Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 180 (2021) 608–617

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry 4.0 and Smart Manufacturing

Early life reliability growth testing with non-constant failure intensity

Nikolaus Haselgruber[a,*], Shawn P. Capser[b], Giorgio I. Vignati[c]

[a]CIS consulting in industrial statistics GmbH, Windgasse 15, A-4170 Haslach, Austria
[b]ESi Inc., 1174 Oak Valley Drive, Ann Arbor, Michigan 48108, USA
[c]ITI s.r.l., Via Michele Novaro 16, 20161 Milano, Italy

Abstract

Early-life reliability is a key characteristic for complex technical products as it addresses the very first time-span of product use. Any unreliability in this phase jeopardizes both customer satisfaction, as well as warranty costs and finally the economic success. Review of several early-life warranty cases shows a mix of classical reliability issues, i.e., problems occurring during usage time, and quality related issues that often manifest themselves during initial inspections by product quality or even the customer, i.e., at usage time $t = 0$. Classical reliability growth models do not explicitly consider these two different sources of unreliability but require zero failures at time zero and provide time-dependent reliability estimates. We present a model which shares the concept of time-dependent reliability measures but considers in addition the initial quality issues. By taking advantage of well-known reliability growth models and basic statistical principles we were able to provide a flexible approach useful for practical application.

Keywords: Reliability growth; early life; warranty; repairable objects; failure rate; product quality; product performance

1. Introduction and motivation

The initial period of the product life cycle is often subject to manufacturing issues that manifest into quality related failures within the early time of the warranty period. To address these issues, continuous improvement strategies target the high impact failure rates that may result from a Pareto of warranty claim data. Unfortunately, there is a considerable lag time from product manufacturing to the time when it experiences its first failure and to getting the claim data from the repair to be recorded in a warranty system. So, continuous improvement activities can be quite slow in demonstrating any reliability improvement. To reduce the time associated with such continuous improvement efforts, this paper proposes the application of reliability growth (RG) methods for early life (EL) reliability improve-

* Corresponding author. Tel.: +43 664 827 44 10.
E-mail address: nh@cis-on.com

10.1016/j.procs.2021.01.283

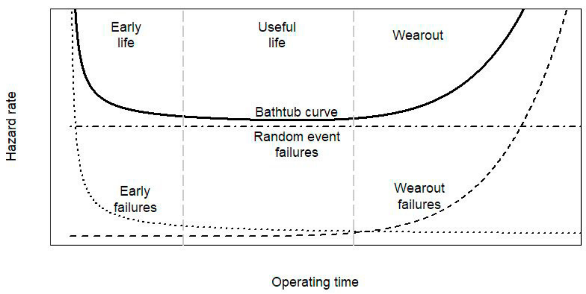

Fig. 1. *Bathtub curve* describing a typical product life cycle.

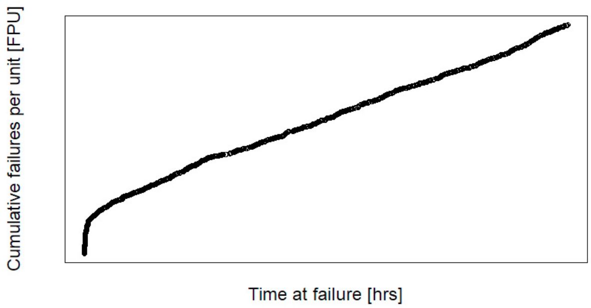

Fig. 2. Cumulative failures per unit versus operating time.

ment of complex, repairable systems, such as automobiles, agricultural equipment and power-generation systems. EL reliability performance is often described as the far-left portion of the *bathtub curve* as indicated in figure 1. Compare, e.g., [1] for further details. It is the period of a decreasing failure (hazard) rate and is typically associated with manufacturing quality issues. For completeness, the typical product life cycle contains two additional periods, one being a constant failure rate performance that is often associated with random-event failures and the far-right portion of the curve is associated with wear-out performance, where the failure rate increases with time. It is the decreasing failure rate performance of the EL period that presents an issue with conventional RG models. A central requirement for a successful and reasonable EL RG application is explainability and interpretability of the model parameters. This is important for the planning phase, where no actual data is available. Furthermore, this directly indicates the magnitude of reliability growth achieved once data is available. Thus, from an artificial intelligence point of view (c.f., [2], [3]), a statistical inferential model is preferred over a predictive black box approach.

1.1. *State of the art in reliability growth modeling*

Conventional RG models, such as the Duane ([4]), Crow ([5], [6]), Crow-AMSAA ([7]), and the PM2 ([8]), are built on the underlying premise that, as long as no improvement measure is implemented, the failure rate of the system is constant with time and the rate of occurrence of new and unique failure modes is also constant with time. In other words, these models are specific to the constant failure rate portion of the bathtub curve. The non-constant, decreasing failure rate of the EL period is an immediate violation of these assumptions and is what drove the need for developing an RG model to manage this specific reliability behavior. The need for such a model provide process engineers, quality engineers, and continuous improvement managers to develop RG test plans and to track the continuous improvement efforts of a product that is naturally in a period of decreasing failure rate, i.e. *reliability growth.*

1.2. *Early life behavior from a warranty perspective*

The proposed application of RG to EL requires prior product warranty analysis to identify the EL and its transition to the useful life (UL) period as well as the start of the wear-out period. The warranty period, where a company focusing for the economical implications, usually spans over both the EL and the first part of the UL. Despite the fact that warranty data can vary a lot based on the market segment, the company being an OEM or a supplier, the presence, or not, of a direct link between the company and the final customer of the product, we can observe the key elements that help describing the way the product performs and support the RG modeling for EL.

