical and physical). Some of these models involve multiple stress factors. For example, Escobar and Meeker [5] discuss an extended Arrhenius relationship that involves both temperature and voltage (V). In this case, $\lambda(T) = \alpha \cdot exp\left(\frac{\beta_1}{T}\right) \cdot exp\big(\beta_2 \cdot log(V)\big) \cdot exp\left(\beta_3 \cdot \frac{log(V)}{T}\right)$. Thomas et al. [17] describe accelerated testing of lithium-ion cells using both temperature and resting state-of-charge as accelerating factors.

Most models relating the onset of failure to stress relate to exposure to a constant stress level. However, in real applications stress can be variable. For example, the product may operate outside and thus be subject to a variable temperature environment. In some of these cases, it might make sense to consider an accumulated damage model that can be responsive to variable stress. One model that was developed is Miner's rule [18]. Miner's rule supposes a product can tolerate only a certain amount of damage, D_*. If a unit experiences damages $D_i \; for \; i = 1,2,\cdots,n$, the total damage on the unit is assumed to be additive. That is, $D_{cum} = \sum_{i=1}^{n} D_i$. The product will fail to operate if $D_{cum} \geq D_*$. This model, which is deterministic, is known to be inadequate for metal fatigue (and other situations) as it does not consider the effect of varying the sequence/order of the $D_i's$ (e.g., see [19]). A number of experts in the physical and engineering sciences have developed other deterministic cumulative damage models for various situations (see e.g., [1] [4]). Statisticians have helped to transform these deterministic models into probabilistic versions (e.g., see [19]). Examples of statistical approaches for modeling the effects of exposure to variable thermal environments are given in [20] and [21].

In general, regardless of the particular combination of probability model / stress relationship that is used, estimates of the model parameters (e.g., α and β in the case of an Exponential/Arrhenius model) need to be obtained. The most straightforward way to do this is via an experiment in which failure time data (either complete or incomplete data) are acquired at several accelerated levels of the stress factor (e.g., temperature). The reliability as a function of time can be predicted by using the estimates of the model parameters. For example, in the case of the Exponential/Arrhenius model, $\hat{R} = exp(-\hat{\lambda}(T_{use}) \cdot t_*)$, where $\hat{\lambda}(T_{use}) = \hat{\alpha} \cdot exp\left(\frac{\hat{\beta}}{T_{use}}\right)$, $\hat{\alpha}$ and $\hat{\beta}$ are the parameter estimates, and T_{use} is the nominal use temperature. One could conceive of this example as being illustrative of level-of-stress acceleration. On the other hand, if $T_{storage}$ is substituted for T_{use}, then the example could be representative of aging-rate acceleration. Nelson [2] discusses a number of other commonly used probability model /stress relationship combinations (e.g., Weibull/Power Law model). In all cases, the underlying models are *fully-parameterized*. The accuracy of a reliability prediction, $\hat{R}$, based on the underlying models depends on the accuracy of all of the assumptions associated with both the probability model and the stress relationship. The precision of the reliability prediction depends on the uncertainty of the estimated model parameters which in turn depends on the specifics of the experiment(s) used to estimate model parameters (including the number of units tested and whether or not sufficiently complete data were acquired). In addition, the accuracy and precision will depend heavily on how far the use condition must be extrapolated from the accelerating conditions. Various statistical methods can be used to estimate the model parameters, predict reliability, and assess model accuracy.

3.4 Parametric/Binomial Approach for Characterizing Reliability via Accelerated Life Testing

In many cases, it is not practical to use a strict nonparametric approach to demonstrate reliability with testing at use conditions as in section 3.2 due to prohibitive requirements on the number of units to be tested. In addition, it may be perceived that full characterizations of the probability and stress models via experimentation are unnecessary due to *existing knowledge* regarding the model parameters. Within this context, a mixed parametric/nonparametric approach has been proposed (and used) at Sandia and elsewhere. It has been claimed to be a means to "demonstrate reliability" to a particular requirement with accelerated testing.

