Improving Reliability Estimates with Bayesian Hierarchical Models

Kassie Fronczyk, Rebecca Dickinson, and Laura Freeman

THE PROBLEM

The reliability of a weapon system is an essential component of its suitability for operational deployment. Yet, in an era of reduced budgets and limited testing, verifying that reliability requirements have been met can be challenging, particularly using traditional analysis methods that depend on a single set of data coming from a single test phase.

Short test periods, high reliability requirements, or few observed failures can result in little confidence in the reliability estimates… DoD [should] employ statistical approaches to capitalize on all available data from multiple test periods and not limit the reliability analysis to a single test period.

In the Department of Defense (DoD), test data are often collected in several phases. The two broad types of testing are developmental testing and operational testing. The primary goal of a developmental test (DT) is to verify that a system meets its design specifications. This testing can occur as contractor testing, government testing, or a mixture of both and is usually carried out in a controlled environment that often lacks the realism of combat scenarios and trained users. The purpose of an operational test (OT), on the other hand, is to determine whether the system is effective and suitable in a combat scenario. OT data are collected under test conditions that replicate, as much as possible, field use.

Reliability is one of the primary aspects of a system's operational suitability. It is important that a system perform as intended under realistic operating conditions for a specified period of time without failure. Reliability requirements for ground vehicles are often based on the mean number of miles between failures. A serious equipment failure that occurs during mission execution and results in the abort or termination of a mission is scored as an Operational Mission Failure (OMF). A less critical failure of a mission-essential component is scored as an Essential Function Failure (EFF). For example, an engine failure would be scored as an EFF if a vehicle took multiple attempts to start but eventually succeeded. If the vehicle could not be started, it would be scored as an OMF.

Requirements are typically written in terms of OMFs. Verifying whether the reliability requirements of a system have been met by looking at only a single test phase, however, can be challenging. Short test periods, high reliability requirements, or few observed failures can result in little confidence in the reliability estimates. The National Academies, in three separate studies (National Research Council 1998,[1] 2004,[2] and 2015[3]), have recommended that DoD employ statistical approaches to capitalize on all available data from multiple test periods and not limit the reliability analysis to a single test period. Despite these recommendations, nearly every published analysis of a major weapon system's reliability limits the assessment to the last test phase, typically because that phase examined the most representative system configuration. In support of the Director, Operational Test and Evaluation (DOT&E), IDA has begun to explore improved techniques for estimating reliability using data from multiple test periods.

BAYESIAN PARADIGM

When combining information from multiple test periods, models need to be carefully selected and evaluated to ensure that they accurately reflect the data and the underlying physical processes. The Bayesian paradigm is tailor-made for these situations because it allows the combination of multiple sources of data and variability to obtain more robust reliability estimates and quantify properly the uncertainty and precision of the estimates. The use of Bayesian methods is becoming increasingly popular because leveraging all of the available information when making decisions under uncertainty makes practical sense. This article uses reliability data from two families of vehicles tracked through multiple phases of testing to illustrate the Bayesian approach of combining information. Applying these methods results in better estimates of system reliability and more precise inferences.

The first case study uses reliability data from the Stryker family of vehicles (FoV), which are armored combat vehicles built for the U.S. Army. The FoV includes 10 system configurations, with two main versions: the Infantry Carrier Vehicle (ICV) (see Figure 1) and the Mobile Gun System (MGS). Our study focuses on the ICV, which provides protected transport and supporting fire for its two-man crew and squad of nine

Source: M1126 Infantry Carrier Vehicle

The ICV serves as the base vehicle for eight additional system configurations.

Figure 1. Stryker Infantry Carrier Vehicle (ICV)

[1] National Research Council. 1998. Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements. Washington, DC: The National Academies Press.

[2] National Research Council. 2004. Improved Operational Testing and Evaluation Methods of Combining Test Information for the Stryker Family of Vehicles and Related Army Systems: Phase II Report. Washington, DC: The National Academies Press.

