Received: 1 February 2018 | Revised: 16 August 2018 | Accepted: 3 September 2018

DOI: 10.1002/qre.2394

SPECIAL ISSUE ARTICLE

WILEY

Analysis of split-plot reliability experiments with subsampling

Rebecca M. Medlin[1] | Laura J. Freeman[1] | Jennifer L.K. Kensler[2] | G. Geoffrey Vining[3]

[1]Institute for Defense Analyses, Alexandria, Virginia

[2]Shell Global Solutions, Houston, Texas

[3]Department of Statistics, Virginia Tech, Blacksburg, Virginia

Correspondence
Rebecca M. Medlin, Institute for Defense Analyses, Alexandria, VA.
Email: rmedlin@ida.org

Abstract

Reliability experiments are important for determining which factors drive product reliability. The data collected in these experiments can be challenging to analyze. Often, the reliability or lifetime data collected follow distinctly nonnormal distributions and include censored observations. Additional challenges in the analysis arise when the experiment is executed with restrictions on randomization. The focus of this paper is on the proper analysis of reliability data collected from a nonrandomized reliability experiments. Specifically, we focus on the analysis of lifetime data from a split-plot experimental design. We outline a nonlinear mixed-model analysis for a split-plot reliability experiment with subsampling and right-censored Weibull distributed lifetime data. A simulation study compares the proposed method with a two-stage method of analysis.

KEYWORDS

design of experiments, nonlinear mixed model, regression with lifetime data, weibull distribution

1 | INTRODUCTION

Reliability is the probability that a product or system will perform a required function without failure under stated conditions (ie, environmental and operating conditions) for a stated period of time.[1] Reliability experiments are used to provide information regarding the expected life of a product or a system under various operating conditions (ie, factors and levels). The data collected in these experiments, however, can be challenging to analyze. Often the reliability or lifetime data follow distinctly nonnormal distributions and tend to include censored observations. Time and cost constraints may also lead to reliability experiments with experimental protocols that are not completely randomized. For example, in many industrial experiments, a split-plot design structure arises because the randomization of the experimental runs is restricted. It may also be cost effective to apply a treatment combination to many units at one time as opposed to each unit individually, which introduces correlation among the observations. The analysis of the lifetime data collected from the experiment should not only account for the nonnormal distribution and censoring but also account for the experimental protocol.

Many of the standard statistical texts on the planning and analysis of experiments focus on normal theory experimental design; see, for example, Montgomery[2] and Myers et al.[3] The focus is on complete (ie, not censored) and normally distributed data. For these reasons, normal theory analysis methods cannot often be directly applied to the analysis of reliability experiments. Random effects, which result from many design of experiment (DOE) choices, only further complicate the analysis of a reliability experiment. These design choices include subsampling, random selection of treatment levels, blocking, and nested data structures.

An early reliability experiment comes from Zelen,[4] in which he describes a reliability experiment designed to determine the effect of voltage and temperature on the lifespan of a glass capacitor. Condra[5] describes experiments focused on improving product reliability for products like ceramic capacitors, circuit boards, and light emitting diodes. McCool and Baran[6] provide an example of using a 2×2 factorial experiment to determine if treating glassy polymers used in dental restoration improves resistance to fracture. Meeker and Escobar[7] provide numerous examples of life test and accelerated life test for products ranging from alloys to Picciotto yarn. Bullington et al[8] use a resolution IV fold-over Plackett-Burman design to determine which of 11 factors influence the lifetime of an industrial thermostat. Fang et al[9] illustrate a 12-run Plackett-Burman design and analysis focused on improving the reliability of a diamond drill bit.

The analysis of lifetime data assuming a completely randomized design (CRD) is well studied.[7,10-14] The collected lifetime data from a CRD are often analyzed using a parametric regression model such as the Weibull. The analysis technique is straightforward, using maximum likelihood estimation, which easily handles both failure and censored observations.

Until recently, analysis methodologies for more complex experimental designs involving multiple error terms have not been a focus of the reliability field. Freeman and Vining,[15,16] Kensler et al,[17,18] and Dickinson[19] recognized the need to develop an analysis methodology for more complex life-test design structures. Specifically, Freeman and Vining,[16] motivated in part by Zelen's description of the execution of the glass capacitor experiment,[4] showed how to properly account for the experimental error when subsampling is present in a reliability experiment. They proposed a two-stage analysis[15] and later a nonlinear mixed model (NLMM) analysis.[16] Kensler et al[18] built on this framework, proposing a two-stage analysis[17] and NLMM analysis[18] for analyzing reliability experiments with random blocks and subsampling. Dickinson[19] extended this work, proposing a two-stage analysis for analyzing a split-plot reliability experiments with subsampling.

Freeman et al[20] provide an excellent overview of the challenges and considerations associated with planning and analyzing reliability experiments. In addition to reviewing the two-stage and NLMM analysis techniques previously noted, they describe a method for calculating the power provided by the design for detecting an effect on a product's lifetime.[21] In design for reliability applications, power is a useful metric for determining the adequacy of the test to support sound design decisions.

In this paper, we extend the NLMM methodology of Freeman and Vining[16] and Kensler et al[18] by developing an NLMM analysis for analyzing reliability data, involving both censored and exact-failure observations, from a split-plot design with subsampling. We focus our work to the Weibull distribution and a Type II, fixed, right-censoring scheme. We show how to properly accounts for the random effects arising from split-plot design with subsampling in the analysis of the failure time data. A simulation study is included to demonstrate the performance of the method.

2 | A SPLIT-PLOT RELIABILITY EXPERIMENT WITH SUBSAMPLING

Consider a reliability experiment in which the experimenter places p items on test, applies a treatment combination, and records either the exact failure time or a censoring time, for each individual item. For example, suppose the experiment is performed to determine the effect of oven temperature and baking time on the lifespan of an electrical component. In this scenario, the treatment combination applied to the p products on test consists of a randomly selected oven temperature and backing time. Additionally, because the experiment involves a factor, oven temperature, which is more difficult to change on every experimental run than the baking time, the temperature-time combinations cannot be completely randomized. Instead, for each setting of oven temperature, all of the baking time settings are run. The execution of this experiment results in a split-plot structure with subsampling (see Figure 1).

