How Much Testing is Needed to Manage Supportability Risks for Beyond-LEO Missions?

Andrew C. Owens[1]
NASA Langley Research Center, Hampton, VA 23681, USA

and

Olivier L. de Weck[2]
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Supportability will be a significantly greater driver of cost and risk for future deep-space crewed missions than it has been in the past. Spares requirements and maintenance risk mitigation in particular present an unprecedented challenge for missions beyond Low Earth Orbit (LEO), since, for the first time in human spaceflight history, crews will be weeks or months away from resupply or a safe return to Earth in the event of an abort. Under these conditions, failure rates are a critical parameter that must be well-understood in order to manage logistics and risk effectively. However, failure rates cannot be measured directly, and can only be estimated based on past experience and test results. Previous research has shown that International Space Station (ISS) operational experience has provided significant benefits to future missions by reducing uncertainty and improving accuracy in failure rate estimates, resulting in significant reductions in mass and risk for beyond-LEO missions. This paper updates and expands on that research and quantifies the potential value of continued testing for future mission supportability. Frequentist and Bayesian models for evaluating, validating, and updating failure rate estimates are described, and are combined with supportability models to examine potential impacts of additional operating experience for future missions in terms of logistics mass reduction. The implications of these results for technology development, system design, and program planning are discussed along with lessons learned and recommendations for future system development. In the end, there is no simple answer to the question of how much testing is required, but the models described in this paper provide a way to evaluate the potential impacts of testing in order to inform test planning.

Nomenclature

α	=	Gamma distribution shape parameter
β	=	Gamma distribution scale parameter
γ	=	Level of statistical significance
ε	=	Error factor
κ	=	K-factor
λ	=	Failure rate (deterministic value)
$\hat{\lambda}$	=	Observed average failure rate
$\bar{\lambda}$	=	Mean failure rate
Λ	=	Failure rate (random variable)
Λ_L^γ	=	Lower $100(1-\gamma)$% failure rate confidence bound
Λ_U^γ	=	Upper $100(1-\gamma)$% failure rate confidence bound
τ	=	Mission endurance
$\chi_a^2(b)$	=	Upper a percentage point of a chi-square distribution with b degrees of freedom

[1] Aerospace Engineer, Space Mission Analysis Branch, MS 462
[2] Professor of Aeronautics and Astronautics and Engineering Systems, Department of Aeronautics and Astronautics, Building 33-410

d	=	Duty cycle
t_o	=	Total accumulated operational time (a.k.a. test time)
$MTBF_L^\gamma$	=	Lower $100(1-\gamma)$% Mean Time Between Failures confidence bound
n_o	=	Total number of failures observed during a test period
q	=	Quantity

AES	Advanced Exploration Systems
CCAA	Common Cabin Air Assembly
CDF	Cumulative Distribution Function
CDRA	Carbon Dioxide Removal Assembly
CFR	Constant Failure Rate
DoD	Department of Defense
ECLSS	Environmental Control and Life Support Systems
FCPA	Fluids Control and Pump Assembly
ISS	International Space Station
LEO	Low Earth Orbit
LoC	Loss of Crew
MADS	Maintenance and Analysis Data Set
MCMC	Markov Chain Monte Carlo
MTBF	Mean Time Between Failures
NRC	National Research Council
OGS	Oxygen Generation System
ORU	Orbital Replacement Unit
PDF	Probability Density Function
POS	Probability of Sufficiency
R&R	Remove and Replace
TCCS	Trace Contaminant Control System
UPA	Urine Processor Assembly
WPA	Water Processor Assembly

I. Introduction

TESTING and operational experience are critical activities for characterizing system supportability in order to inform risk assessment, logistics demands, and overall mission planning, especially for future crewed missions beyond Low Earth Orbit (LEO). Orbital Replacement Unit (ORU) failure rates are key system characteristics, but unlike other parameters such as mass, length, or width they cannot be measured directly and must instead be estimated based on analogy to past experience or evaluation of test results. As a result, failure rates are inherently uncertain, and this uncertainty has significant impacts on supportability risk and the amount of spares required to mitigate it.[1,2] Failure rate uncertainty can be reduced through testing, and previous analysis has shown that improved failure rate estimates resulting from operational experience can provide very significant benefits to future exploration missions. Even considering only a subset of International Space Station (ISS) Environmental Control and Life Support Systems (ECLSS) ORUs, ISS operational experience up to 2018 has enabled significant reductions in the amount of spares mass required for a future Mars mission. In addition, this experience has identified and enabled correction of underestimated failure rates associated with several ORUs. Had these underestimates not been identified, supportability assessments based on those failure rate estimates would have significantly underestimated risk, and therefore underestimated the amount of spares required to mitigate risk.[3] Overall, these results highlight the critical importance of real-world testing and operational experience for reducing risk and logistics requirements for future missions.

This paper updates and expands upon that previous research to investigate the potential value of additional testing and provide models and recommendations to inform future system development and test planning. Frequentist and Bayesian models for evaluating, validating, and updating failure rate estimates as a function of test outcomes are presented and described, along with a series of illustrative examples of the potential impacts of testing on ORU failure rate estimates. These models are then combined with a supportability model to evaluate the potential benefits of additional testing in terms of spares mass reduction, which is applied to a notional future Mars mission as an illustrative example. The models described in this paper provide a quantitative method for evaluating the potential impacts of

additional testing or operational experience in terms of logistics mass requirements, and can help inform test planning and system development decisions.

The remainder of this paper is organized as follows. Section II presents background information on supportability and its importance for future missions, including definitions of key terms. Section III examines the impact of testing on ORU failure rate estimates, describing both a frequentist technique for assessing confidence bounds and intervals based on observed failures as well as a Bayesian technique for updating a failure rate uncertainty distribution based on an initial estimate. Each technique is applied to notional test data, which are discussed as an illustrative example of the implications of test time and test outcomes on ORU failure rate estimates. Section IV then expands on these results to examine the impact of updated failure rate estimates on the amount of spares mass required for notional beyond-LEO exploration missions. The potential benefits of additional system testing to reduce uncertainty, measured in terms of logistics mass reduction, are explored under a range of notional test outcomes. Section V discusses these results and other considerations, including potential issues with the use of error factor to parameterize uncertainty, and recommendations/lessons learned from Department of Defense (DoD) experience with reliable system development. Finally, Section VI presents conclusions.

II. Background

A. Supportability

Supportability is the overarching characteristic of a system that drives the amount of resources required to enable safe and effective system operations during a mission, as a function of several specific characteristics related to reliability and maintainability.[4,5] Key characteristics include ORU failure rates, redundancy, commonality, level of maintenance, and on-demand manufacturing capability, for example. During a mission, ORUs may fail randomly, or reach the end of their useful operating life, and must be replaced. Supportability resources are used to return the system to an operational state. These include physical resources such as spares used to recover from random failures, maintenance items used for scheduled replacements, and consumables, as well as temporal resources such as the amount of crew time required for maintenance activities.[6,7]

Since spares demands are driven by random failures, there is no way to know *a priori* how many spares will be required for each ORU. As a result, supportability analysis is fundamentally an analysis of the tradeoff between risk and resources. Specifically, supportability analysis is used to characterize the relationship between the number of spares provided for each ORU, and therefore the total spares mass, and the probability that those spares will be sufficient to cover all maintenance demands during the mission, known as Probability of Sufficiency (POS).[8,9] Spares are allocated to achieve a POS target (i.e. mitigate risk to a desired level) while minimizing mass.

1. Aleatory and Epistemic Uncertainty

Stochastic maintenance demands are driven by two types of uncertainty. Aleatory uncertainty arises from natural, inherent randomness in a process, while epistemic uncertainty comes from a lack of knowledge about the process.[10] For example, aleatory uncertainty relates to the outcome of a coin flip when the probability of heads or tails is known, while epistemic uncertainty relates to uncertainty in the bias of the coin. In the context of supportability analysis, aleatory uncertainty refers to the distribution of the number of times that an item will fail in a given period of time, given a known failure rate. Epistemic uncertainty refers to uncertainty in the value of the failure rate itself. Unlike other system parameters, failure rates cannot be directly measured. Instead, they must be estimated based on past experience and statistical analysis of test results and operational behavior. A significant amount of epistemic uncertainty typically remains in space systems failure rate estimates, and this uncertainty can have very significant impacts on risk. Specifically, analyses that neglect epistemic uncertainty tend to significantly underestimate POS for long-endurance missions, and therefore underestimate the total spares mass required to mitigate risk to acceptable levels.[1,2,7]

Aleatory uncertainty is an inherent characteristic of the system itself, and only changes when the system changes. For example, identification and removal of failure modes during reliability growth efforts can decrease failure rates and shift the distribution of the number of failures that will occur. However, the design changes made to address these failure modes add additional epistemic uncertainty into the system, since they may introduce new, unknown failure modes. Epistemic uncertainty, in contrast, can be reduced by observing and gathering more information about a system without making changes to it. While updating failure rate estimates based on observed behavior without making design modifications has no impact on the actual (unknown) failure rate, more accurate and precise estimates enable more efficient risk coverage, allowing the same POS values to be achieved with fewer spares.

2. Key Definitions

The following list defines and summarizes several key supportability-related terms:

- *Supportability*: set of system characteristics related to reliability and maintainability that drive the amount of resources required to enable safe and effective system operations.[4,5]
- *Reliability*: the probability that an item will perform its intended function for a given period of time, under a given set of operating conditions.[4,11–13]
- *Failure Rate*: the average rate at which an item fails and requires maintenance, which is a parameter used to define the distribution of the number of failures that will occur for a given item and/or the reliability of that item as a function of time. A deterministic failure rate estimate is denoted using λ, while the random variable representing a failure rate uncertainty distribution is denoted using Λ.
- *Mean Time Between Failures (MTBF)*: the average amount of time that an item will operate before failing, which provides an alternate parameterization of the failure distribution. MTBF is the inverse of failure rate (i.e. $MTBF = \lambda^{-1}$), and is a probabilistic description of the failure characteristics of the item, *not* an item lifetime.
- *K-Factor:* a multiplier on failure rates to account for induced failures (i.e. failures driven by external effects, rather than random failures inherent to the item itself).[14,15]
- *Probability of Sufficiency (POS)*: the probability that the number of spares provided for each ORU is sufficient to cover all failures during a given mission.[8,9]
- *Endurance*: the amount of time that a system must operate and support the crew without resupply.[16] Note that endurance is distinct from mission duration, since long-duration missions (e.g. the ISS) may receive regular resupply and therefore have short endurance.
- *Orbital Replacement Unit (ORU)*: a component or assembly that is designed to be removed and replaced (R&R'd) by the crew if needed during on-orbit operations.