Figure 2 shows the cumulative failures per unit versus operating time. The derivative of this curve is the left part of the bathtub curve that shows the reduction of the failure rate in the EL period. The figure shows the transition from the EL period to the UL, i.e., with rather constant failure rate. The data basis should be at least an annual production lot to minimize possible impacts due to seasonal behavior and to production variations.

The close observation of the initial part of the warranty period shows two distinct phenomena:

1. Problems observed at time zero or extremely short operating time, typically reported by the dealers or by the product distribution network, that we call *initial quality* issues. These problems are found during initial inspections before the delivery to the final customer and are related to functional, assembly and aesthetic issues.
2. Problems observed and reported soon after the final customer started using the product. We call them *initial performance.* They depend on the product configuration, its use and the specific mission.

The warranty system needs to provide specific attributes to distinguish between quality and performance issues.

2. Early life reliability growth modeling

We consider different failure modes as root causes for the unreliability. They are the relevant level to define improvement measurements as a reaction of observed failures. It will be assumed that each observed, reliability-relevant failure can be uniquely assigned to a root cause failure mode. The EL RG activity is organized in test phases $1, \dots, I$, where the objects are not changed during a test phase. The phases are separated by analytical sections where the data of the previous phases are analyzed, improvement measures derived and implemented for the subsequent test phases.

2.1. A reliability growth model

With t as operation test time, it is proposed to model the product failure rate in phase i, $i = 1, \dots, I$ as

$$\lambda_i(t) = \sum_{k=1}^{K_{Qi}} \lambda_{Qik} + \sum_{k=1}^{K_{Pi}} \lambda_{Pik}(t)t. \tag{1}$$

In accordance to [7] we call $\lambda_i(t)$ the *demonstrated* failure intensity. It is composed of the time-independent rate λ_{Qik} and $\lambda_{Pik}(t)$, the cumulative performance rate of failure mode k. $\lambda_i(t)$ and λ_{Qik} are given in failures per unit (FPU), while $\lambda_{Pik}(t)$ is expressed in failures per unit and time. Failure modes that would not be solved directly after phase i but are expected to occur again in the next phase are referred to as *A-modes* (c.f., [5] and section 3.2 below). The remaining failure modes go through the problem resolving process and are assumed to be reduced or even eliminated in the subsequent phases with probabilities q_{Qik} and q_{Pik}, sometimes called *fixed effectiveness factors* (FEF). We denote $a_{Qik} = 1$ if failure mode $k, k = 1, \dots, K_{Qi}$, is an A-mode and $a_{Qik} = 0$ otherwise. Thus, in phase i, we have K_{Qi} unique initial quality and in analogy K_{Pi} unique initial performance failure modes. The *projected* failure rate, i.e., the expected rate after implementation of the defined improvement measures is

$$\lambda_{i^+}(t) = \sum_{k=1}^{K_{Qi}} p_{Qik}\lambda_{Qik} + \epsilon_{Qi} + \sum_{k=1}^{K_{Pi}} p_{Pik}\lambda_{Pik}(t)t + \epsilon_{Pi} \tag{2}$$

where $p_{Qik} = (a_{Qik} + (1 - q_{Qik})(1 - a_{Qik}))$, $p_{Pik} = (a_{Pik} + (1 - q_{Pik})(1 - a_{Pik}))$ and ϵ_{Qi} as well as ϵ_{Pi} account for additionally unseen failures (see also [5]). The projected failure rate $\lambda_{i^+}(t)$ serves as the starting point of phase $i + 1$. The number of required phases I depends on the discrepancy between actual reliability and target, problem solving complexity and further economic facts. In equation 1, we assume Poisson distributed failure counts $N_{Qi} = \sum_{k=1}^{K_{Qi}} N_{Qik} \sim Po(\lambda_{Qi})$ with $\lambda_{Qi} = \sum_k \lambda_{Qik}$ and $N_{Pi}(t) = \sum_k N_{Pik}(t) \sim Po(\lambda_{Pi}(t)t)$ with $\lambda_{Pi}(t) = \phi_i t^{\beta_i - 1}$ such that $N_i = (N_{Qi} + N_{Pi}(t)) \sim Po(\lambda_i(t)t)$ for independent quality and performance failures (see [9] for sums of independent Poisson counts). The assumption of independence is strong in both directions, between initial quality and performance as well as among failure modes. However, empirical investigations of warranty data confirmed that in our application, the initial quality and performance can be treated as independent random variables. In case of failure sequences, we only assign the root cause failure to the corresponding mode which justifies also the assumption of independent counts among failure modes. Note that $\lambda_{Pi}(t)$ is the *cumulative* failure rate over $[0;t]$ which is relevant for the assessment of the EL period for $t = t_{EL}$ rather than the *instantaneous* failure rate $\beta_i\lambda_{Pi}(t)$ which measures the reliability at the age of t.

