Test and Evaluation for Reliability

Laura Freeman and Rebecca Dickinson

The Problem

Reliable systems cost less to operate, are more likely to be available when called upon, and have longer life spans. Unfortunately, we continue to observe systems that fail to meet reliability requirements.

Reliability is a key enabler of suitability and robust reliability leads to reduced life cycle costs.

IDA developed and presented reliability training to the DHS Office of Test and Evaluation (T&E). The organization requested this training after realizing that programs were focusing on availability metrics, when better test programs could be developed around reliability metrics. IDA's training provides information to assist the DHS T&E community in their understanding, review, and assessment of system reliability. We provide an overview of the reliability training we presented to DHS in this article.

The evaluation of system suitability in DHS typically focuses on three components: reliability, availability, and maintainability, often referred to as RAM:

- **Reliability.** The ability of a system to perform a required function under given operating and environmental conditions for a stated period of time
- **Availability.** The probability that the system is operating properly when needed for use
- **Maintainability.** The ability of an item to be retained in, or restored to, a specific condition within a given period of time when maintenance is performed.

For many DHS programs, availability is treated as the primary metric of interest (key performance parameter), and reliability a secondary metric (key system attribute). The focus in this article, however, is on the test and evaluation of reliability. Arguably, reliability is the most informative measure of the three because reliability failures depend on the context of the environment and inform the relevance of the other two measures. It can also be measured more credibly during system development than availability or maintainability. By improving reliability, we improve availability and minimize the impact of maintenance. Note that the definition of availability does not have a mission context; it is strictly a mathematical expression, which can mask underlying reliability problems. A system can achieve high availability despite having poor reliability. Unlike

availability, reliability is a direct expression of the likelihood that a system will complete a mission. What matters to system operators is not whether the system works when it is available, but that it works when it is needed.

Notably, a National Research Council report on reliability growth (National Research Council 2015) recommended that reliability be designated as a key performance parameter, making compliance contractually mandatory and helping to ensure that delivered systems are reliable. However, that recommendation has not yet been adopted.

Despite the importance of acquiring reliable systems, we continue to see systems that fail to meet reliability requirements. The 2015 IDA reliability assessment (Freeman et al. 2016) showed that only about 50 percent of systems under Department of Defense (DoD) oversight meet reliability requirements. This trend has been consistent over time and is continually highlighted by the Director, Operational Test and Evaluation (DOT&E) in the Annual Report to Congress on DoD systems (U.S. Department of Defense 2015; U.S. Department of Defense 2016).

The reasons for failure are complex. Case studies show that a lack of design for reliability effort during the design phase, unrealistic requirements, lack of contractual support, insufficient developmental test time, absence of or disagreement on reliability scoring procedures, and failure to correct significant reliability problems discovered in early testing all contribute to poor reliability outcomes.