B. Supportability Challenges for Beyond-LEO Missions

On missions beyond Low Earth Orbit (LEO), astronauts will be logistically isolated for longer periods of time than ever before, without access to the abort and resupply options that have been used to mitigate risk for the past six decades of human spaceflight. Maintenance resource availability is much more critical under these circumstances, since an inability to repair a system failure could result in Loss of Crew (LoC) when there is no contingency support from Earth. In addition, longer mission endurances mean that more failures are likely to occur during the mission, and there is more uncertainty regarding which items will fail and how many failures will occur. As a result, more spares must be provided to provide the same level of risk mitigation, and maintenance logistics mass requirements will be significantly higher than they have been in LEO. Overall, supportability will be a significantly greater driver of cost and risk for future crewed exploration missions than it has been in the past.[1,4,6,7,17,18]

III. Failure Rate Estimation

Techniques for estimating, evaluating, and updating failure rate estimates based on test results can be generally split into two approaches. In a purely frequentist analysis, the test outcome is used to make statistical observations about the true (unknown) failure rate, for example by defining a confidence interval. The Bayesian approach treats the failure rate as an estimate characterized by an uncertainty distribution (i.e. a random variable) and updates prior estimates using test outcomes. The key data required for either approach are:

- Total accumulated operational time, t_o, also referred to as the test time, which is the total amount of time that the item being tested operates during the test period, accounting for duty cycles and parallel operation of multiple copies.
- Total number of failures observed during the test, n_o.

The following sections describe the methodologies used for frequentist and Bayesian analyses and present illustrative examples. In practice, frequentist methods are used to validate failure rate estimates or generate estimates in the absence of any prior information, while Bayesian methods are used to update and refine failure rate estimates over time. Both methods can also be used to forecast system characteristics as a function of hypothetical test outcomes, as is done here, in order to inform test planning.

Many references describe the following calculations in terms of MTBFs instead of failure rates. However, the methodology presented here works with failure rates instead of MTBFs, since the latter can be non-intuitive and misleading.[7] In some cases MTBFs are reported alongside failure rates to provide context, and MTBFs can be useful

as a reference point for timescales (e.g. when discussing the ratio of test time to MTBF). However, it is important to emphasize that an MTBF is a parameter used to characterize a failure distribution, *not* a statement of the amount of time each ORU will operate before failing. In addition, these techniques assume that the system has a constant failure rate and that no changes occur during the test period. This Constant Failure Rate (CFR) model is a common first-order approach to supportability and reliability modeling.[19]

A. Frequentist Analysis: Statistical Characterization and Validation

Frequentist analysis makes statistical observations about an item's failure rate based only on the outcome of a test. The simplest characterization of this unknown failure rate is the observed average failure rate, $\hat{\lambda}$, which is equal to the number of observed failures divided by the accumulated operational time.[20,21]

$$\hat{\lambda} = \frac{n_o}{t_o} \tag{1}$$

While the observed average failure rate provides a good point estimate of the true failure rate, it provides no information regarding the amount of uncertainty in that estimate. However, test data can also be used to construct confidence bounds and confidence intervals, which enable a much richer understanding of the underlying failure rate.

A lower $100(1-\gamma)$% confidence bound Λ_L^γ is a failure rate value, calculated based on observed test outcomes, such that values of λ above Λ_L^γ are consistent with the observed data at the γ level of significance. In terms of hypothesis testing, a hypothesized failure rates below Λ_L^γ would be rejected as being statistically different from the observed data (i.e. an underestimate) at the γ level of significance. The upper $100(1-\gamma)$% confidence bound Λ_U^γ, similarly, is defined such that values of λ below Λ_U^γ are consistent with the observed data, and hypothesized values above Λ_U^γ would be rejected as being statistically different from the observed data (i.e. an overestimate) at the γ level of significance.[22,23] Put another way, based on the observed evidence,

$$P\left(\lambda \geq \Lambda_L^\gamma\right) = \gamma \tag{2}$$

$$P\left(\lambda \leq \Lambda_U^\gamma\right) = \gamma \tag{3}$$

When time to failure is assumed to follow an exponential distribution (i.e. the CFR model), lower and upper confidence bounds can be calculated using the χ^2 (chi-square) distribution:

$$\Lambda_L^\gamma = \frac{\chi^2_{1-\gamma}(2n_o)}{2t_o} \tag{4}$$

$$\Lambda_U^\gamma = \frac{\chi^2_{\gamma}(2n_o+2)}{2t_o} \tag{5}$$

Here $\chi_a^2(b)$ indicates the upper a percentage point of the χ^2 distribution with b degrees of freedom. Equations 4 and 5 present one-sided confidence bounds, but the same technique can be used to construct a two-sided $100(1-\gamma)$% confidence interval, $\left[\Lambda_L^{0.5\gamma},\ \Lambda_U^{0.5\gamma}\right]$. This interval is constructed such that values of λ within the interval are consistent with the observed results, and hypothesized values outside the interval would be rejected as being statistically different from the observed data at the γ level of significance, i.e.:[20,21]

$$P\left(\Lambda_L^\gamma \leq \lambda \leq \Lambda_U^\gamma\right) = \gamma \tag{6}$$

Figure 1 shows observed average failure rates and 80% confidence intervals as a function of the total accumulated operational time t_o and the number of observed failures during that time n_o. In each case, the solid line indicates the best point estimate for the failure rate, and the data provide 80% confidence that the true failure rate value is within the shaded region between the dotted lines. The upper dotted line corresponds to the upper 90% confidence bound, while the lower dotted line corresponds to the lower 10% confidence bound. Note that for the case in which no failures are observed during the test ($n_o = 0$), the observed average failure rate and lower confidence bound are both equal to

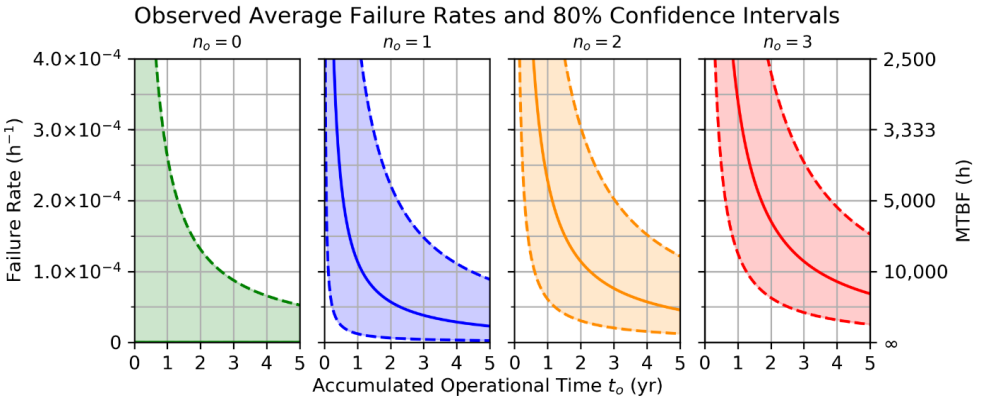

Figure 1: Observed average failure rates (solid line) and 80% confidence intervals (shaded region) as a function of accumulated operational time t_o and number of observed failures n_o. The lower confidence bound and observed average failure rate for the case with no failures (green, far left) are both equal to 0.

0. Accumulated operating times of up to five years are shown, and the y-axis is labeled using both failure rates and MTBFs for context.

As expected, a higher number of failures during a test period results in higher failure rate estimates in terms of both confidence bounds and observed averages. For a given number of observed failures, longer test periods result in lower estimates. The width of the confidence interval also decreases significantly as more test time is accumulated, as expected. However, there are diminishing returns. For example, the upper 90% failure rate confidence bound for an item that has accumulated 1 year of operations with 0 failures is 2.6×10^{-4} h^{-1} (an MTBF of 3,804 h). An additional year of failure-free operations reduces this confidence bound by approximately a factor of two, to 1.3×10^{-4} h^{-1} (an MTBF of 7,609 h). The third failure-free year, however, results in a confidence bound of 8.7×10^{-5} h^{-1} (an MTBF of 11,413 h), a reduction of about 33%.

Testing, reliability evaluation, and mission planning are often more concerned with confirming that an ORU's failure rate is below a given bound than proving a specific value, since lower-than-expected failure rates during a mission are a much more acceptable outcome than higher-than-expected failure rates. The method described above can also be used to examine the amount of accumulated operating time required to demonstrate an upper confidence bound in a given failure rate value, as a function of the desired level of confidence and the number of failures observed during testing. The equation for required operating time is obtained by rearranging equation 5 to obtain

$$t_o = \frac{\chi^2_\gamma(2n_o+2)}{2\Lambda_U^\gamma} \tag{7}$$

One application of this approach would be to examine the failure-free operating time required to demonstrate a desired reliability by setting n_o equal to 0 in equation 7. Any failures during the test period will only reduce the confidence in a given failure rate or increase the failure rate upper bound at the same confidence level. As a result, this is the most optimistic test outcome, and can be used to determine bounds on the amount of test time that would be required to demonstrate a desired failure rate to a desired level of confidence. Curves for several levels of confidence are shown in Figure 2. For example, to empirically demonstrate that the failure rate of an ORU is less than 1×10^{-4} h^{-1} (an MTBF of 10,000 h, or approximately 1.14 years) with 80% confidence, that ORU would need to accumulate 16,094 h (approximately 1.84 years) of operating time without failing.

The upper failure rate confidence bound Λ_U^γ is the inverse of the lower MTBF confidence bound (denoted $MTBF_L^\gamma$), and so test time requirements can also be thought of in terms of the ratio of accumulated operating time to MTBF required to demonstrate a desired level of confidence that the actual MTBF is above that value. The operating time to MTBF ratio at a given level of confidence γ is equal to the product of operating time and Λ_U^γ. Figure 3 shows the operating time to MTBF ratio required to demonstrate a lower MTBF confidence bound as a function of desired confidence level and the number of failures observed during operations n_o. Rearranging equation 7 yields

$$\frac{t_o}{MTBF_L^\gamma} = t_o\Lambda_U^\gamma = \frac{\chi^2_\gamma(2n_o+2)}{2} \tag{8}$$

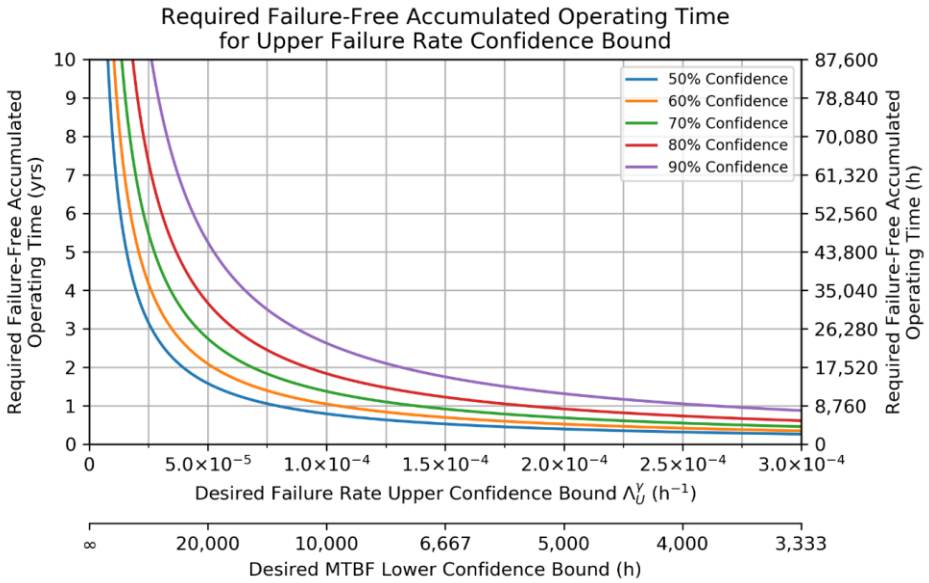

Figure 2: Failure-free accumulated operating time required to demonstrate a specified upper confidence bound on failure rate, Λ_U^γ, as a function of confidence level. A second x-axis shows the associated MTBF for context. The left end of the x-axis indicates higher reliability, while the right end indicates lower reliability.