The basis for this approach is provided by the "Parametric Binomial" or "Lipson" equality (see Lipson and Sheth [22]). This approach provides a mathematical framework to tradeoff the sample size of a test with its duration. That is, with strong assumptions concerning the underlying model, one can claim to "demonstrate reliability" by testing fewer units$\left(N_{\dagger}\right)$, each for a longer duration $(t_{\dagger})$. The Lipson equality, which blends both *parametric* and *nonparametric* statistics, assumes that the age-rate mechanism for the dominant failure mode produces times-to-failure (or cycles-to-failure) that are distributed according to a Weibull model. The assumption of a Weibull model reflects the parametric part of the "Parametric Binomial" equality. The Weibull shape parameter, β, is <u>assumed to be known</u> to be above a lower bound. With these assumptions, one can demonstrate that the fraction of units that will fail the mission life requirement (t_*) is less than $1 - R$ with confidence $C \cdot 100$ %., if $N_{\dagger}$ units are tested without failure each to a duration $t_{\dagger} = t_* \cdot \sqrt[\beta]{\frac{N_*}{N_{\dagger}}}$. Here, from Section 3.2, N_* represents the required number of units to be tested at use conditions without failure to t_* in order to demonstrate a reliability level of R with confidence $C \cdot 100$ %. Note that the benefit (in terms of reducing the sample size) increases as the assumed value of β increases. For example, in the case of an exponential distribution $(\text{i.e.}, \beta = 1)$, $t_{\dagger} = t_* \cdot \frac{N_*}{N_{\dagger}}$. Thus, assuming an exponential model, we must test units sufficiently long without failure such that $t_{\dagger} \cdot N_{\dagger} \geq t_* \cdot N_*$.

This approach has been applied to various situations at Sandia. For example, in a capacitor sampling plan, 6 units per lot over an anticipated production run of 15 lots are planned to be subjected to a 3600VDC stimulus for 20 hours. Assuming an exponential failure time model (or any other Weibull distribution with $\beta \geq 1$ … implying a non-decreasing hazard function) and that no breakdowns are observed in the 1800 unit-hours of testing, we will have 66% confidence that the probability of a unit breaking down during a 1-hour 3600VDC stimulus is no more than 0.0006. In the case of another component, it was conjectured that the time-to-failure is distributed according to a Weibull distribution with $\beta = 3.5$. It is required that the component operates following 14 hours of stress caused by temperature cycling. It was desired to "demonstrate" that the probability of a unit breaking down prior to 14 hours of stress is no more than 0.045. Using the binomial success/failure model, one would need to test $N_* = 15$ representative units without failure to $t_* = 14$ hours to "demonstrate" that the reliability exceeds 0.955 with 50% confidence. Apparently, it was desired to test only $N_{\dagger} = 2$ units. By using the Lipson equality (conservatively using $\beta = 2.0$), it was suggested that the 0.955 reliability could be demonstrated if the two test units functioned after $t_{\dagger} = 38.3$ hours of stress $= t_* \cdot \sqrt[\beta]{\frac{N_*}{N_{\dagger}}}$. Of course, such an assertion requires that the Weibull model holds (with $\beta \geq 2.0$) and that failures due to other types of defectiveness (i.e., infant mortality) are inconsequential.

In general, the validity of the "Parametric Binomial" approach depends on knowledge of the distributional form and the associated parameter (shape parameter in the case of a Weibull distribution). Even if the Weibull model form is accurate, use of an optimistic assessment of β (by assuming a larger value than truth) will result in an invalid inference regarding R. Thus, it is important to recognize that β is difficult to estimate experimentally with meaningful precision. For example, even with a sample size of 30 failure times (complete data), β can only be estimated to within a relative standard error of about 15%. Incomplete data are much less efficient for estimating β. The effect of over-estimating β is illustrated as follows. Suppose that β is assumed to be 2.0 when in fact $\beta = 1.0$. In this case, the test time suggested by the Lipson equality (for the assumed value of β) will need to be multiplied by the factor $\sqrt{\frac{N_*}{N_\dagger}}$ in order to be appropriate for the true value of β. Furthermore, a significant disadvantage of this approach is that "defect-related failures (which might occur in only a small proportion of units) might not show up in the sample" (p.249 in [1]). This is particularly important if $N_\dagger$ is reduced to a very small number, as in the previous example where $N_\dagger = 2$. The historical experience with NW components as fielded is that unreliability is, in fact, dominated by defects due to non-homogeneous production (materials and/or processes). Design robustness to various stressors is of course important, but practitioners of this approach risk radical misunderstanding if they believe that this approach alone can "demonstrate reliability" of the product as fielded.