2.2. Parameter estimation and reliability projection

To generate representative data for the RG analysis, a random sample m_i of the entire population, i.e., the available production, is selected. Under the assumptions stated above, maximum likelihood estimation (MLE) provides

$$\hat{\lambda}_{Qi} = \sum_{k=1}^{\hat{K}_{Qi}} \hat{\lambda}_{Qik} = \frac{\sum_{k=1}^{\hat{K}_{Qi}} \sum_{j=1}^{m_i} n_{Qijk}}{m_i} \tag{3}$$

as initial quality failure rate in phase i, with n_{Qijk} observed initial quality failures of sample j assigned to failure mode k among $\hat{K}_{Qi}$ observed unique failure modes. The cumulative failure rate estimation for initial performance is

$$\hat{\lambda}_{Pi}(t) = \sum_{k=1}^{\hat{K}_{Pi}} \hat{\phi}_{ik} t^{\hat{\beta}_i - 1} = \hat{\phi}_i t^{\hat{\beta}_i - 1}. \tag{4}$$

In this case, available data are time-censored as every single test time is t_{TDi}. To estimate (ϕ_i, β_i), consider the maximum likelihood function (cf., e.g., [10]) $L(\phi_i, \beta_i | t_{ijl}, j = 1, \ldots, m_i, l = 1, \ldots, n_{Pij}) = \prod_j \prod_l \phi_i \beta_i t_{ijl}^{\beta_i - 1} \exp(-\phi_i t_{TDi}^{\beta_i})$ with n_{Pij} observed performance failures in test unit ij and t_{ijl} as cumulative unit test time to event. Setting $\partial(\log L)/\partial\phi_i = 0$, provides the MLE $\hat{\phi}_i = n_{Pi}/(m_i * t_{TDi}^{\hat{\beta}})$ and $\partial(\log L)/\partial\beta_i = 0$ leads to $\hat{\beta}_i = n_{Pi}/(\sum_j \sum_l \log(t_{ijl}/t_{TDi}))$ with $n_{Pi} = \sum_j n_{Pij}$. Thus, the MLE for the rate is $\hat{\lambda}_{Pi}(t) = n_{Pi}/(m_i * t_{TDi})$. In cases where not every tested unit may show a failure, we prefer to apply times to event accumulated over the entire phase, denoted as $t_{il}, l = 1, \ldots, n_{Pi}$. This leads to $L(\phi_i, \beta_i | t_{il}, l = 1, \ldots, n_{Pi}) = \prod_l \phi_i \beta_i t_{il}^{\beta_i - 1} \exp(-\phi_i T_{TDi}^{\beta_i})$ with $T_{TDi} = m_i * t_{TDi}$ as total planned test time in phase i. MLE provides $\hat{\phi}_i = n_{Pi}/(T_{TDi}^{\hat{\beta}})$ and $\hat{\beta}_i = n_{Pi}/(\sum_l \log(t_{il}/T_{TDi}))$. The latter should be bias-corrected by multiplication of $(n_{Pi} - 1)/n_{Pi}$. The estimation of β_i may be difficult due to short test time and low observed failure counts. A poor estimate of the shape can have heavy consequences for the extrapolation of the results from test domain to the EL period. To avoid this, alternatively the shape parameter can be pre-determined as $\tilde{\beta}$, e.g., from historical warranty data, and kept constant for all phases. Then, replace $\hat{\beta}_i$ by $\tilde{\beta}$ and have a conditional maximum likelihood function $L(\phi_i | \tilde{\beta}, t_{il}, l = 1, \ldots, n_{Pi}) = \prod_l \phi_i \tilde{\beta} t_{il}^{\tilde{\beta} - 1} \exp(-\phi_i T_{TDi}^{\tilde{\beta}})$. In this case we reach $\hat{\phi}_i = n_{Pi}/(T_{TDi}^{\tilde{\beta}})$. Note that the MLE does not take the actual times to event into account as $\tilde{\beta}$ has been fixed. Subsequently, the shape is denoted as $\hat{\beta}_i$ which has to be replaced by $\tilde{\beta}$ if the pre-determined shape is applied. Referring to (4), $\hat{\phi}_i$ can be decomposed into failure mode portions as $\hat{\phi}_i = \sum_k^{\hat{K}_{Pi}} \hat{\phi}_{ik}$ with $\hat{\phi}_{ik} = \sum_j n_{Pijk}/\vartheta_i$ where $\vartheta_i = (m_i * t_{TDi}^{\hat{\beta}_i})$ or $\vartheta_i = t_{TDi}^{\hat{\beta}_i}$, dependent on the chosen likelihood function. Thus, the demonstrated failure rate in phase i referring to the test domain is estimated as

$$\hat{\lambda}_i(t_{TDi}) = \hat{\lambda}_{Qi} + \hat{\lambda}_{Pi}(t_{TDi}) t_{TDi}. \tag{5}$$