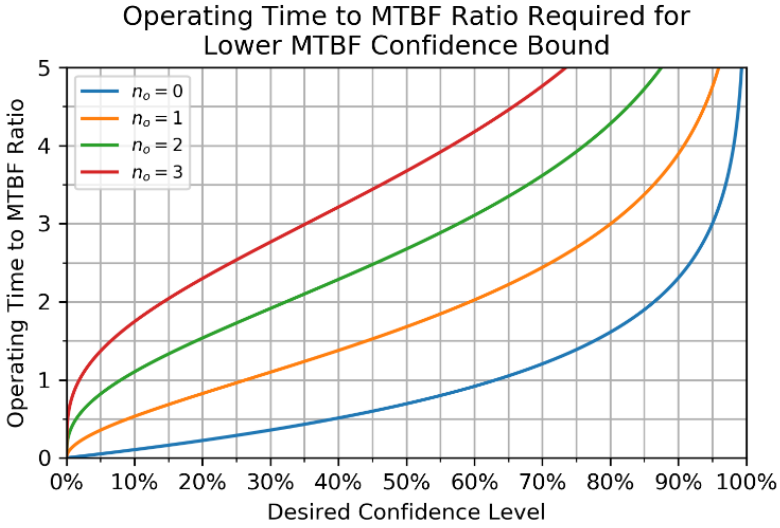

Figure 3: Ratio of accumulated operating time to MTBF required to demonstrate a lower confidence bound on MTBF as a function of the desired confidence level and the number of failures observed during the test n_o.

The failure-free case ($n_o = 0$, shown by the blue line in Figure 3) is often taken as the most optimistic outcome; key values are summarized in Table 1. A failure-free operating time period provides 63% confidence that the true MTBF is longer than that time period; confidence levels of 70%, 80%, and 90% are associated with ratios of 1.20, 1.61, and 2.30, respectively. The amount of operating time required to demonstrate a given MTBF grows exponentially with the level of confidence desired.

As an example, consider a notional ORU with a targeted MTBF of 3 years (26,280 h), which is a failure rate of 3.81×10^{-5} h^{-1}. If a single unit is operated for 3 years (a 1:1 ratio of operating time to MTBF) and does not experience a failure, that test provides 63% confidence that the true MTBF is greater than or equal to 3 years, meaning the failure rate is less than or equal to 3.81×10^{-5} h^{-1}. To obtain 80% confidence that the true MTBF is greater than or equal to that target value, the ORU would need to accumulate failure-free operating time equal to 1.61 times the target MTBF, or 4.83 years. This does not necessarily mean that nearly five years of testing are required to provide that confidence,

Table 1: Key values from the blue curve in Figure 3, relating the ratio of operating time to MTBF to the confidence level provided by that test, assuming no failures are observed.

Operating Time to MTBF Ratio	Confidence in MTBF Lower Bound
0.50	39%
0.92	60%
1.00	63%
1.20	70%
1.50	78%
1.61	80%
2.00	86%
2.30	90%
2.50	92%
3.00	95%

since under the CFR assumption described above that operating time can be accumulated on multiple units at once. With two copies of the ORU testing in parallel, 80% confidence in in the target value could be achieved with 2.42 years of failure-free operations; three copies operating in parallel would enable 80% confidence after 1.61 years of failure-free operations.

These results are not meant to imply that an ORU must achieve these failure-free operating times before it can be deployed. A failure during testing does not mean that the design process must start over. Instead, the equations listed above can be used to update estimates and confidence intervals appropriately. Specifically, when a failure occurs during testing one or several of the following would need to occur:

- More test time with no additional failures would be required to achieve the same failure rate and confidence target,
- The failure rate target can be increased (i.e. MTBF target decreased), and/or
- The confidence target can be decreased.

It is also important to note that this model only characterizes the failure rate / MTBF as it relates to constant, random failures. Failure modes related to wearout or life-limited items would need to be examined using test plans that reach those operating durations, or that use some kind of accelerated life testing to simulate the wearout conditions more quickly. Operating 30 copies of an ORU for $1/10^{th}$ of the target MTBF without a failure would provide 95% confidence in the target MTBF in terms of purely random failures with a constant rate, but it would not provide insights into potential wearout modes that could occur before the target MTBF value. As a result, there is a risk associated with test campaigns that do not involve at least some tests of duration equal to or longer than the target MTBF that must be carefully traded against the cost and schedule impacts of the test campaigns required to reduce it.

In practice, equations 1, 4, and 5 can be used to track the current state of the failure rate estimate over time, both during test campaigns before system deployment and during mission operations. Figure 4 shows an example of what the failure rate estimate and confidence interval might look like over time for a notional ORU. A simulated test was carried out for a single unit over 5 years (43,800 h) of simulated operations. Failures were randomly generated using a simulated true failure rate of 1×10^{-4} h^{-1}, or an MTBF of 10,000 h, indicated by the red horizontal dotted line. The test article was assumed to be replaced immediately with an identical copy each time a failure occurred. The solid black line indicates the observed average failure rate, and the black dotted lines indicate the upper and lower bounds of the 80% confidence interval. Based purely on the observed number of failures at any given point in the test timeline, the solid black line indicates the best point estimate for failure rate, and the data provide 80% confidence that the true failure rate value is somewhere between the two dotted lines. Note that the lower confidence bound and the observed average failure rate are both 0 until a failure occurs. Five failures were observed during the test (i.e. six units were used), as summarized in Table 2.

This notional example shows that the confidence interval and the observed average failure rate converge towards the true value over time, as expected. The observed average failure rate and confidence interval decrease as operational time is accumulated without a failure, but jump sharply upwards when a failure occurs. Over time, the average and confidence interval become more robust to additional failures, and the magnitude of the jump is reduced. At the end of the test period, the observed average failure rate is 1.1×10^{-4} h^{-1}. Even after five years of operations, however, the 80% confidence interval still spans from 5.6×10^{-5} h^{-1} to 2.1×10^{-4} h^{-1}, or an MTBF confidence interval from 4,723 h to 18,005 h – approximately a factor of two in either direction from the true value. These results highlight the fact that significant amounts of uncertainty can remain in failure rate estimates, even after years of operations.

It is important to note that the failure history shown in Figure 4 and described in Table 2 is notional, and represents outputs from a random process. A very different sequence of unit lifetimes could also have occurred, even if the underlying true failure rate remained the same. Observed average failure rates and confidence intervals are estimates developed based on observations from a random process, and as a result they will not necessarily produce the same result if an experiment is repeated. For example, if the notional test described above were repeated, it is possible that the failure count over time would not be the same, and therefore the estimated failure rate and confidence interval would be different. Over time, the estimate would still tend to converge towards the true value, but test planners must

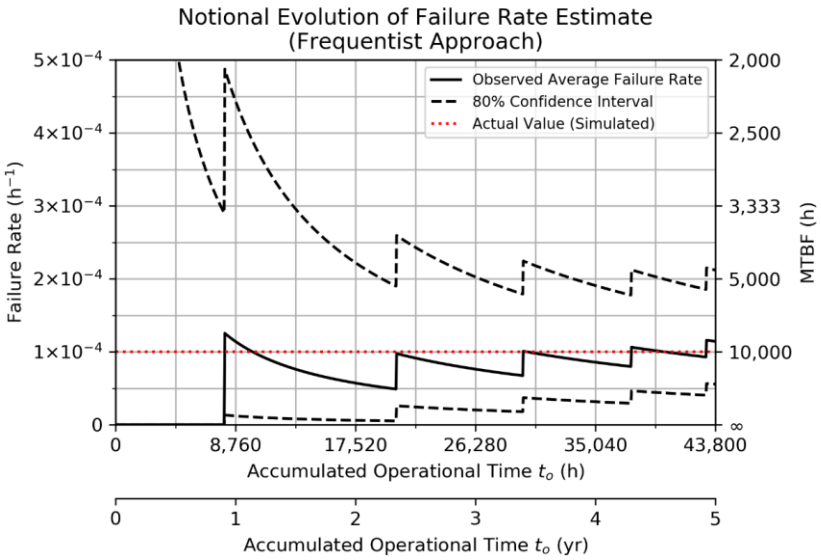

Figure 4: Notional evolution of a frequentist failure rate estimate, showing the observed average failure rate (solid black line) and 80% confidence interval (dotted black lines). Failures, indicated by sharp upward ticks in the failure rate estimate, were randomly generated using a process with an actual failure rate indicated by the horizontal red dotted line.

Table 2: Summary of results from notional simulated test, showing the test time at failure and time since failure for each test unit. Note that the sum time since previous failure and the cumulative test time at failure may not match exactly due to rounding. The 6th test unit was still operational at the end of the test period.

Test Unit	Accumulated Operational Time at Failure (h)	Time Since Previous Failure (h)
1	7,959	7,959
2	20,518	12,559
3	29,750	9,232
4	37,622	7,872
5	43,133	5,510
6	*N/A*	*N/A*

bear in mind that the true failure rate is not known *a priori*, and confidence intervals based on test results are a statistical description of a region that is likely to contain the true failure rate value with some level of confidence, not guaranteed bounds.

Overall, a significant amount of accumulated operational time is necessary to validate high ORU reliabilities (i.e. low failure rates / high MTBFs). As discussed above, this does not necessarily mean that program schedules must incorporate test periods significantly greater than the length of target MTBFs, since the rate at which operational time is accumulated can be accelerated via parallel operation of multiple units. However, these results do highlight the fact that reliability verification and failure rate uncertainty reduction is likely to be a significant driver of program schedule and testing costs if low-failure-rate ORUs are desired with low epistemic uncertainty. Analyses such as the ones presented in this paper can help inform trades between schedule considerations, testing costs, the costs of manufacturing multiple test articles, desired reliabilities and confidence levels, spares mass requirements, risk, and other mission supportability characteristics.

