Simulation-based Bayesian Optimal ALT Designs for Model Discrimination

Ehab Nasir[1], Rong Pan[2]

[1] Industrial Engineer, Intel Corporation, Chandler, Arizona.
[2] Associate Professor, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University. Corresponding Author, (480)965-4259, rong.pan@asu.edu.

ABSTRACT

Accelerated life test (ALT) planning in Bayesian framework is studied in this paper with a focus of differentiating competing acceleration models, when there is uncertainty as to whether the relationship between log mean life and the stress variable is linear or exhibits some curvature. The proposed criterion is based on the Hellinger distance measure between predictive distributions. The optimal stress-factor setup and unit allocation are determined at three stress levels subject to test-lab equipment and test-duration constraints. Optimal designs are validated by their recovery rates, where the true, data-generating, model is selected under the DIC (Deviance Information Criterion) model selection rule, and by comparing their performance with other test plans. Results show that the proposed optimal design method has the advantage of substantially increasing a test plan's ability to distinguish among competing ALT models, thus providing better guidance as to which model is appropriate for the follow-on testing phase in the experiment.

KEYWORDS

Reliability test plans, Hellinger distance, Model selection, Deviance information criterion (DIC), Non-parametric curve fitting.

1. MOTIVATION FOR WORK

Most work of the optimal Accelerated Life Testing (ALT) designs in literature has focused on finding test plans that allow more precise estimate of a reliability quantity, such as life percentile, at a lower stress level (it is usually the use stress level); see, for example, Nelson and

Kielpinski [1] and Nelson and Meeker [2]. Nelson [3, 4] summarized the ALT literature up to 2005 and a significant portion of this article is devoted to the optimal design of ALT planning. More recent discussions of optimal ALT plans and/or robust ALT plans can be found in, e.g., Xu [5], McGree and Eccleston [6], Monroe et al. [7], Yang and Pan [8], Konstantinou et al. [9], Haghighi [10]. In the previous study, the associated confidence intervals of an estimate reflect the uncertainty arising from limited sample size and censoring at test, but do not account for model form inadequacy. However, model errors can be quickly amplified and potentially dominate other sources of errors in reliability prediction through the model-based extrapolation that characterizes ALTs. Implicit in the design criteria used in current ALTs is the assumption that the form of the acceleration model is correct. In many real-world problems this assumption could be unrealistic. A more realistic goal of an initial stage of ALT experimentation is to find an optimal design that helps in selecting a model among rival or competing model forms. The ALT designs that are good for model form discrimination could be quite different from those that are more appropriate for life percentile prediction under a specific model.

Extrapolation in both stress and time is a typical characteristic of ALT inference. The most common accelerated failure time regression models (based, for example, on Lognormal or Weibull fit to the failure time distribution at a given stress level) are only adequate for modeling some simple chemical processes that lead to failure (Meeker and Escobar [11]). However, for modern electronic devices, more sophisticated models with basis in the physics of failure mechanisms are required. These complicated models are expected to have more parameters with possible interactions among stress factors. Therefore, investigating ALT designs with model selection capability is needed more than ever before. Meeker et al. [12] in their discussion of figures of merit when developing an ALT plan emphasizes the usefulness of a test plan's robustness to the departure from the assumed model. For example, when planning a single-factor experiment under a linear model, it is useful to evaluate the test plan properties under a quadratic model. Also, when planning a two-factor experiment under the assumption of a linear model with no interaction, it is useful to evaluate the test plan properties under a linear model with an interaction term. We strongly believe that it is worthwhile to consider

these recommended practices ahead of time when the test plan is being devised in the first place by allowing a design criterion that is capable of model form discrimination.

2. PREVIOUS WORK

A considerable work has been done in the development of experimental designs for discrimination among linear regression models; see, for example, Hunter and Reiner [13], Box and Hill [14], Hill et al. [15], Atkinson and Cox [2]. A comprehensive review of early contributions is given by Hill [17]. More recently, many authors focused on the development of T-optimum criterion (non-Bayesian) for model discrimination (de Leon and Atkinson [18], Atkinson et al. [19]). Dette and Titoff [20] derived new properties of T-optimal designs and showed that in nested linear models, the number of support points in a T-optimal design is usually too small to enable the estimate of all parameters in the full model; Agboto et al. [21] reviewed T-optimality among other new optimality criteria for constructing two-level optimal discrimination designs for screening experiments. These work resulted in sequential experimentation procedures.

Bayesian criteria were also considered in model discrimination. Meyer et al. [22] considered a Bayesian criterion that is based on the Kullback-Leibler information to choose follow-up run after a factorial design to de-alias rival models. Bingham and Chipman [23] proposed a Bayesian criterion that is based on the Hellinger distance between predictive densities for choosing optimal designs for model selection with prior distributions specified for model coefficients and errors. For a comprehensive review on Bayesian experimental design reader is referred to Chaloner and Verdinelli [24].

There are three types of uncertainties involved in the ALT planning – the uncertainty of failure time distribution, the uncertainty of lifetime-stress relationship and the uncertainty of model parameter value (Pascual [25]). Bayesian methods have been proposed for ALT planning to deal with the uncertainty of model parameter (Zhang and Meeker [26]; Yuan, et al. [27]), but, to our knowledge, none has been explicitly targeting the model discrimination of life-stress functions. All of the previous attempts at model discrimination have been in the context of traditional experimental design for linear models, while the failure time regression models used

in ALTs are nonlinear. In particular, failure time censoring is commonly expected in ALT experiments. Nelson [28] (p. 350) has cautioned that the statistical theory for traditional experimental design is correct only for *complete* data, one should not assume that properties of standard experimental designs hold for *censored* and *interval-censored* data as they usually do not hold. For example, aliasing of effects may depend on the censoring structure. In addition, the variance of an estimate of a model coefficient depends on the amount of censoring at all test conditions and on the true value of (possibly all) model coefficients. Thus, the censoring times at each test condition are part of the experimental design and affect its statistical properties. As such, our current work draws its importance from its attempt at contributing to model discrimination literature for accelerated life test planning when censoring is inevitable.

3. PROPOSED METHODOLOGY

3.1. Rationale for Model Discrimination Methodology

Suppose that the objective is to arrive at an ALT test plan that is capable of discriminating among competing acceleration models. Assume that there are two rival models and it is better that the experimental data can help in choosing one of them. Intuitively, a good design should be expected to generate far apart results based on the two competing models, and then the experimenter can select the model based on the actual observations from the experiment. In ALT, the lifetime percentile is typically of interest; therefore the larger the distance (disagreement) in prediction the better our ability to discriminate (distinguish) among these competing models. Therefore, we propose to use the relative prediction performance of each model over the range of its parameters to identify the optimal design. Figure 1 shows how important it is for the experimenter to arrive at the best representative model to reduce prediction errors at use conditions (UCs). For example, if M_1 is the true model but experimenter assumes M_2, then under ALT extrapolation the error in prediction of a quantile of interest at use conditions, $\Delta\hat{\tau}_p(UC)$, is much worse than any predictions at tested conditions.

