Quality Engineering

ISSN: 0898-2112 (Print) 1532-4222 (Online) Journal homepage: www.tandfonline.com/journals/lqen20

Challenges and new methods for designing reliability experiments

Laura J. Freeman, Rebecca M. Medlin & Thomas H. Johnson

To cite this article: Laura J. Freeman, Rebecca M. Medlin & Thomas H. Johnson (2019) Challenges and new methods for designing reliability experiments, Quality Engineering, 31:1, 108-121, DOI: 10.1080/08982112.2018.1546394

To link to this article: https://doi.org/10.1080/08982112.2018.1546394

Published online: 12 Apr 2019.

Submit your article to this journal

Article views: 408

View related articles

View Crossmark data

Citing articles: 2 View citing articles

QUALITY ENGINEERING
2019, VOL. 31, NO. 1, 108–121
https://doi.org/10.1080/08982112.2018.1546394

Challenges and new methods for designing reliability experiments

Laura J. Freeman, Rebecca M. Medlin, and Thomas H. Johnson

Institute for Defense Analyses, Alexandria, Virginia

ABSTRACT
Engineers use reliability experiments to determine the factors that drive product reliability, build robust products, and predict reliability under use conditions. This article uses recent testing of a howitzer to illustrate the challenges in designing reliability experiments for complex, repairable systems. We review research in complex system reliability models, failure-time experiments, and experimental design principles. We highlight the need for reliability experiments that account for various intended uses and environments. We leverage lessons learned from current research and propose methods for designing an experiment for a complex, repairable system.

KEYWORDS
complex system reliability; experimental design; reliability experiments; restricted randomization; power analysis

Introduction

Motivation

The reliability of products and systems is a concern for all fields of engineering. In consumer product industries, poor reliability can undermine consumer confidence in a new product and cause warranty issues for the manufacturer. In our field of defense system testing, reliability is essential in ensuring effective and safe missions. It takes a prominent role in the evaluation of all defense systems.

Reliability is the probability that a product or system will perform a required function without failure under stated conditions (i.e., environmental and operating conditions) for a stated period of time (adapted from ANSI/GEIA Standard 2008). In defense applications, as well as many others, the environmental and operating conditions can change depending on the current scenario, time of year, or user experience, making it challenging to define a single system reliability.

System designs today are increasingly complex and interconnected. These interconnections make the analysis of component-level reliability, paired with reliability block diagrams an incomplete solution to assessing complex system reliability. Relationships between components and subsystems can be challenging to model. Failure modes can arise from unanticipated faulty connections between components. Detailed system designs may not be available to properly model systems as reliability block diagrams. Alternatives include using Bayesian hierarchical models and Poisson processes to model total system reliability based on the arrival of system failures. However, these models ignore valuable information from component and subsystem testing that could be leveraged to improved estimates of overall system reliability.

Methods that incorporate multiple intended uses and operating conditions using the principles of experimental design have the potential to improve assessments of system reliability and enable predictions of future expected reliability. The ability to design experiments for complex systems, paired with recent advances in complex system modeling has the potential to advance assessments of complex system reliability.

Illustrative example: Paladin Integrated Management program

The Paladin Integrated Management (PIM) program self-propelled howitzer (SPH) vehicle that appears in Figure 1 provides ample motivation for why experimental design approaches are needed for evaluating the reliability of complex systems. The PIM SPH must fire ammunition and move to specific locations. The system's use can be defined by miles driven and

CONTACT Laura J. Freeman lfreeman@ida.org Institute for Defense Analyses, Alexandria, VA, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lqen.
This article was presented at the Sixth Stu Hunter Research Conference in Roanoke, Virginia, March 2018.

Figure 1. Paladin Integrated Management self-propelled howitzer.

rounds fired, although its reliability requirement is defined in hours.

For PIM, the vehicle is required to have a 75 percent probability of completing an 18-hour mission without an operational mission failure. The vast majority of systems tested in the Department of Defense (DoD) are repairable systems. For many repairable systems, reliability requirements are specified in terms of a mission duration and probability of completion. We will refer to this quantity as mission reliability. Alternatively, requirements documents will sometimes specify a mean time between failures, for example, PIM mean time between operational mission failures should exceed 62.5 hours. The conversion between the probability and mean uses the homogeneous Poisson process, which assumes that inter-recurrence time (time between failures) are independent and identically distributed exponential.

The translation between time between failures and mission reliability uses the cumulative probability of failure in a mission:

$$F(t) = 1 - e^{-\frac{t}{\theta}},$$

where θ is the mean time between failure (MTBF) for a repairable system. For the HPP, the failure recurrence rate is constant: $\mathrm{v(t)} = \frac{1}{\theta}$. Note that by plugging the mission duration and the MTBF from the example howitzer requirement, we obtain the probability-based requirement:

$$F(18) = 1 - e^{-\frac{18}{62.5}} = 0.25.$$

In words, a system that has a MTBF of 62.5 hours (we will call system reliability), has a 25 percent chance of experiencing a failure in an 18-hour mission, or a 75 percent chance of completing the mission successfully. We show this calculation here because it is commonly used to plan reliability demonstrations in the DoD. However, it is flawed in that it is an oversimplification of mission reliability and depends on the HPP assumption. For most repairable systems, we expect the recurrence rate to change as a function of time, resulting in a nonhomogeneous Poisson process (NHPP), which we discuss in more detail later.

Moreover, the mission reliability relies on the actual use of the system during the 18-hour mission. On Army programs, this use is referred to as the operational mode summary/mission profile (OMS/MP). The original OMS/MP for the PIM SPH specified that in an 18-hour mission, the vehicle would drive 17.4 miles and fire 223 rounds (12.8 rounds/mile). An update to the OMS/MP changed the 18-hour mission to cover 58.8 miles and only fire 104 rounds (1.78 rounds/mile). In the field, usage rates vary depending on the needs of a given mission. Clearly, one would expect this to impact the estimate of overall probability of mission completion. But one might also expect that varying use could affect the system reliability estimate as measured through the recurrence rate.

To support planning future missions, we would like to assess the mission reliability under a variety of use profiles. For PIM SPH, there are two clear factors: rate of firing and rate of driving. Correlated is the amount of idle time in an 18-hour mission. We are interested in determining if the failure recurrence rate changes as a function of firing and driving rates. We also want to understand if there are interaction effects between firing and driving that increase recurrence rate in missions that contain both high rates of firing and driving. These types of analyses motivate the need for an experimental design approach for evaluating PIM SPH reliability.

Article overview

In this article, we summarize current methods for planning reliability demonstrations for complex systems, complex system modeling, and reliability experiments. We highlight important lessons from the experimental design literature that require attention for reliability experiments. We seek to make the case for why one would want to design an experiment to evaluate the reliability of a complex system and show the practical benefits. The literature review also identifies gaps in past reliability experiments literature. It highlights gaps between traditional experimental design research and reliability research.

The next section provides a brief overview of the Weibull distribution, failure-time regression models, and NHPP models for recurrence data. The current methods section provides a summary of reliability demonstrations, complex reliability system models, and current methods for designing reliability experiments. The section on important considerations highlights key concepts of randomization and power calculations to determine sample size. The remainder of the article addresses current challenges and potential solutions for designing reliability experiments for complex, repairable systems. We use the PIM SPH example to frame current approaches, challenges, and potential improvements.