With additional assumptions, the "Parametric Binomial" approach can be further extended to provide additional testing flexibility (and even greater acceleration). Suppose that we first use the "Parametric Binomial" approach to specify a smaller sample that is to be tested longer. However, suppose that $t_\dagger$ is still unacceptably large. If we assume knowledge of the relationship between the life of a product and the stress level, we might be able to reduce the test time (of the reduced sample) to an acceptable level by increasing the stress level. For example, suppose that the life of a product depends on the stress level through the inverse power law, $\tau(S) = {}^{\alpha}/_{S^\gamma}$ with $\gamma > 0$. As in Section 3.2, assume that N_* represents the required number of units to be tested without failure to t_* at the use stress condition (S_*) in order to demonstrate the reliability target, R. Using the "Parametric Binomial" approach, one might demonstrate the reliability target, R, using $N_\dagger$ units tested to $t_\dagger$ using the stress level $S_\dagger = S_*$. If γ is known and the stress level is increased to $S_\ddagger > S_\dagger$, then would could reduce the time on test to $t_\ddagger = t_\dagger \cdot \frac{S_\dagger^\gamma}{S_\ddagger^\gamma}$. Thus, we would have effectively demonstrated the original reliability requirement by testing $N_\ddagger = N_\dagger$ units (under stress $S_\ddagger$) without failure to $t_\ddagger$. This approach could easily be applied to other life-stress relationships (e.g., Arrhenius) associated with acceleration via the aging-rate of a product, or the level of stress under which a product operates.

This approach (which is general, simple, and easy to explain) may be enticing as it apparently offers a rapid and inexpensive way to demonstrate sufficient margin. However, it assumes specific knowledge relating to the acceleration factor as well as the distribution of failure times. By itself, the small set of tests (at accelerated conditions) provides no ability to assess the necessary modeling assumptions or uncertainty.

3.5 Parametric Approach for Characterizing Reliability via Accelerated Degradation

Accelerated degradation tests can be used to study the age-rate degradation of product performance at both use and accelerated conditions over time (see e.g., Chapter 21 in [1] and Chapter 11 in [2]). This form of testing can provide significantly more information than other forms of accelerated tests. However, in order to be effective, such testing must involve one or more measures of performance that relate to the state-of-health of the units. Thus, such an approach requires a high level of knowledge about the product. Failure is usually associated with a certain level of the performance measure. For example, suppose that the performance measure (e.g., resistance in an electrochemical cell) increases as the state-of-health of the product decreases. Then failure occurs when the performance measure increases beyond some critical limit. The goal is to model the progression of the degradation measure across stress conditions and time in order determine a product's useful life. This requires two models. The first, referred to here as a degradation model, expresses the *expected* performance level versus the time and conditions under which a product has been aged. This model can be empirical, chemistry/physics-based, or some combination of both. A wide variety of model forms are possible. In practice, the specific form of the model that is used will depend on the particular technology and set of stress factors. By itself, this model can be used to assess margin. A second model, referred to here as an error model, accounts for the observed variability about the expected performance due to measurement error and the intrinsic differences in performance across test units. The second model can be used to assess the accuracy of the first model as it represents the experimental data. Together, the two models can be used to assess reliability assuming there is a known relationship between reliability and the measure.

For example, Figure 4 shows a fitted degradation model superimposed on experimental data [6]. The semi-empirical degradation model, reflecting the relative change in the discharge resistance of an electrochemical cell, is of the form $\mu(t;T) = 1 + exp\left\{\alpha + \frac{\beta}{T}\right\} \cdot t^{1/2}$, where t represents time and T represents absolute temperature. The ratio defined by dividing the discharge resistance of a cell at time t by the discharge resistance of the cell at $t = 0$ is denoted by $\mu(t;T)$. In this example, cells were aged at various temperatures and were periodically measured. The measurements were not degrading to the units under test. Note, however, that a similar (yet less informative) approach could be used when testing is destructive. Of course, use of nondestructive measurements can reveal the degradation paths of individual units (whereas destructive measurements cannot). For this example, a 30% resistance rise represents the critical performance threshold. Based on the fitted degradation model, the expected cell life is slightly more than six months when aged at 55° C. In this case, the fitted model was developed based on the data acquired at 40° C, 47.5° C, and 55° C. The model suggests a 12.6 year life (on average) for cells aged at the use temperature, 30° C. One can validate/assess the model at the use temperature by comparing the predicted behavior at 30° C (red curve) with the observed data at 30° C (red '+'s). This illustrates that accelerated degradation tests offer an improved ability to assess a model when compared to accelerated life tests (where data at use conditions are not often acquired). Furthermore, one could use the degradation model in combination with the error model (and the distribution of initial resistance values across units) to predict the proportion of cells that meet a resistance requirement for a specified time and temperature (i.e. reliability, presuming a credible requirement has been defined).