Insert Figure 1 here

To distinguish predictive distributions from rival models, the Hellinger distance, as a measure of disagreement between predictive densities, is used in this work.

3.2. Distance (Divergence) Measure of Probability Distributions

There are a substantial number of distance measures applied in many different fields such as physics, biology, psychology, information theory, etc. See Sung-Hyuk Cha [11] and Ullah [35] for a comprehensive survey on distance/similarity measures between probability density functions. From the mathematical point of view, distance is defined as a quantitative measure of how far apart two objects are. In statistics and probability theory, a statistical distance quantifies the dissimilarity between two statistical objects, which can be two random variables or two probability distributions. A measure $D(x, y)$ between two points x, y is said to be a distance measure or simply distance if

I. $D(x, y) > 0$ when $x \neq y$ and $D(x, y) = 0$ if and only if $x = y$,
II. $D(x, y) = D(y, x)$,
III. $D(x, y) + D(y, z) \geq D(x, z)$.

Conditions (I) through (III) imply, respectively, that the distance must be non-negative (positive definite), symmetric and sub-additive (triangle inequality: the distance from point x to z directly must be less than or equal to the distance in reaching point z indirectly through point y.

The choice of a distance measure depends on the measurement type or representation of quantities under study. In this study, the Hellinger distance $(\mathrm{D_H})$ (Deza and Deza [29]) is chosen to measure the distance between the two probability distributions that represent the distributions of $\hat{\tau}_p$ at lower and higher ALT stress test conditions. Computing the distance between two probability distributions can be regarded as the same as computing the Bayes (or minimum misclassification) probability of misclassification (Duda et al. [30], Cha and Srihari [31]). For the discrete probability distributions $P = (p_1 \cdots p_k)$ and $Q = (q_1 \cdots q_k)$, the Hellinger distance (D_H) is defined as:

$$D_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2} \quad (1)$$

This is directly related to the Euclidean norm of the difference of the square root vectors,

$$D_H(P,Q) = \frac{1}{\sqrt{2}}\left\|\sqrt{P} - \sqrt{Q}\right\|_2 \quad (2)$$

For the continuous probability distributions, the squared Hellinger distance is defined as:

$$D_H^2(P,Q) = \frac{1}{2}\int\left(p_x^{\frac{1}{2}} - q_x^{\frac{1}{2}}\right)^2 dx$$

$$= 1 - \int\sqrt{p_x q_x}\ dx \quad (3)$$

Hellinger distance follows the triangle inequality and $0 \leq D_H(P,Q) \leq 1$. The maximum distance of 1 is attained when P assigns probability zero to every set to which Q assigns a positive probability, and vice versa.

3.3. Criterion for Model Discrimination

In Bayesian framework of experimental design, the problem of optimal design can be thought of as finding a design, d^*, such that it maximizes a utility function $\mathrm{U}(\mathrm{d})$ that quantifies the objective of the experiment (which is the model form distinguishability in our case).

Suppose that under design d, the experimental outcome may be generated by one of the following two models:

- Model 1, M_1, with its parameter vector θ_1, its outcome denoted by $Y_1 = (y_{11}, \dots, y_{N1})$
- Model 2, M_2, with its parameter vector θ_2, its outcome denoted by $Y_2 = (y_{12}, \dots, y_{N2})$

Consider, as an initial utility function to be maximized, the difference in prediction of life percentile of interest τ_p at the low stress $\tau_p(S_1)$ of the ALT test setup across all pairs of competing models. Ultimately, interest lies in the prediction of the 1st percentile of life distribution at use condition, $\tau_{0.01}$. Since the lower stress level is the closest to the use stress level, a large difference in prediction at the lower level will give rise to an even larger difference in prediction at the use level (due to the extrapolation error). Therefore, a design that may generate larger difference in the failure time at the lower stress level among rival models is preferable in discrimination sense. However, selection of the lower stress level to optimize the

local utility function may run the risk of not enough fails obtained to sufficiently estimate life distribution percentiles. Therefore, we consider the simultaneous difference in prediction of life percentile of interest, τ_p, at the lower stress $\tau_p(S_1)$ and the higher stress $\tau_p(S_2)$ test setup across all pairs of competing models. This study considers constant-stress ALT plans, where no interaction between stress variables is assumed. It is also assumed that the disperse parameter of log (life) distribution does not depend on stress.

For the two competing models, M_1 and M_2, the pairwise local utilities are as follows:

$$\mathrm{u}_{2|1}\big(\mathrm{d}, \mathrm{M}_1(\theta_1, \mathrm{Y}_1), \mathrm{M}_2(\theta_2, \mathrm{Y}_1)\big) = \mathrm{D}_{\mathrm{S}_1}\left(\hat{\tau}_{\mathrm{p},(\mathrm{M2|Y1})}, \hat{\tau}_{\mathrm{p},(\mathrm{M1|Y1})}\right) + \mathrm{D}_{\mathrm{S}_2}\left(\hat{\tau}_{\mathrm{p},(\mathrm{M2|Y1})}, \hat{\tau}_{\mathrm{p},(\mathrm{M1|Y1})}\right) = \mathrm{u}_{2|1} \quad (4)$$

$$\mathrm{u}_{1|2}(\mathrm{d}, \mathrm{M}_1(\theta_1, \mathrm{Y}_2), \mathrm{M}_2(\theta_2, \mathrm{Y}_2)) = \mathrm{D}_{\mathrm{S}_1}\big(\hat{\tau}_{\mathrm{p},(\mathrm{M1|Y2})}, \hat{\tau}_{\mathrm{p},(\mathrm{M2|Y2})}\big) + \mathrm{D}_{\mathrm{S}_2}\big(\hat{\tau}_{\mathrm{p},(\mathrm{M1|Y2})}, \hat{\tau}_{\mathrm{p},(\mathrm{M2|Y2})}\big) = \mathrm{u}_{1|2} \quad (5)$$

where $D_{S_1}(\cdot)$ and $D_{S_2}(\cdot)$ denote the Hellinger distance at the lower stress and the higher stress, respectively. Equation (4) represents the difference in τ_p prediction of model (M_2) conditional on data from model (M_1) relative to model (M_1) prediction of the same quantity, while Equation (5) represents the difference in τ_p prediction of model (M_1) conditional on data from model (M_2) relative to model (M_2) prediction of the same quantity. That is the relative prediction performance of each model over the range of its parameter vector.