Primer on the Weibull distribution, failure-time regression, and NHPP

Failure-time regression and the Weibull distribution

The Weibull distribution is a popular distribution for modeling failure-time data. A common parameterization of the Weibull probability density function (PDF) is as follows:

$$f(t|\beta^*, \eta) = \frac{\beta^*}{\eta}\left(\frac{t}{\eta}\right)^{\beta^*-1} \exp\left[-\left(\frac{t}{\eta}\right)^{\beta^*}\right], \quad t > 0.$$

Figure 2 shows how the shape parameter, β^*, of the two-parameter distribution provides flexible distribution shapes. Failure-time data can either be right-skewed, which is captured by values of $\beta^* \leq 3.6$, or left-skewed which is reflected by larger values of the shape parameter. When $\beta^* = 1$, the distribution reduces to the exponential distribution. A hypothesis test on β^* provides a convenient framework to determine if the exponential distribution adequately represents the data. The scale parameter, η, also known as the characteristic life of the Weibull distribution, represents the time at which we expect 63.2 percent of the units/products to fail. The Weibull distribution is arguably the most popular distribution for modeling time to failure data because of its flexibility, relationship to the exponential distribution, and ability to mimic multiple failure mechanisms including early failure (infant mortality), random failures, and system wear out.

Maximum-likelihood estimation (MLE) can be used to estimate the parameters of the Weibull distribution from a collection of data with failure times. For N independent, identically distributed failure times, t_i, the likelihood to maximize is as follows:

$$L(\beta^*,\ \eta) = \prod_{i=1}^{N} f(t_i\beta^*, \eta) = \left(\frac{\beta^*}{\eta^{\beta^*}}\right)^{N} \prod_{i}^{N} t_i^{\beta^*-1} \exp\left[-\left(\frac{t_i}{\eta}\right)^{\beta^*}\right].$$

Censoring

Censoring occurs when exact failure times are not observed. The most common form of censoring in failure-time data is right censoring, when the data collection ends before a unit on test fails. Right censored assumes the unit will fail sometime in the future (to the right). Right-censored data can be type I (time-censored) when the test stops based on a maximum time duration or type II (failure-censoring) when the test stops because a desired number of failures have occurred. Data are left-censored if an item fails before the first inspection of a product, or interval-censored if a nonconstant monitoring process is used and a failure occurs in between inspection opportunities.

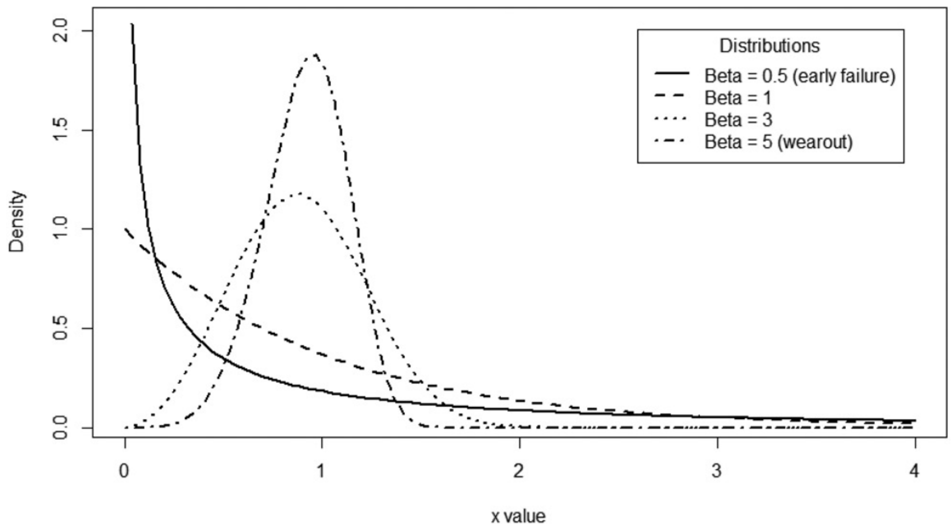

Figure 2. Comparison of the Weibull distribution for various values of the shape parameter ($\eta = 1$).

Censoring is accounted for in the likelihood by calculating the area under the PDF for the region of censoring.

An important consideration for planning failure-time tests is the relationship between the characteristic life and censoring. Figure 3 compares the proportion of the data expected to be right-censored for two different scale parameters with a common shape parameter, $\beta^* = 3$. This concept is important when thinking about how to design reliability experiments. Typically in both planning and practice, the effect of the experimental conditions is translated through the scale parameter. For example, if the PDFs in Figure 3 represent failure times of capacitors under a set voltage, then decreasing the voltage might result in the increase in the scale parameter from 1 to 1.5. As a result, if the experiment is not well planned, we could see disproportionate numbers of failures under different conditions and/or insufficient data to estimate the change in the scale parameter.

Failure-time regression

Failure-time data from an experimental design can be analyzed using a parametric regression model from the log-location scale class of models. These include the lognormal and Weibull distributions. Assuming the experimental design has $\mathrm{i} = 1, 2, 3, \ldots, \mathrm{k}$ unique design points and x_i^T is the i^{th} unique row of the model matrix **X**, these regression models have the form:

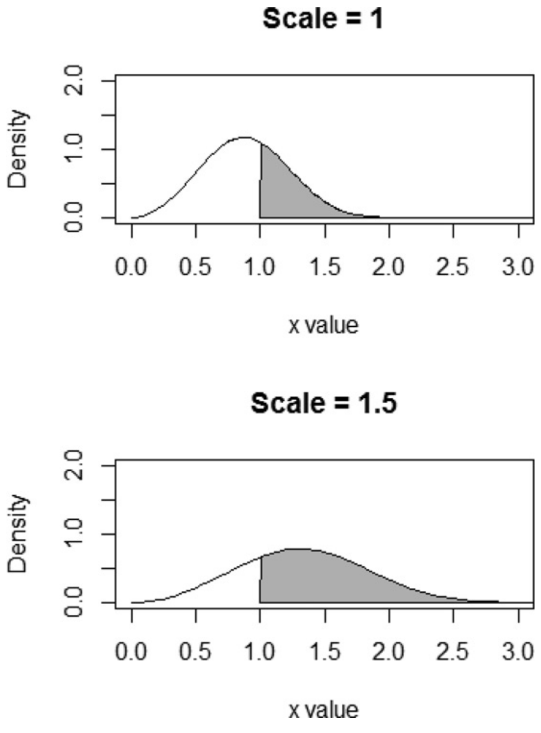

Figure 3. Censoring rate and the Weibull scale parameter ($\beta = 3$).

$$y_i = \log(t_i) = \mu_i + \sigma\epsilon_i$$

where $\mu_i = x_i^T\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is a vector of model coefficients. The Weibull model assumes $\mathrm{t_i} \sim$ *Weibull* $(\mu_i,\ \sigma)$, $\mu_i = \log(\eta_i)$, η_i is the scale parameter of the Weibull distribution, $\sigma = \frac{1}{\beta^*}$, β^* is the shape parameter of the Weibull distribution, and $\epsilon_i \sim SEV(0, 1)$. In this model, μ_i depends on the explanatory variables x_i and σ is constant.