2.1 | The model

For the split-plot experimental setup shown in Figure 1, let $t_{k(ij)}$ denote the lifetime of an individual electrical component. The notation $t_{k(ij)}$ indicates the lifetime of the k^{th} observational unit ($k = 1, \dots, p = 8$) within the j^{th} subplot ($j = 1, \dots, n = 8$) of the i^{th} whole-plot ($i = 1, \dots, m = 4$). If the failure times of the components follow a Weibull distribution, then

$$f_W(t_{k(ij)}|\beta, \eta_{ij}, \delta_i, \epsilon_{ij}) = \frac{\beta}{\eta_{ij}}\left(\frac{t_{k(ij)}}{\eta_{ij}}\right)^{\beta-1} \exp\left[-\left(\frac{t_{k(ij)}}{\eta_{ij}}\right)^{\beta}\right]$$
$$F_W(t_{k(ij)}|\beta, \eta_{ij}, \delta_i, \epsilon_{ij}) = 1 - \exp\left[-\left(\frac{t_{k(ij)}}{\eta_{ij}}\right)^{\beta}\right], \quad (1)$$

where $\eta_{ij} > 0$ is the characteristic life (or scale parameter) for experiment ij and $\beta > 0$ is the common shape parameter.

To explain why some components fail quickly, while others survive for much longer periods of time, we can express the Weibull log-scale parameter $\mu_{ij} = log(\eta_{ij})$, as a function of the experimental factors. For the split-plot design,

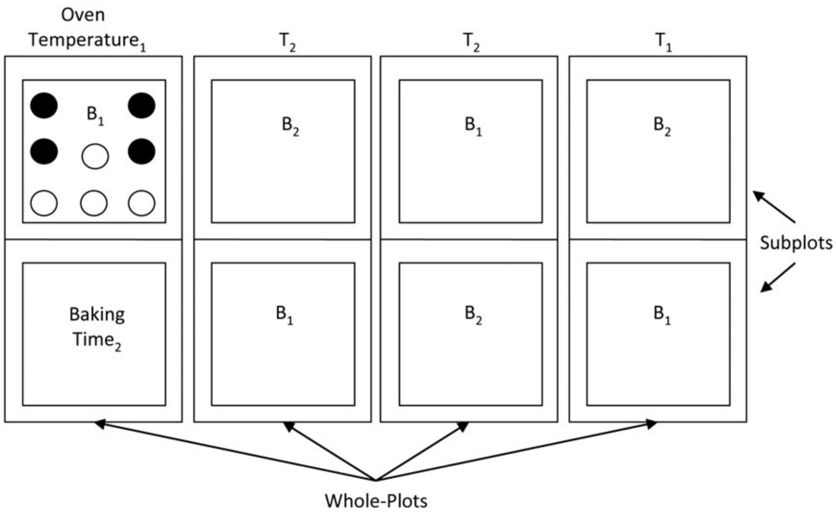

FIGURE 1 A split-plot design with one whole-plot factor (Oven Temperature), one subplot factor (Baking Time) and $P = 8$ subsamples per test stand. A darkened circle represents a failed component with a known failure time and a clear circle represents a right-censored component

the expression that incorporates the dependence of the log-scale parameter on the whole-plot and subplot factors is as follows:

$$\mu_{ij} = \log(\eta_{ij}) = \theta_0 + \theta_1 x_{T,i} + \delta_i + \theta_2 x_{B,ij} + \theta_3 x_{TB,ij} + \epsilon_{ij}, \quad (2)$$

where θ_0, θ_1, θ_2, and θ_3 are the fixed effect regression coefficients; δ_i is the random error associated with the i^{th} whole-plot; and ϵ_{ij} is the random error associated with the j^{th} subplot within the i^{th} whole-plot, and δ_i and ϵ_{ij} are mutually independent. Assuming the random errors follow normal distributions, then

$$\begin{aligned} f_\delta(\delta_i) &= \frac{1}{\sqrt{2\pi\sigma_\delta^2}} e^{\frac{-\delta_i^2}{2\sigma_\delta^2}} \\ f_\epsilon(\epsilon_{ij}) &= \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{\frac{-\epsilon_{ij}^2}{2\sigma_\epsilon^2}}. \end{aligned} \quad (3)$$

Note too, additional coefficients can be added to the model to accommodate a split-plot design having more than one whole-plot factor and more than one subplot factor.

In the construction of our failure-time model, we chose to model the impact of the experimental factors through the Weibull scale parameter, η. Because the Weibull shape parameter, β reflects the specific failure mechanism, assuming a constant β is a reasonable as long as the applied treatment combinations are not expected to change the failure mechanism. In this paper, we focus on the Weibull distribution, but the methodology can be extended to include other common reliability distributions.

3 | A REVIEW OF THE TWO-STAGE ANALYSIS

Dickinson[19] extended the work of Freeman and Vining[15] and Kensler et al[17] and proposed a two-stage analysis method for analyzing reliability data from a split-plot design with subsampling. In the simulation study, we compare the results of our proposed Weibull NLMM analysis to the results of the two-stage method. We include the necessary details of the two-stage method below for understanding.

3.1 | Stage one

The analysis in stage one accounts for the subsamples and reduces the dimensionality of the problem by mapping the failure times to a set of Weibull parameters (β, μ_{ij}) for each experimental unit. Under normal theory experimental design, the stage one analysis could be equated to calculating an average response, say $\bar{y}$, for each experimental unit. Using the Weibull distribution to model the failure times within each experimental unit, the first stage yields a common estimate of the shape parameter β and a separate estimate of the log-scale parameter μ_{ij} for each of the mn experimental units. The estimation of the Weibull parameters is based on the method of maximum likelihood estimation.

3.2 | Stage two

The analysis in stage two incorporates the experimental design by relating the treatment levels of the experimental factors to the maximum likelihood estimates of the log-scale parameter through a linear model. This allows one to test for factor effects. The specification of the linear model in stage two is dependent on the experimental design. For the split-plot design, the stage two model is

$$\hat{\boldsymbol{\mu}} = \log(\hat{\boldsymbol{\eta}}) = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{4}$$

where $\hat{\boldsymbol{\mu}}$ is the $mn \times 1$ vector of the estimated log-scale parameters in state one, $\boldsymbol{X}$ is the model matrix for the treatment effects, $\boldsymbol{\theta}$ is a vector of model coefficients including the intercept, $\boldsymbol{Z}$ is a $mn \times m$ incidence matrix, δ is a $m \times 1$ vector of random whole-plot errors and ϵ is a $mn \times 1$ vector of random subplot errors. The stage two analysis assumes $\delta_i \sim_{\text{iid}} \mathcal{N}(0, \sigma_\delta^2)$, $\epsilon_{ij} \sim_{\text{iid}} \mathcal{N}(0, \sigma_\epsilon^2)$, and that the δ's and ϵ's are independent.

Restricted maximum likelihood (REML) can be used to estimate and make inference on the parameters in Equation 4. When the OLS-GLS equivalence conditions hold, parameter estimates can be found using the ordinary least-squares (OLS) solution and the split-plot variance components can be estimated using expected mean squares.[22] Exact tests with known degrees of freedom can also be constructed for all terms in the model.