B. Bayesian Analysis: Updating an Initial Estimate

The frequentist analysis described above assumes no prior knowledge or estimates of failure rates. The Bayesian approach, on the other hand, starts with a prior failure rate estimate in the form of a random variable Λ which is updated based on observed system behavior. The probability distribution associated with Λ represents epistemic uncertainty associated with the failure rate value. A lognormal distribution, characterized by a mean failure rate $\bar{\lambda}$ and

error factor ε, is typically used to represent failure rate uncertainty. The error factor, defined as the ratio of the 95th and 50th percentiles of the distribution, is used as an indicator of the level of uncertainty in the estimate,[8,10,24,25] though there are drawbacks to this parameterization, described below. The expected value and variance of a lognormal uncertainty distribution are[26]

$$\mathrm{E}[\Lambda] = \bar{\lambda} \tag{9}$$

$$\mathrm{Var}[\Lambda] = \bar{\lambda}^2 \left(e^{\left(\frac{\ln \varepsilon}{1.645}\right)^2} - 1 \right) \tag{10}$$

Unfortunately, as described further in Section IV.A, use of the lognormal distribution prevents closed-form evaluation of POS for a given spares allocation. However, the gamma distribution can be used to closely approximate the lognormal distribution by matching the expected value and variance with the following shape and scale parameters:[3,6,26]

$$\alpha = \frac{\mathrm{E}[\Lambda]^2}{\mathrm{Var}[\Lambda]} = \left(e^{\left(\frac{\ln \varepsilon}{1.645}\right)^2} - 1 \right)^{-1} \tag{11}$$

$$\beta = \frac{\mathrm{E}[\Lambda]}{\mathrm{Var}[\Lambda]} = \bar{\lambda}^{-1} \left(e^{\left(\frac{\ln \varepsilon}{1.645}\right)^2} - 1 \right)^{-1} \tag{12}$$

Lognormal and gamma distributions can both be used to represent an uncertain failure rate in a Bayesian analysis. However, the lognormal distribution is not a conjugate prior for the parameter of a Poisson process – the model used to represent random failures, described in greater detail in Section IV.A – and therefore Bayesian updates of lognormal failure rate distributions must be performed numerically. Markov Chain Monte Carlo (MCMC) simulation[27] can be used, as was done in previous work,[26] but this approach can be computationally expensive. The gamma distribution, on the other hand, is a conjugate prior for the parameter of a Poisson distribution, and therefore enables simple, closed-form Bayesian updates. Given a gamma-distributed prior failure rate estimate with parameters α_{prior} and β_{prior}, the parameters of the posterior, updated estimate are[25]

$$\alpha_{post} = \alpha_{prior} + n_o \tag{13}$$

$$\beta_{post} = \beta_{prior} + t_o \tag{14}$$

The mean, variance, error factor, and selected percentiles of the posterior distribution can be calculated using these parameters, along with other values of interest.

Figure 5 shows an example of how a Bayesian failure rate estimate might update over time, using the same randomly-generated failure history used for the frequentist example above, described in Table 2. The blue line and shaded region indicate the mean and 80% credible interval of the Bayesian estimate, while the black solid and dotted lines show the observed average failure rate and 80% confidence interval (the same results shown in Figure 4) for context. The Bayesian $100(1-\gamma)$% credible interval is analogous to the frequentist confidence interval, and represents the central $100(1-\gamma)$% region of the posterior Probability Density Function (PDF), between the $\frac{1-\gamma}{2}$ and $1-\frac{1-\gamma}{2}$ quantiles of the posterior distribution.[27] In this case, the prior estimate and observed evidence result in a posterior failure rate uncertainty distribution with 80% of the probability mass within the blue shaded region; that is, based on the failure rate uncertainty distribution, there is a 10% chance that the failure rate is above the blue shaded region, and a 10% chance that it is below that region. As with the frequentist results presented in Figure 4, the Bayesian estimate decreases as time passes without a failure and jumps upwards when a failure occurs, gradually converging on the true value. The estimate also becomes more robust as more data are collected. For this example, the initial mean failure rate estimate was assumed to be accurate (1×10^{-4} h^{-1}) and the initial error factor was set to 4. After five years of testing and five observed failures, the estimated mean failure rate is 1.12×10^{-4} h^{-1} with an error factor of 1.91 – a slight overestimate, but with significantly decreased uncertainty. Put another way, the variance in the failure rate estimate was reduced from 1.03×10^{-8} h^{-2} to 2.09×10^{-9} h^{-2}, a reduction of approximately 80%.

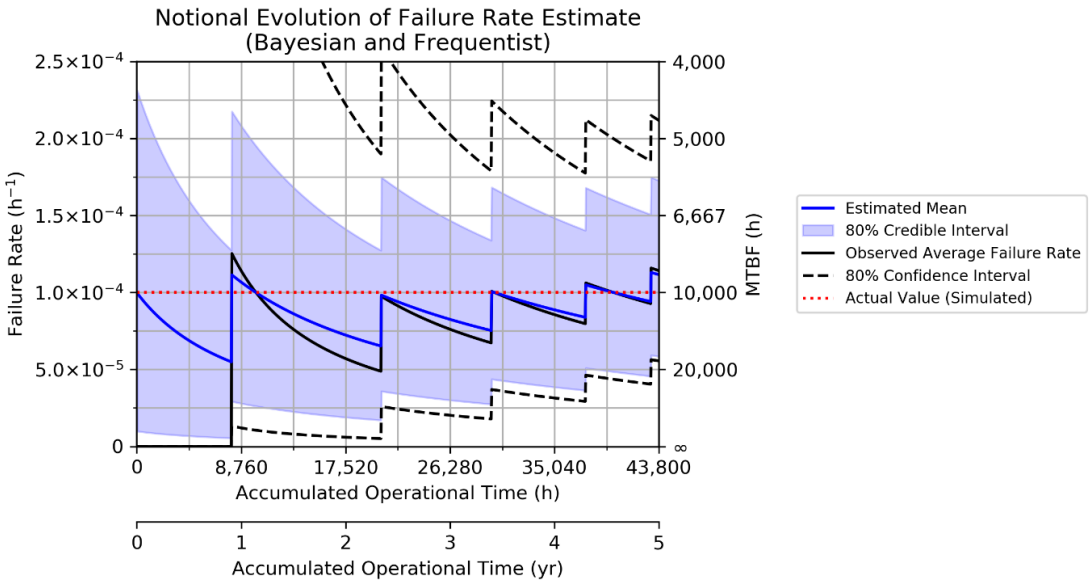

Figure 5: Notional evolution of a Bayesian failure rate estimate, showing the estimated mean failure rate (solid blue line) and 80% credible interval (blue shaded area). The initial mean failure rate was assumed to be accurate, and the initial error factor is equal to 4. The frequentist observed average failure rate (solid black line) and 80% confidence interval (dotted black lines) are also shown for context. Failures, indicated by sharp upward ticks in the failure rate estimate, were randomly generated using a process with an actual failure rate indicated by the horizontal red line.

The key benefit from the Bayesian approach is that the use of an initial estimate can enable faster reduction in uncertainty. As shown in Figure 5, for example, the upper end of the 80% credible interval is typically significantly lower than the upper end of the 80% confidence interval. This decreased uncertainty results from the extra information provided by the prior estimate. However, prior failure rate estimates should be used with caution, since they introduce subjectivity into the analysis that incurs additional risk unless it is corrected by sufficient real-world experience. For example, approximately 15% of ISS ORU failure rate estimates (including approximately 18% of ISS ECLSS ORUs) have shifted upwards when they are updated based on operational experience, indicating that those items turned out to be less reliable in practice than was initially estimated.[1,26] These underestimated failure rates can significantly impact supportability logistics planning and risk assessment. For example, if spares allocations for a 1,200-day Mars mission had been planned using the initial ISS ECLSS ORU failure rate estimates but actual spares demands during that mission had been more in line with updated failure rate estimates, the actual risk of insufficient spares would have potentially been an order of magnitude higher than expected.[26] Overall, prior failure rate estimates must be clearly justified and carefully examined in the context of whatever available evidence (past experience, initial test outcomes, etc.) are available, and updated estimates should be continually validated against a purely frequentist analysis each time new data become available.

To investigate the impact of inaccurate prior failure rates, Figure 7 and Figure 6 repeat the analysis shown in Figure 5 with over- and underestimated prior failure rate estimates, respectively. In each case, the prior mean failure rate estimate is assumed to be off by a factor of 5, but the prior error factor is the same (i.e. 4). When the initial failure rate estimate is higher than the actual value – that is, when the initial estimate is pessimistic – the estimate gradually converges to the true value. The 80% credible interval also remains within the 80% confidence interval and contains the true value across the entire test timeline. At the end of the test, the mean failure rate estimate is 1.30×10^{-4} h^{-2}, 16% higher than the baseline case. The error factor is 1.91, the same error factor that was obtained when the initial mean failure rate estimate was accurate. This is because, under the gamma model, error factor is only impacted by the number of observed failures. This does not mean, however, that the same level of uncertainty is present. The variance at the end of the test is 2.85×10^{-9} h^{-2}, an increase of approximately 36% from the case where the initial mean failure rate estimate is accurate. These results, as well as the ones described below, indicate that error factor may not be a good metric for representing uncertainty.

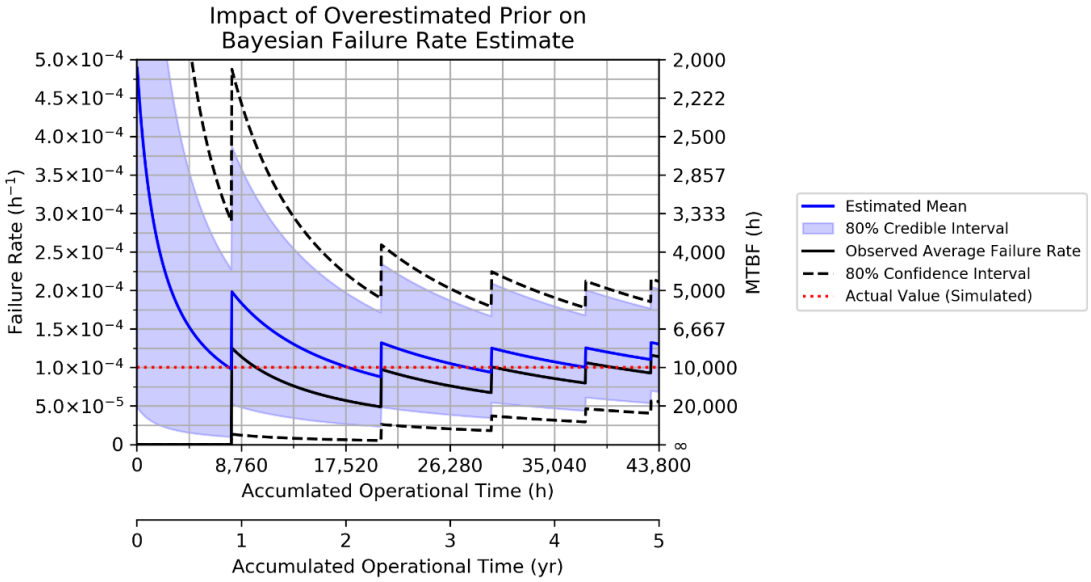

Figure 7: Notional evolution of a Bayesian failure rate estimate starting from an overestimated prior. The initial mean failure rate was overestimated by a factor of 5, with an initial error factor of 4. Note that the scale of the y-axis here is different from that of Figure 5 and Figure 6.