In this article, we will focus on translating factor effects through the scale parameter, η_i. As previously noted, this common assumption has the effect of increasing or decreasing the characteristic life. However, it is possible for the factors to affect the shape parameter, β^*. This is of particular concern for accelerated life tests (ALTs), where the nature of the acceleration can actually change the physics of the failure mechanism.

The model coefficients $\boldsymbol{\beta}$ and scale parameter σ are often estimated using MLE. For a sample of N independent observations with right-censored and exact failure observations, the log-likelihood to maximize is

$$\log\left[L(\mu_i, \sigma)\right] = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \log\left[f\left(t_{ij}|\mu_i,\ \sigma\right)\right]\delta_{ij}$$
$$+\log\left[1-F\left(\left(t_{ij}|\mu_i,\ \sigma\right)\right)\right]\left(1-\delta_{ij}\right),$$

and is dependent on $\boldsymbol{\beta}$ through $\mu_i = x_i^T\boldsymbol{\beta}$. Here, t_{ij} is the j^{th} failure time under the i^{th} test condition, where $j = 1,2,3, \ldots n_i$, and the total sample size is $\sum_{i=1}^{k} n_i = N$. The censoring indicator $\delta_{ij} = 1$ for an exact failure time and $\delta_{ij} = 0$ for a right-censored observation. Similar terms can be added to the likelihood for left-censored or interval-censored data; for example, see Meeker and Escobar (1998).

Nonhomogeneous Poisson process

A critical assumption for failure-time regression models is that the failure times are independent and identically distributed. While individual components of complex systems may be used until failure, the full system is typically repairable making it more sensible to model the system as a recurrence process. The Poisson process is a parametric model commonly used to analyze recurrence data. The HPP assumes a constant recurrence rate, $\nu(t) = \frac{1}{\theta}$, where θ is the mean time between failures.

A NHPP is more flexible, in that it allows for a nonconstant recurrence rate. A common recurrence rate, because of its relationship to the Weibull model, is the power-model recurrence rate

$$\nu(t) = \frac{\beta^*}{\eta}\left(\frac{t}{\eta}\right)^{\beta^*-1}, \; \beta^* > 0, \; \eta > 0 \; .$$

Note that when $\beta^* = 1$, the rate is a constant and the model reduces to the HPP, which provides a convenient framework for conducting a hypothesis test if an HPP is appropriate for the data.

The likelihood for the power-model NHPP is as follows:

$$L(\beta,\eta) = \left(\frac{\beta^*}{\eta^{\beta^*}}\right)^r \prod_i^r t_i^{\beta^*-1} \exp\left[-\left(\frac{t_t}{\eta}\right)^{\beta^*}\right],$$

where t_t is the total observation time, r is the exact number of failures, t_i are the recurrence times (not differences), so $t_1 < t_2 < \ldots < t_r < \; t_t$, and β^* and η are the shape and scale parameters, respectively. Note that the primary difference between this likelihood and the likelihood for the independent failure times is in how the time on the system factors into the likelihood. The total operating time is accounted for by t_i's and t_t. This likelihood already includes the concept of censoring, because of recurrence the data are only considered censored for the period of testing between the last observed failure (t_r) and the end of test (t_t). Similar to failure-time regression, covariates can be incorporated into the NHPP model through either the shape or scale parameter. A common incorporation, and one that we will use later, is as follows:

$$\eta_i = \; x_i^T \boldsymbol{\beta}.$$

Current practices

Traditional repairable system planning and analysis

The DoD uses reliability demonstrations under a set of fixed conditions as the primary test planning tool. Meeker and Escobar (1998) describe reliability demonstration tests as a simple hypothesis that tests: "do the data provide enough evidence to reject the null hypothesis that reliability is smaller than the target." This is contrast to reliability life tests (experiments) or ALT which seek to characterize reliability as a function of use conditions or factors that impact reliability.

The operating characteristic (OC) curve is the primary tool for displaying and investigating the properties of acceptance sampling plans (Test and Evaluation of System Reliability, Availability, and Maintainability: A Primer 1982). Central to the development of these curves is the balancing of consumer risk and producer risk (Montgomery 2013). Consumer risk is the probability that a bad system (below the required reliability) will be accepted, whereas producer risk is the probability that a good system (above the required reliability) will be rejected. The goal is to determine whether a system performs consistently enough that both the consumer and producer are protected.

In DoD acceptance testing, we control consumer risk in ensuring that the requirement has been met. Producer risk tends to be less tightly controlled and often a function of the cost of testing and available resources. Therefore, we construct these curves by first solving the consumers risk equation for allowable failures ranging from 0 to some fixed number (often no more than 5–10 because of limited test resources). Consider the howitzer example, which must have a 75 percent probability of completing an 18-hour mission without failure. Using the exponential assumption, we can generate reliability demonstration tests for duration-based test requirements. To construct an OC curve, we determine the total test duration for a fixed number of failures using the chi-squared lower confidence bound:

$$Test \; Duration = \frac{\chi^2_{\alpha/2, \; 2c+2}\theta}{2},$$

where c is the fixed allowable number of failures and θ is the required mean time between failures. For example, the howitzer with a 62.5 mean requirement and a consumer risk of 20 percent, the minimum test duration for a test that passes the requirement with 4 failures is 420 hours. The full OC curve is generated by varying the true mean time between failure parameter, θ and calculating the probability of passing using the Poisson distribution. In words, we need to calculate the probability of seeing c or fewer failures given a known failure rate, which is exactly the cumulative distribution function (CDF) of the Poisson distribution. Figure 4 shows the OC curves for passing a reliability demonstration with variable lengths. Notice that the curves fixed the consumer risk at 0.20 percent for the 75 percent mission reliability requirement. Tests that allow more failures do a better job of controlling producer risk.

While a reliability demonstration provides a straightforward method for determining a minimal

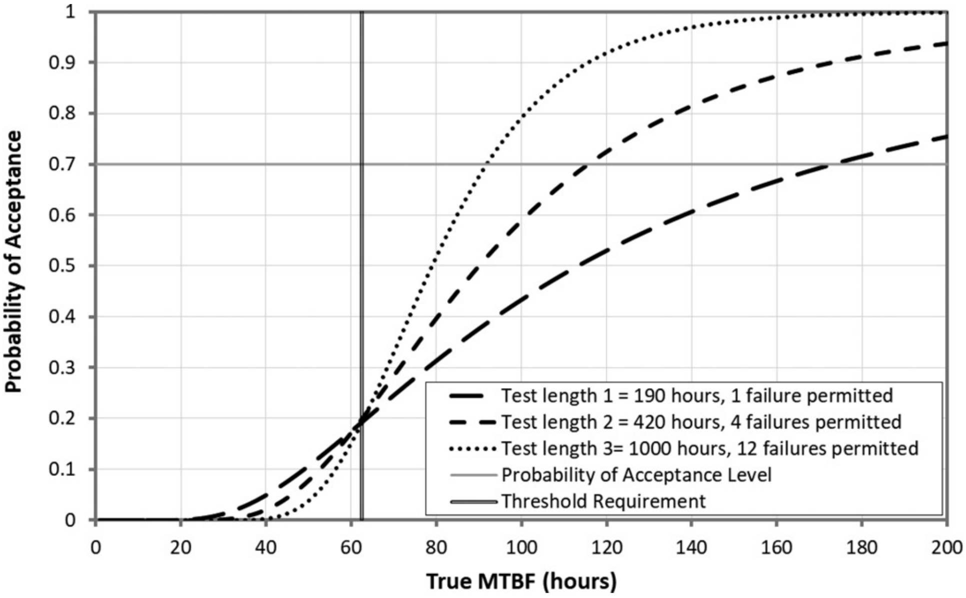

Figure 4. Operating characteristic curves for PIM SPH. Excel Spreadsheet available at testsicence.org.