[3] National Research Council. 2015. Reliability Growth: Enhancing Defense System Reliability. Washington, DC: The National Academies Press.

infantry soldiers. The ICV serves as the base vehicle for the eight remaining system configurations.[4] The vehicles share a common chassis and are outfitted with additional components specific to the mission of each vehicle. This analysis heavily leverages the common chassis of the vehicles to support combining information from all of the configurations. The reliability data, at the OMF level, used in this study come from two test phases: one DT and one OT.

The second case study is based on a notional future combat family of vehicles and data collected from multiple testing phases, as would be common for a program like Stryker. For this example, we will assume a family of vehicles similar to Stryker with four vehicles of various configurations that go through a series of three test phases with a corrective action period between each phase. Unlike the Stryker case study, for this notional example, we assume more detailed failure data are available, specifically EFFs and OMFs, as opposed to only OMFs. Because all OMFs are, by definition, EFFs, using all failures in the analysis provides a more robust reliability estimate.

For both cases, the goal is to characterize the reliability of the entire family of vehicles. In the Stryker study, we have OT data, but these data are limited; therefore, we need to leverage the commonalities of the vehicles and the DT data. For the notional future combat vehicle, we have assumed detailed information is available about failures from the three phases of testing and can pool information across the phases and four vehicles. A Bayesian framework that requires only slight modifications from one FoV to the other provides a mechanism to make the most use of this additional information.

STATISTICAL MODELS FOR COMBINING DATA: BAYESIAN RELIABILITY

A standard reliability analysis employed by the DoD test community considers each test phase (and each system configuration, such as vehicle type) independently and uses the exponential distribution to empirically model the miles between failures. Reliability is expressed in terms of the mean number of miles between a failure (MMBF), and is estimated as

$$\widehat{\text{MMBF}} = \frac{\text{Total Miles Driven}}{\text{\# of Failures}}.$$

Although this approach is standard for nearly every ground vehicle program in the Department, it ignores valuable information on individual vehicles in different phases of testing. Although frequentist statistical methods similar to the standard reliability analysis described previously (and illustrated in the Stryker analysis) could be used, a Bayesian approach provides a natural framework for combining multiple sources and types of information.

[4] The Antitank Guided Missile Vehicle (ATGMV), Commander's Vehicle (CV), Engineer Squad Vehicle (ESV), Fire Support Vehicle (FSV), Medical Evacuation Vehicle (MEV), Mortar Carrier Vehicle (MCV), Reconnaissance Vehicle (RV), and the Nuclear, Biological and Chemical Reconnaissance Vehicle (NBC RV). The NBC RV was excluded from the study because of its different acquisition timeline.

Bayesian methods are valuable for their logical integration of prior information and their practical convenience for modeling and estimation. Before looking at the test data, we construct a prior distribution – or starting assessment – for the parameters in the empirical model that we plan to construct. We use the data to revise our starting assessment and derive the updated assessment (i.e., the posterior distribution) for the parameters in the empirical model.

The reliability of the FoV is defined as a function of the failure rate parameter, λ (i.e., the mean time between failure (MTBF) or MMBF is $1/\lambda$). The exponential distribution is often used as the underlying assumption for the data's distribution, and a common choice of a prior distribution to describe the possible values of λ is the gamma distribution. The gamma distribution restricts the value of the failure rate to positive values and provides computational ease. Table 1 shows the Bayesian models for the Stryker and notional future combat vehicle FoV side by side to highlight the similarities and differences. For the Stryker analysis, we construct the statistical model such that each vehicle variant has its own failure rate, which is estimated by the data, and a single parameter to capture a common downgrade across vehicles from DT to OT. On the other hand, in the future combat vehicle example, the statistical model is written to capture the fact that the program has the ability to fix specific failure modes between phases (i.e., the corrective action periods). The statistical model, therefore, includes a separate estimate for each failure mode in each of the postulated test phases and fix effectiveness factors specific to each failure mode.