A disadvantage of the two-stage analysis is that analysts cannot easily perform inference using standard software on all quantities of interest, such as t_p (the p^{th} quantile), which is a function of both β and μ. This occurs because the covariances between $\hat{\beta}$, estimated in stage one, and $\hat{\boldsymbol{\theta}}$, $\hat{\sigma}_\delta$ and $\hat{\sigma}_\epsilon$, estimated in stage two, are unknown. In reliability testing, t_p, is of particular interest because it allows one to estimate how long a particular percentage of the population will last before failing. Dickinson[19] outlined a two-stage parametric bootstrap sampling procedure, which can be used to construct confidence intervals about all quantities of interest.

Another disadvantage of the two-stage method is that it is susceptible to producing biased estimates of the shape parameter β, which again, impacts the estimation and inference that is made on t_p. Wang et al[23] proposed a bias correction factor for grouped (ie, subsamples) lifetime data. This correction factor can be applied to the two-stage method, which results in an improved analysis.[19]

The two-stage method is intuitive, but the steps to improve on the analysis, that is, performing inference on all quantities of interest and correcting for the bias in the estimate of β, are cumbersome. For this reason, we propose an alternative analysis: a Weibull nonlinear mixed-model analysis (NLMM).

4 | NONLINEAR MIXED-MODEL ANALYSIS

The Weibull NLMM analysis allows for the estimation of all of the model parameters in a single step. Because the analysis is based on the joint-likelihood of all the model parameters, inferences can be made on all function of the parameters. In our simulation study, we will show that this analysis also reduces the bias in the estimate of β.

The joint-likelihood is for the Weibull NLMM for a split-plot design with subsampling is as follows:

$$L(\Theta) = \prod_{i=1}^{m}\left[\int_{-\infty}^{\infty}\left(\prod_{j=1}^{n}\left[\int_{-\infty}^{\infty} G_w(t_{k(ij)}|\delta_i,\epsilon_{ij}) f_\epsilon(\epsilon_{ij})d\epsilon_{ij}\right]\right) f_\delta(\delta_i)d\delta_i\right] \tag{5}$$

where

$$G_w(t_{k(ij)}|\delta_i,\epsilon_{ij}) = \prod_{k=1}^{p} f_w(t_{k(ij)}|\delta_i,\epsilon_{ij})^{r_{k(ij)}} F_w(t_{k(ij)}|\delta_i,\epsilon_{ij})^{1-r_{k(ij)}} \tag{6}$$

and

$$r_{k(ij)} = \begin{cases} 1 \text{ if } t_{k(ij)} \text{ is a failure} \\ 0 \text{ if } t_{k(ij)} \text{ is right-censored} \end{cases}.$$

Integration is required to evaluate and maximize the likelihood expression in Equation 5. The likelihood must be approximated using numerical techniques, because the expression involves two intractable integrals. The method of Gauss-Hermite quadrature is available because the random effects follow a normal distribution. This method yields an approximate closed-form solution of the total likelihood, allowing for the derivation of an asymptotic variance-covariance matrix, which is necessary for inference.

Gauss-Hermite quadrature approximates integrals of the form:

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dv,$$

as a weighted sum

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dv \approx \sum_{s=1}^{d} f(x_s)w_s,$$

where the w_s are weights and x_s are evaluation points, which are designed to provide an accurate approximation to the integral. The weights and evaluation points are of the form:

$$x_s = i^{th} \text{ zero of } H_d(x)$$

$$w_s = \frac{2^{d-1}d!\sqrt{\pi}}{d^2[H_{d-1}(x_s)]^2},$$

where $H_d(x)$ is the Hermite polynomial of degree d. Tables of x_s and w_s for values of d, the number of quadrature points, are available in the *Handbook of Mathematical Functions*.[24] A more convenient option, however, is to make use of the function *ghq()* found in the R library glmmML, which calculates the weights and evaluation points.

For the total joint likelihood expression in Equation 5, two iterations of Gauss-Hermite quadrature are required.

Applying Guass-Hermite quadrature, the approximate log-likelihood is

$$\mathcal{L}(\Theta) \approx \sum_{i=1}^{m} \log \left\{ \pi^{\frac{-(n+1)}{2}} \left[\sum_{s_2=1}^{d_2} W_{s_2} \left(\prod_{j=1}^{n} \left[\sum_{s_1=1}^{d_1} G_w \left(t_{k(ij)} | \sqrt{2\sigma_\delta^2} q_{s_2}, \sqrt{2\sigma_\epsilon^2} q_{s_1} \right) W_{s_1} \right] \right) \right] \right\}, \quad (7)$$

where d_1, d_2, W_{s_1}, W_{s_2}, q_{s_1}, and q_{s_2} are the number of quadrature points, the weights, and the evaluation points. Further details of the derivation can be made available by the corresponding author.

The closed-form approximate log-likelihood in Equation 7 can be maximized through standard maximization techniques to obtain the maximum likelihood estimates, $\hat{\Theta}^t = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\beta}, \hat{\sigma}_\delta^2, \hat{\sigma}_\epsilon^2)$. The estimated asymptotic variance-covariance matrix, $\hat{\Sigma}_{\hat{\Theta}}$, is found by taking the inverse of the observed information matrix, which allows for inferences to be performed on functions of parameters, including t_p.

5 | SOFTWARE RECOMMENDATIONS

The Weibull NLMM solution can be found using both R software and the SAS NLMIXED procedure. Using R to approximate and to optimize the log-likelihood involves an extensive amount of programming. In comparison, the SAS NLMIXED procedure is much more convenient because it only requires the specification of the conditional log-likelihood of the data given the random effects and the defining of the random effects and their distributions. A background on the NLMIXED procedure along with examples are provided in Chapter 15 of *SAS for Mixed Models*[25] and details on the syntax and computation algorithms behind the procedure are provided in Chapter 63 of *SAS/STAT9.3 User's Guide*.[26] Example R and SAS code can be made available upon request by the corresponding author.

6 | INFERENCE

A general approach for constructing hypothesis tests and confidence intervals is to use Wald's method or the normal-approximation method. Meeker and Escobar[7] provide a comprehensive discussion of Wald inferences on reliability data from a completely randomized design. These methods can be applied to the case of split-plot experiments with subsampling by using the variance-covariance matrix from the NLMM analysis. Confidence intervals for the parameters $\Theta = (\theta_0, \theta_1, \theta_2, \theta_3, \beta, \sigma_\delta, \sigma_\epsilon)$ follow:

- A $100(1-\alpha)\%$ confidence interval for the regression coefficients $\theta_0, \theta_1, \theta_2$ and θ_3 is

$$\hat{\Theta} \pm z_{(1-\alpha/2)} \hat{se}_{\hat{\Theta}}, \quad (8)$$

where $\hat{se}_{\hat{\Theta}} = \sqrt{Var(\hat{\Theta})}$ is the square root of the appropriate diagonal element of the variance-covariance matrix.