test length, it fails to provide a methodology for determining how much of that test should be firing, driving, and idling. Additionally, standard applications of the OC curve method by DoD programs aggregate reliability data across multiple vehicles, which is inappropriate. However, generating separate OC curves for each system tends to result in tests that are larger than programs can afford and also fail to capitalize on the commonality across vehicles. However, lacking better methods from research, this methodology is the standard practice for scoping reliability tests for operational tests of DoD systems.

Current research

Experimental design in the reliability literature

For simple systems, reliability experiments have been used for years to determine product reliability. Most of the examples employ classical experimental design techniques. An early example comes from Zelen (1959), in which he considers a replicated 2 × 4 factorial experiment to determine the effect of voltage and temperature on the lifespan of a glass capacitor. Condra (2001) uses classical experimental design to improve reliability for products like ceramic capacitors, circuit boards, and light-emitting diodes. Condra emphasizes the need to select factors and levels for reliability experiments that affect reliability outcomes including design structure, manufacturing process, and material selection. McCool and Baran (1999) provide an example of using a 2 × 2 factorial experiment to determine if treating glassy polymers used in dental restorations improves resistance to fracture. Bullington et al. (1993) use a resolution IV fold-over Plackett–Burman design to determine which of 11 factors influence the lifetime of an industrial thermostat. Meeker and Escobar (1998) provide numerous examples of life tests and ALTs for products ranging from alloys to yarn.

Classic statistical texts on the planning and analysis of experiments focus on normal theory experimental design; see, for example, Montgomery (2017) and Myers, Montgomery, and Anderson-Cook (2009). The focus is on complete (i.e., not censored) normally distributed data. Meeker and Escobar recommend easy adaptations to classical experiments to directly account for the skewed response and censoring. For example, they recommend using more than two levels and increasing replication in cases where censoring may result few failures. These approaches are straightforward, but may result in designs that cost more than necessary. Another approach used by Zelen (1959) is to use type II censoring to ensure a sufficient number of failures. However, this approach can fail if it takes longer for failures to occur than anticipated in the planning phase, resulting in either a longer test, or insufficient data. Literature on generalized linear models provides insight to designing experiments for skewed responses (see Myers et al. 2012), but fails to incorporate censoring.

Monroe et al. (2011) note these limitations in the motivation for their research on using generalized linear models for designing ALTs. In addition to challenges with non-normal data and censoring, they note challenges in feasible design regions and increased uncertainty in extrapolations as reasons for limited research in experimental designs for ALTs.

Many researchers account for the skewed nature of the response variable and focus on estimating the hazard function at a given time or a particular quantile of the data. Meeker and Nelson (1977) focus on estimating distribution quantiles with sufficient precision for the Weibull distribution under time-based (type I) censoring. Meeker, Escobar, and Hill (1992) emphasize the importance of ensuring that the sample size in the experiment will deliver the desired precision by estimating the hazard function at a given time. These papers provide useful guides for selecting sample size for life tests under a fixed set of conditions, but they do not easily extend to experiments with two or more experimental factors.

Noticing these disconnects, recent papers on reliability tests including ALTs have started to include the anticipated regression models in the design selection. Meeter and Meeker (1994) develop optimal and compromise ALTs under the assumption of nonconstant scale parameters. Escobar and Meeker (1995) develop ALT plans for tests with two or more experimental factors, including practical considerations such as avoiding accelerating variables that interact. Zhang and Meeker (2006) develop Bayesian optimum and compromise test plans for ALTs. Monroe et al. (2010) discuss methods for evaluating the sensitivity of an optimal ALTs for two factors with interactions. These papers focus on determining the optimal setting for factors and allocation of test resources across those factor levels. However, they often fail to provide concrete recommendations on the required sample size or a discussion of power analysis.

While minor gaps exist in the current reliability experiments research, experimental design research for complex repairable systems is essentially nonexistent. Meeker and Escobar (1998) and Condra (2001) discuss modeling system reliability with reliability block diagrams and/or fault trees. Alternatively, these texts offer the use of the HPP or NHPP to model the reoccurrence data through the inter-arrival times of failures. Reliability growth models leverage the NHPP process over time to account for the test-analyze-fix-test cycle (National Research Council 2015). In this case, the arrival rate of failures is a function of time. Cook and Lawless (2007) include covariates in a variety of different recurrence models.

Restricted randomization designs

Fisher (1937) identified randomization, replication, and blocking (local control of error) as core tenets of experimental design. Randomization is essential in that it breaks the link to potential confounding variables, allowing randomized experiments to provide statements about the cause of the experimental result. Common design strategies for dealing with restrictions in randomization include blocking and split-plot designs. Jones and Nachtsheim (2009) highlight the prevalence of split-plot designs in industrial applications provides a straightforward overview of their value. Goos (2002) provides a comprehensive review of designing optimal experiments for various analysis models with blocking and split-plot experimental structures.

Restricted randomizations are extremely common in reliability experiments with accelerating factors (e.g., temperature) that can be applied to a batch of products (e.g., in a large oven) in a cost-effective manner. Additionally, as the split-plot and blocking literature has shown us for classical experiments, restrictions in randomization have a significant effect on error variance, which often translates to lower statistical power, for example, see Bisgaard (2000).

Freeman and Vining (2010), Kensler, Freeman, and Vining (2014), and Medlin et al. (2018) highlight the importance of accounting for restrictions in randomization in reliability models. They develop two-stage analysis methods that account for subsampling, blocking, and split-plot designs structures, respectively. In the first stage, they use failure-time regression models and find maximum-likelihood estimates (MLEs) for a common Weibull shape parameter, β^* along with independent scale parameters (η_i) for each of the experimental units. The second stage then takes into account the appropriate experimental protocol (e.g., blocking or split-plot) by assuming the log of the scale parameter is normally distributed. This assumption of normality leverages asymptotic normality from the first-stage MLEs. The treatment levels of the experimental factors are estimated using MLEs for a linear model of the log-scale parameter.

Alternatively, Freeman and Vining (2013), Kensler, Freeman, and Vining (2015), and Medlin et al. (2018) provide a nonlinear mixed model (NLMM) solution. The Weibull NLMM analysis allows for the estimation of all of the model parameters in a single step. However, to account for the random effect, the likelihood includes one or more intractable integrals, complicating the analysis. The authors use Gauss–Hermite quadrature to approximate the integral, see Liu and Pierce (1994). This method yields an approximate closed-form solution of the total likelihood, allowing for the derivation of an asymptotic variance–covariance matrix, supporting inference.