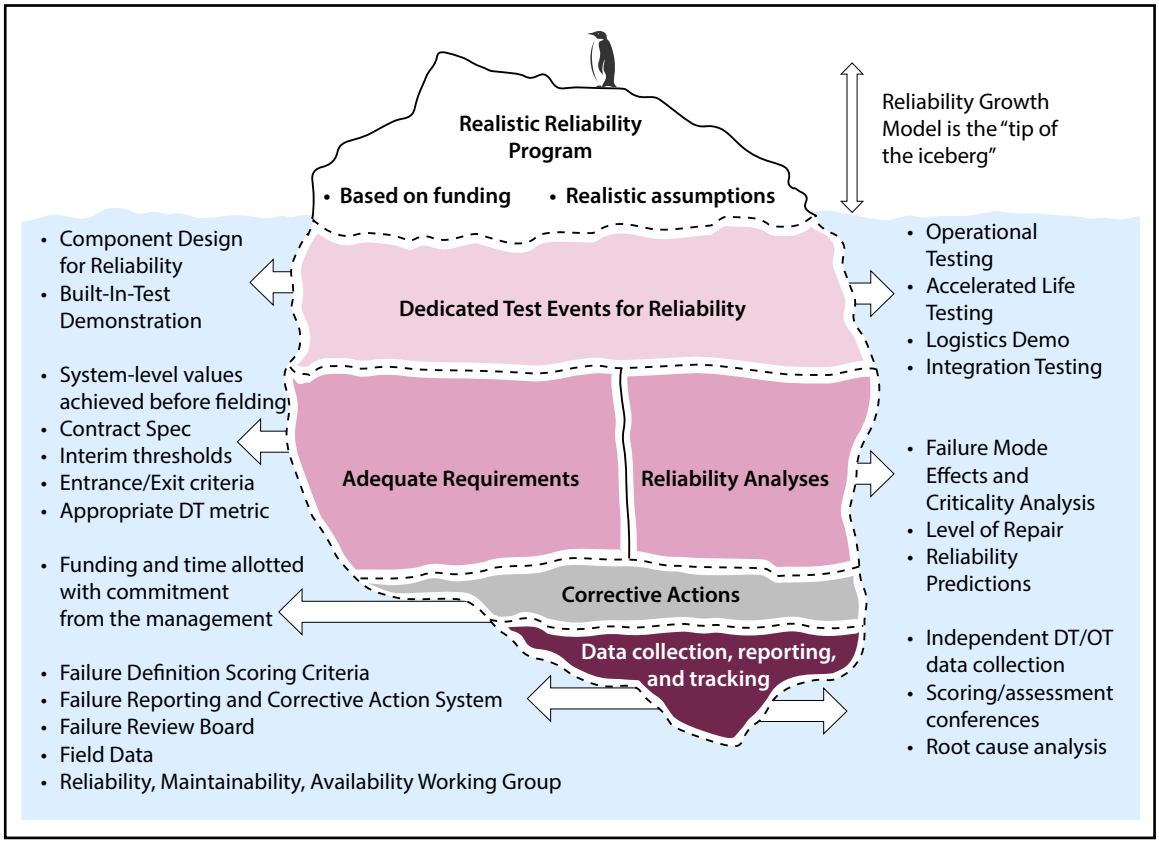
Figure 1 shows that a successful reliability program requires many levels of effort, beginning early in the program with writing adequate requirements.

To ensure success, it is important to understand all of the aspects of a good reliability program. As discussed below, IDA researchers have developed training that spans the full range of successful reliability program activities, including developing requirements, implementing a design for reliability program, and testing and evaluating reliability. We have also applied methods to assess reliability more efficiently. For example, IDA often leverages Bayesian methods for combining reliability data for systems with multiple test phases and for systems with common base platforms to maximize the information.

Defining Reliability Requirements

A first step toward producing reliable systems is to ensure that the requirements are appropriate. Appropriate requirements should be attainable, testable, and grounded in operational relevance:

- **Attainable.** Do similar technologies have comparable requirements? Is there adequate schedule time and funding to reach the requirement? Do the contracting documents contain a reliability specification?
- **Testable.** High requirements necessitate long tests. For example, it requires a much longer test to evaluate a requirement of 99 percent probability of completing a two-hour



Note: A well-run reliability program requires a dedicated engineering effort. Failure to take any piece of the iceberg seriously could cause the entire reliability program to "sink."

Figure 1. Successful Reliability Program

mission, compared to a requirement of 95 percent. Testers should discuss whether a 4 percent increase in probability of mission completion is meaningful.

- **Operational Relevance.** The requirement rationale should be based on what is required for the users to accomplish a mission in the anticipated operational conditions.

Requirements should also be linked explicitly to the cost of acquisition and sustainment over the lifetime of the system. While it may cost more to build reliable systems in the near term, the future savings potential is too great to ignore. As systems evolve, the requirements may need to be updated as the system engineering becomes more fully understood, but all changes in these requirements should be considered in the context of the mission impact.