- A $100(1-\alpha)\%$ confidence interval for the positive valued parameters, β, σ_δ^2 and σ_ϵ^2, is

$$\left(\hat{\Theta}/w, \hat{\Theta}w\right), \quad (9)$$

where $w = \exp\left[z_{(1-\alpha/2)} \hat{se}_{\hat{\Theta}}/\hat{\Theta}\right]$ and $\hat{se}_{\hat{\Theta}} = \sqrt{Var(\hat{\Theta})}$ is the square root of the appropriate diagonal element of the variance-covariance matrix.

An advantage over the two-stage method is the ability to more easily obtain inferences about functions of the parameters, including t_p. Confidence intervals can be constructed for functions of the parameters, say $g_1 = g_1(\Theta^t)$. Quantities of particular interest may include the log-scale parameter μ and the p^{th} quantile, t_p. Meeker and Escobar[7] provide the general formulas needed for deriving $Var(\hat{g}_1)$ and $Var(\log(\hat{g}_1))$ for positive valued functions. Confidence intervals for the specific functions of interest, μ and t_p follow:

- A $100(1-\alpha)\%$ confidence interval for the log-scale parameter μ, estimated by $\hat{\mu} = \hat{\theta}_0 + \hat{\theta}_1 x_A + \hat{\theta}_2 x_B + \hat{\theta}_3 x_C$, is

$$\hat{\mu} \pm z_{(1-\alpha/2)} \hat{se}_{\hat{\mu}}, \quad (10)$$

where $\hat{se}_{\hat{\mu}} = \sqrt{Var(\hat{\mu})}$.

- A $100(1-\alpha)\%$ confidence interval for the positive valued quantile t_p, estimated by $\hat{t}_p = \exp\left(\hat{\mu} + \frac{\Phi_{sev}^{-1}(p)}{\hat{\beta}}\right)$, where $\Phi_{sev}^{-1}(p)$ is the p^{th} quantile for the standardized smallest extreme value distribution, is

$$\left[\hat{t}_p/w, \hat{t}_p w\right], \quad (11)$$

where $w = \exp\left[z_{(1-\alpha/2)} \hat{se}_{\hat{t}_p}/\hat{t}_p\right]$ and $\hat{se}_{\hat{t}_p} = \hat{t}_p \left\{ Var(\hat{\mu}) - 2\frac{\Phi_{sev}^{-1}(p)}{\hat{\beta}^2} Cov(\hat{\mu}, \hat{\beta}) + \left(\frac{\Phi_{sev}^{-1}(p)}{\hat{\beta}^2}\right)^2 Var(\hat{\beta}) \right\}^{1/2}$.

7 | AN ILLUSTRATIVE EXAMPLE: ELECTRICAL COMPONENT LIFE-TEST

Hicks and Turner[27] describe a factorial experiment used to study the effect of four oven temperatures (580, 600, 620, and 640 °F) and three baking times (5, 10, and 15 minutes) on the life of an electrical component. The experiment is run as a split-plot design because complete randomization of the oven temperature, a hard-to-change factor, is impractical and expensive. Note, the execution of a completely randomized design would have required assigning one of the four temperatures at random to an oven and heating it to temperature; then selecting a baking time at random, and placing an electrical component in the oven to bake for the time selected. After which, the entire procedure must be repeated until all of the lifetime data are collected. The split-plot design minimizes the number of required experimental settings of oven temperature.

In this example, subsamples are introduced because the experimental protocol consists of heating an oven to temperature and inserting nine components; three baked for 5 minutes, three for 10 minutes, and three for 15 minutes. Table 1 summarizes the data from the unreplicated factorial split-plot experiment. Note that, in this example, all of the life times are exact observations (ie, not censored).

Hicks and Turner[27] use the normal distribution to model the individual component lifetimes and treat the subsamples as three distinct replications of the split-plot design. The standard ANOVA-based approach to the analysis indicates that oven temperature is a significant treatment. Like most DOE textbooks, a primary purpose of their text is to provide readers with the fundamental concepts of design and their approach emphasizes the basic philosophy and analysis of a split-plot design.

In the analysis of data collected from an experimental design, it is important to take into account the experimental protocol. Failure to do so can lead to incorrect conclusions and/or misrepresenting the true experimental error. We use the lifetime data in Table 1 to demonstrate the both the NLMM method and the two-stage method of analysis, accounting for the earlier described experimental protocol.

Table 2 provides the estimates, standard errors, and 95 percent confidence intervals for the parameters using the proposed Weibull NLMM method of analysis. We estimate $log(\sigma_\delta)$ and $log(\sigma_\epsilon)$ which improves the convergence of the NLMM. By the invariance property of ML estimators, we find $\sigma_\delta = 0.103$ and $\sigma_\epsilon = 0.003$. Oven temperature is significant at the $\alpha = .05$ level. However, caution must be exercised when testing the effects for significance because this is an unreplicated split-plot design.

Table 3 compares the results the NLMM analysis to the results of the two-stage analysis. The methods produce different estimates of the shape parameter β. The two methods give similar estimates of the coefficients θ.

In this example, the experimenters are also interested in the time by which the first 10 percent of the components fail. Table 4 includes the estimates and 95% confidence

TABLE 1 Electrical component life-test results[27]

Whole-Plot	Oven Temperature, T, °F	Baking time, B, min 5	10	15
1	**580**	217	233	175
		188	201	195
		162	170	213
2	600	158	138	152
		126	130	147
		122	185	180
3	620	229	186	155
		160	170	161
		167	182	182
4	640	223	227	156
		201	181	172
		182	201	199

TABLE 2 NLMM analysis results for the electrical component life-test

The NLMM Solutions			
Parameter	**Estimate**	**Standard error**	**95% Confidence intervals**
θ_0	5.231	0.021	(5.189-5.272)
θ_1	0.056	0.027	(0.003-0.109)
θ_2	−0.023	0.026	(−0.074-0.282)
θ_3	−0.053	0.034	(−0.121-0.015)
β	8.716	1.115	(6.783-11.200)[a]
$\log(\sigma_\delta)$	−2.272	0.216	(−2.696,-1.849)
$\log(\sigma_\epsilon)$	−5.765	6.200	(−17.918,6.388)

[a]Interval based on a log transformation.