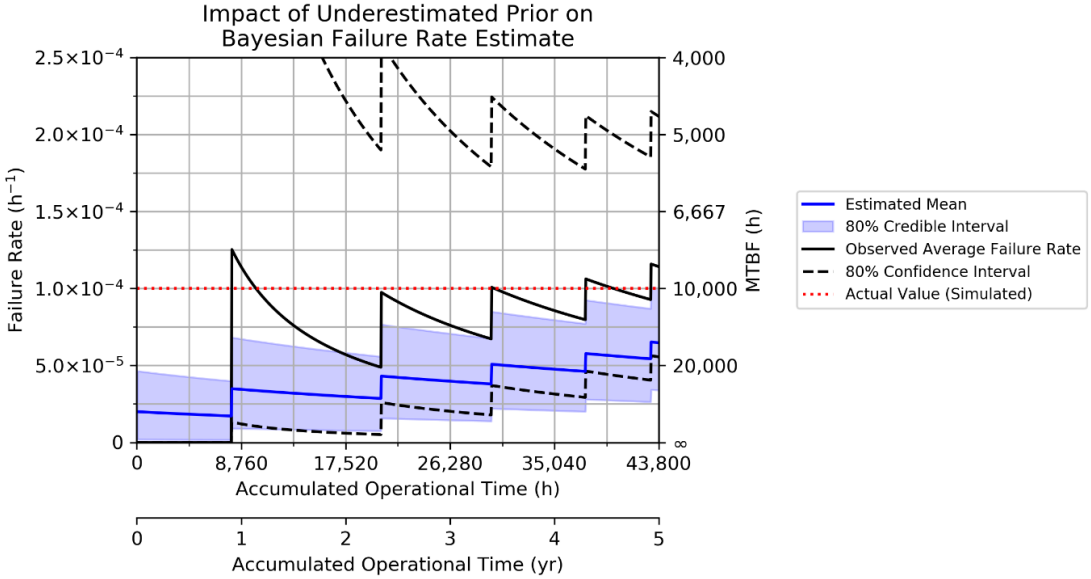

Figure 6: Notional evolution of a Bayesian failure rate estimate starting from an underestimated prior. The initial mean failure rate was underestimated by a factor of 5, with an initial error factor of 4.

However, when the initial failure rate estimate is lower than the actual value – that is, when the initial estimate is optimistic – and the error factor is kept the same, then the Bayesian update process corrects the initial underestimate much more slowly. The mean failure rate estimate remains significantly below the observed average, and the 80% credible interval is no longer contained within the 80% confidence interval, indicating disagreement between the Bayesian estimate and observed system behavior. At the end of the test period, the estimated failure rate is 6.48×10^{-5} h^{-1}, just 58% of the result from the case where the initial mean estimate is accurate. As with the overestimated case, the error factor at the end of the test is 1.91. The variance, however, is 7.03×10^{-10} h^{-2}, 66% less than the baseline case.

Error factor is defined as a ratio of two values in the distribution, and as a result estimates with the same error factor can have significantly different amounts of uncertainty – as measured by variance – depending on the mean failure rate. Equation 9 shows that, given a fixed error factor ε, the variance is proportional to the square of the mean failure rate. Pessimistic failure rate estimates (i.e. overestimated failure rates) have much higher variance than optimistic failure rates (i.e. underestimated failure rates) for the same error factor. As a result, a model that uses the same error factor across the entire range of mean failure rate estimates will implicitly assign higher confidence to more optimistic failure rate estimates by giving them lower variances. The net effect is that more optimistic failure rate estimates will be more robust and resistant to change, even when Bayesian updates are carried out in response to evidence that strongly suggests that the failure rate estimate is too low, as is the case here. Underestimated failure rates have significant negative impacts on supportability risk for long-endurance missions, and therefore this issue is a potentially significant source of risk that should be addressed. The implications of these results are discussed further in Section V.A.

Bayesian analysis is a useful method for incorporating prior information in order to more rapidly reduce failure rate uncertainty. However, analysists, system designers, and mission planners should be cognizant of the risks associated with prior failure rate estimates. If these estimates are overly optimistic, they may lead to significant underestimation of risk and logistics requirements. While Bayesian estimates will tend to converge towards the true value as more data are gathered, the rate of convergence may be relatively slow (depending on initial uncertainty estimates), as was the case with the results shown in Figure 6. Ideally, these priors would be based on early test results or analysis of past missions as much as possible, in order to provide an empirical basis for estimation. However, a significant amount of experience may be required to adjust the estimated failure rate to the true value.

As with the frequentist analysis described in Section III.A, Bayesian updating can also be accelerated by operating multiple units in parallel. Accelerated testing efforts can be targeted, allocating additional testing resources to specific ORUs if early results indicate that they may be less reliable than expected. For example, as discussed in greater detail in previous work,[26] the ISS Urine Processor Assembly (UPA) Fluids Control and Pump Assembly (FCPA) has experienced significant reliability challenges, and its current failure rate estimate is significantly higher than the initial estimate. While years of testing were required to arrive at the new value, the first FCPA failure occurred after an operating time equal to only 12% of the predicted MTBF – much earlier than expected. Similarly, in the example shown here differences between the observed average failure rate / confidence interval and Bayesian mean failure rate / credible interval can indicate that the observed behavior is significantly different from what was expected. Indicators such as these can be used to flag potential failure rate underestimates, and more resources can be directed towards gathering operating time on that ORU, for example by operating a second unit in parallel or increasing the duty cycle. As with all system decisions, however, the costs associated with accelerating operational data gathering and failure rate evaluation (including test time, procurement of additional units, etc.) should be balanced against the supportability costs and risks associated with the ORU in question.

IV. Spares Mass Impacts of Updated Failure Rate Estimates

The models described in Section III examine changes in failure rate estimates at the ORU level, but from a system development and mission planning perspective it is critical to understand the impacts of those changes on maintenance logistics requirements. This section examines changes in total spares mass requirements as a function of potential improvements in failure rate estimates resulting from testing. Section IV.A briefly presents a spares mass model based on one used in previous research. Section IV.B then describes a method for using that model along with the Bayesian model described in Section III.B to forecast potential spares mass reduction for a future Mars mission as a function of test duration and a range of hypothetical test outcomes. These models are used to characterize the potential value of additional ISS operations for testing and uncertainty reduction, building upon previous analysis of the value of past ISS operational experience.[26]

A. Spares Mass Model Description

The spares mass model used in this analysis is the same one that was used to evaluate the impact of ISS experience in previous work.[26] A brief summary of the model is presented here, and readers interested in a more detailed description are directed to that paper. In addition, Owens[6] presents extensive background (pp. 33-77), derivations and model descriptions (pp. 168-173 and 182-194), verification (pp. 221-227), and validation (pp. 234-262). The model evaluates the total mass of spare parts required to achieve a given POS requirement, given a set of ORUs and a mission endurance. Time to failure for each ORU is assumed to follow an exponential distribution (i.e. the CFR model), and the number of failures that an ORU will experience in a given time period follows a Poisson distribution with a

parameter equal to the expected number of failures – that is, the failure rate multiplied by the quantity q, K-factor κ, duty cycle d, and mission endurance τ.[9,19] In the standard CFR model, the failure rate is assumed to be a deterministically known value, λ. However, in this model epistemic uncertainty is captured by representing failure rates using the random variable Λ, as described in Section III.B above. When the parameter for a Poisson distribution is itself a random variable, the result is a mixed Poisson distribution.[26,28] A lognormally-distributed failure rate Λ results in a Poisson-lognormal failure distribution with parameter $\tau q \kappa d \Lambda$, but unfortunately there is no closed-form Cumulative Distribution Function (CDF) for Poisson-lognormal probabilities, and therefore evaluation of POS for a given number of spares using a lognormal failure rate would require Monte Carlo simulation or other numerical approximation.[29]

However, the lognormal distribution can be approximated using a gamma distribution with the same expected value and variance, as described in Section III.B. A Poisson distribution with a gamma-distributed parameter is known as a gamma-Poisson distribution, and has the same CDF as a negative binomial distribution. Therefore the POS associated with each ORU i given a number of spares n_i is given by

$$POS_i(n_i) = \sum_{k=0}^{n_i} \binom{k+\alpha_i-1}{k} \left(\frac{\beta_i}{\beta_i+\tau q_i \kappa_i d_i}\right)^{\alpha_i} \left(1 - \frac{\beta_i}{\beta_i+\tau q_i \kappa_i d_i}\right)^{k} \quad (15)$$

where α_i and β_i are the shape and scale parameters of the gamma failure rate uncertainty distribution and q_i, κ_i, and d_i are the quantity, K-factor, and duty cycle for ORU i.[3,6,26,28,30,31] The negative binomial approximation has been validated by comparing POS values to those calculated using Poisson-lognormal model (via large-scale Monte Carlo simulation) and showing close agreement between the two approaches.[6]

Equation 15 relates the number of spares provided for each ORU to the POS associated with that ORU. The overall system POS is the product of the POS for each ORU. Many different spares allocations will result in a system POS greater than the specified requirement, but only one will do so while minimizing total spares mass. Discrete optimization – specifically marginal analysis[9,32] with a branch and bound search algorithm,[33] described in greater detail by Owens[6] (pp. 185-194) – is used to identify this optimal spares allocation. Total spares mass is then calculated by multiplying the number of spares provided for each ORU by the mass of that ORU and summing the result.

B. Mars Mission Case Study

The value of additional test time is assessed by examining the spares mass required to achieve a given POS target for a notional 1,200-day Mars mission as a function of the amount of additional accumulated operational time and a hypothetical test outcome. POS requirements of 0.99, 0.995, and 0.999 are examined, and the system being evaluated consists of the same 50 ISS ECLSS ORUs examined in previous analysis of the value of past ISS experience.[26] This includes major components from the Water Processor Assembly (WPA), UPA, Oxygen Generation System (OGS), Carbon Dioxide Removal Assembly (CDRA), Common Cabin Air Assembly (CCAA), and Trace Contaminant Control System (TCCS). Current system characteristics, including mean failure rate and error factor (for both initial and current failure rate estimates), K-factor, mass, duty cycle, and quantity were gathered from the ISS Maintenance and Analysis Data Set (MADS).* Spares mass requirements are assessed using the model described above, and hypothetical future failure rate estimates are generated as a function of test outcome using the Bayesian update model described in Section III.B.