At the time of designing an experiment, the experimental outcome is yet to observe and the true model form and its parameter vector are unknown. Therefore,

a) The utility $u_{i|j}(\cdot)$ of a design is assessed by its expectation with respect to the sampling distribution of the data $p(y_1|\theta_1, d)$, and $p(y_2|\theta_2, d)$, and the prior distribution of the parameter vectors $\pi(\theta_1)$ and $\pi(\theta_2)$. That is calculating the pre-posterior expectation.

$$E(u_{2|1}) = \iint u_{2|1}\, p(y_1|\theta_1, d)\, \pi(\theta_1|d) dy_1 d\theta_1 \quad (6)$$

$$E(u_{1|2}) = \iint u_{1|2}\, p(y_2|\theta_2, d)\, \pi(\theta_2|d) dy_2 d\theta_2 \quad (7)$$

Equation (6) gives an expression of the expected pre-posterior prediction difference in τ_p of M_2 conditional on data from model M_1 relative to model M_1 prediction of the same quantity. The reverse is true for Equation (7).

b) Since it is not known which of the two models (M_1) or (M_2) is the true model, a weighted sum of expected utilities $E(u_{i|j})$ is obtained as the desired global utility function $U(d)$ to be maximized. The weighing is achieved by priors assigned to the models, $\pi(M_1)$ and $\pi(M_2)$ respectively.

$$U(d) = \sum_{\substack{i,j=1,2 \\ i \neq j}} \pi(M_i).E(u_{i|j})$$

$$= \pi(M_1).E(u_{2|1}) + \pi(M_2).E(u_{1|2}) \quad (8)$$

Equation (8) can be interpreted as a measure of model distinguishability between two models. The larger the value of $\mathrm{U(d)}$, the dissimilar the two models are to each other. Extending (8) to account for situations where more than two models are to be distinguished among is straightforward.

As can be seen from Equations (6)-(8), arriving at an optimal design d^* that maximizes (8) is a nontrivial task due to the high dimensional integration and optimization required. There is no closed form solution to (8). Numerical evaluation of the multiple integral for a given choice of design will be needed, which in itself a formidable task given the fact that the integration is defined over the data space and parameter space. The obtained estimate of $U(d)$ must then be maximized over the design variable, d, which is in often cases a multidimensional vector. We use a Monte Carlo simulation-based approach to find the optimal design, d^*. As non-sequential ALT designs are always performed off-line, the relatively heavy computation requirement of this approach is not a critical issue.

Figure 2 presents a high level flow of the proposed methodology. For a candidate design, it is evaluated by the utility function (8) according to the assumed model and the true model. The numbered steps of this process are explained below.

Step 1 - Fail data Y_1 are generated from the acceleration model M_1 (the assumed true model). For those failure times that exceed the test censoring time, they are replaced by the censoring time.

Step 2 – Given model M_1 is assumed, the failed data are combined through Bayes's theorem with the prior info available on parameters $\pi_{M_1}(\theta)$ to produce the posterior estimates of the parameter, $P_{M1}(\theta|Y_1)$. Repeat it for model M_2.

Step 3 - Posterior distribution of the predicted life percentile of interest τ_p is obtained using Gibbs sampler at both high and low stress conditions, which are Steps 3a and 3b. Same steps are repeated on same data set Y_1 using rival model M_2.

Step 4 - The Hellinger distances between the prediction distributions under the true model M_1 and the rival model M_2 are obtained for both the lower and higher stress levels.

Step 5 - The sum of the Hellinger distances is denoted as the local utility.

If there are more than one rival model, this process will be repeated for models M_2 through M_m. Local utilities are then weighted by model priors and summarized into a global utility. This utility value gives the overall performance of model discrimination of a candidate design.

An R program is written to automate the process. It first generates random fail data according to the true acceleration model. Then, it calls WinBUGS to perform Bayesian inference by Markov chain Monte Carlo (MCMC). Using WinBUGS, a stream of samples from the posterior distribution of the life percentile under an assumed acceleration model will be generated and they are feedback to the R program to compute the Hellinger distance and the utility value. Eventually, multiple candidate designs will be evaluated and a response surface model is used to fit their utility values. The best design can be found by maximizing the fitted model.

Insert Figure 2 here

4. MODEL SELECTION CRITERION

In this section, the tools that are used for validating the obtained optimal designs are introduced. It can be shown that these designs are indeed optimal under desired optimality criterion as they maximize the proportion of times (the recovery rate) in which the true, data-generating, model is selected under an appropriate model selection rule.

The Deviance Information Criterion (DIC) was introduced by Spiegelhalter et al. [32] as an easily computed and rather universally applicable Bayesian criterion for posterior predictive model comparison. It compromises between data fit and model complexity, like many other non-Bayesian criteria. It generalizes Akaike's information criterion (AIC) that appears as a special case under a vague prior (negligible prior information), and Bayesian information criterion (BIC), also known as Schwarz criterion. DIC is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by MCMC simulation, because it can be directly computed using the MCMC samples. Claeskens and Hjort (Ch. 3.5) [33] show that the DIC is large-sample equivalent to the natural model-robust version of the AIC.

Define the following

- Deviance is defined as $Dev(\theta) = -2\log[p(y|\theta)] + C$, where y are the data, θ are vector of model unknown parameters, $p(y|\theta)$ is the likelihood function and C is a constant term that cancels out when comparing models.
- Expectation is defined as $\overline{Dev} = E_\theta[Dev(\theta)]$. This measures how well a model fits the data, the larger its value, the worse the fit.
- Effective number of model parameters is defined as $p_D = \overline{Dev} - Dev(\bar{\theta})$, where $\bar{\theta}$ is the expectation of θ. The larger p_D, the easier for the model to fit the data.

Finally, DIC is defined as a classical estimate of fit plus twice the effective number of parameters, i.e.,

$$\begin{aligned} DIC &= Dev(\bar{\theta}) + 2p_D \\ &= \overline{Dev} + p_D \end{aligned} \tag{9}$$

When comparing models, models with smaller DIC are preferred to models with larger DIC. Models are penalized both by the value of $\overline{Dev}$, which favors a good fit, but also (in common with AIC and BIC) by the effective number of parameters p_D. Since $\overline{Dev}$ decreases as the number of parameters in a model increases, the p_D term compensates for this effect by favoring models with a smaller number of parameters.

5. METHODOLOGY ILLUSTRATION

In this section, we use a real industrial example to demonstrate the proposed methodology. The R and WinBUGS codes of this example had been submitted to the publisher's website as the supplementary material.

5.1. Description of Design Problem

Reliability engineer is interested in studying the intermetallic growth of Au-Al interface in a semi-conductor assembly. It is known that fail mechanism of interest is activated by temperature stress so an accelerated life test is desired in order to estimate the device lifetime. However, there is an uncertainty as to whether the relationship between log(life) and the stress (possibly transformed) is linear or exhibit some curvature as indicated by an early look-ahead data set. As a result, current interest lies in an accelerated life test plan that is capable of discriminating between linear and quadratic acceleration models in temperature stress. There are also constraints imposed by available budget for testing (test units), and stress-lab equipment availability and capability as shown below.