Simulation studies (Medlin et al. 2018) analyzing the two approaches conclude that:

- The two-stage analysis and the NLMM produce similar estimates of the scale parameter coefficients.
- The NLMM has high type I error rates compared to the two-stage analysis.
- The two-stage shape parameter estimates have higher bias than the NLMM estimates.

The conclusion of the simulation is that accounting for the experimental protocol matters and incorporating it into the analysis is more challenging than when responses are modeled by the Weibull distribution. The results can be increased bias in the shape parameter or inflated type I error rates for the factor significance on the scale parameter, which are an important considerations when interpreting results or planning future reliability experiments.

Randomization considerations also prove to be important for repairable system reliability experiments. As we discuss further in the next section, the assignment of conditions to systems is problematic. One option is to assign the treatment at the system level. However, without replication, this confounds the treatment with the system. This may also not be possible when the number of treatments exceeds the number of systems.

Statistical power considerations

Power analysis provides a methodology for determining if replication is sufficient for answering the objectives of the experiment. Discussion of power is common in classic experiment design evaluation, but it is not as prevalent in reliability research.

Numerous papers by Meeker and Nelson (1977), Meeker, Escobar, and Hill (1992), Meeter and Meeker (1994), Escobar and Meeker (1995), and Zhang and Meeker (2006) develop test designs based on the precision around a quantile estimate or hazard functions. These papers use Monte Carlo analysis for a more detailed assessment of design properties. The Zhang and Meeker (2006) paper takes a Bayesian approach. Focusing on precision is appropriate for quality control applications because the focus is on ensuring that products under any use conditions meet a specified lower bound. However, for many reliability experiments, power to detect the effect of a factor on the product lifetime is the exact reason the test is conducted. In design for reliability applications, power is a useful metric for determining the adequacy of the test.

Monte Carlo simulation is a flexible and accurate approach for estimating power, among other design properties, but in some cases it is convenient to have an approximation. Johnson et al. (pending publication) present a closed-form approach for calculating power for failure-time reliability experiments, based on the noncentral chi-squared approximation to the distribution of the likelihood ratio statistic. Their closed-form approximation of power for both the Weibull and lognormal distributions allows a researcher to quickly compare multiple designs and accommodate trade-space analyses between power, effect size, model formulation, sample size, censoring rates, and design type. Code for using the approximations as well as comparisons to simulated power is available at https://test-science.shinyapps.io/survpow/.

Complex system reliability

Meeker and Escobar (1998) provide an introduction to system reliability concepts. They use reliability block diagrams with components arranged in various configurations. Systems can be arranged in series, parallel, with redundancy at component or system level. More complex structures include bridge systems to transition loads when components fail, and k-out-of-s systems that require only k of the systems to remain operational for the system to continue working. These probability-based methods allow for system-level analyses of reliability. However, the quality of the assessment depends on the correct specification of the block diagram, adequate data on all components, and components that do not interact in unexpected ways. Fault trees are another method for showing system reliability that focus on modeling probability of failure (failure perspective) instead of reliability (success perspective). The block diagram math can get complicated as system complexity increases, and as previously noted, it may not even be possible to write down a credible reliability block diagram model.

Wilson et al. (2006) highlight the increasing complexity of systems that statisticians are routinely asked to evaluate the reliability. They highlight that in addition to increasing system complexity, often many different sources of information are available ranging from component and subsystem testing to full system tests. They also state that standard analysis methods could be improved by incorporating all available information to get the best understanding of full system reliability. They show how traditional system reliability models specified by reliability block diagrams and fault trees can be converted into Bayesian networks. This conversion is the key step in being able to include multiple types of information into an overarching system assessment.

Dickinson et al. (2015) demonstrate the benefits of using a statistical model to combine information across multiple testing events and similar systems for the Stryker family of vehicles. They employ both frequentist and Bayesian inference techniques to show that when available information is combined, reliability estimates are more accurate and precise than the traditional DoD analysis methods, which tend to use only the data available from a single test event to assess system reliability.

Anderson-Cook et al. (2007) illustrates how Bayesian models can be used to model missile reliability for a missile with numerous components nested in various subsystems. They focus on incorporating pass-fail data from component and full system tests along with subject matter expertise. They use reliability block diagrams to guide the model specification. Anderson-Cook et al. (2008) expand this methodology to include continuous data collected as part of quality assurance measurements, which can account for aging.

In a different approach to complex system reliability analysis, Gilman, Fronczyk, and Wilson (2018) use the data itself to develop a surrogate hierarchical component-system reliability model based on failures observed on subsystems of the Joint Light Tactical Vehicle (JLTV). They use a Bayesian hierarchical model to combine time between failure data from three phases of testing and across common vehicle components. The hierarchical models allow them to make assessments of the system reliability and the reliability of "components" or failure modes. They allow for changes in the reliability due to corrections to failure modes between tests. This flexible framework allows for component and system-level modeling, even when the underlying system architecture is unknown (as was the case in this example). However, a limitation is that the model can be dependent on the number of observed failure modes.

Despite the growing complexity and flexibility of models available to assess complex system reliability, most of the current research neglects how to be design tests to gather the information for these models. In a 2009 paper, Anderson-Cook, Graves, and Hamada look at how the cost of component, subsystem, and full system tests can be factored into selecting the best place to invest test resources to improve complex system reliability. Using a generic system configuration, they develop an algorithm that evaluates current information, develops multiple future allocations, simulates new data sets, updates the model, and makes a recommendation based on the allocation that reduces uncertainty the most for a fixed cost. This simulation approach is intuitive, but time-consuming.

Gilman, Fronczyk, and Wilson (2018) also look at how previous test data can be used to scope a future test using assurance testing for the JLTV. They design an assurance test that leverages information from the three previous testing phases to help determine the duration of testing required to meet a requirement. Gilman, Fronczyk, and Wilson (2018) show that using an assurance test that leverages previous data, they can greatly reduce test requirements for the JLTV. For systems with lots of detailed system structures, component-level failure data, and/or previous test data, assurance tests can dramatically reduce the resources required to show a requirement has or has not been met.

Designing experiments for complex systems

Challenges

Many complex systems are designed to operate in multiple environments under a variety of operating conditions, which makes it challenging to estimate a single system reliability. Instead, it may be appropriate to model failure recurrence as a function of use. In current work, we consider the problem of estimating reliability for a complex system for varying use profiles.

Complicating considerations for complex systems include the repairable nature of the systems, the lack of independence between failures, the varying levels in the severity of failures, and nonconstant use profiles. The repairable nature poses problems because some subsystems or components may be replaced when they break, while others may be repaired. This leads to variability in the total operating time in components of the system as well as varying states of repair. Ideally, operating time would be tracked at the component level, but most systems do not yet incorporate that level of tracking for reliability monitoring purposes.

Independence poses a challenge on two levels; first strictly on the modeling side, the arrival of our failures is not independent and therefore should not be treated as independent and identically distributed. This drives analysis complexity. As previously discussed, recurrence models and Bayesian hierarchical models are two strategies that have been used to account for dependent failures. There is also the challenge of induced failure modes. It is not uncommon for a failure in one subsystem to put a heavier burden on another subsystem and induce a failure that would likely not have occurred otherwise. For example, in our PIM SPH system, driving over rough terrain may induce a failure in the howitzer. Consequently, it is not possible to bin individual failures into "driving" failures versus "firing" failures and separate the analysis.