Since the outcome of future testing is not (and cannot be) known, this research examines two cases. In the first, the number of failures observed during the test for each ORU is assumed to be equal to the expected number of failures for that ORU. That is, for each ORU i, the accumulated operational time $t_{o,i}$ and number of failures $n_{o,i}$ are calculated as a function of the total additional test time t_o:

$$t_{o,i} = q_i d_i t_o \quad (16)$$

$$n_{o,i} = \bar{\lambda}_i q_i \kappa_i d_i t_o \quad (17)$$

The second scenario assumes that 50% more failures than expected are observed during the test by adding a coefficient of 1.5 to the right side of equation 16. Effectively, the first scenario represents the case where current failure rate estimates are generally correct, while the second represents the pessimistic case where all failure rates are

* Data used in this paper are current as of April 24, 2019.

underestimated by 50%. In each scenario, updated failure rate estimates are determined using equations 13 and 14 and used to assess spares mass requirements as a function of additional test time.

It is important to note that this case study only examines hypothetical futures, based on assumptions about how many failures are experienced by each ORU in the coming years. These results are not a prediction of the future; they are simply a mechanism for examining potential futures in order to inform decision-making. *There is no way to know what the outcome of a test will be before the test is executed.* Tests are activities performed to gather unknown information; if the outcome were known ahead of time, there would be no reason to perform the test. Even when there is high confidence in a predicted outcome – for example, if an ORU's predicted failure rate is low enough that there should not be any observed failures during the test – a test can be an extremely valuable or even necessary activity to provide an empirical basis for validating those predictions. Predictions are not validated until the test is executed, and results based on those predictions remain hypothetical.

In addition, this case study does not account for changes to the system that may occur in the coming years, such as upgrades to add/remove items, change level of repair, change mass, or improve reliability by removing failure modes. The models described in Section III are designed to evaluate and estimate failure rates for items with constant failure rates. Different modeling approaches can be used to forecast potential reliability growth as a result of testing with the intent of identifying and removing failure modes by making changes to the system, but those models are beyond the scope of this paper. Reliability growth modeling has been discussed in previous work,[34] and readers are directed to papers and reports by Duane,[35] Crow,[36,37] the DoD,[12,38] and the National Research Council (NRC)[39] for more details about those methods. Future work will combine the uncertainty reduction models described here with these reliability growth modeling approaches in order to develop an integrated methodology for test planning and system reliability evaluation.

1. Baseline: Expected Failures Observed

Figure 8 shows the spares mass required for a 1,200-day Mars mission as a function of the amount of operational experience accumulated before the mission. The impact of past experience is shown in terms of the change in spares mass from the initial failure rate estimates to current estimates, and the impact of future experience is forecast under the assumption that the number of failures observed for each ORU in the coming years is the expected number of

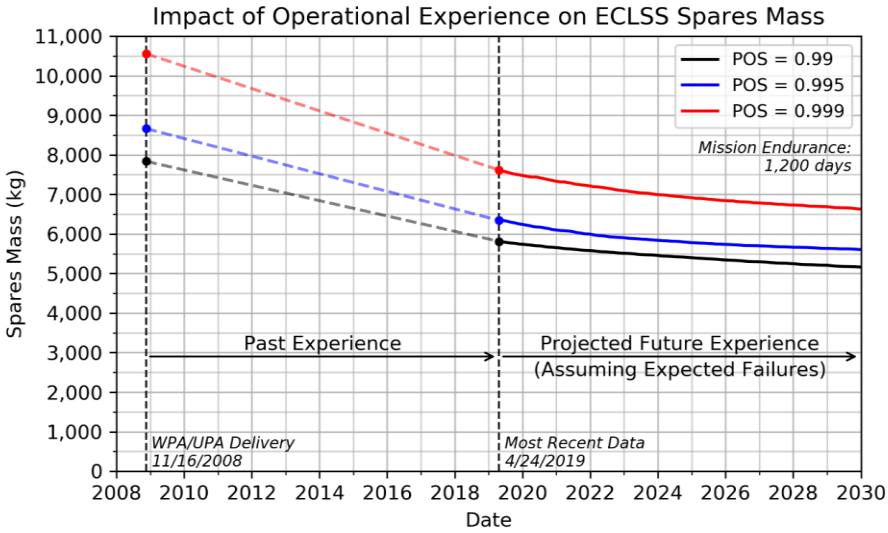

Figure 8: Impact of past and projected future operational experience on ECLSS spares mass for the baseline case, assuming the expected number of failures are observed for all ORUs in future years.

Table 3: Differences in spares mass requirements from initial and current failure rate estimates.

POS	Spares Mass (kg)		Mass Savings
	Using Initial Estimates	Using Current Estimates	
0.99	7,836	5,811	2,025 *(26%)*
0.995	8,663	6,341	2,322 *(27%)*
0.999	10,557	7,615	2,942 *(28%)*

failures. For the purposes of this analysis, the initial failure rate estimates are associated with November 16, 2008, the date that the WPA and UPA were delivered to the ISS aboard STS-126 (ULF2).[40,41] Current estimates are based on the most recent data compiled from MADS, from April 24, 2019. Note that the dotted line connecting initial and current estimates is meant only to indicate a connection between points at the same POS level, and does not represent actual spares mass values at any intermediate points. Changes in spares mass requirements probably do not follow a linear trend during that time period. Efforts are currently underway to collect and organize the data required to evaluate these intermediate values.

The spares mass values for initial and current failure rate estimates shown here, summarized in Table 3, are lower than the ones presented in previous work.[26] This is primarily due to the fact that the previous analysis did not appropriately take duty cycle into account during spares assessment. Specifically, spares were mistakenly allocated assuming that the listed failure rate was relative to calendar time, rather than operational time on that specific unit, and as a result failure rates were overestimated for the subset of ORUs that have a duty cycle of less than one. When this error is corrected, the spares mass results for both initial and current failure rate estimates are lower, and the mass savings resulting from that past experience are slightly smaller. However, that past operational experience still enables multiple tons of spares mass reduction, with more than a 25% reduction in total spares mass at all POS values examined here. While not quantitatively examined here, the benefits related to risk reduction described in that previous paper are also expected to still be present when duty cycle is accounted for, since those are driven primarily by the significantly underestimated initial failure rate values in a handful of ORUs. Past ISS experience has still enabled the correction of those underestimates, thereby helping to mitigate that source of risk.

Figure 9 shows a more detailed view of the data on the right side of Figure 8, showing the amount of spares mass required as a function of additional accumulated operating time under the assumption that each ORU fails the expected number of times in future years. Key values are summarized in Table 4. Focusing on the 0.995 case as a baseline, if the current mean failure rate estimates are correct then the uncertainty reduction facilitated by five additional years of

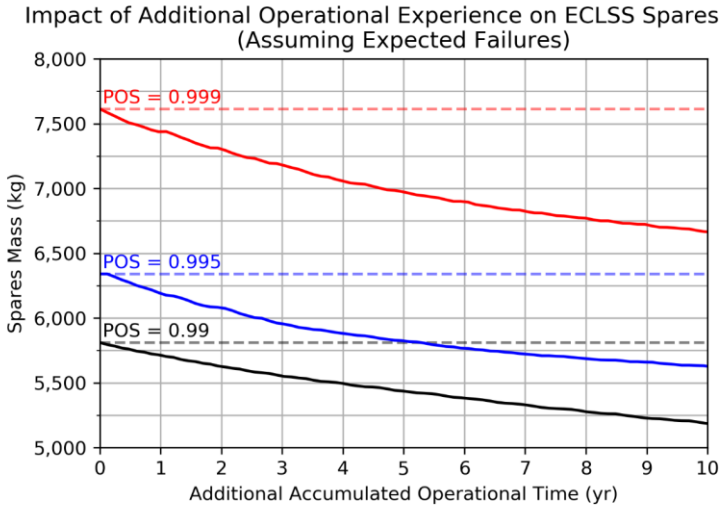

Figure 9: ECLSS spares mass required for a 1,200-day mission as a function of POS value and projected changes in failure rate estimates from additional accumulated operating time, assuming the expected number of failures are observed for all ORUs in future years.

Table 4: Key spares mass values and percent changes as a function of additional accumulated operational time and POS requirement, assuming the expected number of failures are observed for all ORUs in future years.

Additional Operational Time (yr)	Spares Mass (kg) POS = 0.99		POS = 0.995		POS = 0.999	
0	5,811		6,341		7,615	
5	5,437	*(-6.4%)*	5,826	*(-8.1%)*	6,972	*(-8.4%)*
10	5,186	*(-10.8%)*	5,626	*(-11.3%)*	6,663	*(-12.5%)*

operations would enable a 515 kg reduction in ECLSS spares mass requirements, or approximately 8.1%. Five more years of operations beyond that, or a total of 10 years (e.g. ISS operations through mid-2029), are forecast to result in a 715 kg reduction in ECLSS spares mass from their current estimated values, or approximately 11.3%. Similar trends are seen at other POS levels, with higher spares mass reduction amounts forecast (in terms of both mass and percentage) when the POS requirement is higher. These impacts are not linear with time, however; the information gained via additional testing has diminishing returns.

2. Pessimistic: 50% More than Expected Failures Observed

The results shown in Section IV.B.1 represent the expected outcome of future testing, based on current estimates. However, past experience shows that tests do not always produce the expected outcome. Figure 10 and Table 5 show results for a more pessimistic scenario in which all ORUs experience 50% more failures than would be expected, given current mean failure rate estimates. This is considered a conservative scenario because it seems unlikely (based on examination of changes in ISS ORU failure rates to date presented by Stromgren et al.[8] and Owens and de Weck[26]) that *all* failure rates are currently underestimated by 50%.

The results of this conservative case indicate that, even if all ORUs are less reliable than expected, the reduction in uncertainty provided by testing still enables spares mass reductions. The mean failure rate estimate for each ORU goes up, since the number of failures observed is higher than expected. However, the amount of uncertainty in the estimate is reduced, and therefore the amount of risk associated with epistemic uncertainty is smaller and POS targets can be achieved using fewer spares. The amount of mass reduction is smaller than the case in which the current mean failure rate estimates are correct. For the 0.995 case, for example, 5 years of testing would enable a reduction of 167 kg (2.6%), and 10 years would enable a reduction of 283 kg (4.4%). More importantly, however, this testing would identify and correct the underestimated failure rates, reducing the risk of incorrect/insufficient spares allocations. This second point is particularly important since it is the specific spares allocation, not the mass of that allocation, that

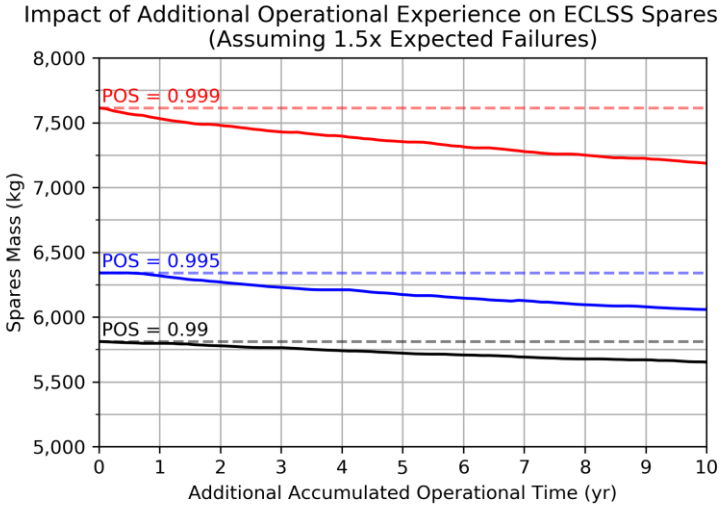

Figure 10: ECLSS spares mass required for a 1,200-day mission as a function of POS value and projected changes in failure rate estimates from additional accumulated operating time, assuming the number of failures observed for all ORUs in future years is 50% higher than expected.