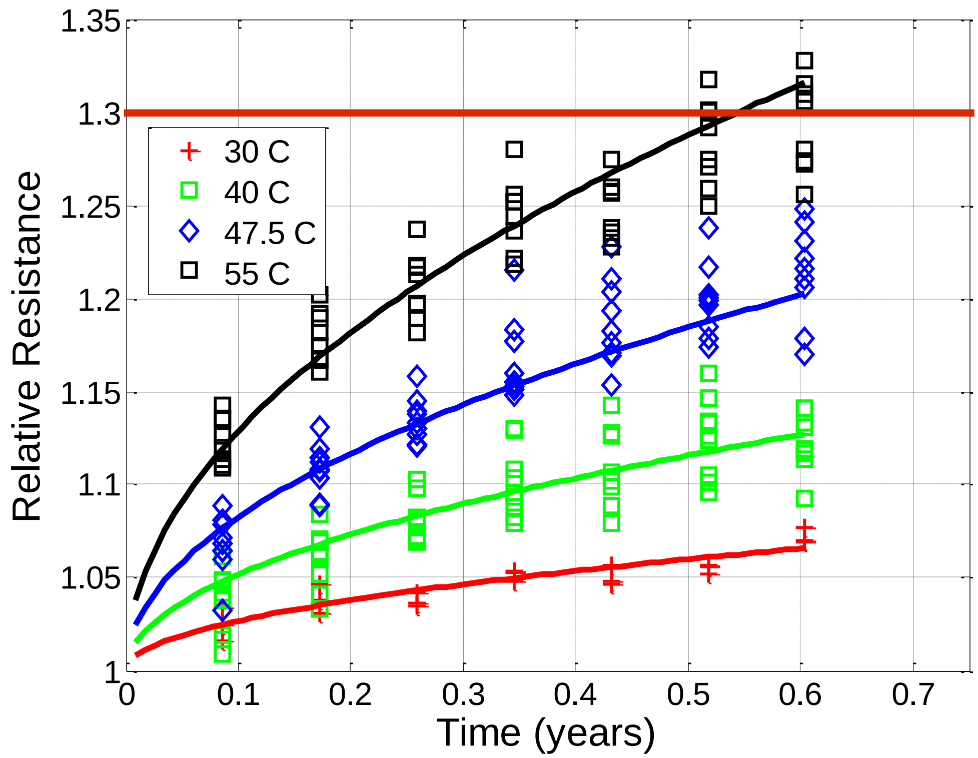

Figure 4. Fitted Degradation Model with Experimental Data

To summarize, the value of degradation tests is clearly stated in [23] … "For the same amount of test time, degradation data always contain more information than the failure-time data and thus provide more precise estimates, particularly when there are few or no failures. Degradation data have another advantage in that they provide much more information for assessing the adequacy of an acceleration model." While the data obtained from an accelerated degradation experiment have much intrinsic value, a credible performance requirement is needed in order to turn the degradation information into a reliability estimate.

3.6 Bayesian Approaches for Characterizing Reliability via Accelerated Testing

There are Bayesian analogs for characterizing reliability with accelerated testing. For example, Hamada et al. [23, pp.239-241] illustrate how one might use a combined Weibull/Arrhenius model to assess the lifetime of a hypothetical mechanical component subjected to temperature stress. In this case a prior distribution is assigned to the Weibull shape parameter (β) directly. The characteristic life (α) is defined in terms of the Arrhenius parameters which are also assigned priors. Hamada et al. [24] also gives a Bayesian perspective on accelerated degradation testing. The relationship between the Bayesian approach and other approaches depends on the particular prior (or set of priors) that is chosen. For example, if the prior is sufficiently concentrated, the result of the Bayesian approach will likely resemble the result of the "Parametric Binomial" approach where parameters are assumed to be known. If the priors are sufficiently diffuse, the Bayesian approach might resemble the various parametric approaches that were discussed in Sections 3.3 and 3.5. It would seem that the Bayesian approach can provide added value if the priors are based on information previously acquired from testing components that are similar to the component of interest (when considering materials and failure mechanisms).