The varying level of failure severity may best be explained through an example. Consider a family-owned automobile; failure of the brake system is a much more serious problem than failure of the radio. The DoD has a formal process associated with scoring the severity of failures. The DoD categorizes the severity of failures as nonessential function failures, essential function failures, or mission failures. Nonessential function failures are failures that are non-critical (e.g., car radio) and repairs are often delayed until the next scheduled maintenance period. Essential function failures and mission failures occur on critical subsystems (e.g., car brakes), the difference between them is that mission failures must be repaired before the mission can be completed. This adds another layer of complication because it requires tracking the operational status of systems at the component level to get an accurate reflection of time in use between failures.

Nonconstant use profiles reflect that components and subsystems of repairable systems are not necessarily used continuously during system operation. Subsystems can be used on demand, for example, the windshield wipers in your car; or subsystems can be used continuously with either fixed or variable loads. For example, the engine of your car is used continuously while the car is running, but the load varies over time. Instrumented data collection on system components could ensure that usage rates and failure rates are tracked in a sensible fashion, but is not yet common for DoD systems.

Finally, from a design perspective, the design methodology should account for all of these modeling challenges, while allowing for multiple use profiles. We hypothesize that an experimental design strategy at the system and subsystem level could aid in assessing the impact of the individual tasks on system reliability when multiple operational profiles will be employed by the user. The goal is to be able to produce predictions of reliability (failure intensity) across a variety of use profiles by choosing the factors in the design space to parametrically span use profiles.

Data collected during PIM testing

The PIM systems provide an interesting example of how we might reconsider test designs for complex repairable systems. The change in the OMS/MP clearly illustrates the need to be able to predict reliability under varying use profiles. It also provides actual data collected under varying use conditions. Table 1 shows the results of the three phases of PIM testing. The first two phases of testing were "developmental" in nature, and the third was "operational." Developmental testing (DT) implies a lower level of fidelity in the operational realism of the environments, missions, and users. Between test phases, repairs were made to a subset of the failure modes found in previous testing. The result is that we cannot identify if changes in the recurrence rate are due to changes in the use of the system, changes to the system design, or the fixing of failure modes previously identified. However, we can still use the data to illustrate the potential value of modeling recurrence rate as a function of covariates.

The failures summarized in the table are essential function failures, which are failures of any of the PIM SPH subsystems that are essential for mission completion. The requirement and test scope were based on operational mission failures, which are a subset of the essential function failures. During the three phases of testing, only two mission failures occurred in DT1, four occurred in DT2, and three occurred in the limited user test (LUT). Unfortunately, time associated with failures was only tracked in the more operational LUT. Note that the LUT was 222.7 hours in duration, which allows for one failure under a traditional reliability demonstration test mentality. Clearly, there is a statistical advantage in assessing reliability by using all failures (or at least the essential ones) in that there are more data to compare across conditions. The authors have advocated for the DoD to change their test design methodology and rationale to reflect seeing differences in use profiles, but for that to occur, the

Table 1. Summary of essential function failures from three phases of PIM testing.

Test phase	Vehicle	Failures	Miles	Hours	Rounds	Rounds/Miles
DT1	Vehicle 1	24	66.4		555	8.36
	Vehicle 2	21	67.3		445	6.61
DT2	Vehicle 1	21	316.9		680	2.15
	Vehicle 2	22	254		743	2.93
LUT	Vehicle 1	9	431.5	109.9	624	1.45
	Vehicle 2	16	432.6	112.8	623	1.44

statistical community must develop practical approaches that can be implemented in the place of current practices.

Figure 5 shows the distributions of the failures for the number of rounds fired between failure versus the number of miles driven between failure. Clearly, the distribution of the failures for the LUT is different from the earlier two phases of testing. During the LUT, the vehicle achieved longer durations between failures in terms of both miles driven and rounds fired. There are smaller differences in the failure distributions between DT1 and DT2.

We used parametric recurrence analysis using the power-model NHPP to investigate the failure intensity in terms of miles driven between failure and rounds fired between failures. Unfortunately, since failure-time data were not collected until the third phase of testing, we cannot use that outcome to investigate the failure recurrent rate as a function of use profile. We use cumulative miles by vehicle and cumulative rounds by vehicles as potential predictors of changes in the scale parameter of the intensity function given by:

$$\nu(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}, \quad \beta > 0, \;\; \eta > 0$$

Table 2 summarizes the parametric NHPP models for this data set. The miles model and rounds-fired model were constructed separately, both using backward selection. The best model for predicting the mean miles between failure is the NHPP with miles driven. The best model for predicting the mean rounds between failures is the NHPP with rounds fired. The p-values are from the likelihood ratio test (Meeker and Escobar 1998).

The intensity function for the power-model NHPP is:

$$\nu(t) = \frac{\beta^*}{\eta}\left(\frac{t}{\eta}\right)^{\beta^*-1}, \quad \beta^* > 0, \;\; \eta > 0$$

In the "miles model" intensity function for the NHPP, $\eta = \; -0.215 + 0.258 * \text{Miles in Mission}$ and t in the intensity function is cumulative miles on the system. Similarly, for the rounds model, $\eta = 30.5 - 0.032{*}\text{Rounds in Mission}$ and t in the intensity function is cumulative rounds fired on the system. The miles model has a positive coefficient indicating that more miles driven in a mission increase the scale parameter, which is captured by the miles coefficient in Table 2. The rounds model has a negative coefficient (albeit not significantly different from zero) indicating more rounds fired decrease the scale parameter.

Clearly, these models are flawed. Because the system was repaired and updated between phases of testing, it is not possible to determine what portion of the change in failure intensity is due to system

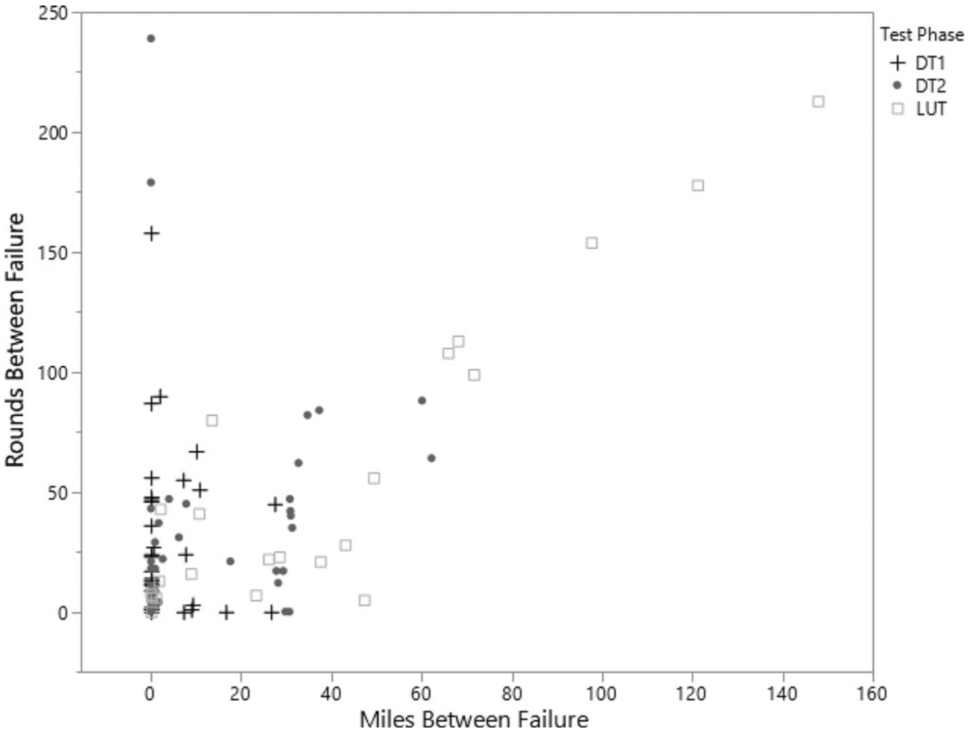

Figure 5. Distribution of failures observed by test phase.