- Bake stress chambers are available for 42 days (1,008 hours maximum test time).
- Two types of bake ovens are available with different temperature range capabilities.
  - The lower stress bake oven can be set to run temperature range from 60°C to 115°C.
  - The higher stress bake oven can be set to run temperature range from 100°C to 250°C.
  - The equipment's tolerance is estimated at +/-5°C.
- Experimental budget allows for no more than 20 runs.

The engineer's objective is to determine a feasible test plan, including the stress level settings and the test unit allocation at each stress level, so as to discriminate between the two competing acceleration models.

5.2. Competing Acceleration Models

Based on the past experience with similar fail mechanism, the reliability engineer believes that Weibull distributions would adequately describe Au-Al intermetallic growth life in a

semiconductor package, which implies a smallest extreme value (SEV) distribution for the log-life. That is, if T is assumed to have a Weibull distribution,$\mathrm{T}\sim\mathrm{WEIB}(\alpha,\beta)$, then $\log(t)\sim SEV(\mu,\sigma)$, where $\sigma=\frac{1}{\beta}$ is the scale parameter and $\mu=\log(\alpha)$ is the location parameter. The Weibull CDF and PDF can be written as

$$F(t|\alpha,\beta)=\Phi_{SEV}\left(\frac{\log(t)-\mu}{\sigma}\right)=1-exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{10}$$

$$f(t|\alpha,\beta)=\frac{1}{\sigma t}\phi_{SEV}\left(\frac{\log(t)-\mu}{\sigma}\right)=\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right],\ t>0 \tag{11}$$

In above parameterization, $\beta>0$ is the shape parameter and $\alpha>0$ is the scale parameter as well as the 0.632 quantile.

The Arrhenius life-temperature relationship was expected to describe the acceleration behavior.

$$t(Temp)=A\,.\exp\left(\frac{E_a}{K\times Temp}\right), \tag{12}$$

where,

- $t(Temp)$ is the life characteristic related to temperature.
- A is constants, and (E_a) is the activation energy of the chemical reaction in electron volts.
- $Temp$ is temperature in Kelvin (^{o}C+273.15).
- K is Boltzmann's constant $(8.617385\ E^{-5}\ eV/K)$

However, due to the complexity of the material, engineer would like to consider two possible life-stress relationships, namely, the linear relationship M_1 and the quadratic relationship M_2.

The M_1 model can be expressed in the linearized form by taking the logarithmic of both sides of (12) as

$$\mu_1=\beta_0+\beta_1 x \tag{13}$$

By standardizing the accelerating variable, the above model can be expressed as

$$\mu_1 = \gamma_0 + \gamma_1 \xi\,, \tag{14}$$

where the standardized variables are expressed a $\xi = \frac{(x - x_{low})}{(x_{high} - x_{low})}$, $\xi \in [0, 1]$.

New coefficients are related to previous ones through $\gamma_0 = \beta_0 + \beta_1 x_{low}$ and $\gamma_1 = \beta_1(x_{high} - x_{low})$. Thus, we have $\mu_{1\,low} = \gamma_0$ and $\mu_{1\,high} = \gamma_0 + \gamma_1$.

The M_2 model is an extension of M_1 by adding a quadratic term as

$$\mu_2 = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{15}$$

By standardizing the accelerating variable, the above model (15) can be expressed as

$$\mu_2 = \gamma_0 + \gamma_1 \xi + \gamma_2 \xi^2\,, \tag{16}$$

where $\gamma_0 = \beta_0 + \beta_1 x_{low} + \beta_2 x^2{}_{low}$, $\gamma_1 = \beta_1\left(x_{high} - x_{low}\right)$, and $\gamma_2 = \beta_2(x^2{}_{high} - x^2{}_{low})$. Similarly, we have $\mu_{2\,low} = \gamma_0$ and $\mu_{2\,high} = \gamma_0 + \gamma_1 + \gamma_2$.

For both models, for Type-I censored data (time censoring), the probability of obtaining a censored observation at time t_c is given by

$$\Pr(\mathrm{t} > \mathrm{t}_c) = exp\left[-\left(\frac{t_c}{\alpha}\right)^{\beta}\right], \qquad t_c > 0 \tag{17}$$

5.3. Prior Distributions Elicitation

Engineer assumed an equal weight for both models to begin with. That is, $\pi(M_1) = \pi(M_2) = 0.5$. For model M_1, Equation (14) shows parameter vector θ as$(\gamma_0, \gamma_1, \sigma)^T$, and for model M_2, Equation (16) shows parameter vector θ as$(\gamma_0, \gamma_1, \gamma_2, \sigma)^T$. One would need to specify a prior distribution for each of the parameters or $p_{M1}(\theta)$ and $p_{M2}(\theta)$. We would initially use the parameters in their original units (before transformation) to relate to the engineer's prior knowledge. Standardization is applied once prior distributions in original units have been effectively solicited from engineers.

Given historical learning and previous experience with similar fail mechanism, the reliability engineer believes that appropriate independent prior distributions on the parameters can be specified as follows: for the activation energy, a uniform distribution that gives an equal

likelihood for values that range from 1.0 to 1.05 eV would be appropriate to use. Note that in the case of the quadratic model M_2 this parameter may no longer directly correspond to the activation energy of the chemical reaction. Not much was known about the intercept, and the quadratic coefficient in M_2 so both were given a vague (diffuse) normal distribution with mean of 0.0 and low precision of $1.0E^{-6}$ ($\sigma = 1000\ or\ \sigma^2 = 1E^{+6}$). A positive density support was assumed for the Weibull shape parameter as gamma distribution with shape of 2 and scale of 1.

5.4. Construction of Optimal Design

The optimization algorithm is Monte Carlo simulation-based, in which the optimal design d^* is arrived at by evaluating the design criterion in (9) for each of the candidate designs, and selecting the design that maximizes the design criterion (utility function of interest). The optimization steps are summarized as follows:

1. For a given experimental run budget, N, and the number of stress-factors to study, k, construct a Latin hypercube design ($LHD\ (N, k)$).
2. Over the design grid, for each candidate design d randomly simulate fail data from the joint density $(\theta_i, y_i)_{d,M_i}$ of each of the rival models M_i $(i = 1,2)$.

$$(\theta_i, y_i)_{d,M_i} \sim p_{d,M_i}(\theta, Y) = p(\theta)_{M_i}.p_{d,M_i}(y|\theta) \tag{18}$$

   That is, independently generate random fail data using the competing acceleration models (using (14) for model M_1 and (15) for model M_2). Consider all possible combinations of sample sizes (unit allocation) at each stress factor-level combinations. Computational time can be reduced if units are allocated at increments >1 to each of the stress levels.
3. Simulated experiments (failure times) are compared against a predetermined test duration t_c to determine if a test unit failure time is censored.
4. Calculate the relative prediction performance of each model over the range of its parameters. This is done by using a Gibbs sampler (WinBUGS) to compute posterior predictions of, $\tau_p(x_S)$, the 100 p^{th} quantile of the lifetime distribution at both the higher and lower stress conditions ($S = S_{High}, and\ S = S_{Low}$). A typical reliability interest is when $p = 0.01$, so in the case of models M_1 and M_2, the outcome of this step is the posterior distribution of the predicted quantile values for each model given the same data set; i.e.,

$\hat{\tau}_{0.01,(M2|Y1)}$, $\hat{\tau}_{0.01,(M1|Y1)}$, $\hat{\tau}_{0.01,(M1|Y2)}$ and $\hat{\tau}_{0.01,(M2|Y2)}$ at both the higher and lower stress conditions.