The projection of $\hat{\lambda}_i(t_{TDi})$ to the EL period is according to equation (2). For more details, see [5]. Beside the unsolved and the expected portion of unsuccessful improvements, in addition a bias correction to account for new, so far unseen failures is taken into account. This requires the effectiveness factors to be known. We estimate for the initial quality $\hat{\epsilon}_{Qi} = \bar{q}_{Qi}(\hat{K}_{Qi}/m_i)\hat{\beta}_{Qi}$ with average FEF $\bar{q}_{Qi} = (1/\hat{K}_{Qi}) \sum_k q_{Qik}$ and $\hat{\beta}_{Qi} = (\hat{K}_{Qi} - 1)/\sum_k(\log(m_i/m_{ik}))$ as estimated shape of a non-homogeneous Poisson process with count function $\phi_{Qi} m^{\beta_{Qi}}$, i.e., on unit basis. m_{ik} is the number of inspected units until the failure mode k has been initially observed. Thus, we estimate the projected quality failure rate (FPU) as

$$\hat{\lambda}_{Qi^+} = \sum_{k=1}^{\hat{K}_{Qi}} p_{Qik}\hat{\lambda}_{Qik} + \hat{\epsilon}_{Qi} \tag{6}$$

For the initial performance we apply $\hat{\epsilon}_{Pi} = \bar{q}_{Pi}(\hat{K}_{Pi}/t_{TDi})\hat{\beta}_{Pi}$. Here, $\hat{\beta}_{Pi} = (\hat{K}_{Pi} - 1)/\sum_k(\log(t_{TDi}/\min_j t_{ijk}))$ is the estimated shape of a non-homogeneous Poisson process with count function $\phi_{Pi}t^{\beta_{Pi}}$, i.e., on time basis, and $\min_j t_{ijk}$ is the first cumulative observation time of failure mode k in phase i. For details on the parameter estimation, we refer again to [5] and [10]. The projected performance rate of phase i for the test domain is

$$\hat{\lambda}_{Pi^+}(t_{TD}) = \sum_{k=1}^{\hat{K}_{Pi}} p_{Pik}\hat{\phi}_{ik}t_{TDi}^{\hat{\beta}_i - 1} + \hat{\epsilon}_{Pi}. \tag{7}$$

The extrapolation to the EL period results in the projected performance failure rate as

$$\hat{\lambda}_{Pi^+}(t_{EL}) = \hat{\lambda}_{Pi^+}(t_{TD}) * (t_{EL}/t_{TDi})^{\hat{\beta}_i}. \tag{8}$$

This provides the estimated projected failure rate (FPU), referring to the EL period, as

$$\hat{\lambda}_{i^+}(t_{EL}) = \hat{\lambda}_{Qi^+} + \hat{\lambda}_{Pi^+}(t_{EL})t_{EL}. \tag{9}$$

The Poisson assumption of the failure counts leads to $\hat{\lambda}_i(t_{EL}) \sim \Gamma(N_i(t_{EL}); 1)$, with $N_i(t_{EL}) = N_{Qi} + N_{Pi}(t_{EL})$ where N_{Qi} is estimated by $\hat{\lambda}_{Qi}$ and $N_{Pi}(t_{EL})$ by $\hat{\lambda}_{Pi}(t_{EL})t_{EL}$. There are several alternatives to identify the probability distribution of $\hat{\lambda}_{i^+}(t_{EL})$. A reasonable way would be stochastic simulation. Alternatively, we can interpret the scaling factors in expressions (6) and (7), i.e., p_{Qik} and p_{Pik} as probabilities of Bernoulli variates to describe the likelihood that failure mode k appears again after introduction of the improvement measures. Under consideration of a Poisson filter (cf., e.g., [9]) on the failure counts N_{Qik} and N_{Pik} the Poisson distribution holds also for the failure counts after projection, i.e., N_{Qi^+k} and $N_{Pi^+k}(t_{EL})$. Thus, the Gamma distribution would even hold for the projection case, if the failure counts of the bias correction, $N_{\epsilon_{Q}i}$ and $N_{\epsilon_{P}i}$ would be Poisson counts independent from N_{Qi} and $N_{Pi}(t_{EL})$, too. However, since the one data basis for estimation is a subset of the other, they are obviously not fully independent. Stochastic simulation showed that the violation of this requirement has negligible consequences. Thus, we assume

$$\hat{\lambda}_{i^+}(t_{EL}) \dot{\sim} \Gamma(N_{i^+}(t_{EL}); 1), \tag{10}$$

with $N_{i^+}(t_{EL}) = N_{Qi^+} + N_{Pi^+}(t_{EL})$, where N_{Qi^+} is estimated by $\hat{\lambda}_{Qi^+}$ and $N_{Pi^+}(t_{EL})$ by $\hat{\lambda}_{Pi^+}(t_{EL})t_{EL}$. Consequently, we estimate a $100 * C\%$-confidence interval for $\lambda_{i^+}(t_{EL})$ due to time-censoring approximately by

$$[\underline{\lambda}_{i^+}; \overline{\lambda}_{i^+}] = [\Gamma^{-1}((1-C)/2; \hat{\lambda}_{Qi^+} + \hat{\lambda}_{Pi^+}(t_{EL})t_{EL}; 1); \Gamma^{-1}(1-(1-C)/2; \hat{\lambda}_{Qi^+} + \hat{\lambda}_{Pi^+}(t_{EL})t_{EL} + 1; 1)], \tag{11}$$

where $\Gamma^{-1}(C; a; b)$ is the C-quantile of the Gamma distribution with shape parameter a and scale parameter b. Expression (11) can be simply applied to the demonstrated failure rate if the corresponding estimates are replaced. The growth between two subsequent phases can be seen statistically significant if the corresponding confidence intervals are non-overlapping. Depending on the number of phases and the required growth statement, in case of multiple comparisons the confidence level per interval should be adapted correspondingly.