Table 2. Power-model NHPP coefficient estimates.

	Coefficient	SE	p-Value
Miles model			
η Intercept	−0.215	0.64	
η Miles	0.258	0.019	<.0001
β^*	0.801	0.11	.094
Rounds Model			
η Intercept	30.5	14.9	
η Rounds	−0.032	0.017	0.078
β^*	0.782	0.094	0.0374

upgrades versus different use. Instead of increased driving resulting in a higher reliability, it is far more likely that many of the failures were corrected as testing progressed and the overall reliability improved. Note that the later phases of testing also have the most miles driven. However, these fits could be used to guide the selection of test designs from a statistical power prospective as illustrated previously.

While these models are flawed, the intensity function provides a useful illustration of the benefit of designing experiments for complex system. Consider the full intensity function for miles:

$$v(miles) = \frac{\beta}{-0.215 + 0.258{*}\mathrm{Miles}}\left(\frac{Cum\ \ Miles}{-0.215 + 0.258{*}\mathrm{Miles}}\right)^{\beta-1}$$

This function can now be used to estimate the recurrence rate for future tests or operations and incorporate the cumulative use of the system.

Improving reliability designs for repairable systems

If we want to better understand the impact of different use profiles on the expected mission reliability, we need an experimental design that orthogonally varies these conditions across multiple system configurations and test phases. However, often programmatic constraints overshadow the data needs or experimental design choices. For example, for the PIM program, the OMS/MP changed in the middle of the system design and testing to accommodate current use of the howitzer in real-world scenarios. It is important that the selection of the factors and levels reflects potential use cases for the system, but also provides enough spread to potentially impact the recurrence rate.

Table 3 provides a proposal for how one might change the operational profiles for PIM in a way that affects the reliability estimate. It is a two-factor, two-level full factorial with a center point. This simple design has the advantage of spanning the old OMS/MP and the new OMS/MP but also adding additional points so firing rate and driving rates are orthogonal and we can estimate their effect on the failure rate.

Table 3. 2 × 2 Factorial design with center point design proposal.

Mission	Type	Miles (coded)	Rounds (coded)
1	Original OMS/MP	17.4 (−1)	223 (1)
2	Low tempo	17.4 (−1)	104 (−1)
3	Mid-tempo	38 (0)	164 (0)
4	Current OMS/MP	58.5 (1)	104 (−1)
5	High tempo	58.5 (1)	223 (1)

These design missions were selected based on the two existing OMS/MPs and discussion with system experts on what other use conditions may occur. From an experimental design perspective, it may seem ideal to execute missions that are 100 percent one activity. However, such a mission would not occur in reality and would undermine the credibility of the test design with operators of the system. Additionally, it could fail to detect failure modes induced by one activity on another system. For example, if firing the howitzer caused a misalignment in the chassis, it may go undetected if the operators do not actually attempt to drive the howitzer.

This overly simplified design proposal ignores many important considerations that need to be addressed to support the actual execution of such a design. First, how many failures would be required under each operating condition? The power analysis tools Johnson et al. have developed for simple systems provide insights for these complex systems, but to truly optimize sample size, we need a better understanding of the interdependencies that exist and how to account for them in our models and therefore our power calculations.

Next, we have learned from research in restricted randomization designs that order matters. If we continue to only use one OMS/MP during a phase of testing, we cannot decouple the system configuration changes from the failure recurrence rate. However, incorporating multiple use profiles into one test phase brings its own challenges. Should use profiles be assigned to specific vehicles? Or should there be a rotation through use profiles in a counterbalanced fashion, similar to the designs used to control for learning effects in human research? If we opt to assign a specific use profile to a specific vehicle, we need a sufficient number of vehicles to have replication across the use condition. Otherwise, we are simply confounding the individual vehicle with the use profile. If we rotate vehicles through use profiles, we must decide how many missions are conducted before rotating? We also must be able to account for the cumulative wear and tear on the system.

We must also carefully consider how we use data from DT and operational testing in one model. The

DT is not necessarily "operationally realistic." It may, however, give us estimates of failure rates for specific activities. In the PIM data set, we face the challenge that hours were not collected in DT. We need models that can capture the changes in failure rates due to changes in operational realism and fixes to the system.

Conclusions

The desire to predict failure time or failure recurrence as a function of use conditions is common across multiple applications of reliability experiments. Recent research summarized in this article shows that while there are many similarities to conventional experimental design methods, key differences exist that should be accounted for in the experimental design. These key differences include, censoring, the distribution shape, and the challenge of manipulating and modeling multiple factors, especially when restrictions in randomization exist.

In his original works on experimental design, Fisher highlighted randomization, replication, and local control of error. Those concepts are especially important when considering the design of complex, system reliability experiments. Additionally, it is important to consider coverage of the operational use conditions and orthogonality of the test conditions over test phases that span multiple system configurations.

While research is expanding in the design of reliability experiments for simple products, there are many open design and analysis questions for complex systems. As suggested earlier in this article, the increasing availability of networked instrumentation may very soon help us solve these problems by tracking data at the component/subsystem level and recoding operating conditions at that level in real time. As that technology evolves, the statistical community needs to be ready with defensible design strategies and modeling techniques that will allow us to leverage all of the reliability data and make inferences at the component, subsystem, and system levels.

About the authors

Laura J. Freeman is an Assistant Director of the Operational Evaluation Division at the Institute for Defense Analyses. In that position, she established and developed an interdisciplinary analytical team of statisticians, psychologists, and engineers to advance scientific approaches to DoD test and evaluation. Dr Freeman has a BS in Aerospace Engineering, a MS in Statistics, and a PhD in Statistics, all from Virginia Tech. Her PhD research was on design and analysis of experiments for reliability data.

Rebecca M. Medlin is a Research Staff Member in the Operational Evaluation Division at the Institute for Defense Analyses. She supports the Air Warfare Test and Evaluation Task providing expertise in statistics. Her areas of emphasis include design of experiments and reliability analysis. She received her PhD in statistics from Virginia Tech.

Thomas H. Johnson is a Research Staff Member in the Operational Evaluation Division at the Institute for Defense Analyses, Alexandria, VA, where he supports the Live Fire Test and Evaluation Task providing expertise in statistics. His areas of emphasis include sample size determination, sensitivity experiments, and acceptance sampling plans. He received a BS degree from Boston University, and MS and PhD degrees from Old Dominion University, all in Aerospace Engineering.