TABLE 3 Comparing the two-stage and NLMM analysis results for the electrical component life-test

Parameter	Two-Stage Estimate	SE	NLMM Estimate	SE
θ_0	5.227	0.070	5.231	0.021
θ_1	0.014	0.094	0.056	0.027
θ_2	−0.023	0.026	−0.023	0.026
θ_3	−0.053	0.035	−0.053	0.034
β	10.220	1.374	8.716	1.115
σ_δ	0.133	0.073	0.103	0.022
σ_ϵ	0.071	0.021	0.003	0.019

TABLE 4 Comparing the two-stage and NLMM point estimates and 95% confidence intervals for $t_{.10}$ for the electrical component life-test data[a]

Oven temperature		Baking time		Two-Stage Estimate $\hat{t}_{.10}$	Two-Stage Bootstrap 95% CI[b]	NLMM Estimate $\hat{t}_{.10}$	NLMM Wald 95% CI
580	(−1)	5	(−1)	142.98	(133.30-165.90)	132.45	(115.07-152.45)
580	(−1)	10	(0)	147.39	(141.23-166.98)	136.51	(122.88-151.65)
580	(−1)	15	(1)	151.94	(141.22-176.26)	140.69	(123.26-160.59)
600	(−0.33)	5	(−1)	149.55	(142.83-169.50)	142.49	(127.58-159.15)
600	(−0.33)	10	(0)	148.75	(144.09-166.29)	141.73	(129.39-155.23)
600	(−0.33)	15	(1)	147.97	(141.29-167.54)	140.96	(127.03-156.42)
620	(0.33)	5	(−1)	156.31	(149.16-177.120	153.13	(137.68-170.32)
620	(0.33)	10	(0)	150.10	(145.24-167.84)	147.06	(134.47-160.83)
620	(0.33)	15	(1)	144.15	(137.55-163.30)	141.23	(127.36-156.60)
640	(1)	5	(−1)	163.48	(151.90-189.48)	164.75	(144.58-187.72)
640	(1)	10	(0)	151.49	(144.54-171.69)	152.68	(137.99-168.94)
640	(1)	15	(1)	140.37	(130.58-163.03)	141.50	(124.19-161.22)

[a]Coded design units listed in parenthesis.
[b]Intervals based on a bootstrap sample size 10 000.

intervals for t_{10} for both the two-stage and NLMM analyses. The differences in the estimates of $\hat{t}_{10}$ reflect the differences in the estimate of $\hat{\beta}$.

8 | SIMULATION STUDY

In this section, we explore the performance of the Weibull NLMM method for analyzing a split-plot reliability experiment with subsamples. We evaluate the proposed methodology against its ability to estimate model parameters, preserve the nominal Type I error rate, and detect significant factor effects (ie, power). We also compare the NLMM performance to the two-stage method.

8.1 | Simulation setup

A factorial split-plot design with one whole-plot factor and one subplot factor, each having two levels (low= -1 and high= 1), was used to generate the lifetime data for this study. We assume a design structure having two replications of the whole-plot factor and one replication of the subplot factor, giving a total of eight experimental units. This design supports a first-order model with interaction. In addition, we assume each experimental unit contains p subsamples. To incorporate censoring, the simulation records the first r failures per each experimental unit, resulting in failure censoring or Type II censoring.

The failure times for each experimental unit were simulated using a Weibull distribution with a common shape parameter β and a log-scale parameter

$$\begin{aligned}\mu_{ij} &= \log(\eta_{ij}) = \theta_0 + \theta_1 w_i + \delta_i + \theta_2 s_{ij} + \theta_3 ws_{ij} + \epsilon_{ij}\\ i &= 1, \cdots, m = 8 \text{ (whole-plots)}\\ j &= 1, \cdots, n = 2 \text{ (subplots per whole-plot)},\end{aligned} \tag{12}$$

where $\delta_i \sim \mathcal{N}(0, \sigma_\delta^2)$, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and δ_i and ϵ_{ij} are mutually independent. Here, w denotes the whole-plot factor, s denotes the subplot factor, and ws denotes the cross-product.

The parameters and settings explored in this study include the following:

- Weibull shape parameter: β
- Whole-plot variance σ_δ^2
- Subplot or test stand variance σ_ϵ^2
- Products per test stand: p
- Failure per test stand: r.

Computations were carried out using five values for the Weibull shape parameter, β (.5, 1, 3, 5 and 10), three values for the total variance, $\sigma_t^2 = \sigma_\delta^2 + \sigma_\epsilon^2$ (.01, .05, .10), two variance ratios, $\rho = \sigma_\delta^2/\sigma_\epsilon^2$ (2 and 4) and six combinations of sample size and failures per test stand, (p, r) ((8,4), (8,6), (8,8), (12,6), (12,9) and (12,12)). For all the simulations, we fixed the intercept θ_0 at 5. The true values of θ_1, θ_2, and θ_3 change depending on the desired nominal power.

Additionally, recall that the NLMM uses two iterations of Guass-Hermite quadrature. These computations become very time consuming because the total number of evaluation points is the product of the number of quadrature points. For this reason, we set the number of quadrature points at $d_1 = d_2 = 5$, and the total number of simulations performed for each simulation scenario at N = 1000.

8.2 | Simulation results

The results of our simulation study provide a clear contrast between the NLMM and two-stage analysis method in their ability to estimate model parameters, preserve the type I error rate and determine factor significance. Figures 2 to 7 highlight outcomes for several specific settings explored. Across all settings explored, we found the following statements to be true:

- The NLMM provides the lowest bias estimates for most of the model parameters $(\beta, \theta_0, \sigma_\delta^2, \sigma_\epsilon^2)$.
    a. Both methods provide similar estimates of θ_1, θ_2, and θ_3
- The NLMM inflates the nominal empirical type I error rate; the two-stage method does not.
- The NLMM method has a higher a higher empirical power; the two-stage method has a lower empirical power.

For each of the above statements, we find that increasing sample size and decreasing the censoring rate improves the performance of each method. We also find that the performance of each method is affected by the variance components σ_δ^2 and σ_ϵ^2, as well as the shape parameter β, which also affects the variance of the Weibull distribution. Because these components affect the variability of failure times, they also affect the power of the test.

We recommend using the NLMM analysis if accurate estimation of the parameters and/or functions of the parameters is the objective of the reliability experiment. Because the two-stage method preserves the type I error rate, we recommend using it if the objective of the reliability experiment is to determine the significant factors that increase or decrease product lifetime.

8.2.1 | Parameter estimates

To generalize the results for all values of β, we plot the sample mean of the estimates in terms of the pivotal quantity, $\hat{\beta}/\beta$. In Figure 2, we compare the NLMM estimates to the two-stage estimates for the sample size and failure combinations $(8, 4)$ and $(12, 6)$. For both methods of analysis, the estimate of β is biased above the true value. The NLMM method has the least bias in estimating β over the simulation cases considered. Additionally, Figure 2 shows that the bias in the estimate of β decreases as the sample size and number of observed failures increase. Note that Figure 2 was generated using the variance components $\sigma_\delta^2 = .067$ and $\sigma_\epsilon^2 = .033$. The results are consistent for other combinations of σ_t^2 and ρ.