4. DISCUSSION

Various methods for characterizing reliability using accelerated tests were described and illustrated in Sections 3.3 thru 3.6. It is interesting to compare the assumptions and attributes of the various approaches and identify potential pitfalls associated with their use.

The general methods for characterizing reliability are associated with various assumptions. When any of these methods are used to characterize performance, a tacit assumption is that the test units are representative of the totality of production (an assumption that is impossible to justify if units come only from pre-production lots). To the extent that they are not representative represents a significant risk! That is where techniques such as Highly Accelerated Stress Screening (HASS) may be useful (but beyond the scope of this document).

Each of the methods assumes a specific form for an underlying probability distribution. Any use of data acquired at accelerated conditions for making quantitative inferences about performance at use conditions requires additional assumptions regarding the relationship between stress level and performance. The "Parametric/Binomial" approach (Section 3.4) also assumes/asserts specific values for parameters of the distributions. The other approaches allow for estimation of the parameters via the experimental data. The other approaches also allow for the quantification of the statistical uncertainty regarding the model parameters (or functions of the model parameters).

A desirable attribute of an accelerated testing approach is the ability to assess model assumptions. The most direct way to assess assumptions is with experimental data acquired during the accelerated testing. One might also rely on data acquired from other related experiments. Such experiments might definitively reveal the dominant failure modes/mechanisms that occur. Other experiments might provide the basis for assuming a specific stress/performance relationship (e.g., activation energy). <u>In the case of the "Parametric/Binomial" approach, there is little if any ability to assess assumptions.</u> The "Parametric/Accelerated-Life" approach (Section 3.3) provides some ability to assess model assumptions (at least at conditions where experimental data were acquired). However, in general when using this approach, data are not acquired at use conditions so that it is not possible to validate the models assumptions at use conditions. The "Parametric/Accelerated-Degradation" approach (Section 3.5) provides improved ability to assess model assumptions (including at conditions where experimental data were acquired).

Another desirable attribute is robustness to changes in requirements and/or use conditions. Ideally, one would like the test results to be applicable to perturbations of the requirements/conditions. This can be helpful if requirements and/or use conditions change and become more demanding. In the case of the "Parametric/Binomial" approach, there is no ability to assess margin or make inferences concerning performance for more demanding requirements.

Inferences drawn from the "Parametric/Binomial" approach are specific to a particular requirement. With the other approaches, there is at least the possibility of using the fitted models to make inferences concerning performance at more demanding requirements/conditions.

There are various risks associated with using any approach for accelerated testing. Meeker and Escobar [25] and Meeker et al. [23] identify a number of "pitfalls" when using accelerated testing to make predictions about product performance at use conditions. Many of the pitfalls arise from an incomplete knowledge of the failure mechanisms (e.g., multiple (unrecognized) failure modes, and multiple factors affecting degradation). This is an important concern for accelerated testing of MC-level components in particular, which are a complex combination of materials in different geometries and environments. In some of these cases, it is difficult to imagine that a simple model can be used to accurately predict performance.

Other pitfalls arise from the fact that test units may not fully represent the population due to changes in production or some other special aspect of the test units (e.g., modification of the test unit to enable measurement). Still others arise from the fact that laboratory tests do not always reflect use conditions. In many cases, test units are often only exposed to isothermal conditions, when in fact, the thermal environment can be highly variable (even unknown). <u>There is significant risk of inaccuracy in cases when model parameters are asserted or assumed (e.g., "Parametric/Binomial") based on published or previously used values for similar products/materials.</u> For example, Nelson [2 p. 253] comments that in many cases acceleration factors (based on assumed model parameter values) "are company tradition and their origins may have been long forgotten."

Perhaps the biggest risk has to do with the intrinsic nature of an accelerated test. That is, accelerated tests are performed at conditions more extreme than use conditions. Statistical models (whether physically or empirically based) used to predict behavior at use conditions must extrapolate to use conditions. Inaccuracies in the models will naturally be magnified when extrapolated to use conditions. Use of extreme levels of acceleration exacerbates the effects of these inaccuracies. Even if the models are accurate, statistical uncertainty in the estimates of the model parameters will result in significant uncertainty in predictions at use conditions. Again, extreme levels of acceleration amplify the problem. Independent of whether or not the models are accurate, computer codes can provide predictions (and in some cases, estimates of uncertainty) at use conditions without regard to the level of acceleration. <u>Thus, indiscriminate acceptance of the output from these codes without a careful assessment of model accuracy is a significant risk.</u>