5. Use the Hellinger distance measure, D_H, to calculate pairwise local utilities $(u_{2|1})$ and $(u_{1|2})$ as in (7) and (8), which are reproduced below for convenience:

   For model M_2 conditional on data from model M_1

   $$u_{2|1} = D_{H_{S_{High}}}\left(\hat{\tau}_{0.01,(M2|Y1)}, \hat{\tau}_{0.01,(M1|Y1)}\right) + D_{H_{S_{Low}}}\left(\hat{\tau}_{0.01,(M2|Y1)}, \hat{\tau}_{0.01,(M1|Y1)}\right)$$

   For model M_1 conditional on data from model M_2

   $$u_{1|2} = D_{H_{S_{High}}}\left(\hat{\tau}_{0.01,(M1|Y2)}, \hat{\tau}_{0.01,(M2|Y2)}\right) + D_{H_{S_{Low}}}\left(\hat{\tau}_{0.01,(M1|Y2)}, \hat{\tau}_{0.01,(M2|Y2)}\right)$$

6. Since it is unknown which of the two models is the true data generating model, we combine the Monte Carlo samples of local utilities $u_{2|1}$ and $u_{1|2}$ to obtain the desired total observed utility function $u(d) = u_{2|1} + u_{1|2}$ to be maximized for an optimal design.
7. Approximate the pre-posterior global utility $U(d) = E[u(d)]$ by fitting a smooth surface to the combined Monte Carlo sample generated in Step 6 as a function of selected design.
8. The optimal design d^* is found by maximizing the fitted surface (the maximum pre-posterior Hellinger distance between predictive densities).

A direct application of Monte Carlo simulation to find the optimal design will require very large scale simulations and it will be computationally inefficient due to the large number of iterations needed and the duplication of effort in neglecting valuable information already generated at a nearby design points. Therefore, to reduce computational cost, in Step 7 and Step 8 the non-parametric surface fitting approach, originally proposed by Müller and Parmigiani [34] and Müller [35], is used for finding optimal designs.

5.5. Results for Discriminating Linear vs. Quadratic ALT models

Table 1 lists the temperature stress ranges that were used in the planning of the ALT experiment. The surface fitting smoothing approach for finding optimal design requires the simulation of experiments (d_i, θ_i, y_i) on a design grid. Full grid of the three temperature ranges can be used in the simulation. However, we instead use a modified Latin Hypercube design to replace the full grid and reduce computational cost at no loss of coverage and to allow available

experimental budget. Table 2 shows the design grid created using a modified Latin Hypercube design ($mLHD$) for the available budget of 20 experimental runs.

Table 1. Temperature Stress Range used in Experiment

Bake Stress	Temperature Range in ^{o}C (Oven tolerance ± 5 ^{o}C)	
	Lower	Upper
T_{High}	150	180
T_{Middle}	115	145
T_{Low}	80	110

Table 2. *mLHD* Grid with 12 Runs and 8 Corner Augmentations

Run #	Low Temp Oven Setup	High Temp Oven Setup		Run Source
	Temp°C (low)	Temp°C (Mid)	Temp°C (High)	
1	85	140	155	mLHD
2	80	125	150	mLHD
3	90	115	170	mLHD
4	100	120	160	mLHD
5	95	145	165	mLHD
6	100	130	180	mLHD
7	110	135	160	mLHD
8	95	130	165	mLHD
9	110	120	175	mLHD
10	90	125	150	mLHD
11	105	135	175	mLHD
12	80	140	170	mLHD
13	80	115	150	AUG-C1
14	110	145	180	AUG-C2
15	110	115	150	AUG-C3
16	110	115	180	AUG-C4
17	80	115	180	AUG-C5
18	80	145	150	AUG-C6
19	80	145	180	AUG-C7
20	110	145	150	AUG-C8

Following the simulation optimization algorithm steps, the optimal design under criterion (8) for discriminating between linear and quadratic acceleration models in single accelerating variable (temperature) is summarized in Figure 3.

Figure 3 displays the pre-posterior expected value of the utility function ($Log[U(d)]$) as a function of the stress magnitude and percent unit allocation to each of the three stress levels used in planning of the experiment. The utility function is maximized when

- The higher temperature level is set at $180°\mathrm{C}$ with an approximated unit allocation of 12%.
- The middle temperature level is set at $130°\mathrm{C}$ with an approximated unit allocation of 55%.
- The lower temperature level is set at $100°\mathrm{C}$ with an approximated unit allocation of 33%.

Insert Figure 3 here

5.6. Some Remarks on the Optimal Model-Discrimination Test Plan

The Bayesian model-discrimination test plan is compared to some conventional test plans and the Meeker and Hahn's 4:2:1 compromise ALT plan. Although the primary objective of some of these plans (model estimation accuracy) is quite different than ours (model discrimination), pointing out similarities and dissimilarities between them is of an added value in our judgment. Nelson [28] pointed out that "a good plan should be multi-purpose and robust and provide accurate estimates." Assumptions used in the stress setup and unit allocation for each plan are:

- The same prior distributions are given to same parameters across all models.
- All plans use three levels of stress (temperature) in the range of $(150°\mathrm{C} - 180°\mathrm{C})$ for high temp, $(115°\mathrm{C} - 145°\mathrm{C})$ for middle temp, and $(80°\mathrm{C} - 110°\mathrm{C})$ for low temperature stress. All plans share the same fixed experimental budget (sample size).
- Stress setup and unit allocation are determined as follows
    - **Model-discrimination plan**: *unequally* spaced test levels with *unequal* allocation that puts more units at the middle of the test range. Optimal design setup used: highest temp of $(180°\mathrm{C})$ with 12% allocation, intermediate temp of $(130°\mathrm{C})$ with 55%

allocation, and lower temperature of $(100°C)$, slightly above the intermediate value in the low temp range, with 33% allocation.

- **Good compromise plan**: *equally* spaced test levels with *unequal* allocation that puts more units at the extremes of the test range and fewer in the middle. We've used 50% at lower level, 30% at higher level, and remaining 20% at the middle level. For the equal spacing of stress levels $180°C$ was selected as highest possible, $110°C$ as lowest, and $145°C$ as the intermediate stress.
- **Best traditional plan**: *equally* spaced test levels with *equal* allocation. Typically, the highest possible stress needs to be selected, which is $180°C$. The lowest test stress is selected to minimize std. error of ML estimate of log mean life at design stress, which is $110°C$. The intermediate stress at an equal space is then $145°C$. Equal allocation puts approximately 33.33% of units at each stress level.
- **Meeker and Hahn's 4:2:1 compromise plan**: Borrowing from the best traditional plan, $180°C$, $145°C$ and $110°C$ are set as the high, intermediate and low test stress levels, respectively. Allocation of samples follows $\frac{4}{7}$ *or* (57%) to low stress, $\frac{2}{7}$ *or* (29%) to middle stress, and $\frac{1}{7}$ *or* (14%) to high stress.