In Figure 3, we show the mean estimates of the regression coefficients $\hat{\theta}$, relative to the true value of θ. These

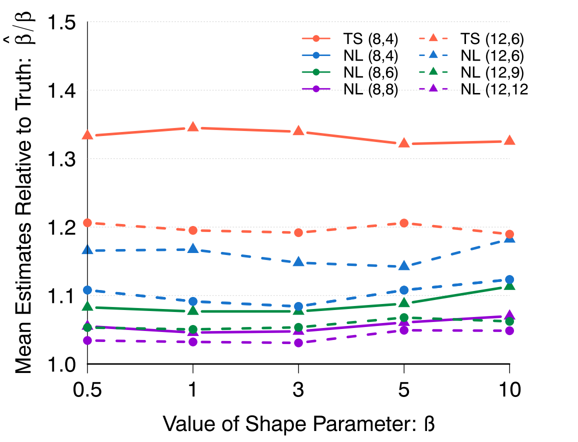

FIGURE 2 Mean estimates of $\hat{\beta}$ relative to β for both the NLMM method, denoted by NL, and the two-stage method, denoted by TS [Colour figure can be viewed at wileyonlinelibrary.com]

specific results are based on the total variance $\sigma_t^2 = 0.1$ and the variance ratio $\rho = 2$. In all cases considered, the NLMM method performs best in the estimation of the intercept θ_0. We found that the bias in the estimate depends on the number of products and failures per test stand (p, r) and the true value of the shape parameter β. Although not shown, the estimate of the intercept does not depend on the variance components σ_δ^2 and σ_ϵ^2.

The ability to estimate θ_1, θ_2, and θ_3 is nearly invariant to the modeling method, and the differences between the estimates of the NLMM method and the two-stage method are within the bounds of the simulation error. The bias of the estimates depends on the number of products and failures per test stand (p, r), the true value of the shape parameter β and the variance components σ_δ^2 and σ_ϵ^2.

In Figures 4 and 5, we include the estimation of the whole-plot and subplot variance components for the variance ratio $\rho = 2$, where $\sigma_\delta^2 = 0.067$ and $\sigma_\epsilon^2 = 0.033$. Compared to the two-stage method, the NLMM method provides a much better estimate of σ_δ^2 and σ_ϵ^2. The results for all other simulation cases are comparable. The bias in the NLMM estimates depend on the number of products and failures per test stand (p, r), the true value of β.

8.2.2 | Type I error rate

To determine if the NLMM method is able to control the nominal type I error rate, we generated lifetime data using the zero model. That is, using Equation 12, we set $\theta_0 = 5$ and $\theta_1 = \theta_2 = \theta_3 = 0$. Figure 6 depicts the empirical error for testing $H_0 : \theta_1 = 0$ versus $H_A : \theta_1 \neq 0$, $H_0 : \theta_2 = 0$ versus $H_A : \theta_2 \neq 0$, and $H_0 : \theta_3 = 0$ versus $H_A : \theta_3 \neq 0$, under a nominal type I error rate of $\alpha = 0.05$. In this particular figure, the settings of the variance components are $\sigma_\delta^2 = .067$ and $\sigma_\epsilon^2 = .033$. Notice that the two-stage method preserves the nominal

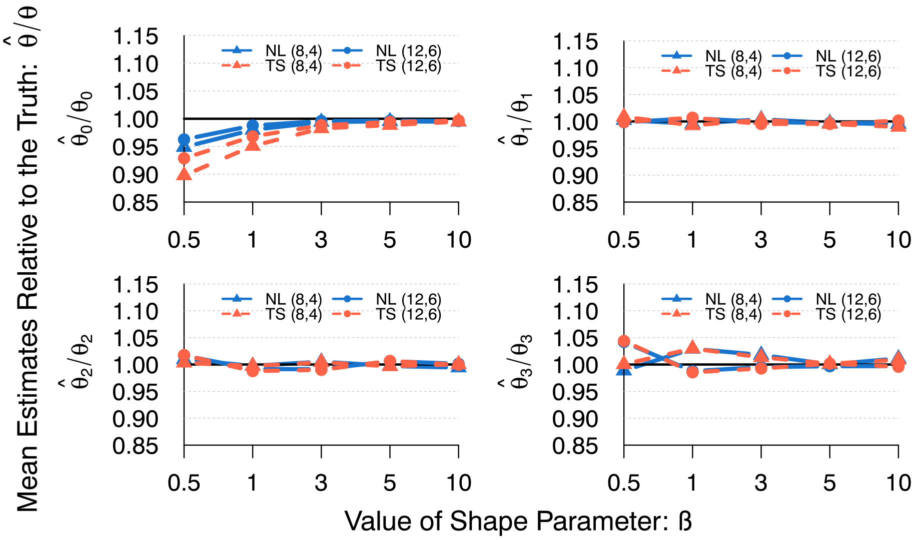

FIGURE 3 Mean estimates of regression coefficients $\hat{\theta}$, relative to the true value of θ where $\sigma_\delta^2 = 0.067$ and $\sigma_\epsilon^2 = 0.033$ [Colour figure can be viewed at wileyonlinelibrary.com]

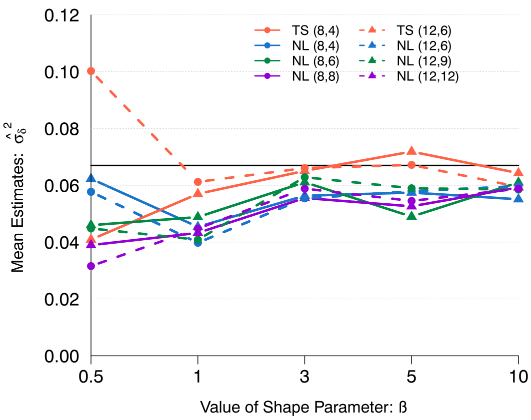

FIGURE 4 Mean estimates of the whole-plot variance $\hat{\sigma}_\delta^2$ where $\sigma_\delta^2 = 0.067$ and $\sigma_\epsilon^2 = 0.033$ [Colour figure can be viewed at wileyonlinelibrary.com]