In other cases, extreme acceleration activates failure modes not observed at use conditions. In such cases the effects can vary, but in general results in an inaccurate extrapolation (often anti-conservative). For example Bernstein et al. [26] studied the degradation behavior of parachute materials using an accelerated aging experiment using tensile strength as a degradation measure. During this study, tensile strength data were obtained from nylon parachute specimens thermally aged at 138, 124, 109, 99, 95, 80, 64, 48, and 37°C for up to 5.5 years. An Arrhenius plot derived from the experimental data reveals significant nonlinear behavior indicating a change in mechanism (see Figure 5). The y-axis of this plot (shift factor)) is proportional to the degradation rate. The x-axis of this plot represents inverse absolute temperature. In this case if the high temperature data were used to predict life at use conditions, the prediction of life would be "too high by many orders of magnitude" [26]. That is, projecting the line from the highest three temperatures (on the left-side of the plot) to use conditions (on the right side of the plot) would underestimate the shift factor (degradation rate). <u>Thus, it is very risky to use extreme levels of acceleration to predict performance at use conditions.</u> Also, the value of a degradation study (having a wide range and large number of acceleration levels) is clearly illustrated by this example.

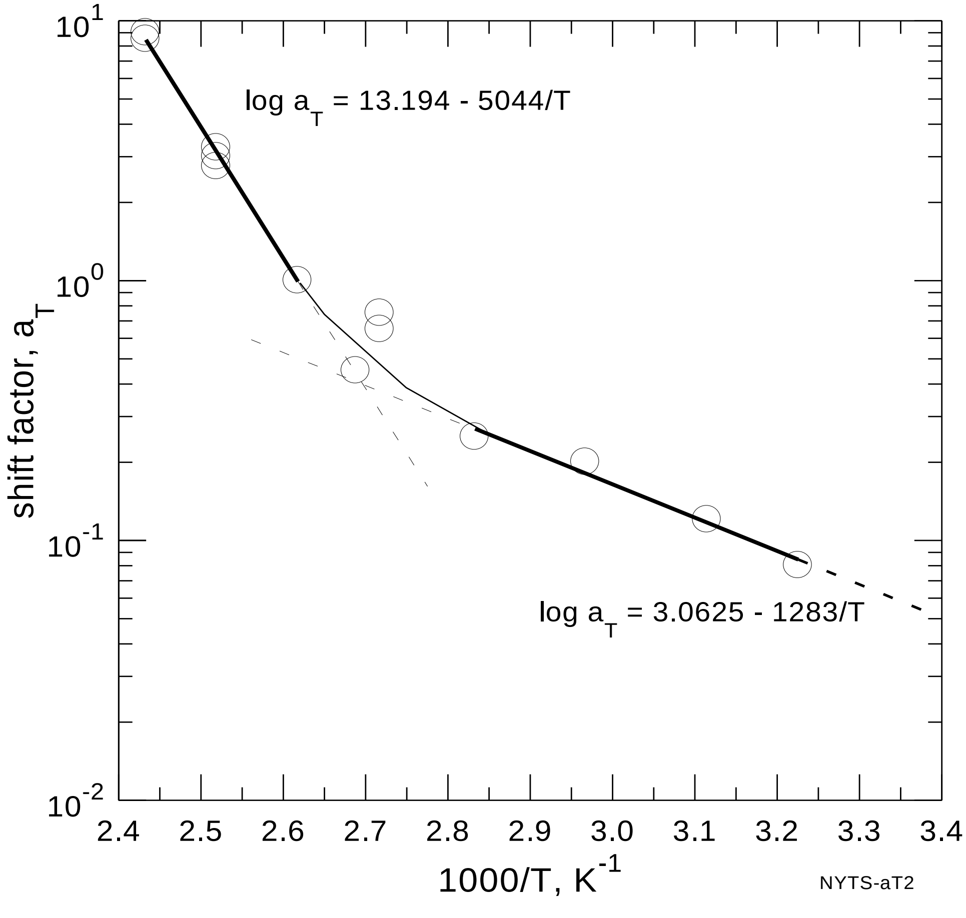

Figure 5 – Nylon thermal-oxidative shift factor plot. (Copied from Figure 22 in [26])

