Institution of MECHANICAL ENGINEERS
Journal of RISK AND RELIABILITY

Original Article

Proc IMechE Part O:
J Risk and Reliability
2015, Vol. 229(1) 46–51

DOI: 10.1177/1748006X14549394
pio.sagepub.com
SAGE

A simulation-based approach to determine the evaluation adequacy of system-of-systems operational test configurations

Robert N Tamburello[1] and Jeffrey W Herrmann[2]

Abstract
A system-of-systems is defined as "a set or arrangement of systems that results from independent systems integrated into a larger system that delivers unique capabilities." Given practical resource constraints, it is rare that the full-field configuration of the system-of-systems can be exercised during an operational reliability demonstration test. However, as we consider various potential operational test configurations for a given system-of-systems during the reliability test program planning process, it is critical to understand how testing a configuration that is smaller than the full-field configuration decreases the adequacy of the test by reducing the accuracy of the system-of-systems' reliability estimate that is based on the test results. Thus, it is useful to assess the adequacy of potential system-of-systems' operational test configurations before adopting one. We present a novel simulation-based method that can be employed to assess the adequacy of a given test configuration for any type of system-of-systems. To illustrate how this simulation-based method can be used to aid in the identification of the best alternative from among a group of potential operational test configuration alternatives, we include an example application using a notional air defense system-of-systems. Trade-offs with respect to cost, schedule, and accuracy are addressed within the context of this application.

Keywords
System-of-systems, reliability, operational test, test configuration, stochastic simulation, Monte Carlo, evaluation adequacy, test and evaluation, reliability test program planning

Date received: 11 January 2014; accepted: 6 August 2014

Introduction

Test and evaluation activities provide valuable information about the behavior of a system and its reliability, but the information is limited by the resources available for test and evaluation. Evaluation risk is the risk that the reliability evaluation will be insufficient to characterize the actual reliability behavior of a system that is being developed. This risk corresponds to the residual uncertainty (after the test and evaluation activities are complete) about the system's reliability in its intended operational environment. The evaluation risk associated with a reliability test program (RTP) can result from multiple causes, including the following.[1]

The RTP may include only a subset of the operational conditions. For example, for a military vehicle that has a "worldwide" mission profile, an RTP that does not include testing in cold regions would increase evaluation risk. An RTP that includes only a subset of a system's missions would also increase evaluation risk. Such cases can lead to an inaccurate evaluation of system reliability.

If the RTP includes insufficient test time, key failure modes may fail to appear, which increases evaluation risk. If the activities in the RTP collect limited information about the failures that occur, the evaluation risk increases. For example, state-of-the-art vehicle test systems can record system parameters such as engine speed, road speed, orientation (roll, pitch, and yaw), ambient external temperature, Global Positioning

---

[1] US Army Evaluation Center, Aberdeen Proving Ground, MD, USA
[2] Department of Mechanical Engineering, Institute for Systems Research, University of Maryland, College Park, MD, USA

Corresponding author:
Robert N Tamburello, US Army Evaluation Center, 2202 Aberdeen Boulevard, ATTN: 2nd Floor (ISMED), Aberdeen Proving Ground, MD 21005-5001, USA.
Email: robert.n.tamburello.civ@mail.mil

System (GPS) location, and many more for near-real-time or post-test analysis. Such information can be highly valuable during failure mode root cause analysis. Evaluation activities that do not collect this information make it difficult to understand the root causes and increase evaluation risk.

Per the Defense Acquisition Guidebook,[2] a system-of-systems (SoS) is defined as "a set or arrangement of systems that results from independent systems integrated into a larger system that delivers unique capabilities." A reliability demonstration test, also known as an Initial Operational Test (IOT), is an important step in evaluating the SoS reliability. Given practical resource constraints, however, it is rare that the full-field configuration of an SoS can be exercised during a reliability demonstration test. Thus, when considering various potential operational test configurations for a given SoS during the RTP planning process, it is critical to assess their adequacy before adopting one.