3. An early life reliability growth planning and tracking example

We consider a technical product which consists of up to 37 subsystems. Some of them are mandatory, i.e., required in every product unit. The remaining subsystems are optional. This causes some variability in the population with respect to the product configuration. Field data analysis shows that the EL reliability of the current product violates the cumulative failure rate target of $\theta(t_{EL}) = 2$ FPU by approximately 60%. To bring this in line with the target, the RG activity should be carried out in two phases:

- Phase 1 for determination of initial quality and performance, classification of failures, decision on improvement measures and their implementation
- Phase 2 for proving the effectiveness of the implemented measures and final EL reliability demonstration

The EL period is $t_{ED} = 1000h$ (cf., section 1), the test domain should not be longer than $t_{TD} = 500h$ in both phases. We denote an assumption for planning, e.g., derived from long-term experience, as $\tilde{\square}$ and, as usual, an estimation based on observed data as $\hat{\square}$. The current product has on average $\tilde{\lambda}_Q = 0.65$ initial quality failures per unit and a cumulative initial performance failure $\tilde{\lambda}_P(t_{EL}) = 0.0025$ failures/h. $\tilde{\lambda}_P(t_{EL})$ is not constant over time but shows $\tilde{\beta} = 0.8$ which is assumed to be identical for both phases. The plan should both allow to demonstrate the target in phase 2 with confidence of $C = 90\%$ and confirm a significant RG. An improvement by a factor $q = 2$ is aspired between the two phases to be able to demonstrate the required target with reasonable effort.

3.1. Determination of optimal test duration and selection of test units for EL RG planning

We set up an optimization routine to find the total test times for the phases $T_{TDi},\ i = 1, 2$ in order to fulfill the stated requirements. The number of units to be tested in phase i is $m_i = T_{TDi}/t_{TD}$ and the number of total expected initial quality failures is $\tilde{n}_{Qi} = m_i * \tilde{N}_{Qi}$ with $\tilde{N}_{Qi} = \tilde{\lambda}_Q * q^{-(i-1)}$. The number of expected performance failures in phase i is $\tilde{n}_{Pi}(t_{TD}) = m_i * \tilde{N}_{Pi}(t_{TD})$ with $\tilde{N}_{Pi}(t_{TD}) = \tilde{\phi}_{Pi} * t_{TD}^{\tilde{\beta}}$, where $\tilde{\phi}_{Pi} = \tilde{\lambda}_P(t_{EL}) * q^{-(i-1)} * t_{EL}^{-\tilde{\beta}}$. To extrapolate the failure behavior to the early life period, it is $\tilde{n}_{Pi}(t_{EL}) = m_i * \tilde{N}_{Pi}(t_{EL})$ with $\tilde{N}_{Pi}(t_{EL}) = \tilde{\phi}_{Pi} * t_{EL}^{\tilde{\beta}}$. Thus, it is $\lambda_i(t_{EL})t_{EL} \sim \Gamma(\tilde{N}_{Qi} + \tilde{N}_{Pi}(t_{EL}); 1)$. Under consideration of time-censored data the reliability target is demonstrated if $\overline{\lambda}_2(t_{EL}) = \Gamma^{-1}(C; \tilde{N}_{Q2} + \tilde{N}_{P2}(t_{EL}) + 1; 1) \leq \theta(t_{EL})$. To show a significant reliability increase over the phases, we request non-overlapping confidence intervals. For adequate multiple comparison we compare the lower $1 - \sqrt{C}$-confidence bound of phase 1, $\underline{\underline{\lambda}}_1(t_{EL})$, with the upper $\sqrt{C}$-confidence bound of phase 2, $\overline{\overline{\lambda}}_2(t_{EL})$. The final optimization criterion should be $\max(0, \overline{\overline{\lambda}}_2(t_{EL}) - \underline{\underline{\lambda}}_1(t_{EL})) + \max(0, \underline{\lambda}_2(t_{EL}) - \theta(t_{EL}))$. The criterion has to be minimized with minimum total test duration per phase (T_{TD1}, T_{TD2}). The optimization uses the `R` function `optim` with method `L-BFGS-B` ([11]). A valid solution suggests $5056h$ in phase 1 and $10064h$ in phase 2 requiring $m_1 = 11$ and $m_2 = 21$ test units in phases 1 and 2, respectively. Figure 3 shows the result where the solid step function indicates the expected failure rate per phase, where the gray fields show the confidence intervals with adapted confidence level and the dotted horizontal line represents the EL reliability target to be demonstrated with phase 2. The black triangle marks the upper 90%-confidence bound in phase 2 which is relevant for target demonstration, i.e., without adapted confidence level. Note that the rate in phase 1 is the expected demonstrated failure rate while that of phase 2 can be seen as the projected failure rate of phase 1 which should be demonstrated with the tests of phase 2.