Some remarks on the obtained optimal model-discrimination test plan are:

1. The test plan allocates the larger proportion of units to the intermediate stress level (~55%). This is favorable for test robustness and for generating more failure observations. This plan will be most sensitive for detecting nonlinearity of the relationship (minimize variance of the estimate of the quadratic coefficient).
2. The test plan allocates more test units to the lower stress level (~33%) than to the higher stress level (~12%). This is favorable for more accurate extrapolation with respect to stress, as suggested by optimum plans.
3. The test plan sets the high temperature value to the highest possible in its allowable range, this is known to be a good practice when interest lies in minimizing the standard error of the estimate of any percentile at the design stress (a very common objective of ALTs).

4. The test plan does not set the lower temperature value to the lowest possible in its allowable range as suggested by the optimum test plan (effective if the design stress is close to the test range), but rather it chooses an intermediate value. One drawback to having to test at the lowest extreme of the test range is the longer test time needed for units to fail.

5.7. Recovery Rates of Different Test Plans

Optimal model-discrimination designs are expected to maximize the proportion of times in which the true, data-generating, model is selected under an appropriate model selection criterion. We have chosen to use the DIC model selection rule as explained in Section 4. Other methods of model selection such as BF (Bayes Factor) or BIC (Bayesian Information Criterion) could have also used. The following definitions are used in creating the plan comparison, as shown in Figure 4.

- True Model: The acceleration model from which fail data were generated; that is, Equation (14) for the linear model, and Equation (16) for the quadratic model.
- Assumed Model: The actual acceleration model fitted to the simulated data.
- % Recovery Rate: The fraction of times the true model recovered (correctly identified) under the DIC-based model selection.

Insert Figure 4 here

Figure 4 clearly illustrates the superior recovery rate of the proposed model-discrimination plan. As the sample size increases, this recovery rate converges to 90%, while the recovery rates of other plans are still below 50%. As noted previously, the primary objectives of these test plans could be different from model-form discrimination; therefore, the apparent superiority of our test plan w.r.t. to the recovery rate under DIC should come as no surprise. By the comparison across different plans, it is demonstrated that the proposed methodology is effective in recommending stress setup and unit allocation for model discrimination. In general, model-discrimination plan tends to allocate a higher percentage of units to the middle stress level, which is intuitively appealing for detecting curvatures in acceleration models and for generating more failures.

6. CONCLUSIONS

This paper presents a simulation-based Bayesian approach to the accelerated life test planning with the objective of differentiating competing acceleration models. It is different from the previous research of Bayesian methods for ALT planning, which is concerned with the model parameter uncertainty only. We propose a design criterion that is based on the Hellinger distance measure between the predictive distributions of a life percentile of interest under different acceleration models. Therefore, when facing the model form uncertainty, the experimental results from this type of test plan can better assist the experimenter in choosing the right model. Our approach was applied to a real-world application, where there was uncertainty as to whether the relationship between log mean life and the stress variable is linear or exhibits some curvature. Both the stress-factor setup and the unit allocation at three stress levels were optimized and the obtained optimal test plan was validated by its recovery rate of correct model using simulated data. Comparing to other conventional test plans, such as the three stress-level good compromise plan, the best traditional plan and the well-known 4:2:1 compromise ALT test plan, our test plan has the advantage of substantially increasing the desirable ability of distinguishing among competing model forms, thus provides better guidance as to which model is appropriate for the follow-on product testing.

The main limitation of the proposed approach is the intensive computation required for the point-wise evaluation of utility function. This has been eased by the use of a modified Latin Hypercube sampling scheme, followed by the application of curve-fitting optimization approach. The simulation-based Bayesian approach described in this paper could be extended to model-discrimination ALT planning problems with more than one accelerating variable and more complicated acceleration models.

ACKNOWLEDGEMENT

This research is partially funded by NSF CMMI program, grant no. 0928746.

REFERENCES

[1] W. Nelson, and T. J. Kielpinski, Theory for optimum censored accelerated life tests for normal and lognormal life distributions, *Technometrics* 1976*,* 18 (1): 105-114.

[2] W. Nelson, and W. Q. Meeker, Theory for optimum accelerated censored life tests for Weibull and extreme value distributions, *Technometrics* 1978, 20 (2): 171-177.

[3] W. Nelson, A bibliography of accelerated test plans, *IEEE Transactions on Reliability* 2005, 54 (2): 194-196.

[4] W. Nelson, A bibliography of accelerated test plans, part II – references, *IEEE Transactions on Reliability* 2005, 54 (3): 370-373.

[5] X. Xu, Robust prediction and extrapolation designs for censored data, *Journal of Statistical Planning and Inference* 2009, 139: 486-502.

[6] J. M. McGree, J. A. Eccleston, Investigating design for survival models, *Metrika* 2010, 72: 295-311.

[7] E. M. Monroe, R. Pan, C. M. Anderson-Cook, D. C. Montgomery, C. M. Borror, Sensitivity analysis of optimal designs for accelerated life testing, *Journal of Quality Technology* 2010, 42: 121-135.

[8] T. Yang, R. Pan, A novel approach to optimal accelerated life test planning with interval censoring, *IEEE Transactions on Reliability* 2013, 62 (2): 527-536.

[9] M. Konstantinou, S. Biedermann, A. Kimber, Optimal designs for two-parameter nonlinear models with application to survival models, *Statistica Sinica* 2014, 24: 415-428.

[10] F. Haghighi, Optimal design of accelerated life tests for an extension of the exponential distribution, *Reliability Engineering and System Safety* 2014, 131: 251-256.

[11] W. Q. Meeker, L. A. Escobar, A Review of Recent Research and Current Issues in Accelerated Testing, *International Statistical Review* 1993, 61 (1): 147-168.

[12] W. Q. Meeker, G. Sarakakis, A. Gerokostopoulos, More pitfalls of accelerated tests, Journal *of Quality Technology* 2013, 45 (3): 213-222.

[13] W. G. Hunter, A. M. Reiner, Designs for discriminating between two rival models, *Technometrics* 1965, 7: 307-323.

[14] G. E. P. Box, W. J. Hill, Discrimination among mechanistic models, *Technometrics* 1967, 9: 57-71.

[15] W. J. Hill, W. G. Hunter, D. W. Wichern, A joint design criterion for the dual problem of model discrimination and parameter estimation, *Technometrics* 1968, 10: 145-160.