When evaluating an SoS, two additional situations can increase evaluation risk.[1] In the first, the "bootstrap" calculation of SoS reliability using only system-level failure data does not include failure modes due to the integration of or interactions between the systems in the SoS or "emergent" behaviors.[3,4] Second, for a large SoS, it may be infeasible to exercise the full-field configuration during the reliability demonstration test (or at any point during the course of the RTP). In this case, some subset of the SoS will be exercised; however, using this test configuration may increase the inaccuracy of the reliability estimate. In addition, considering only the minimum test configuration may obscure interactions due to scale, integration issues, and similar concerns.

In this context, an RTP is considered "adequate" if the associated evaluation risk is acceptable to the decision authority.

This article will consider the problem of using a test configuration that is not as large as the field configuration. The organization designing the RTP would like to know the relative evaluation risk of different test configurations in the presence of uncertainty about the system-level failure rates. (The need to test the interactions between the systems and uncover those failure modes is beyond the scope of this article.) The evaluation risk considered here is the likelihood of an inaccurate reliability estimate. In particular, this article presents a simulation-based approach for assessing the adequacy of a test configuration. This information can be used, along with information about the time and cost of the tests, to manage the evaluation risk and select the most appropriate test configuration. Note that the approach presented here does not calculate the economic value of the information,[5] but it could be extended in that direction.

The following work is related to estimating SoS reliability. For systems consisting of multiple non-repairable components configured in series, the Lloyd–Lipow (Lindstrom–Madden) methodology can be used to develop system-level reliability estimates.[6] For series-parallel systems, the Maximus method can be used to generate point estimates and confidence bounds on system-level reliability.[7] Coit[8] also provides a method for calculating confidence intervals for systems with active redundancy by using component-level failure data, and no assumption regarding the time-to-failure distributions for the components is required.

The next section of this article presents the problem statement, and the subsequent section presents the details of the simulation-based approach. After this, this article presents an illustrative example to demonstrate the utility of the approach. This article concludes with some discussion of future research directions.

Problem statement

In this problem, we are given information about the size of the SoS field configuration, possible test configurations, and the uncertainty in the system-level failure rates. Our goal is to understand the trade-off of test configuration size and cost versus accuracy of the SoS reliability estimate.

We assume that for reliability purposes, the SoS is a series of blocks, one block for each type of system. Each block is a k-out-of-n (KN) structure for which at least k out of the n systems must be operational in order for that block to be considered operational. Of course, all the blocks must be operational in order for the SoS to be considered operational.

Let L be the number of distinct types of systems in the SoS. For each system type i, $i = 1,..., L$, let n_i be the number of systems of each type in the field configuration of the SoS, and let k_i be the number of such systems that must be operational for the SoS to be operational.

For every type of system, the time-to-failure of a system of that type is exponentially distributed; let λ_i be the failure rate for a system of type i. The uncertainty about the failure rates is expressed as a set of A scenarios; in scenario a, the type i system failure rate equals λ_{ia}, $a = 1,..., A$.

In practice, various methods can be used to identify possible values for the system-level failure rates; the values selected should be vetted by subject matter experts (e.g. system engineers, testers and evaluators) to ensure the validity of the simulation results. Using a set of possible scenarios will enable practitioners to determine, for specific applications, whether the overarching results of this method are sensitive to the actual reliability of the SoS.

At time t, let $R_i(t)$ be the reliability of one system of type i at time t, let $R_{KN_i}(t)$ be the reliability of the ith KN structure in the SoS, and let $R_{SoS}(t)$ be the reliability of the SoS at time t[9,10]

$$
\begin{aligned}
R_i(t) &= e^{-\lambda_i t} \\
R_{KN_i}(t) &= P(X \geqslant k_i; n_i, R_i(t)) \\
&= \sum_{j=k_i}^{n_i} \binom{n_i}{j} [R_i(t)]^j [1 - R_i(t)]^{n_i - j} \\
R_{SoS}(t) &= \prod_{i=1}^{L} R_{KN_i}(t)
\end{aligned}
$$

Let T_0 represent the time at which the SoS reliability will be calculated. There are B possible test configurations. For $b = 1,..., B$, let m_{ib} represent the number of systems of type i that will be tested in test configuration b, and let T_b represent the amount of time spent testing these systems. It will be useful if one of the possible test configurations is the field configuration; although testing that configuration may be infeasible, it can be useful to include it as a benchmark.