It is also important to define failures and the scoring criteria to be used, early on in the program in a Failure Definition Scoring Criteria (FDSC). This process is essential for contractual verification at various intermediate system development points, but often is not done until much later in the program's lifecycle. Establishing consistent scoring criteria early on and for all phases of testing also makes it easier to combine

data analytically from different test phases to improve the precision of the estimated reliability parameters.

Requirements, contracting specifications, and reliability growth programs often focus only on a mission-level reliability requirement that includes only failures discovered during mission execution that result in an abort or termination of the mission in progress. A majority of failures that occur during testing, however, do not lead to mission aborts. Bell and Bearden (2014) note that reliability metrics limited to mission aborts are important, but exclude a large portion of failure modes that drive maintenance and cost and reduce system availability. A comprehensive reliability program should establish requirements on measures that include all failures of mission essential components that drive maintenance costs and degrade system availability, regardless of when the failure is discovered.

Design and Redesign for Reliability

Reliability must be designed into a system from its initial conceptualization. Finding failure modes and fixing them after system specifications are determined can provide a marginal improvement in reliability, but the largest gains are realized by designing the system with reliability as a key goal.

During the design and redesign stage, key engineering activities supporting a reliability growth program include the following:

- Allocating reliability to system components and subsystems
- Developing a reliability block diagram and predictions for completing system configurations
- Updating the FDSC
- Analyzing failure modes, mechanisms, and effects
- Refining system environmental loads and expected use profiles
- Dedicating test events for reliability (e.g., accelerated life testing, maintainability, and built-in test demonstrations).

In the early production of a system, reliability testing should shift from the subsystem level to the testing of the full system. It is essential to incorporate operational realism into the testing as early as possible to flesh out failure modes that will be discovered only in an operational environment. A test, analyze, fix, and test strategy should be used to identify and eliminate design weakness inherent to these intermediate system prototypes. A system's rate of growth generally depends on the following:

- The rate at which failure modes surface
- The turnaround time for analyzing and implementing corrective actions
- The fraction of the initial failure rate addressed by corrective actions (i.e., management strategy)
- The fix effectiveness factor—percent decrease in a failure mode due to a corrective action.

Implementing a design for reliability approach early in system

development is a key recommendation issued in a report by the Defense Science Board (U.S. Department of Defense 2008, 23–24):

> The single most important step necessary to correct high suitability failure rates is to ensure programs are formulated to execute a viable system engineering strategy from the beginning .... *No amount of testing will compensate for deficiencies in RAM [Reliability, Availability, Maintainability] program formulation* [emphasis added].