STRYKER FOV: ANALYSIS AND RESULTS

The reliability requirement for Stryker is that each vehicle has a mean of at least 1,000 miles between OMFs. Frequentist and Bayesian inference techniques were both employed to compare and contrast different approaches to combining

Table 1. Bayesian Reliability Models for Stryker and Future Combat Vehicle

Stryker	Future Combat Vehicle
$t_{DT} \sim exp(\lambda_i)$ $t_{OT} \sim exp(\lambda_i/\eta)$	$t_{T_1} \sim exp(\lambda_{ij})$ $t_{T_2} \sim exp(\lambda_{ij}(1-\rho_{1_j}))$ $t_{T_3} \sim exp(\lambda_{ij}(1-\rho_{1_j})(1-\rho_{2_j}))$
i=1, 2, 3, 4 (vehicle variants)	$i = 1, 2, 3, 4$(vehicle variants) $j = 1, 2, ..., 26$(failure modes)
$\lambda_i \sim gamma(a, b)$	$\lambda_{ij} \sim gamma(a, b)$
$\eta \sim beta(1, 1)$	$\rho_1 \sim beta(1, 1)$ $\rho_2 \sim beta(1, 1)$
$a \sim gamma(.001, .001)$	$a \sim gamma(.001, .001)$
$b \sim gamma(.001, .001)$	$b \sim gamma(.001, .001)$

DT and OT data. Figure 2 illustrates the results of the traditional analysis, the frequentist analysis,[5] and the Bayesian analysis. All three analyses use an exponential distribution to model the miles between failures, as discussed previously.

The mean miles between operational mission failures (MMBOMF) estimates reported in Figure 2 under the traditional analysis do not leverage the DT data or the relationships among the various types of Stryker vehicles. Notice that the CV vehicle stands out as potentially having an optimistically high MMBOMF of 8,494 miles. This estimate is based on a single failure and a combination of all the individual operating distances for each of the six CV vehicles. None of the six CV vehicles, however, used in OT traveled more than 2,000 miles. To claim that any one vehicle's MMBOMF is greater than 8,000 miles when no single vehicle traveled that far is questionable. Furthermore, if we consider that the estimate of MMBOMF in DT for the CV was less than 2,200 miles, we can conclude that it is highly unlikely that we would see such large improvements in the reliability between late DT and OT since no major changes were made to the system configuration. The MMBOMF estimate based on the traditional analysis approach is therefore highly suspect.

Confidence intervals for the FSV and the RV are also extremely wide because of the limited number of failures observed in OT. Because no failures were recorded for the MEV in OT, only a lower confidence bound can be estimated.