The algorithm presented in the next section evaluates one test configuration in one scenario. The results across all the test configurations and scenarios can be used to compare the test configurations and make an informed decision about which test configuration to adopt.

Approach

In order to assess the adequacy of one test configuration in one scenario, the algorithm presented in this section will conduct a number of trials. Each trial will simulate the test results, estimate the system failure rates, and calculate the corresponding SoS reliability estimate.

General assumptions

1. Individual systems are assumed to fail independently (all system states are independent).
2. Failures at the system-level occur according to a homogeneous Poisson process. Hence, the time between failures for each individual system in the SoS is assumed to follow an exponential distribution.
3. Individual systems within each k-out-of-n structure, that is, of the same type, are identical and share the same underlying constant Poisson failure rate.
4. For the SoS, if an individual system fails during operation, no repair is attempted (i.e. the SoS will operate continuously until an SoS-level abort occurs).
5. The SoS can be modeled as a series of KN structures, one for each type of system.

Model formulation

Let λ_i be the common Poisson failure rate for each system of type i in this scenario.

Let m_i represent the number of systems of type i in this operational test configuration of the SoS.
Let T represent the length of the operational event during which the SoS test configuration will be exercised.
Let Q represent the desired number of simulation trials.

Given the systems' failure rates and the operational test configuration of the SoS, the following algorithm generates the average relative absolute error of the SoS reliability estimate by simulating the test results of that operational test configuration.

Algorithm

1. Calculate the a priori value for the SoS reliability $R_{SoS}(T_0)$.
2. For $\delta = 1,..., Q$, perform Steps 3–5 (this constitutes a trial).
3. Simulate an operational test of duration T that exercises the test configuration to obtain system-level reliability data as follows:
   (a) For $i = 1,..., L$, and $j = 1,..., m_i$, draw a random number of failures F_{ij} for each individual system from the Poisson distribution with mean $\lambda_i T$.
   (b) For $i = 1,..., L$, calculate point estimates of the mean time between failure (MTBF) for system type i as $\hat{\theta}_i = (m_i T)/(\sum_{j=1}^{m_i} F_{ij})$.
4. For each $i = 1,..., L$, based on the estimated MTBF, estimate the reliability of a type i system and then calculate the SoS reliability estimate for this trial

$$
\begin{aligned}
R_i(T_0) &= e^{-\frac{T_0}{\hat{\theta}_i}} \\
R_{KN_i}(T_0) &= P(X \geqslant k_i; n_i, R_i(T_0)) \\
&= \sum_{j=k_i}^{n_i} \binom{n_i}{j} [R_i(T_0)]^j [1 - R_i(T_0)]^{n_i - j} \\
R_{SoS}(T_0) &= \prod_{i=1}^{L} R_{KN_i}(T_0)
\end{aligned}
$$

5. Determine the relative absolute error for this trial $|R_{sos}(T_0) - R_{est}(T_0)|/R_{sos}(T_0)$.
6. Calculate the average relative absolute error associated with this SoS test configuration and scenario across all Q trials.

Illustrative example

To illustrate the utility of the proposed method, this section describes a hypothetical but realistic example of an SoS whose size is based on an Army missile defense system. The reliability numbers are only examples. This example illustrates the essential trade-off: increasing the test effort (by using a larger test configuration and more test time) can improve the expected accuracy of the test