As table 1 summarizes, we expect approximately 7 initial quality failures in each phase as well as 15 to 16 initial performance failures. The projection of the performance to the EL period results in approximately 27 failures per phase, given that the stated assumptions are correct. All $E(n.)$ are in absolute failure counts, $E(N.)$ in FPU. Furthermore, the m_i test samples have to be selected from the available production lot in order to cover as much variability in

Table 1. Expected number of failures and rates according to the early life reliability test plan

Expected failures	Phase 1 ($m_1 = 11$)	Phase 2 ($m_2 = 21$)
$E(n_Q)$	7.15	6.83
$E(N_Q)$	0.65	0.33
$E(n_P(t_{TD}))$	15.79	15.08
$E(N_P(t_{TD}))$	1.44	0.72
$E(n_P(t_{EL}))$	27.50	26.25
$E(N_P(t_{EL}))$	2.50	1.25
$E(N(t_{EL}))$	3.15	1.60

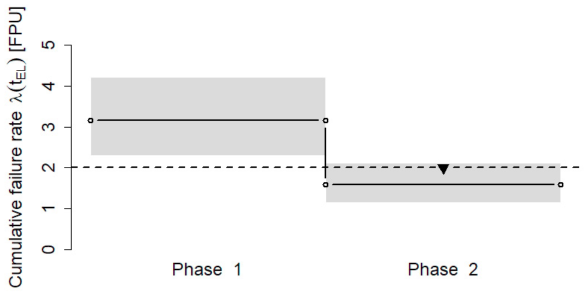

Fig. 3. Optimized EL RG test plan for two phases with non-overlapping confidence intervals and a target-conform upper bound in phase 2.

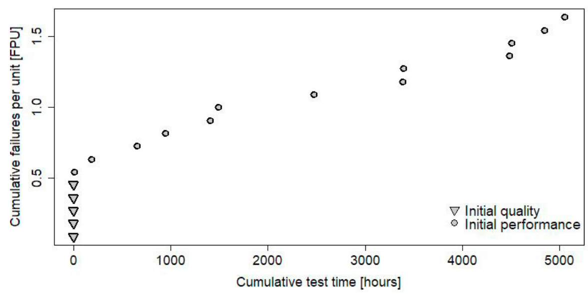

Fig. 4. Cumulative FPU versus cumulative test time showing the initial quality and initial performance behaviors of phase 1.

configuration, i.e., as much subsystems as possible and being representative for the population. A second optimization routine provided representative samples which contain all available subsystems.

3.2. Data collection, management and analysis for EL RG tracking of phase 1

The RG tracking subsumes the execution of the test plan, including data collection and management, failure assessment, derivation of improvement measures and their effectiveness as well as actual parameter estimation to calculate the demonstrated and the projected failure rate per phase. Here the 11 units were tested in the field by real-world customers under various anticipated usage conditions. Each unit was operated for a total of $t_{TD1} = 500h$ hours. Thus, the anticipated cumulative test time for phase 1 is $T_{TD1} = 5500h$. The test protocol is such that it includes separate opportunities for observing product failures:

1. Initial inspection: These time-independent failures are deemed as *initial quality* and tracked on a per unit basis.
2. Testing period: Conducted as a physical test of the complete product, failures are tracked based on cumulative time and deemed as *initial performance.*
3. Final inspection: At the conclusion of the physical test a final inspection is performed on the test units.

The results of the tests performed on the given units are provided in table 2 which contains the tracking data from both the initial inspection and the performance testing. The minimum information necessary for tracking the test units include the date and assignment of the failure, the operating time on the unit when the failure occurred, and the classification for the failure mode type, i.e. initial quality or initial performance.

Crow ([7]) presented the Crow-AMSAA model *Test-Find-Test.* It is used to track and analyze RG tests, where failures are observed but any corrective actions are delayed until after the test phase is complete. This model permits classification of failure modes into categories, *A-modes* or *BD-modes.* A-modes are failures that are not going to be addressed as part of a problem resolution process, while BD-modes are specific to those failure modes which will be

Table 2. Failures observed in phase 1 testing showing classification of failure modes as *initial quality* or *initial performance*.

Failure date	Unit id	Mode id	Unit test time to event [h]	Cumulative time to event [h]	Failure type	Classification
Mar-01-2019	19	35	0	0	Q	BD
Mar-16-2019	19	29	127.9	1407.1	P	BD
Mar-19-2019	71	3	0	0	Q	A
Mar-28-2019	51	4	0.6	6.8	P	BD
Mar-28-2019	51	20	0	0	Q	BD
Apr-06-2019	68	35	0	0	Q	BD
Apr-08-2019	68	21	16.6	182.5	P	BD
Apr-15-2019	17	35	0	0	Q	BD
May-10-2019	69	16	135.2	1487.6	P	A
May-19-2019	80	7	59.2	651.2	P	BD
May-22-2019	51	35	440.7	4848.1	P	BD
May-22-2019	80	29	86.1	947.4	P	BD
Jun-10-2019	30	15	308.5	3393.0	P	BD
Jun-14-2019	69	21	410.0	4510.5	P	BD
Jun-19-2019	80	35	308.2	3390.0	P	BD
Jun-22-2019	30	7	407.9	4486.7	P	BD
Jun-27-2019	75	17	224.9	2473.5	P	BD
Jul-08-2019	80	21	459.3	5052.2	P	BD