5. SUMMARY AND RECOMMENDATIONS

Accelerated testing involves testing products at higher than normal levels of stress. The degree of acceleration is reflected by how far the test conditions are from normal use conditions. When the degree of acceleration is large, it is risky to extrapolate test results in a quantitative way to use conditions due to model inaccuracies, etc. However, in a qualitative context, such tests (deemed highly accelerated) can be useful to find and fix problems during design and preproduction (HALT). Due to schedule and budget constraints, there is pressure to use accelerated testing at extreme levels of stress, where the relationship between those levels and normal levels of stress is unknown, in order to make quantitative inferences. The extreme levels might be a result of using a "canned", "cookie-cutter" approach (e.g., "Parametric Binomial") where the relationship between stress and life is assumed (model <u>and</u> parameter values) in order to minimize the number of units to test and/or minimize the amount of time units need to be on test. Use of these approaches requires assumptions that may be difficult to justify so that the use of extreme levels in such cases may be without a sufficient basis. Once test data are acquired, computer software will rapidly provide quantitative predictions (right or wrong) based on the user's assumptions. Due to the nature of accelerated data, which are often incomplete (either right or left or interval censored), model assumptions and parameter uncertainties are often difficult to assess. <u>Thus, quantitative inferences based on such experiments and data analysis are difficult to substantiate and can lead to incorrect inferences that mislead decision-makers.</u>

Rather than using a "canned" approach for planning and analysis, it is recommended that engineers follow a highly collaborative approach that will lead to experiments and analyses that are <u>tailored</u> for the specific product. Expertise from a variety of sources is needed. Key is the product engineer who is knowledgeable about the intended use of the product as well as the use environments/stresses that are expected during the products life. Collaboration between the product engineer and materials experts is needed to identify the dominant failure modes/mechanisms that can occur within normal use. Rationale for selection of a particular model should be developed based upon careful consideration of failure mechanisms that are known or suspected. A relevant process for accelerating the identified failure modes (manifested in terms of one or more controllable stress factors) needs to be established. A test engineer should be consulted in order to identify relevant measurements. A statistician should be consulted in order to assist in the planning of efficient experiments and to analyze the experimental data (including assessment of model adequacy and uncertainty). There may be some synergistic benefits accrued by coordinating the planning and analysis of HALT tests (find and fix) with tests used to make quantitative inferences [27]. The planning process should recognize the extensive literature associated with accelerated tests and, if possible, focus on the measurement of informative degradation data (rather than the relatively uninformative failure data). While in some cases it may not be possible to identify and develop a relevant degradation measure, it is well worth some additional time and effort to try. If it is not possible to use a degradation approach, it is worthwhile to consider the additional information gained by conducting tests to failure versus testing to a requirement.

In summary, accelerated testing can be very useful as a qualitative tool for finding and fixing problems during design and preproduction (HALT). With sufficient knowledge (e.g., well understood failure mechanisms and reasonable modeling assumptions), accelerated testing can also be useful for making quantitative inferences regarding performance at use conditions. Note that modeling assumptions will inherently involve some level of risk. Naturally, the assumptions become more critical (and risky) as accelerated conditions become further removed from the use conditions. Additional risk is incurred as the assumptions are farther removed from knowledge. Decision makers need to understand the assumptions, the level of knowledge, and the concomitant risks.

REFERENCES

[1] Meeker, W.Q., and Escobar, L. A., *Statistical Methods for Reliability Data*, John Wiley, 1998.

[2] Nelson, W. B., *Accelerated Testing, Statistical Models, Test Plans, and Data Analysis*, 2004, John Wiley.

[3] Nelson, W. B., "A Bibliography of Accelerated Test Plans," *IEEE TRANSACTIONS ON RELIABILITY,* **54**, 2005, pp. 194-197.

[4] Nelson, W. B., "A Bibliography of Accelerated Test Plans Part II - References," *IEEE TRANSACTIONS ON RELIABILITY,* **54**, 2005, pp. 370-373.

[5] Escobar, L. A. and Meeker, W. Q., "A Review of Accelerated Test Models," *Statistical Science,* **21**, 2006, pp. 552-577.

[6] Thomas, E. V., Bloom, I., Christophersen, J. P., and Battaglia, V.S., "Statistical Methodology for Predicting the Life of Lithium-ion Cells via Accelerated Degradation Testing," *Journal of Power Sources*, **184**, 2008, pp. 312-317.

[7] Hobbs, G. K., *Accelerated Reliability Engineering: HALT and HASS*, John Wiley, 2000.

[8] McLean, H. W., *HALT, HASS, and HASA Explained, Accelerated Reliability Techniques*, 2009, American Society for Quality, Quality Press.