[16] A. C. Atkinson, and D. R. Cox, Planning experiments for discriminating between models, *Journal of Royal Statistics Society Series B* 1974, 36: 321-348.

[17] P. D. H. Hill, A review of experimental design procedures for regression model discrimination, *Technometrics* 1978, 20 (1): 15-21.

[18] P. D. Leon, A. C. Atkinson, Optimum experimental design for discriminating between two rival models in the presence of prior information, *Biometrika* 1991, 78: 601-608.

[19] A. C. Atkinson, A. N. Donev, R. Tobias, *Optimum Experimental Designs with SAS*, Oxford, England: Oxford University Press 2007.

[20] H. Dette, S. Titoff, Optimal discrimination designs, The Annals of Statistics 2009, 37 (4): 2056-2082.

[21] V. Agboto, W. Li, and C. Nachtsheim, Screening designs for model discrimination, *Journal of Statistical Planning and Inference* 2010, 140 (3): 766-780.

[22] R. D. Meyer, D. S. Steinberg, G. E. P. Box, Follow-up designs to resolve the confounding in multifactor experiments, *Technometrics* 1996, 38: 303-313.

[23] D. R. Bingham, and H. A. Chipman, Optimal designs for model selection, *Technical Report 388*, University of Michigan and University of Waterloo 2002.

[24] K. Chaloner, I. Verdinelli, Bayesian experimental design: a review. *Statistical Science* 1995, 10: 273-304.

[25] F. G. Pascual, Accelerated life test plans robust to misspecification of the stress-life relationship, *Technometrics* 2006, 48 (1): 11-25.

[26] Y. Zhang, W. Q. Meeker, Bayesian methods for planning accelerated life tests, *Technometrics* 2006, 48 (1): 49-60.

[27] T. Yuan, X. Liu, W. Kuo, Planning simple step-stress accelerated life tests using Bayesian methods, *IEEE Transactions on Reliability* 2012, 61 (1): 254-263.

[28] W. Nelson, *Accelerated Testing*, John Wiley & Sons: New York, 1990.

[29] E. Deza, M. M. Deza, *Dictionary of Distances*, Elsevier 2006.

[30] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd ed. Wiley 2001.

[31] S.-H. Cha, and S. N. Srihari, On measuring the distance between two histograms, *The Journal of the Pattern Recognition Society* 2002, 35: 1355-1370.

[32] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. van der Linde, Bayesian measures of model complexity and fit, *Journal of Royal Statistics Society Series B* 2002, 64: 583-639.

[33] G. Claeskens, N. L. Hjort, *Model Selection and Model Averaging*, Cambridge 2008.

[34] P. Müller, G. Parmigiani, Optimal design via curve fitting of Monte Carlo experiments, *Journal of the American Statistical Association* 1995, 90: 1322–1330.

[35] P. Müller, Simulation-based optimal design, in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, 2000, 459–474.

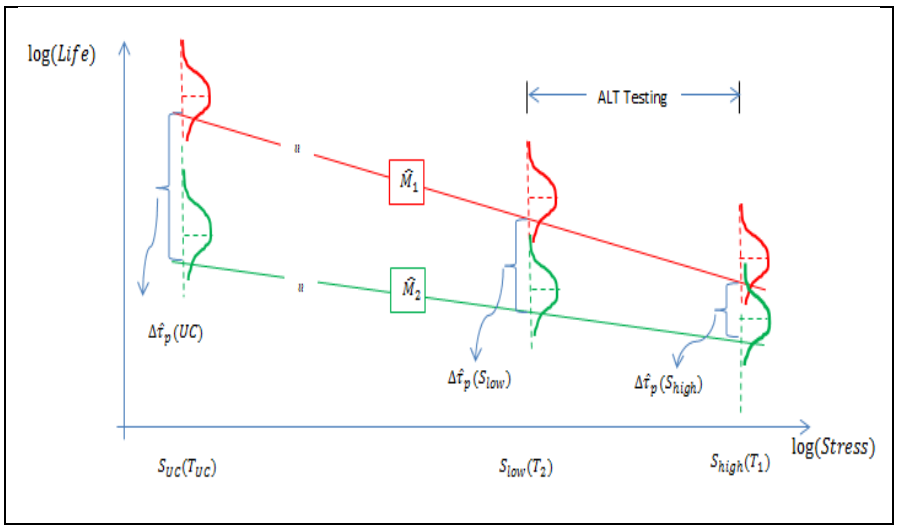


Figure 1. $\widehat{M}_1$ versus $\widehat{M}_2$ at UCs: importance of identifying correct model