investigated as part of a closed-loop problem resolution process with the intent of introducing the validated solution at the end of the test phase, but prior to the start of the next phase. Observed failures would be investigated by a *failure review board* which provides the classification per failure mode as shown in the last column of table 2. Finally, five failures associated with initial quality have been observed and assigned to three different modes (id 3, 20, 3 times 35) with classification (A, BD, BD). The cumulative number of units that were inspected before each of the BD-failure modes were first observed, was 4 and 1, respectively. The FEFs for both BD modes were assumed to be 90%. Similarly, 13 failures associated with the initial performance were observed and assigned to 8 unique failure modes (id 4, 2 x 7, 15, 16, 17, 3 x 21, 2 x 29 and 2 x 35). All modes except id 16 were classified as BD modes with FEFs (id 4: 75%, 7: 99%, 15: 95%, 17: 99%, 21: 75%, 29: 75% and 35: 95%).

3.3. Parameter estimation and results for phase 1

Prior to any statistical analysis a graphical view of the observed data should be made. Figure 4 is a plot of the cumulative failures per unit (FPU) versus the cumulative test time of all units (in hours). The figure shows a vertical step in the FPU at $t = 0$ which is a result of the initial quality inspections. Immediately following the curve shows a fairly straight-line behavior for the initial performance during the actual testing of the $m_1 = 11$ units. This figure represents the demonstrated performance of the system, prior to the introduction of any corrective actions. The summary statistics for the BD failure modes are summarized in table 3 where $\hat{\beta}_{BD}$ corresponds to $\hat{\beta}_{Qi}$ and $\hat{\beta}_{Pi}$, respectively.

Table 3. Summary statistics for BD failure modes.

	BD failures observed	$\hat{\beta}_{BD}$	$\bar{q}$
Initial quality	4	0.29	0.90
Initial performance	12	0.41	0.88

From the observed failure counts and the respective FEF of each failure mode the individual failure per unit statistics can be calculated for each BD mode, demonstrated and projected. Figure 5 compares the two BD failure modes associated with initial quality and figure 6 is in analogy for the initial performance. The plots show that each failure rate was reduced significantly through the corrective action process.

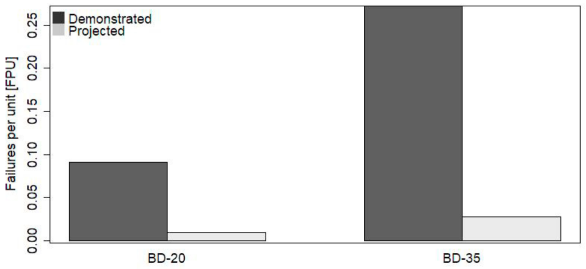

Fig. 5. Demonstrated and projected initial quality failure rate for BD modes in phase 1.

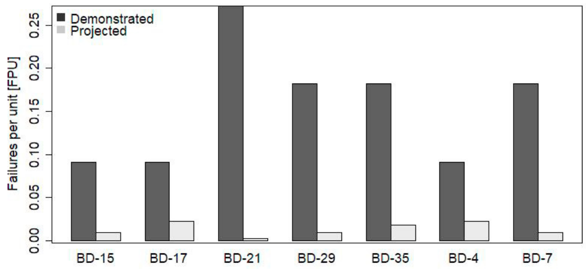

Fig. 6. Demonstrated and projected initial performance failure rate for BD modes in phase 1.

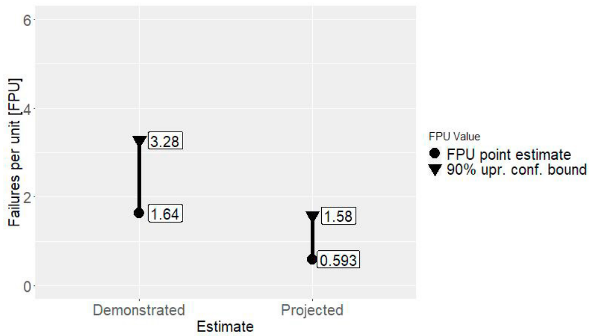

Fig. 7. Product FPU for phase 1 EL RG testing as derived for the test domain.

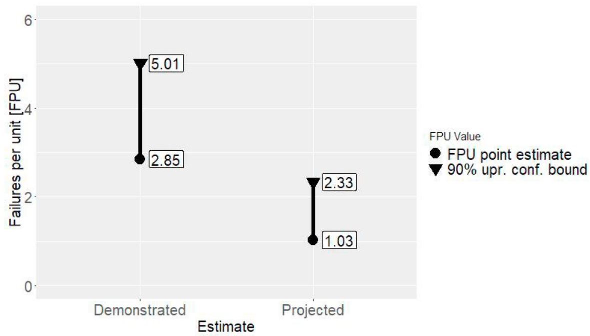

Fig. 8. Product FPU for phase 1 EL RG testing as derived for the warranty domain.

It is the accumulation of these improvements from demonstrated to projected rate that provides the RG for the EL performance. Table 4 summarizes the demonstrated and projected failure per unit statistics for the complete product, separated for initial quality and initial performance.