Resourcing for Reliability Test Events

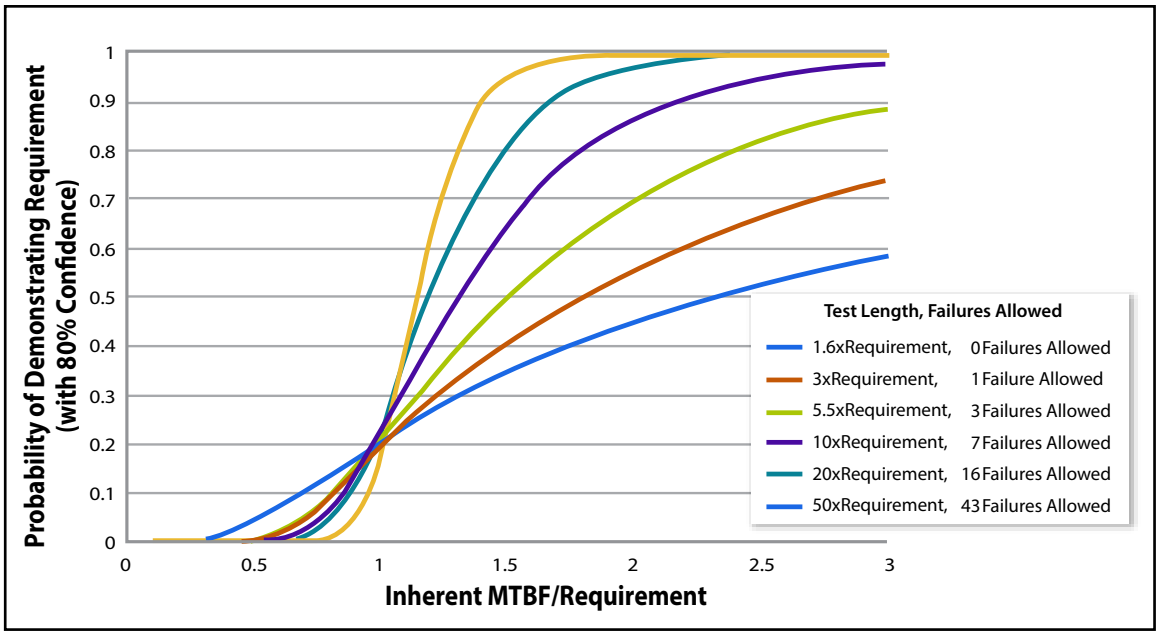
Test Length

A challenge in demonstrating whether a system meets its reliability requirement in operational testing is planning a long enough test. While tests are generally not scoped with respect to the reliability requirement, sufficient data should be captured throughout all test phases to determine the reliability of the system as it compares to the requirements.

To prove with statistical confidence that a system has achieved its reliability requirement, the observed failure rate for that system must be less than the requirement by some design margin. The size of that margin is determined by the inherent reliability of the system, as well as the precision of the estimated failure rate. Demonstrating with confidence that the threshold is met is a tradeoff between test length (longer tests allow for more precise estimates) and the underlying designed-in (inherent) reliability of the system.

Operating Characteristic (OC) curves are a helpful tool for determining whether test length is adequate for demonstrating the requirement. They describe the relationship between test lengths, requirements, and producer and consumer risk. Producer risk is the probability that a good system (above threshold reliability) will be rejected, which is a risk to the contractor. Consumer risk is the probability that a bad system (below threshold reliability) will be accepted, which is a risk to the Government. The curves are used to impute the underlying inherent reliability a system must achieve to demonstrate the requirement for a specified levels of producer risk and consumer risk.

If the inherent reliability of the system is close or equal to the reliability requirement, more testing will be needed to demonstrate the requirement with a high probability of success. This concept is illustrated in Figure 2, which shows a normalized presentation of several OC curves. In the construction of these curves, the consumer risk level is fixed at 20 percent (or 80 percent confidence). This means that a system with an inherent reliability equal to or below the requirement would have, at most, a 20 percent chance of demonstrating the requirement. If the system was designed to achieve a reliability twice that of the requirement, then a test duration of 10 times the requirement would provide a high probability (87 percent power) of the system successfully demonstrating the requirement in a test and a low risk of failing the test (13 percent producer risk). If the system was designed to



Note: OC curves are a useful tool for determining if a test period will be adequate. For a given test length, a system with a designed-in (inherent) reliability greater than that of the requirement has a higher probability of demonstrating the requirement than a system with an inherent reliability close to or equal to that of the requirement.

Figure 2. Normalized OC Curves

achieve a reliability 1.5 times that of the requirement, a test duration of 20 times the requirement would be necessary to provide a comparable level of producer risk.

The operational test duration for many systems is not long enough to demonstrate reliability requirements with statistical confidence. For systems with high reliability requirements, a greater emphasis must be placed on ensuring that the developer designs high reliability into the initial system from the beginning.

It may also be necessary to use test data from all available sources to make a reliability assessment. When system reliability is poor, even a short test might be adequate to prove that the system did not meet its reliability requirement.