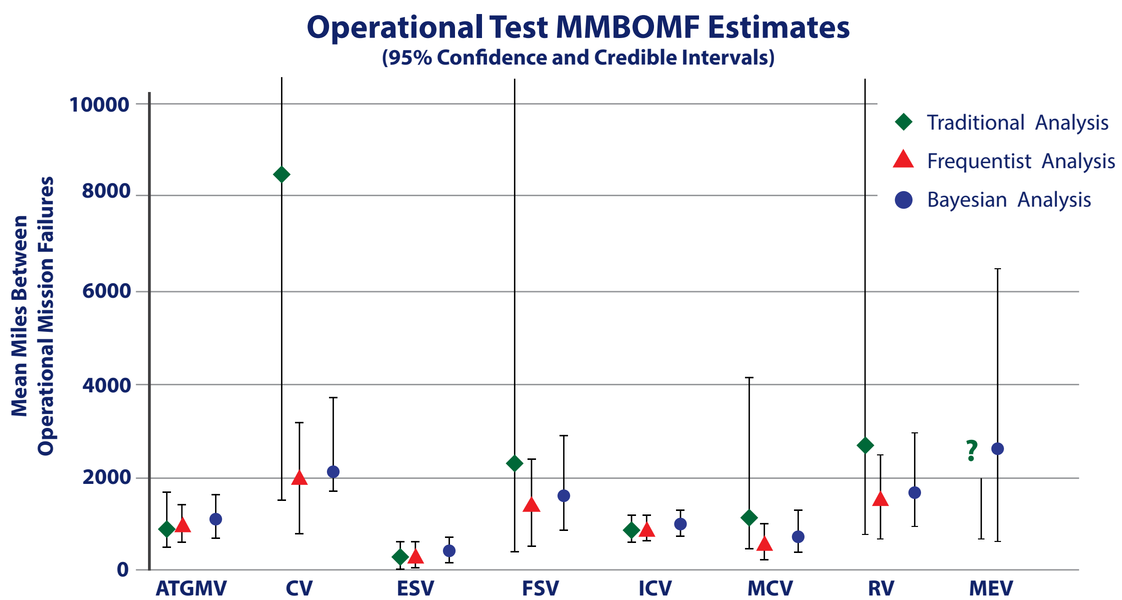

Figure 2. Stryker FoV: Comparisons of the OT MMBOMF Vehicle Variant Estimates for the Traditional Analysis, Frequentist Analysis, and Bayesian Analysis Using the Exponential Distribution

5 An exponential regression model was used. Test phase and vehicle variants were included as explanatory variables so that individual reliability estimates for each of the vehicles within each test phase could be estimated.

Using a statistical model to formally account for differences in performance across test phases and vehicle variant has a large practical impact on the reliability results. In Figure 2, the two model-based analyses (i.e., frequentist analysis and Bayesian analysis) provide a more realistic estimate of CV reliability and improve the overall precision of the estimates of system reliability for the vehicles that exhibited a small number of failures in OT. The tighter confidence intervals are obtained by leveraging the failure information from the other variants and DT data. One clear advantage of using the Bayesian analysis in this example is that we can obtain a point estimate for the reliability of the MEV. The reliability estimate for the MEV is driven by the information that we have for the seven other vehicles.

The Stryker example demonstrates that when we combine the available information across two test phases, the reliability estimates are more accurate and precise than estimates based solely on OT data. We also obtain inferences for vehicles on which no OT data are available. The analysis considers only OMFs since this analysis allows for a direct comparison between standard DoD analysis and the analysis that combines information across the DT and OT phases. However, further improvements in reliability estimates might be achieved by leveraging information from EFFs and/or failure modes. In the following example, we leverage information from OMFs and EFFs.

FUTURE COMBAT VEHICLE: ILLUSTRATION OF ANALYSIS AND RESULTS

For the notional future combat vehicle example, we assume very detailed failure information exists for the four vehicles tested in three test phases. In other words, the data for EFFs are available in addition to OMFs, and each EFF and OMF is attributed to a specific failure mode (e.g., brakes, fuel system, and suspension). The Bayesian model in Table 1 allows for a separate reliability estimate for each observed failure mode that arises across the test phases. Also, by using the information learned in the analysis about the individual failure modes, we can estimate the reliability for each vehicle. This reliability estimate provides a much richer source of information than the estimate derived in the equation on page 30, which simply takes the total number of miles driven by all four vehicles in each phase and divides by the total number of failures from the phase to determine reliability for the FoV.

In the Department of Defense, reliability requirements are typically written at the family level for these types of programs and in the language of OMFs. However, this analysis focuses at the vehicle level and includes all EFFs. By analyzing all EFFs and capitalizing on the information that is known about each of the failure modes, we are more likely to identify a larger portion of failures that cause system downtime, which will lead to greater improvements in reliability, availability, and maintainability and reduced operating and maintenance costs. Furthermore, by breaking out the failures by vehicle, we more accurately determine the reliability of the FoV.

Figure 3 shows the estimated mean miles between essential function failures (MMBEFF) for each vehicle across the three test phases. The statistical model links the reliability estimates through each of the three phases by estimating the reliability as a function of the successful fixes between phases. If a program is using the corrective action period to fix some fraction of the observed failure modes, then the MMBEFF across the three phases of test should increase. In fact, the model assumes that this increase must occur (see Table 1). As seen in Figure 3, for the four vehicles, the MMBEFF increases from around 50 miles to around 60 miles from Phase 1 to Phase 2 and then gains another 30 miles from Phase 2 to Phase 3.