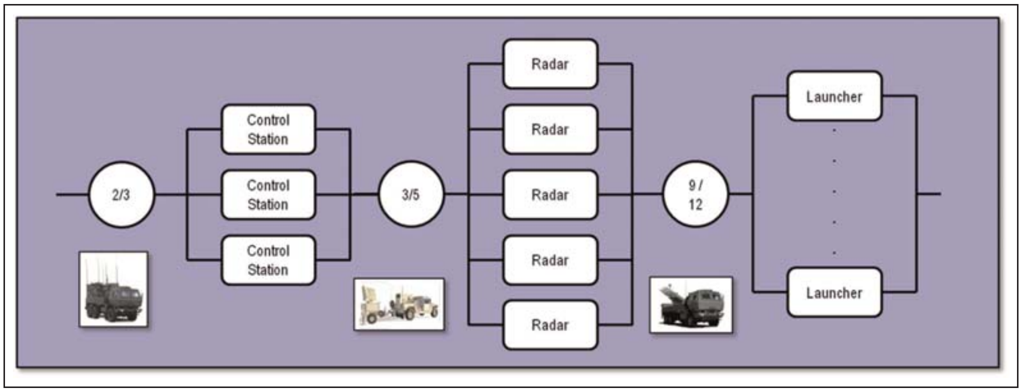

Figure 1. Reliability block diagram for the full field configuration of the SoS.

Table 1. Possible system-level failure rates and the associated SoS reliability.

Scenario	λ_1	λ_2	λ_3	R_{sos} (24 h)
1	0.0048	0.0073	0.0048	0.9039
2	0.0062	0.0089	0.0053	0.8556

Table 2. Test configuration cases (the test lengths are in hours).

Test configuration case	Test configuration	Operational test length (h)
1	1-1-1	100
2	1-1-1	300
3	1-1-1	500
4	1-1-1	1000
5	2-2-2	100
6	2-2-2	300
7	2-2-2	500
8	2-2-2	1000
9	2-3-6	100
10	2-3-6	300
11	2-3-6	500
12	2-3-6	1000

results and also increases cost and the time required to complete the tests.

For the scenarios that we consider, SoS reliability is defined as the probability that the SoS will complete a 24-h mission without incurring an SoS-level abort. The SoS is operational if and only if at least two of the three control stations are operating, at least three of the five radars are operating, and at least 9 of the 12 launchers are operating, as shown in Figure 1. The SoS reliability requirement is chosen to be 90%, that is, $P(T > 24\,\mathrm{h}) \geqslant 0.90$.

For this example, we assume that the system-level failure rates are uncertain. For each, however, a range of possible values are given. We will assess the adequacy of different test configurations for two scenarios: in Scenario 1, every system-level failure rate is at its lower bound; in Scenario 2, every system-level failure rate is at its upper bound. These rates and the corresponding actual SoS reliability at 24 h are given in Table 1. Thus, the SoS meets the reliability requirement in Scenario 1 (the low failure rates), but it does not in Scenario 2 (the high failure rates).

Table 2 lists the 12 test configuration cases that we considered. Each case has a configuration that lists the number of systems (of each type) to test and the length of the test in hours.

Execution of simulation cases

We will use the terms *test configuration case* (or simply *case*) and *trial* for our simulation. Each trial requires the specification of the SoS design, SoS-level reliability, and the test configuration. For each case, we used the algorithm presented earlier to determine the average relative absolute error associated with that case and how often the SoS reliability estimate was at least 0.90, the requirement. In Scenario 1, the actual SoS reliability is greater than 0.90, so this indicated how often the reliability estimate was “correct.” In Scenario 2, the actual SoS reliability is less than 0.90, so this indicated how often the reliability estimate was “incorrect.”

Simulation results

For this example, the algorithm presented earlier was run for 1000 trials for each case in each scenario. Tables 3 and 4 show the results for all 12 cases and both scenarios. The full-field configuration was also assessed as a benchmark. For Scenario 2, the frequency of “correct” estimation is calculated from the frequency of “incorrect” estimation.

As expected, exercising more distinct test assets for longer periods of time improves the accuracy of the SoS reliability estimates (and the likelihood of a “correct” estimate) because the system-level MTBF point estimates are more accurate. In Scenario 1, because the

Table 3. Simulation results for Scenario 1 (the SoS reliability equals 90.43%).