Table 5: Key values and percent changes in spares mass as a function of additional accumulated operational time and POS requirement, assuming the number of failures observed for all ORUs in future years is 50% higher than expected.

Additional Operational Time (yr)	Spares Mass (kg) POS = 0.99		POS = 0.995		POS = 0.999	
0	5,811		6,341		7,615	
5	5,721	*(-1.6%)*	6,174	*(-2.6%)*	7,354	*(-3.4%)*
10	5,653	*(-2.7%)*	6,058	*(-4.4%)*	7,187	*(-5.6%)*

matters form a risk perspective. It is possible to carry sufficient spares mass but still carry the wrong types/quantities of spares and thereby not mitigate risk to desired levels.[26]

The two scenarios investigated here represent relatively simple hypothetical futures in which all ORUs exhibit the same behavior – they either fail the expected number of times, or they fail 50% more than expected. In reality, different ORUs will likely behave in different ways, with some turning out to be more reliable than expected and others turning out to be less reliable. The net impacts of those results will depend on the mass and failure rate of each ORU. While those more complex cases are not presented here, the methodology described in this paper can be used to examine them if ORU-specific failure counts are provided.

V. Discussion

ORU failure rate estimates and total spares mass requirements can both change significantly as a result of information gained through testing and operational experience. Underestimated failure rates, which can be a significant and potentially undetected source of risk for long-endurance missions, can be detected and corrected through testing. Overestimated failure rates, which may result in higher spares mass than is actually required, can also be corrected. Empirical validation of failure rate estimates improves overall confidence in supportability risk assessments, and general reduction in epistemic uncertainty allows risk targets to be achieved using fewer spares and less spares mass overall. When an initial reliability estimate is made, one of three things is true, and in each case testing provides value:

1) The initial estimate is accurate, in which case testing validates that estimate, mitigating the risk of an incorrect estimate, and removes epistemic uncertainty, allowing the same POS levels to be achieved with fewer spares;
2) The initial estimate is too low, in which case testing reveals and corrects the underestimated failure rate and mitigating the risk of insufficient spares that would have been present if the underestimated failure rate went undetected, or allows designers to make changes to improve reliability; or
3) The initial estimate is too high, in which case testing reveals and corrects the overestimated failure rate, allowing the same POS levels to be achieved with fewer spares.

In all cases, information provided by testing and operational experience improves risk assessment and logistics planning by providing an empirical basis for the analysis.

However, these benefits are only achieved when a significant amount of operational time is accumulated. Tests that characterize, validate, and/or correct failure rate estimates are not demonstration activities in which a system is turned on and operated for a few weeks or months to prove that it performs all required functions. Systems need to accumulate long operating times, on the same scale of or longer than the estimated/targeted MTBFs that are being examined, in order to provide strong statistical validation of those MTBFs and the associated failure rates. These long test durations do not necessarily imply long schedule delays, however, since multiple units can be tested in parallel to accelerate the rate at which data are gathered.

In addition, testing provides significant benefits beyond the mass reduction described here. Experience on the ISS to date has been extremely valuable for identifying and helping to correct unexpected failure modes, and future experience is expected to continue providing these benefits. Long-duration testing activities help system designers upgrade systems to improve performance and reliability, and help analysts and mission planners accurately understand risks and logistics requirements for future missions. Extended testing does exhibit diminishing returns, and test campaign planning should carefully consider the potential benefits of a test (using the models presented above, ideally under a variety of potential outcomes) alongside the costs of test activities, including the test articles themselves as well as personnel, facilities, and any other implications related to schedule changes or opportunity costs associated with testing an existing system rather than developing and testing a new one.

A. Issues with Using Error Factors to Characterize Uncertainty

One key finding in this paper, described in more detail in Section III.B, is the potentially misleading relationship between error factor and uncertainty. While error factor is commonly used to describe the amount of uncertainty in a failure rate estimate, different estimates can have significantly different variances as a function of the mean value of the estimate. As a result, if the same error factor is used for an optimistic estimate vs. a pessimistic estimate, the optimistic estimate (i.e. low failure rate) will implicitly have significantly less uncertainty and will be slower to adjust in the face of opposing evidence during Bayesian updating. If Bayesian updating is carried out for a population of ORUs by assigning the same error factor to all initial estimates, underestimated failure rates will be more robust to change than overestimated failure rates. Put another way, this modeling approach may result in a systematic bias towards underestimating failure rates within the population, thereby leading to underestimation of supportability risks

and the number of spares or total spares mass required to mitigate them. Shorter-endurance systems such as the ISS may not be as impacted, but underestimated failure rates are a major source of risk for long-endurance missions.

Future work will continue to investigate this effect and attempt to identify uncertainty measures that provide a more uniform response to Bayesian updating across a wider range of failure rates. Potential candidates include using the same variance or using the same ratio of variance to expected value. Ideally, however, the amount of uncertainty associated with an initial failure rate estimate would be specific to that estimate and based on the information used to generate the estimate. Less time is required to demonstrate higher failure rates (as shown by Figure 2 and Figure 3), and therefore high failure rate estimates should in general be less uncertain than low failure rate estimates. Interestingly, applying the same error factor to all ORUs results in the opposite effect, artificially inflating confidence in low failure rate estimates relative to high failure rate estimates. ORU-specific initial uncertainty estimates would allow analysts to capture differences in uncertainty between different ORUs, and ISS operational experience has provided and will continue to provide a wealth of data that can be used to inform these estimates for future systems.

B. DoD Lessons Learned and Recommendations

Lessons learned and associated recommendations from other industries can be useful for space system development activities as well. The DoD and defense contractors have extensive experience in developing and fielding complex systems in environments where supportability characteristics can have significant impacts on operational capabilities and risks, and have faced challenges in doing so. Between 2006 and 2011, half of all major defense systems failed to meet reliability targets, and as a result experienced significant cost overruns, schedule delays, and reduced operational capability. In response, the Office of the Secretary of Defense initiated an effort to improve the reliability outcomes of defense system development projects. The NRC was asked to study and evaluate these efforts, and summarized their findings in a 2015 report entitled "Reliability Growth: Enhancing Defense System Reliability."[39] One of the key findings of this report, which has important implications for future space system development and testing strategies, is summarized by the following quote (from p. 6):

> "High reliability early in system design is better than extensive and expensive system-level developmental testing to correct low initial reliability levels. The former has been the common *successful* strategy in non-DoD commercial acquisition; the latter has been the predominantly *unsuccessful* strategy in DoD acquisition."[39]

Overall, successful strategies for attaining high reliability early in system design tend to have two elements. First, reliability is emphasized during the initial design process, and reliability targets are explicitly codified as measurable system/subsystem requirements. Second, components, assemblies, and subsystems are tested early and often to identify failure modes and inform design improvements. Critically, this testing occurs before the design is complete, since it is typically more difficult and expensive to modify a finished design. The NRC made several recommendations in its report, many of which are highly applicable to NASA and space system development in general, including:[39]

- Reliability goals should be specified up-front using measurable, testable requirements, and reliability should be a key figure of merit, clearly related to system/mission cost, risk, and performance during trade studies and concept selection.
- Reliability growth modeling should be used to set realistic goals, establish test plans, and track progress during system development.
- Initial reliability estimates should be performed using modern techniques such as physics-of-failure models, or generated based on previous system performance. Older techniques based on MIL-HDBK-217 ("Military Handbook: Reliability Prediction of Electronic Equipment") are considered invalid, inaccurate, and misleading, and as a result have been rejected for use in reliability prediction by the Army and by General Motors.[42] Failure rate estimates should be validated and updated using empirical test results as they become available.
- Management practices should incentivize and allocate specific funding for the design and testing activities required to achieve and demonstrate reliability targets during early system development. Programs should not plan on "testing in" reliability during full systems testing after the design is complete, since that approach is typically more expensive, less efficient, slower, and tends to not achieve reliability goals.
- All data associated with testing, operational maintenance, and reliability analysis should be collected and archived for the purposes of validating system characteristics and informing future development efforts. Detailed reliability and test data is necessary to enable empirical risk-informed system investment, design,

development, and operational decisions. Contracts should be structured to require this data collection and ensure that the end user of a system receives these datasets in addition to the system itself.

The concept of "testing in" reliability described in the 4th bullet point above – that is, attempting to identify and remove failure modes during full systems testing after the design is complete rather than emphasizing reliability during design – is relevant regardless of whether the full system is considered operational or it is considered a testbed. Effectively, this recommendation is based on the fact that it is difficult to make changes to a system after it is complete and integrated. Full-system testing is itself a valuable activity for identifying system failure modes such as interface issues or negative interactions between elements. However, it is not an efficient approach for identifying and addressing reliability issues within systems or subsystems. More targeted test activities prior to system integration can focus on identifying and correcting failure modes before they make their way into the final design, and can help avoid costs and schedule delays associated with retrofits and redesigns. By the time systems are integrated, all subsystems should have statistically-supported confidence that they have achieved their reliability goals so that full-system testing can focus on identifying and managing interface/interaction issues rather than lower-level reliability issues.

While the case study presented in Section IV examined the value of additional operational experience in terms of full-system testing aboard the ISS, full-system testing alone is not necessarily the most efficient way to gain that experience. However, the ISS is already operating as a valuable ECLSS testbed, and therefore it makes sense to examine continued operations. In addition, the ISS is currently the only available platform for long-duration microgravity operations, and therefore has the additional benefit of closely recreating the operating environment of future exploration systems. However, in some cases targeted subsystem-level ground testing, or on-orbit testing as a payload rather than as part of the integrated station ECLSS, may be a more efficient approach for evaluating reliability.

The final bullet point is also worth emphasizing, especially for spaceflight activities. Without data collection and analysis, there is no way to make informed decisions regarding system reliability and supportability. The analyses presented here and in many past papers examining the supportability challenges of future spaceflight would not be possible without the MADS database and other datasets recording maintenance actions, failure rate estimates, and other data collected over the past few decades of ISS operations and system development. The spaceflight industry is typically severely data-limited when it comes to reliability evaluation, and every data point counts. Testing and operations designed to support future exploration system development have the opportunity to provide significant value, but much of that value may be lost if there is not proper documentation of the results of those activities and dissemination of lessons learned.