Similar to the Stryker example, we investigate the gain over a traditional analysis. Figure 4 shows MMBEFF estimates and intervals for the four vehicles across the three test phases using the Bayesian hierarchical model and the traditional exponential analysis, separated by vehicle and phase. The Bayesian analysis always provides a tighter interval estimate, meaning those results are more certain and precise, which is a direct result of leveraging information from all vehicles and all phases of test. The Bayesian analysis also shows distinct growth for each of the four vehicles, while the traditional analysis reveals growth in reliability across phases for only two of the four vehicles: vehicle 3 and vehicle 4.

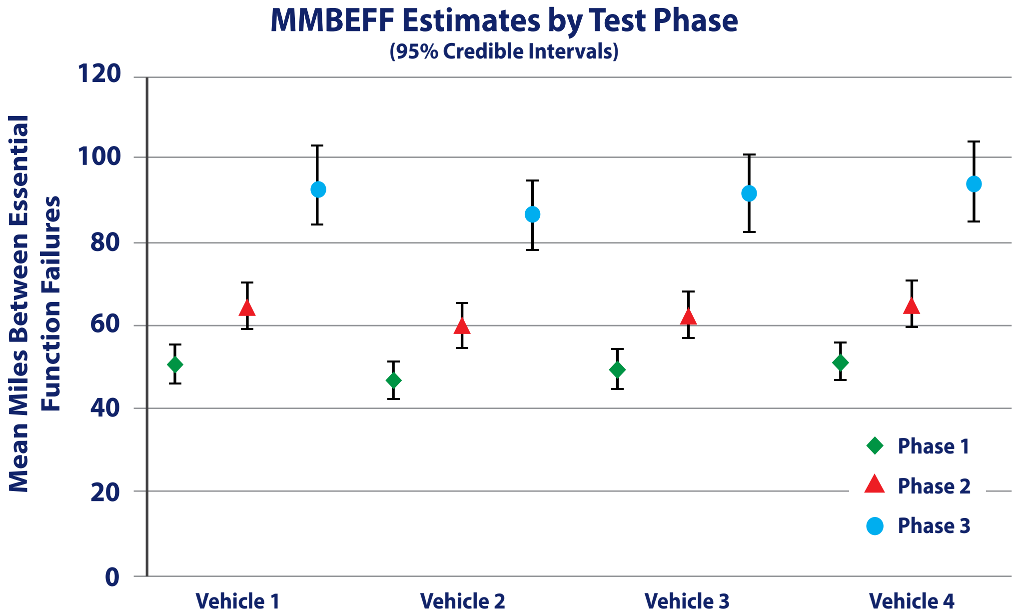

Figure 3. MMBEFF Estimates and 95 Percent Credible Intervals for Each Future Combat Vehicle and All Three Phases of Test

In Figure 4, the traditional analysis does not show any evidence of growth in the first two phases of testing for three of the four vehicles (only vehicle 3 shows some marginal growth). The Bayesian analysis assumes that growth will occur between each phase as a result of the model specification, so the results are more definitive with respect to growth between periods; however, this assumption is not required. Future sensitivity analyses of the results on the model specification will be important for understanding the influence of the model and assumptions used. Nevertheless, these results reveal the strength of these methods for analyzing data and capitalizing on all the data available to provide more accurate insight into system reliability over time.