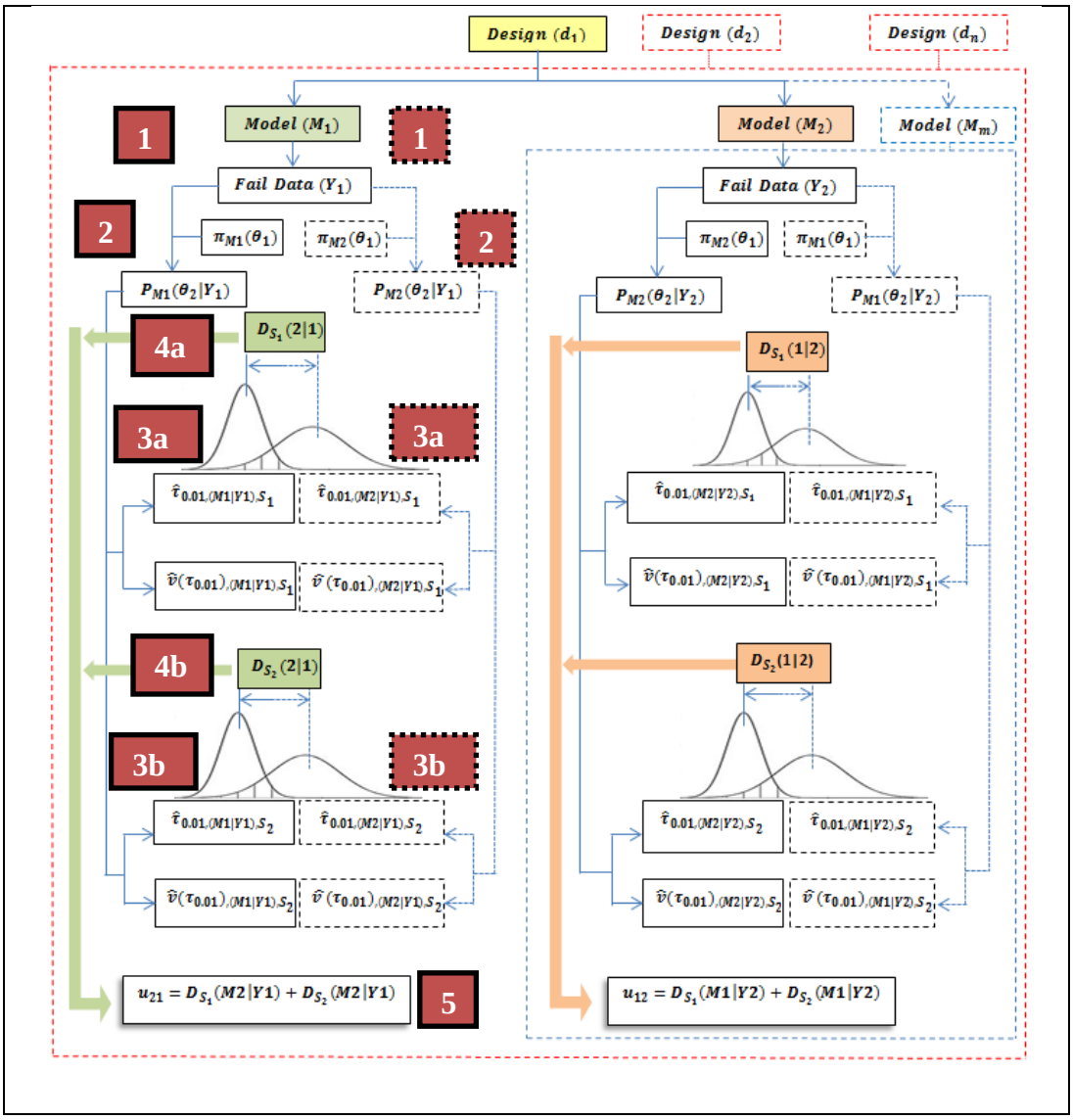
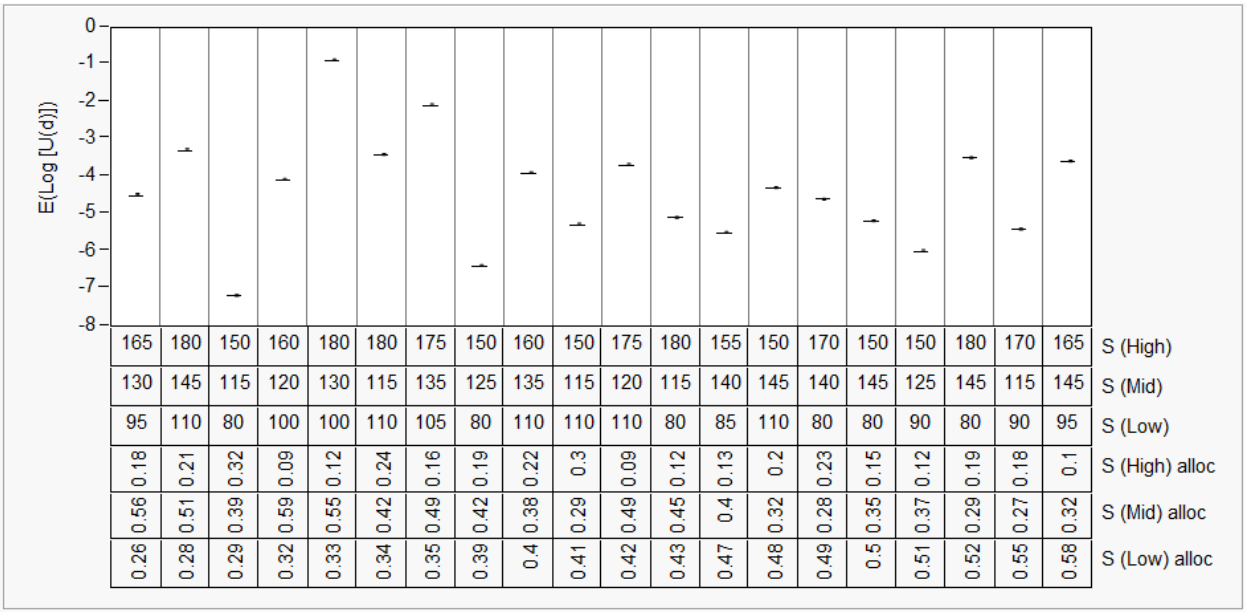


Figure 2. High level methodology flow chart



S (High)	165	180	150	160	180	180	175	150	160	150	175	180	155	150	170	150	150	180	170	165
S (Mid)	130	145	115	120	130	115	135	125	135	115	120	115	140	145	140	145	125	145	115	145
S (Low)	95	110	80	100	100	110	105	80	110	110	110	80	85	110	80	80	90	80	90	95
S (High) alloc	0.18	0.21	0.32	0.09	0.12	0.24	0.16	0.19	0.22	0.3	0.09	0.12	0.13	0.2	0.23	0.15	0.12	0.19	0.18	0.1
S (Mid) alloc	0.56	0.51	0.39	0.59	0.55	0.42	0.49	0.42	0.38	0.29	0.49	0.45	0.4	0.32	0.28	0.35	0.37	0.29	0.27	0.32
S (Low) alloc	0.26	0.28	0.29	0.32	0.33	0.34	0.35	0.39	0.4	0.41	0.42	0.43	0.47	0.48	0.49	0.5	0.51	0.52	0.55	0.58

Figure 3. The pre-posterior expected Log(U) as a function of stress condition and unit allocation

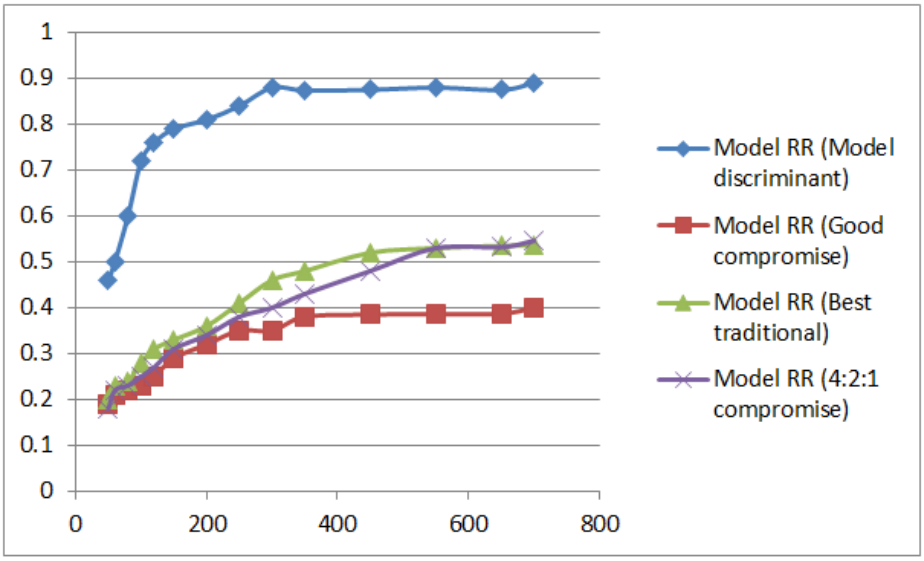


Figure 4. Plots of recovery rate versus sample size for different test plans