[9] Doganaksoy, N., review of HALT, HASS and HASA Explained: Accelerated Reliability Techniques, by Harry McLean, Milwaukee: ASQ Quality Press, pp. 489-490, 2000, *Technometrics*.

[10] Bazovsky, I. *Reliability Theory and Practice*, 2004, Dover Publications, Mineola, N.Y.

[11] Knight, C. R., Jervis, E.R., and Herd, G. R., 1955, *Terms of Interest in the Study of Reliability*, ARINC Monograph No. 2, Aeronautical Radio, Inc., Washington.

[12] Wright, D. L. and Bierbaum, R. L., "Nuclear Weapon Reliability Evaluation Manual," SAND2002-8133, 2002.

[13] Lawless, J. F., *Statistical Models and Methods for Lifetime Data*, 1986, John Wiley, New York.

[14] Abernethy, R. B., Breneman, J. E., Medlin, C. H., and Reinman, G. L., *Weibull Analysis Handbook*, 1983, Pratt and Whitney, West Palm Beach, Florida.

[15] Kaplan, E. L., and Meier, P., "Nonparametric Estimation from Incomplete Observations," Journal of the American Statistical Association, **53**, 1958, pp. 457-481.

[16] Newcomer, J. T., “DLT Testing of the MC4682 Lot 4 Capacitor (OUO),” Memo to Distribution, June 21, 2012.

[17] Thomas, E. V., Case H. L., Doughty D.H., Jungst R. G., Nagasubramanian G., and Roth E. P., “Accelerated Power Degradation of Li-Ion Cells,” *Journal of Power Sources*, **124**, 2003, pp. 254-260.

[18] Miner, M. A., “Cumulative Damage in Fatigue,” *Journal of Applied Mechanics*, 12, 1945, A159-A164.

[19] Birnbaum, Z. W., Saunders, S. C., “A Probabilistic Interpretation of Miner’s Rule,” *SIAM J. Appl. Math.*, 16, 1968, pp. 637-652.

[20] Chan, V. and Meeker, W. Q., “Estimation of Degradation-Based Reliability in Outdoor Environments,” Technical Report, Department of Statistics, Iowa State University, July 2001.

[21] Thomas, E. V., Bloom, I., Christophersen, J. P., and Battaglia, V.S., “Rate-based Degradation Modeling of Lithium-ion Cells,” *Journal of Power Sources,* **206**, 2012, pp. 378-382.

[22] Lipson, C., Sheth, N., *Statistical Design and Analysis of Engineering Experiments*, 1973, McGraw-Hill Book Company, New York.

[23] Meeker, W. Q., Sarakakis, G., and Gerokostopoulos, A., “More Pitfalls of Accelerated Testing,” Preprint.

[24] Hamada, M. S., Wilson, A. G., Reese, C. S., and Martz, H. F., *Bayesian Reliability*, 2008, Springer, New York.

[25] Meeker, W. Q. and Escobar, L. A., “Pitfalls of Accelerated Testing,” *IEEE Transactions on Reliability*, **47**, pp.114-118, 1998.

[26] Bernstein, R., Derzon, D. K., Whinery, L. D., Shedd, M. M., and Gillen, K. T., “Parachute Aging Studies; Nylon and Kevlar (OUO),” SAND2008-6540, 2008.

[27] Collins, D. H., Freels, J. K., Huzurbazar, A. V., Warr, R. L., and Weaver, B. P., “Accelerated Test Methods for Reliability Prediction,” Journal of Quality Technology, **45** (3), 2013, pp. 244-259.

Distribution

1	MS0124	Marcey Hoover	08040
1	MS0330	James Lavin	00426
1	MS0428	JF Nagel	00400
1	MS0472	Timothy Edwards	02134
1	MS0557	Troy Savoie	01557
1	MS0829	Neil Lapetina	00413
1	MS0829	Janet Sjulin	00410
1	MS0829	Jennifer Franklin	00415
1	MS0829	Edward Thomas	00415
1	MS0830	Elmer Collins	00413
1	MS0888	Lysle Serna	01818
1	MS0889	Jill Glass	01818
1	MS0958	Michael Kelly	01833
1	MS1071	Carol Sumpter	01740
1	MS1078	Michael Daily	01740
1	MS9007	Alix Robertson	08225
1	MS9014	Dawn Skala	08238
1	MS0899	Technical Library	09536 (electronic copy)

Sandia National Laboratories