CONCLUSIONS

The Bayesian approach to reliability analysis provides a formal framework to combine information from multiple sources and attain appropriate uncertainty quantification. The two examples discussed in this article illustrate the advantages of

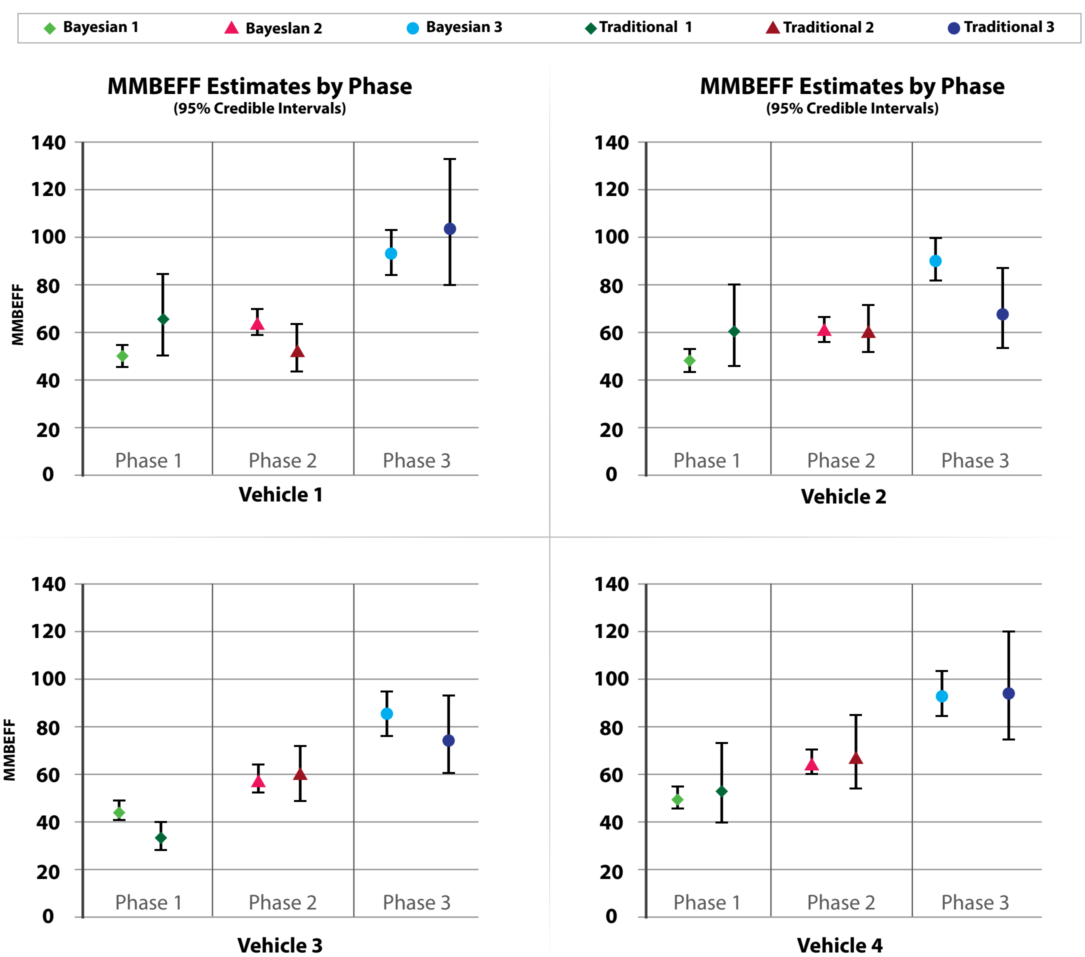

Figure 4. Comparisons of the MMBEFF for Four Vehicles Across the Three Phases of Test, for the Bayesian Analysis and Traditional Analysis Using the Exponential Distribution

using data from multiple phases of testing and leveraging data from systems with common infrastructures. The results are better estimates of system reliability and more precise inferences. Further improvements in reliability estimates are achieved by leveraging information from EFFs. By exploiting all available information and tools, we can obtain rich inferences for very complex problems.

Dr. Fronczyk is a Research Staff Member in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in statistics and stochastic modeling from the University of California, Santa Cruz.

Dr. Dickinson is a Research Staff Member in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in statistics from the Virginia Polytechnic Institute and State University.

Dr. Freeman is an Assistant Director in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in statistics from the Virginia Polytechnic Institute and State University.