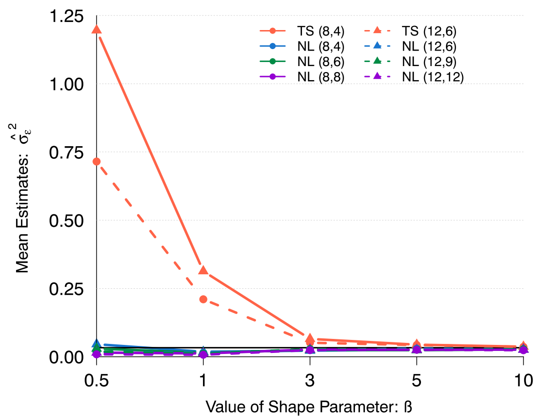

FIGURE 5 Mean estimates of the subplot variance $\hat{\sigma}_\epsilon^2$ where $\sigma_\delta^2 = 0.067$ and $\sigma_\epsilon^2 = 0.033$ [Colour figure can be viewed at wileyonlinelibrary.com]

type I error rate, but that the NLMM method does not. This trend is consistent for all cases considered in the simulation study. For this particular set of conditions, as the true value of β increases the empirical error rate increases from 0.10 to 0.45. Although not shown, the empirical error rate increases as the values of σ_δ^2 and σ_ϵ^2, increase; the two-stage method does not depend on the variance components.

8.2.3 | Power

Lastly, we consider the power of a test. The power of a test is the probability of rejecting the null hypothesis H_0 for a given value of the noncentrality parameter λ,

$$Power = P(F_{obs} > F_{(1-\alpha)}, df_1, df_2, \lambda).$$

For the split-plot design assumed in our simulation study, F-tests with known degrees of freedom can be constructed for testing all of the model coefficients. For our split-plot design, the appropriate error degrees of freedom for testing the whole-plot, subplot, and interaction effect is $df_1 = 1$ and $df_2 = 2$. Note that the denominator degrees will not always match for the test of each effect and careful thought must always be given to determining the appropriate degrees of freedom. The noncentrality parameters for testing each of the model effects are the following:

- Whole-Plot Effect, $\lambda_w = \frac{8\theta_1^2}{\sigma_\epsilon^2 + n\sigma_\delta^2}$,
- Subplot Effect, $\lambda_s = \frac{8\theta_2^2}{\sigma_\epsilon^2}$,
- Whole-Plot by Subplot Effect Interaction, $\lambda_{ws} = \frac{8\theta_3^2}{\sigma_\epsilon^2}$.

Under the assumption that $H_0 : \theta_1 = 0$ is true, $P(F_{obs} > F_{0.95,1,2,\lambda_w=0} = 18.51282) = 0.05$. Now, to find the value of θ_1 that yields a nominal power of say 0.80, we must find

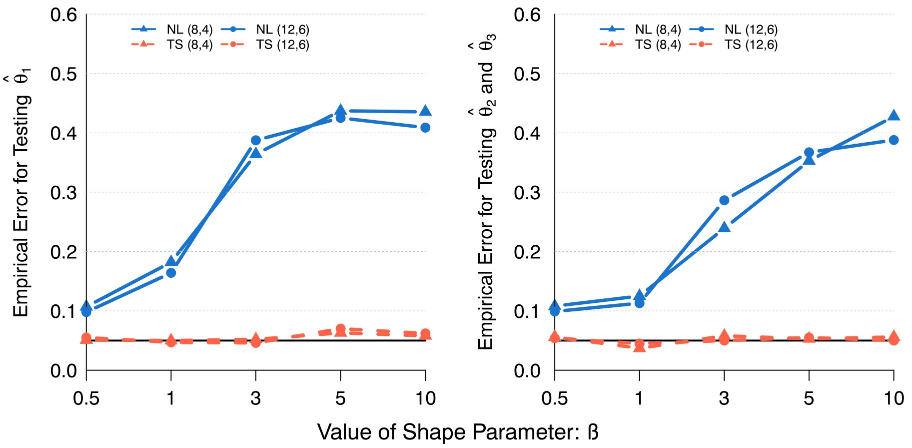

FIGURE 6 A comparison of the empirical error rate for the NLMM and two-stage analysis for the parameter settings $\sigma_\delta^2 = 0.067$ and $\sigma_\delta^2 = 0.033$ [Colour figure can be viewed at wileyonlinelibrary.com]

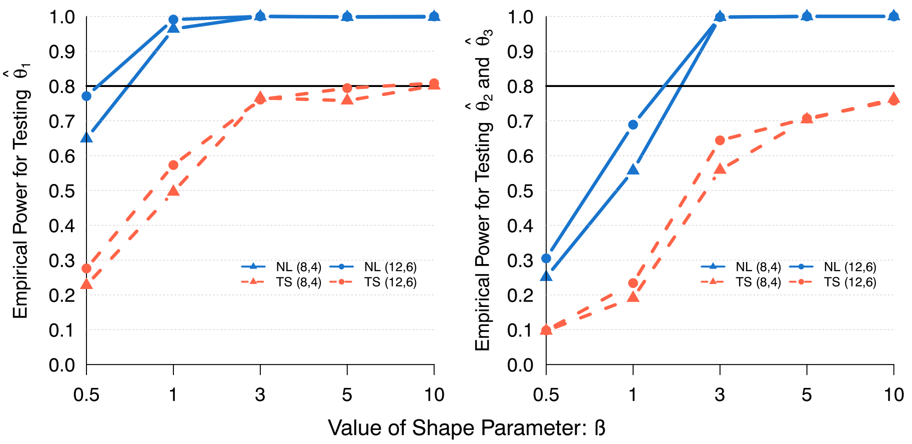

FIGURE 7 A comparison of the empirical power rate for the NLMM and two-stage analysis for the nominal power 0.80 and parameter settings $\sigma_\delta^2 = 0.067$ and $\sigma_\epsilon^2 = 0.033$ [Colour figure can be viewed at wileyonlinelibrary.com]

the value of λ_w such that the probability of rejecting H_0 is 0.80. That is, we must find the value of λ_w such that $P(F_{obs} > 18.51282|\lambda_w) = 0.80$). Using a bisection search algorithm, the value of λ_w that makes this statement true is $\lambda_w = 32$. By substitution and letting $\sigma_\delta^2 = 0.067$ and $\sigma_\epsilon^2 = 0.033$, we find $\theta_1 \pm 0.817$. Similarly, $\theta_2 = \pm 0.363$ and $\theta_3 = \pm 0.363$. Notice, to have more power in detecting a nonzero whole-plot effect, the size of the effect must be large compared with the effect size required for a subplot effect and whole-plot by subplot interaction effect. For all simulations, we generated lifetimes using the positive coefficient solutions.

Figure 7 reports the empirical power of a test on θ_1, θ_2 and θ_3 under a nominal power of 0.80. The NLMM method tends to overstate the power, whereas the two-stage method tends to understate the power. This trend is consistent for all cases considered in the simulation study. As shown, in the figure, both methods are dependent on the shape parameter. In the same way, both methods are also affected by the variance components σ_δ^2 and σ_ϵ^2.