Configuration	Test time *T* (h)	Average relative absolute error	Frequency of "correct" estimation
1-1-1	100	0.2298	0.327
	300	0.1187	0.389
	500	0.0906	0.383
	1000	0.0578	0.440
2-2-2	100	0.1633	0.338
	300	0.0791	0.416
	500	0.0593	0.439
	1000	0.0395	0.477
2-3-6	100	0.1022	0.375
	300	0.0532	0.460
	500	0.0406	0.454
	1000	0.0281	0.500
3-5-12	100	0.0743	0.417
	300	0.0400	0.469
	500	0.0293	0.481
	1000	0.0212	0.509

Table 4. Simulation results for Scenario 2 (the SoS reliability equals 85.74%).

Configuration	Test time *T* (h)	Average relative absolute error	Frequency of "correct" estimation
1-1-1	100	0.2625	0.778
	300	0.1414	0.737
	500	0.1079	0.761
	1000	0.0744	0.807
2-2-2	100	0.1810	0.745
	300	0.1004	0.769
	500	0.0748	0.825
	1000	0.0504	0.873
2-3-6	100	0.1260	0.769
	300	0.0703	0.800
	500	0.0556	0.844
	1000	0.0363	0.906
3-5-12	100	0.0969	0.778
	300	0.0546	0.843
	500	0.0398	0.902
	1000	0.0281	0.960

SoS reliability is just greater than 90%, it is more difficult to get a "correct" estimate than it is in Scenario 2.

During RTP planning, the emphasis should not be placed on the particular average relative absolute error values obtained for each trial via simulation; rather, practitioners should consider the marginal improvement that may be realized by increasing the operational test duration or augmenting the SoS test configuration. The sensitivity of average relative absolute error to the test configuration and operational test duration is a tool to compare distinct options, ideally, while concurrently considering the cost to execute each option. In addition, using this approach during the RTP planning stage will enable practitioners to document the rationale for selecting a particular option and communicate that rationale to decision authorities.

The iron triangle of cost, schedule, and accuracy

The results in Tables 3 and 4 show clearly how increasing the test effort improves the accuracy of the reliability estimate. In practice, larger test configurations and longer tests can increase costs by tens of millions of dollars.

For instance, in this example, suppose that the cost of the smallest test configuration (1-1-1) is US$12 million, the cost of the second test configuration (2-2-2) is US$24 million, and the cost of the largest test configuration (2-3-6) is US$36 million. Furthermore, the cost of every 100 h of test time is US$1.25 million. Then, case 1 costs US$13.25 million and case 12 costs US$48.5 million. Figure 2 plots the cost and the average relative absolute error for all 12 cases. Moreover, based on the test facility's work schedule, every 100 h of test time takes 1.25 weeks of testing. The longest test times would require 12.5 weeks of testing.

Now, if the test organization will recommend only an operational test plan for which the average relative absolute error is expected to be less than or equal to 10% in both scenarios, only cases 4, 7, 8, 10, 11, and 12 should be considered. Among these six cases, case 4 has the least cost (US$24.5 million).

By establishing the maximum acceptable threshold for the average relative absolute error associated with SoS reliability, practitioners are certifying that any SoS test configuration for which the average relative absolute error exceeds that threshold is inadequate. Without going through this process, or at least a variant of this process, practitioners struggle to effectively articulate this information to program managers. Concordantly, program managers are less likely to allocate the necessary resources to conduct an adequate operational test for a complex SoS.

Concluding remarks

The novel simulation-based approach described herein is an appropriate technique to assess the evaluation adequacy of a given SoS test configuration. This methodology offers the potential for practitioners to understand how various operational test lengths and SoS test configurations affect the accuracy of the SoS reliability estimate. Given existing data on the systems that constitute the SoS (or similar systems, based on subject matter expertise and engineering judgment), this approach provides information for selecting the test configuration that balances the objectives of cost, time, and adequacy. This approach can be extended to a variety of similar situations, and the underlying assumptions invoked herein may be relaxed to handle more complex situations.