Table 4. Demonstrated and projected failure per unit (FPU) after 5500h of testing in phase 1.

	Demonstrated FPU	Projected FPU
Initial quality $\hat{\lambda}_{Q1(+)}$	0.4545	0.1753
Initial performance $\hat{\lambda}_{P1(+)}(500h)$	1.1818	0.4176
EL reliability $\hat{\lambda}_{1(+)}(500h)$	1.6363	0.5929

The initial quality and initial performance statistics can be pooled to make an estimate of the overall system EL FPU performance. In addition, statistical confidence limits are included to capture the uncertainty around the point estimates. For the example presented, the overall system FPU performance, demonstrated and projected, is given in figure 7. As described by equation (8) above, the results obtained for the test domain have to be translated into the warranty domain, i.e., the EL period. The estimates presented in figure 7 were therefore adjusted for the warranty domain and are presented in figure 8. The non-overlapping confidence intervals show that a statistically significant impact was made on the product failure rate (at 90% confidence). Clearly, this demonstration is based on the projected impact of the corrective actions and their respective estimates in their FEFs.

3.4. Step to the next growth phase

The statistically significant improvement shown in figure 8 provides an expectation for the test units at the start of phase 2. The reader should be reminded that having the improvement measures introduced into phase 2, does not mean that the failure mode cannot reappear in phase 2. The corrective actions are often introduced with a FEF that is less than 1, meaning that there still is the opportunity for the failure to occur. It should also be noted that the selection of phase 2 units proceeds in the exact same manner as for phase 1. An investigation has been made to the potential build configurations of the population from which phase 2 units were selected. After the introduction of the improvement measures derived from phase 1, the second phase of testing was executed with the previously determined total of $m_2 = 21$ units. The test results in a total of 17 failures which were shared across 11 unique failure modes. The split of these 17 failures were such that 5 were initial quality and 12 were related to initial performance. This permitted a final demonstration of the product FPU to be $\overline{\lambda}_2(t_{EL}) = 1.875$ with 90% statistical confidence. Since $\overline{\lambda}_2(t_{EL}) < \theta(t_{EL}) = 2$ FPU, the aspired RG could be demonstrated and the improvement program was successful. The entire computational effort behind this work has been done with the statistical programming language `R`, version 4.0.2. For details, see [11]. Some of the graphs have been produced with `ggplot2`, as described in [12].

4. Summary

The model presented in this paper allows for separate consideration of initial quality and initial performance failures which ensures some increased flexibility over standard RG models. Due to its generic formulation it can be applied for different product types and variable number of test phases. The results express the root causes of the unreliability clear and allow to assess the actual effectiveness of defined improvement measures. The RG model is based on standard probabilistic concepts such as the Poisson distribution. Hence, parameter estimation procedures are simple to execute, i.e. also efficient to implement. At some points, assumptions and approximations, such as identical test units, identical shape parameters in all phases, unbiased effectiveness factors and independence of specific random variables, reduce the model complexity. Their validity should be checked in every application case and kept in mind when interpreting the results. The challenges associated with the practical application is largely due to the management of the test program and the problem resolution process. Key aspects for a successful implementation are the management of the test units according to build date and availability of configurations as well as interpretation of documented failure data, in particular if tests are executed by customers. Furthermore, deviations of actual test time accumulation from plan as well as the timing of the problem resolution process are issues to be handled.

References

[1] Ebeling, C.E. (1997) "Reliability and Maintainability Engineering." *McGraw-Hill*.
[2] Rudin, C. (2019) "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." arXiv:1811.10154. *Nature Machine Intelligence, Vol 1, May 2019, 206-215*.
[3] Emmert-Streib, F., O. Yli-Harja and M. Dehmer (2020) "Explainable artificial intelligence and machine learning: A reality rooted perspective." DOI: 10.1002/widm.1368 *WIREs Data Mining and Knowledge Discovery, Wiley*.
[4] Duane, J.T. (1964) "Learning Curve Approach to Reliability Monitoring." *IEEE Transactions on Aerospace*, Vol. 2, No. 2.
[5] Crow, L.H. (1983) "Reliability Growth From Delayed Fixes." *1983 Proceedings Annual Reliability and Maintainability Symposium*.
[6] Crow, L.H. (2011) "Planning a Reliability Growth Program Utilizing Historical Data." *2011 Reliability and Maintainability Symposium*.
[7] Crow, L.H. (2004) "An Extended Reliability Growth Model for Managing and Assessing Corrective Actions." *2004 Proceedings Annual Reliability and Maintainability Symposium*.
[8] MIL-HDBK-189C (2011) "Department of Defense Handbook Reliability Growth Management." *U.S. Army Materiel System Analysis Activity (AMSAA)*.
[9] Davenport, W.B. (1970) "Probability and Random Processes." *McGraw-Hill*.
[10] Rigdon, S.E. and A.P. Basu (2000) "Statistical Methods for the Reliability of Repairable Systems." *Wiley Series in Probability and Statistics*.
[11] R Core Team (2020). "R: A language and environment for statistical computing." *R Foundation for Statistical Computing, Vienna, Austria*. Available online at https://www.R-project.org/.
[12] Wickham, H. (2016) "ggplot2: Elegant Graphics for Data Analysis." *Springer-Verlag New York*.