9 | CONCLUDING REMARKS

Freeman and Vining[15,16] provided a new framework for thinking about the design and analysis of reliability experiments. Their work focused on properly accounting for the experimental error when subsampling is present in a reliability experiment. Kensler et al[17,18] extended the work to include the proper analysis of a reliability experiment with random blocks and subsampling. In this paper, we extend the NLMM method for analyzing a split-plot reliability experiment with subsampling, involving both censored and exact-failure observations. We show how to properly account for the random effects introduced by the split-plot design structure and the subsampling in the analysis of failure time data.

Based on the results of our simulation study, we recommend using the NLMM analysis if the objective of the readability experiment is accurate estimation of the parameters and/or functions of the parameters. We recommend using the two-stage method if the objective of the reliability experiment is to determine the significant factors that increase or decrease product lifetime. In any case, accounting for the experimental protocol using the two-stage analysis or the NLMM analysis is better than completely ignoring the experimental protocol.

ACKNOWLEDGEMENTS

This research was supported by the Office of the Secretary of Defense, Director of Operational Test and Evaluation (OSD DOT&E) and the Test Resource Management Center (TRMC) within the Science of Test research consortium.

DISCLAIMER

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense or the US Government.

REFERENCES

1. Standard A. 0009 Reliability Program Standard for Systems Design, Development, and Manufacturing; 2008.
2. Montgomery DC. *Design and Analysis of Experiments*, 9th ed. Hoboken, New Jersey: John Wiley & Sons; 2017.
3. Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed. Hoboken, New Jersey: John Wiley & Sons; 2009.
4. Zelen M. Factorial experiments in life testing. *Technometrics*. 1959;1(3):269-288.
5. Condra L. *Reliability Improvement with Design of Experiment*. Boca Raton, Florida: CRC Press; 2001.
6. McCool JI, Baran G. The analysis of 2×2 factorial fracture experiments with brittle materials. *J Mater Sci*. 1999;34(13):3181-3188.
7. Meeker WQ, Escobar LA. *Statistical Methods for Reliability Data*. Hoboken, New Jersey: John Wiley & Sons; 1998.
8. Bullington RG, Lovin S, Miller DM, Woodall WH. Improvement of an industrial thermostat using designed experiments. *J Qual Technol*. 1993;25(4):262-270.
9. Fang J, He Z, Zhang Z, He S. Reliability improvement of diamond drill bits using design of experiments. *Qual Eng*. 2018;30(2):339-350.
10. Lawless JF. *Statistical Models and Methods for Lifetime Data*. Hoboken, New Jersey: John Wiley & Sons; 2011.
11. Hamada M. Using statistically designed experiments to improve reliability and to achieve robust reliability. *IEEE Trans Reliab*. 1995;44(2):206-215.
12. Wu CF, Hamada MS. *Experiments: Planning, Analysis, and Optimization*, Vol. 552. Hoboken, New Jersey: John Wiley & Sons; 2011.
13. Nelson W. *Accelerated Life Testing: Statistical Models Test Plan and Data Analysis*. New York, USA: J. Wiley; 1990.
14. Rigdon SE, Englert BrandonR, Lawson IA, Borror CM, Montgomery DC, Pan R. Experiments for reliability achievement. *Qual Eng*. 2012;25(1):54-72.
15. Freeman LJ, Vining G. G. Reliability data analysis for life test experiments with subsampling. *J Qual Technol*. 2010;42(3):233-241.
16. Freeman LJ, Vining GG. Reliability data analysis for life test designed experiments with sub-sampling. *Q Reliab Eng Int*. 2012;29(4):509-519.
17. Kensler JL, Freeman LJ, Vining GG. A practitioner's guide to analyzing reliability experiments with random blocks and subsampling. *Qual Eng*. 2014;26(3):359-369.
18. Kensler JL, Freeman LJ, Vining GG. Analysis of reliability experiments with random blocks and subsampling. *J Qual Technol*. 2015;47(3):235-251.
19. Dickinson RM. Statistical methods for improving and maintaining product reliability. *PhD thesis*; 2014. Virginia Polytechnic Institute and State University.
20. Johnson TH, Medlin RM, Freeman LJ. Challenges and New Methods for Designing Reliability Experiments. UnderReview; 2018.
21. Johnson TH, Medlin RM, Freeman LJ. Power Approximations for Failure-Time Regression Models. Submitted for Review; 2018.
22. Vining GG, Kowalski SM, Montgomery DC. Response surface designs within a split-plot structure. *J Qual Technol*. 2005;37(2):115-129.
23. Wang G, Niu Z, He Z. Bias reduction of MLEs for Weibull distributions under grouped lifetime data. *Qual Eng*. 2015;27(3):341-352.
24. Abramowitz M, Stegun IA. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Mineola, NY: Courier Dover Publications; 2012.
25. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. *SAS for Mixed Models*, 2nd ed. Cary, NC: SAS Institute Inc.; 2006.
26. SAS Institute Inc. *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute Inc.; 2011.
27. Hicks CR, Turner KV. *Fundamental Concepts in the Design of Experiments*, 5th ed. Bethesda, Maryland: Oxford University Press, Inc.; 1999.

AUTHOR BIOGRAPHIES

Rebecca Medlin is a Research Staff Member in the Operational Evaluation Division at the Institute for Defense Analyses. She supports the division by providing expertise in statistics. Her areas of statistical expertise include designed experiments, reliability analysis, and industrial statistics. She focuses on the operational test of radars and electronic warfare systems. Medlin has a PhD in statistics from Virginia Tech.

Laura Freeman is an assistant director of the Operational Evaluation Division at the Institute for Defense Analyses. Her areas of statistical expertise include designed experiments, reliability analysis, and industrial statistics. She focuses on operational tests for Air Force fixed wing aircraft. Freeman has a PhD in statistics from Virginia Tech.

Jennifer Kensler is a statistical consultant at Shell Global Solutions. Her areas of statistical expertise include designed experiments, reliability analysis, and industrial statistics. Kensler has a PhD in statistics from Virginia Tech.

Geoffrey Vining is a Professor of Statistics at Virginia Tech. He is a Fellow of the ASQ and a recipient of the ASQ Shewhart Medal and the European Network for Business and Industrial Statistics (ENBIS) Box Medal.

How to cite this article: Medlin RM, Freeman LJ, Kensler JLK, Vining GG. Analysis of split-plot reliability experiments with subsampling. *Qual Reliab Engng Int*. 2019;35:738–749. https://doi.org/10.1002/qre.2394