Test Assets

Testing one system for 100 hours is not the same as testing 10 systems for 10 hours each. Testing numerous systems, each for a short time, prevents the surfacing of failures that would be observed only after the system has been exposed to a sufficient amount of testing, and testing only one system makes it impossible to observe variations in reliability that might occur between different systems of the same configuration. The number of assets required for a test depends primarily on the system under test, whether it is a single-use system (e.g., a disposable chemical agent detector), a repairable system (e.g., a new border patrol vehicle), or a one-off system (e.g., a new aircraft carrier). Test asset planning considerations should

include the following:

- How users will employ the system in operation (e.g., a representative unit might require five vehicles)
- Whether to test all variants of the system if there is more than one
- Whether additional assets are required to test under different environmental conditions
- Availability of assets due to cost constraints.

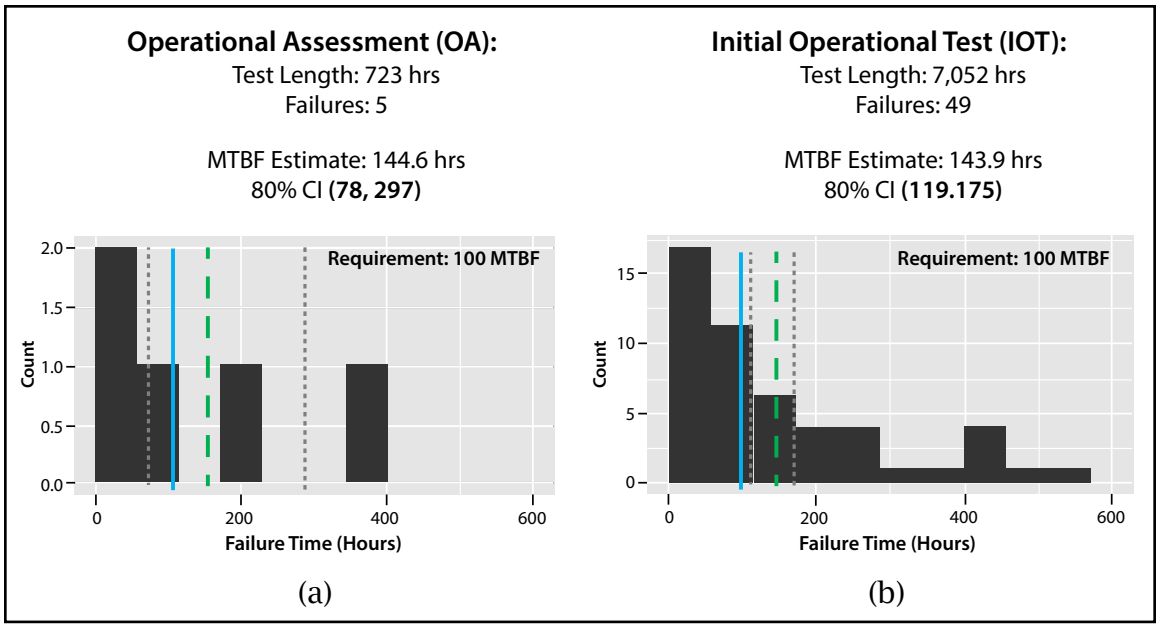
Monitoring, Evaluating, and Reporting Reliability

Reliability should be monitored and reported throughout the acquisition process to evaluate whether a program is on track to meeting its reliability requirements. It should not stop there; monitoring should continue for the duration of the system usage. During development testing, the system configuration typically changes as a result of corrective actions being made. A common monitoring approach is to compare demonstrated reliability to the anticipated reliability of the growth planning curve. If the analysis indicates that the system is not growing in accordance with the plan, it is important to update the growth strategy using more realistic inputs, consider whether additional resources/testing are necessary to reach goals and, if reliability is extremely poor, redesign the system.