VI. Conclusions

Supportability will be a significant driver of logistics mass and risk for future missions, and testing and operational experience are critical activities for estimation and verification of key system supportability parameters. This paper has described and demonstrated a series of models for understanding the evolution of failure rate estimates as a function of accumulated operating time and the number of observed failures, as well as models for understanding the impacts of those changes on overall spares mass requirements. The results of the case studies and examples presented above indicate that additional testing has the potential to provide significant benefits to future exploration systems, but that a significant amount of testing may be required.

In the end, there is no simple answer to the question of how much testing is required to manage supportability risks for beyond-LEO missions. The operating context for these missions is significantly more challenging than past experience, and therefore an accurate understanding of failure rates and system supportability characteristics is more critical. Testing is a key element of risk and logistics mass management for these future systems, and long test durations (possibly on the order of years) may be required to provide a high degree of confidence in failure rate estimates and associated risk/logistics assessments. However, test time incurs costs of its own, both in terms of financial cost, lengthened schedules, and the opportunity costs. Test campaign planning must carefully balance these mass, schedule, cost, and risk (both technical and programmatic risk) considerations in order to make the most effective use of available resources. This planning must also consider the availability of key test platforms such as the ISS. The ability to continuously operate spacecraft systems in a relevant microgravity environment and observe their behavior for years at a time has provided and will continue to provide essential data for system reliability assessment and supportability planning. The availability of on-orbit test platforms is a critical consideration during program schedule development. The models presented in this paper can provide valuable inputs to inform test planning and system development activities.

In general, a process that emphasizes reliability from the earliest phases of system design, defines measurable/testable requirements related to subsystem and system reliability, and supports and funds that activities

required to execute and respond to those tests tends to be more effective in developing reliable systems than processes that neglect reliability until after the design phase. This approach does require increased up-front investment, but can significantly reduce lifecycle costs and risks during operations. Good design decisions are facilitated by good design practices, and a programmatic emphasis on reliability and supportability can help ensure that future systems achieve their goals and mitigate risks on long-endurance beyond-LEO missions without excessive logistics requirements.

Acknowledgments

This research was initiated while the lead author was a doctoral candidate at MIT, and was supported by a NASA Space Technology Research Fellowship (grant number NNX14AM42H) during that time. It has been supported by the Advanced Exploration Systems (AES) ECLSS Supportability Study since the lead author completed his PhD and began working at NASA Langley Research Center full-time. The authors would like to thank Bill Cirillo, Chel Stromgren, Jordan Klovstad, Kandyce Goodliff, and Kevin Earle for their thoughts and input on this work.

References

[1] Stromgren, C., Goodliff, K. E., Cirillo, W., and Owens, A., "The Threat of Uncertainty - Why Using Traditional Approaches for Evaluating Spacecraft Reliability Are Insufficient for Future Human Mars Missions," *AIAA SPACE 2016*, Long Beach, CA: American Institute of Aeronautics and Astronautics, 2016.

[2] Owens, A. C., de Weck, O. L., Stromgren, C., Goodliff, K., and Cirillo, W., "Accounting for Epistemic Uncertainty in Mission Supportability Assessment: A Necessary Step in Understanding Risk and Logistics Requirements," *47th International Conference on Environmental Systems*, Charleston, SC: International Conference on Environmental Systems, 2017.

[3] Owens, A., and de Weck, O., "International Space Station Operational Experience and its Impacts on Future Mission Supportability," *48th International Conference on Environmental Systems*, Albuquerque, NM: International Conference on Environmental Systems, 2018.

[4] Cirillo, W., Aaseng, G., Goodliff, K., Stromgren, C., and Maxwell, A., "Supportability for Beyond Low Earth Orbit Missions," *AIAA SPACE 2011 Conference & Exposition*, Long Beach, CA: American Institute of Aeronautics and Astronautics, 2011.

[5] National Aeronautics and Space Administration, *NASA Systems Engineering Handbook*, National Aeronautics and Space Administration, 2007.

[6] Owens, A. C., "Multiobjective Optimization of Crewed Spacecraft Supportability Strategies," Doctoral Thesis, Massachusetts Institute of Technology, 2019.

[7] Owens, A., De Weck, O., Stromgren, C., Goodliff, K. E., and Cirillo, W., "Supportability Challenges, Metrics, and Key Decisions for Future Human Spaceflight," American Institute of Aeronautics and Astronautics, 2017.

[8] Anderson, L., Carter-Journet, K., Box, N., DiFilippo, D., Harrington, S., Jackson, D., and Lutomski, M., "Challenges of Sustaining the International Space Station through 2020 and Beyond: Including Epistemic Uncertainty in Reassessing Confidence Targets," *AIAA SPACE 2012 Conference & Exposition*, Pasadena, CA: American Institute of Aeronautics and Astronautics, 2012.

[9] Sherbrooke, C. C., *Optimal inventory modeling of systems: multi-echelon techniques*, Boston: Kluwer Academic, 2004.

[10] Stamatelatos, M., and Dezfuli, H., *Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners*, Washington, DC: National Aeronautics and Space Administration, 2011.

[11] Birolini, A., *Reliability Engineering: Theory and Practice*, Berlin, Heidelberg: Springer, 2004.

[12] Department of Defense, *Department of Defense Handbook: Reliability Growth Management*, United States Department of Defense, 1981.

[13] National Aeronautics and Space Administration, *NASA Technical Standard: Planning, Developing and Managing an Effective Reliability and Maintainability (R&M) Program*, National Aeronautics and Space Administration, 1998.

[14] Fisher, W. F., and Price, C. R., *Space Station Freedom External Maintenance Task Team Final Report*, Houston, TX: National Aeronautics and Space Administration, 1990.

[15] Anderson, L., Harrington, S., Omeke, O., and Schwaab, D., "Improving the Estimates of International Space Station (ISS) Induced 'K-Factor' Failure Rates for On-Orbit Replacement Unit (ORU) Supportability Analyses," American Institute of Aeronautics and Astronautics, 2009.

[16] Do, S., "Towards Earth Independence - Tradespace Exploration of Long-Duration Crewed Mars Surface System Architectures," Doctoral Thesis, Massachusetts Institute of Technology, 2016.

[17]Jones, H. W., Hodgson, E. W., and Kliss, M. H., "Life Support for Deep Space and Mars," *44th International Conference on Environmental Systems*, Tucson, AZ: 2014.
[18]Agte, J. S., "Multistate analysis and design : case studies in aerospace design and long endurance systems," Doctoral Thesis, Massachusetts Institute of Technology, 2011.
[19]Ebeling, C. E., *An introduction to reliability and maintainability engineering*, Long Grove, Ill: Waveland Press, 2010.
[20]Epstein, B., "Estimation from Life Test Data," *Technometrics*, vol. 2, Nov. 1960, p. 447.
[21]Department of Defense, *Department of Defense Handbook for Reliability Test Methods, Plans, and Environments for Engineering, Development, Qualification, and Production*`, United States Department of Defense, 1996.
[22]Cox, D. R., and Hinkley, D. V., *Theoretical statistics*, Boca Raton: Chapman & Hall/CRC, 2000.
[23]Mack, C., *Essentials of statistics for scientists and technologists*, New York: Plenum Pub. Corp., 1975.
[24]Vitali, R., and Lutomski, M. G., "Derivation of Failure Rates and Probability of Failures for the International Space Station Probabilistic Risk Assessment Study," *International Conference on Probabilistic Safety Assessment and Management*, Berlin: 2004.
[25]Dezfuli, H., Kelly, D., Smith, C., Vedros, K., and Galyean, W., *Bayesian Inference for NASA Probabilistic Risk and Reliability Analysis*, Washington, DC: National Aeronautics and Space Administration, 2009.
[26]Shynk, J. J., *Probability, random variables, and random processes: theory and signal processing applications*, Hoboken, NJ: Wiley, 2013.
[27]Hamada, M. S., Wilson, A. G., Reese, C. S., and Martz, H. F., eds., *Bayesian reliability*, New York, NY: Springer, 2008.
[28]Karlis, D., and Xekalaki, E., "Mixed Poisson Distributions," *International Statistical Review*, vol. 73, Jan. 2007, pp. 35–58.
[29]Shaban, S. A., "Poisson-Lognormal Distributions," *Lognormal distributions: theory and applications*, E.L. Crow and K. Shimizu, eds., New York: M. Dekker, 1988, pp. 195–210.
[30]Slay, F. M., Bachman, T. C., Kline, R. C., O'Malley, T. J., Eichorn, F. L., and King, R. M., *Optimizing Spares Support: The Aircraft Sustainability Model*, McLean, VA: Logistics Management Institute, 1996.
[31]McLachlan, G. J., and Peel, D., *Finite mixture models*, New York: Wiley, 2000.
[32]Fox, B., "Discrete Optimization Via Marginal Analysis," *Management Science*, vol. 13, Nov. 1966, pp. 210–216.
[33]Hillier, F. S., and Lieberman, G. J., *Introduction to Operations Research*, Boston: McGraw-Hill, 2001.
[34]Owens, A., and de Weck, O., "Limitations of Reliability for Long-Endurance Human Spaceflight," *AIAA SPACE 2016*, Long Beach, CA: American Institute of Aeronautics and Astronautics, 2016.
[35]Duane, J. T., "Learning Curve Approach to Reliability Monitoring," *IEEE Transactions on Aerospace*, vol. 2, 1964, pp. 563–566.
[36]Crow, L. H., "Planning a Reliability Growth Program Utilizing Historical Data," *Reliability and Maintainability Symposium*, Lake Buena Vista, FL: IEEE, 2011, pp. 1–6.
[37]Crow, L. H., *Reliability Analysis for Complex, Repairable Systems*, Aberdeen Proving Ground, MD: U.S. Army Materiel Systems Analysis Activity, 1975.
[38]Department of Defense, *Department of Defense Handbook: Reliability Growth Management*, United States Department of Defense, 2011.
[39]National Research Council, *Reliability Growth: Enhancing Defense System Reliability*, Washington, D.C.: National Academies Press, 2015.
[40]Williams, D. E., and Gentry, G. J., "International Space Station Environmental Control and Life Support System Status: 2008 - 2009," *39th International Conference on Environmental Systems*, Savannah, GA: SAE International, 2009.
[41]Takada, K. C., Ghariani, A. E., and Van Keuren, S., "Advancing the Oxygen Generation Assembly Design to Increase Reliability and Reduce Costs for a Future Long Duration Mission," *45th International Conference on Environmental Systems*, Bellevue, WA: International Conference on Environmental Systems, 2015.
[42]Peter, A., Das, D., and Pecht, M., "Appendix D: Critique of MIL-HDBK-217," *Reliability Growth: Enhancing Defense System Reliability*, National Academies Press, 2015, pp. 203–245.