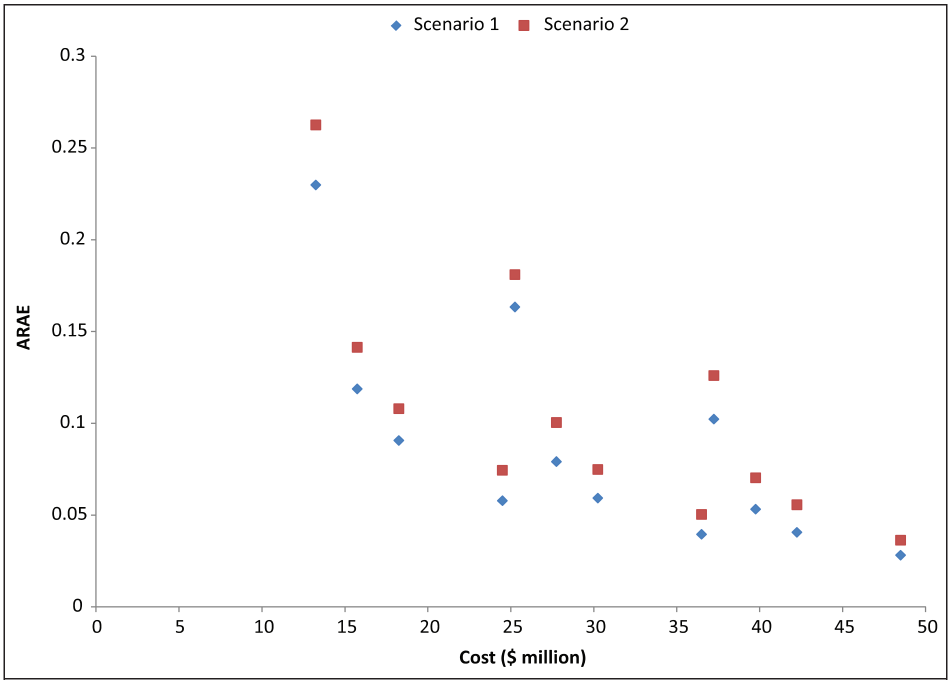

Figure 2. Trade-off between ARAE and cost for all 12 cases in both scenarios.
ARAE: average relative absolute error.

Assessing the reliability of an SoS involves issues other than how the inaccuracy of the system-level reliability estimates affects the SoS reliability estimate. There may be failure modes due to the integration of or interactions between the systems in the SoS and "emergent" behaviors. These must be considered when creating an RTP as well.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Tamburello RN. *Evaluation-focused reliability test program planning methodology*. PhD Dissertation, University of Maryland, College Park, MD, 2013.
2. US Department of Defense. *Systems engineering guide for system of systems*. Washington, DC: US Department of Defense, 2008.
3. Corning PA. The re-emergence of "emergence": a venerable concept in search of a theory. *Complexity* 2002; 7(6): 18–30.
4. Goldstein J. Emergence as a construct: history and issues. *Emergence* 1999; 1(1): 49–72.
5. Lawrence DB. *The economic value of information*. New York: Springer, 1999.
6. Lloyd DK and Lipow M. *Reliability: management, methods, and mathematics*. Englewood Cliffs, NJ: Prentice Hall, 1962.
7. Ellner PM, Easterling R, Kniss J, et al. *Handbook for the calculation of lower statistical confidence bounds on system reliability*. Ad-Hoc Methodology Working Group on Nuclear Weapons Reliability Assessment, Report #1, February 1980. McLean, VA: Maximus, Inc.
8. Coit DW. System-reliability confidence-intervals for complex-systems with estimated component reliability. *IEEE T Reliab* 1997; 46(4): 487–493.
9. Bain LJ and Engelhardt M. *Statistical analysis of reliability and life-testing models: theory and methods*. 2nd ed. New York: Marcel Dekker, 1991.
10. Modarres M, Kaminskiy M and Krivtsov V. *Reliability engineering and risk analysis: a practical guide*. 2nd ed. New York: CRC Press, 2009.