During operational testing, the system configuration is usually fixed, and a primary evaluation goal is to determine whether the system meets its reliability requirement. When reporting a reliability estimate, such as a mean time between failures (MTBF), it is important to include the corresponding statistical confidence intervals. Confidence intervals permit an assessment of the certainty in a result, showing how sure we are about system reliability. Figure 3 highlights the importance of bounding the certainty. In this example, both versions of the system "demonstrated" the system MTBF requirement of 100 hours, but there is more information from one test than the other. From the Operational Assessment, we can state that the system demonstrated the requirement but not with statistical confidence. From the Initial Operational Test, we can state that the system met the requirement with statistical confidence.

There is no single appropriate way to analyze reliability, despite the common misconception that one should simply divide the test duration by the number of failures. Several areas of consideration to address when reporting on reliability are as follows:

- Is the system sufficiently reliable to conduct its mission?
- What was the demonstrated reliability (point estimate and confidence interval)?
- Did the system meet the requirement? Is it a statically significant difference? Is the difference meaningful in an operational context?
- How does the system's reliability compare to the legacy system? Did an upgrade improve reliability or did it degrade reliability?



Note: Confidence intervals quantify the certainty about a reliability estimate, such as the MTBF: (a) demonstrated requirement, but not with statistical confidence; (b) met requirement with statistical confidence.

Figure 3. Confidence Intervals

We noted earlier that it is not always possible or cost effective to collect all of the data on system reliability in a single test. For such cases, using a range of additional sources of relevant information may provide a better assessment of the system reliability. Integrating multiple sources of information, including component, subsystem, and full system data, as well as possible previous test data or subject matter expert opinions, to inform a reliability assessment is not trivial. The Bayesian paradigm is tailor made for this situation. It allows for the combination of multiple sources of data and variability to obtain more robust reliability estimates and uncertainty quantification. For recent examples and discussion on combining information using a Bayesian framework, we recommend Dickinson et al. (2015), Fronczyk and Freeman (2016), and Wilson and Fronczyk (2017).

Conclusion

Reliability is a key enabler of suitability and robust reliability leads to reduced life cycle costs. Although reliability design and growth testing can be expensive and require careful planning, the return on investment can also be high if properly executed. Using quantitative methods, IDA researchers have improved the estimation of the test durations required for confident evaluation of system reliability. IDA training is available for the community on topics spanning all aspects of reliability programs, including developing requirements, implementing a design for reliability program, and testing and evaluating reliability.

References

Bell, Jonathan L., and Steven D. Bearden. 2014. "Reliability Growth Planning Based on Essential Function Failures." Paper presented at the 2014 Annual Reliability and Maintainability Symposium (RAMS), Colorado Springs, CO, January 27–30.

Dickinson, Rebecca M., Laura J. Freeman, Bruce A. Simpson, and Alyson G. Wilson. 2015. "Statistical Methods for Combining Information: Stryker Family of Vehicles Reliability Case Study." *Journal of Quality Technology* 47 (4) (October): 400–415.

Freeman, Laura J., Allison L. Goodman, Matthew R. Avery, Jonathan L. Bell, Robert M. Hueckstaedt, Douglas A. Peek, and Max W. Roberts. 2016. 2015 *Reliability Assessment*. IDA Document D-8152. Alexandria, VA: Institute for Defense Analyses, November.

Fronczyk, K. M., and L. J. Freeman. 2016. "Improving Reliability Estimates with Bayesian Statistics." *The ITEA Journal of Test and Evaluation* 37 (4) (December).

National Research Council. 2015. *Reliability Growth: Enhancing Defense System Reliability*. Washington, DC: National Academies Press.

U.S. Department of Defense. 2008. *Report of the Defense Science Board Task Force on Developmental Test and Evaluation*. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, May.

U.S. Department of Defense. 2015. *FY 2015 Annual Report*. Washington, DC: Director, Operational Test and Evaluation.

U.S. Department of Defense. 2016. *FY 2016 Annual Report*. Washington, DC: Director, Operational Test and Evaluation.

Wilson, Alyson G., and Kassandra M. Fronczyk. 2017. "Bayesian Reliability: Combining Information." *Quality Engineering* 29 (1): 119–129.

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

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