References

Anderson-Cook, C. M., T. L. Graves, M. Hamada, N. Hengartner, V. E. Johnson, C. S. Reese, and A. G. Wilson. 2007. Bayesian stockpile reliability methodology for complex systems. *Military Operations Research* 12 (2): 25–37. doi:10.5711/morj.12.2.25.

Anderson-Cook, C. M., T. L. Graves, and M. S. Hamada. 2009. Resource allocation for reliability of a complex system with aging components. *Quality and Reliability Engineering International* 25 (4):481–94. doi:10.1002/qre.987.

Anderson-Cook, C. M., T. Graves, N. Hengartner, R. Klamann, A. C. K. Wiedlea, A. G. Wilson, G. Anderson, and G. Lopez. 2008. Reliability modeling using both system test and quality assurance data. *Military Operations Research* 13 (3):5–18.

Bisgaard, S. 2000. The design and analysis of 2k–p × 2q–r split plot experiments. *Journal of Quality Technology* 32 (1):39–56. doi:10.1080/00224065.2000.11979970.

Bullington, R. G., S. Lovin, D. M. Miller, and W. H. Woodall. 1993. Improvement of an industrial thermostat using designed experiments. *Journal of Quality Technology* 25 (4):262–70. doi:10.1080/00224065.1993.11979472.

Condra, L. 2001. *Reliability improvement with design of experiment*. New York, NY: CRC Press.

Cook, R. J., and J. Lawless. 2007. *The statistical analysis of recurrent events*. New York, NY: Springer Science & Business Media.

Dickinson, R. M., L. J. Freeman, B. A. Simpson, and A. G. Wilson. 2015. Statistical methods for combining information: Stryker family of vehicles reliability case study. *Journal of Quality Technology* 47 (4):400–15. doi:10.1080/00224065.2015.11918142.

Director Operational Test and Evaluation. 1982. *Test and evaluation of system reliability, availability, and maintainability: a primer*. Washington, DC.

Escobar, L. A., and W. Q. Meeker. 1995. Planning accelerated life tests with two or more experimental factors. *Technometrics* 37 (4):411–27. doi:10.1080/00401706.1995.10484374.

Fisher, R. A. 1937. *The design of experiments*. Edinburgh/ London, UK: Oliver and Boyd.

Freeman, L. J., and G. G. Vining. 2010. Reliability data analysis for life test experiments with subsampling. *Journal of Quality Technology* 42 (3):233–41. doi:10.1080/00224065.2010.11917821.

Freeman, L. J., and G. G. Vining. 2013. Reliability data analysis for life test designed experiments with sub-sampling. *Quality and Reliability Engineering International* 29 (4): 509–19. doi:10.1002/qre.1398.

Gilman, J. F., K. M. Fronczyk, and A. G. Wilson. 2018. Bayesian modeling and test planning for multiphase reliability assessment. *Quality and Reliability Engineering International*. doi:10.1002/qre.2406.

Goos, P. 2002. *The optimal design of blocked and split-plot experiments*. Vol. 164. New York, NY: Springer Science & Business Media.

Jones, B., and C. J. Nachtsheim. 2009. Split-plot designs: What, why, and how. *Journal of Quality technology* 41 (4):340–61. doi:10.1080/00224065.2009.11917790.

Kensler, J. K., L. J. Freeman, and G. G. Vining. 2014. A practitioner's guide to analyzing reliability experiments with random blocks and subsampling. *Quality Engineering* 26 (3):359–69. doi:10.1080/08982112.2014.887101.

Kensler, J. K., L. J. Freeman, and G. G. Vining. 2015. Analysis of reliability experiments with random blocks and subsampling. *Journal of Quality Technology* 47 (3): 235–51. doi:10.1080/00224065.2015.11918130.

Liu, Q., and D. A. Pierce. 1994. A note on gauss—Hermite quadrature. *Biometrika* 81 (3):624–9. doi:10.1093/biomet/81.3.624.

McCool, J. I., and G. Baran. 1999. The analysis of 2 × 2 factorial fracture experiments with brittle materials. *Journal of Materials Science* 34 (13):3181–8. doi:10.1023/A:1004629822772.

Medlin, R. M., L. J. Freeman, J. K. Kensler, and G. G. Vining. 2018. Analysis of split-plot reliability experiments with subsampling. *Quality and Reliability Engineering International*. doi:10.1002/qre.2394.

Meeker, W. Q., and L. A. Escobar. 1998. *Statistical methods for reliability data*. New York: John Wiley & Sons, Inc.

Meeker, W. Q., L. A. Escobar, and D. A. Hill. 1992. Sample sizes for estimating the Weibull hazard function from censored samples. *IEEE Transactions on Reliability* 41 (1): 133–8. doi:10.1109/24.126687.

Meeker, W. Q., and W. Nelson. 1977. Weibull variances and confidence limits by maximum likelihood for singly censored data. *Technometrics* 19 (4):473–6. doi:10.1080/00401706.1977.10489588.

Meeter, C. A., and W. Q. Meeker. 1994. Optimum accelerated life tests with a non-constant scale parameter. *Technometrics* 36 (1):71–83. doi:10.1080/00401706.1994.10485402.

Monroe, E. M., R. Pan, C. M. Anderson-Cook, D. C. Montgomery, and C. M. Borror. 2010. Sensitivity analysis of optimal designs for accelerated life testing. *Journal of Quality Technology* 42 (2):121–35. doi:10.1080/00224065.2010.11917811.

Monroe, E. M., R. Pan, C.M. Anderson-Cook, D. C. Montgomery, and C. Borror. 2011. A generalized linear model approach to designing accelerated life test experiments. *Quality and Reliability Engineering International* 27 (4):595–607. doi:10.1002/qre.1143.

Montgomery, D. C. 2017. *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons.

Montgomery, D. C. 2013. *Introduction to statistical quality control*. Hoboken, NJ: John Wiley & Sons.

Myers, R. H., D. C. Montgomery, and C. M. Anderson-Cook. 2009. *Response surface methodology*. Hoboken, NJ: John Wiley & Sons.

Myers, R. H., D. C. Montgomery, G. G. Vining, and T. J. Robinson. 2012. *Generalized linear models: With applications in engineering and the sciences*. Vol. 791. New York, NY: John Wiley & Sons.

National Research Council. 2015. *Reliability growth: Enhancing defense system reliability*. Washington, DC: The National Academies Press.

Reliability Program Standard for Systems Design, Development and Manufacturing. 2008. ANSI/ GEIA Standard. GEIASTD0009. SAE International.

Wilson, A. G., T. L. Graves, M. S. Hamada, and C. Shane Reese. 2006. Advances in data combination, analysis and collection for system reliability assessment. *Statistical Science* 21 (4):514–31. doi:10.1214/088342306000000439.

Zelen, M. 1959. Factorial experiments in life testing. *Technometrics* 1 (3):269–88. doi:10.1080/00401706.1959.10489862.

Zhang, Y., and W. Q. Meeker. 2006. Bayesian methods for planning accelerated life tests. *Technometrics* 48 (1): 49–60. doi:10.1198/004017005000000373.