INSTITUTE FOR DEFENSE ANALYSES

IDA Support to DTE&A Initiative: Improving the Technical Rigor in DTE&A Assessments

John S. Hong, Project Leader
James M. Gilmore
Lance E. Hancock
Olivia S. Sun

September 2021
IDA Publication D-22772
Log: H 2021-000293

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-19-D-0001, Project AX-1-3100 "DTE&A Initiative," for the AX / Dir, DTE&A / Director, Developmental Test Evaluation and Assessments. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments

The authors would like to thank IDA committee, Dr. Stephen M. Ouellette (chair), Dr. Jonathan L. Bell, Dr. Leon R. Hirsch, Dr. Kyle A. Morrison, and Mr. Christopher A. Martin for providing technical review of this effort.

For More Information

John S. Hong, Project Leader
jhong@ida.org, (703) 845-2564

Stephen M. Ouellette, Director, SED
souellet@ida.org, 703-845-2443

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-22772

IDA Support to DTE&A Initiative: Improving the Technical Rigor in DTE&A Assessments

John S. Hong, Project Leader

James M. Gilmore

Lance E. Hancock

Olivia S. Sun

Executive Summary and Introduction

At the beginning of FY 2021, the Director, Developmental Test, Evaluation, and Assessment (D,DTE&A) decided to pursue 18 initiatives to improve the effectiveness of the office. The initiatives span a broad array of topics including both policy and practice. The latter category includes, but was not limited to: greater use of statistical techniques for planning tests and evaluating their results; methods and capabilities needed for testing of autonomous systems enabled by artificial intelligence; greater use of modeling and simulation in the planning and conduct of test and evaluation; approaches enabling earlier involvement of the developmental test community in the Defense Department's acquisition programs (also known as Shift Left); and increasing the rigorous technical content of Developmental Test, Evaluation, and Assessments (DTE&A)'s evaluations and assessments.

DTE&A requested IDA develop a set of best practices that would help the office achieve the latter initiative; i.e., to prepare rigorous assessments. In response, this paper provides IDA's suggestions for best practices DTE&A could employ consistent with achieving the goals DTE&A has stated for the initiative, which are the following:

- To help increase the technical rigor in DTE&A assessments to better influence and inform senior leaders making critical acquisition decisions.
- To develop specific best practices for conducting independent quantitative analyses of information collected during tests and assessing system technical performance and integration maturity in a mission context.

The best practices discussed in this paper include, but are not limited to, the use of specific analytical approaches (both quantitative and qualitative), as well as the collection and assessment of measures and metrics. DTE&A assessments already accomplished employ a number of these practices, at least in part. For example, the Office's analyses of reliability test data have employed statistical techniques. The best practices provide practical guidance, examples, and references that should reinforce and help expand the benefits of incorporating rigor in DTE&A's assessments. We group the best practices into five categories: (1) General Rigor-Related Considerations; (2) Considering Operational Context and Properly Characterizing Test Results; (3) Using Statistical Methods; (4) Assessing and Using Software Modeling and Simulation; and, (5) Assessing Risks and Readiness. Grouped using these categories, the 14 specific best practices discussed in this paper include the following:

General Rigor-Related Considerations

1. General considerations for improving technical rigor when conducting assessments

Considering Operational Context and Properly Characterizing Test Results

1. Incorporating operational context and including sensitivity analyses in assessments
2. Soliciting user feedback on the implications of test results and analyses
3. Characterizing the implications of test results

Using Statistical Methods

1. Using statistical techniques to analyze data
2. Using statistical methods to analyze reliability
3. Using statistical techniques to assess test sufficiency

Assessing and Using Software and Modeling and Simulation

1. Choosing software development metrics
2. Incorporating technical rigor in assessments of modeling and simulation (M&S)
3. Conducting M&S using operational vignettes

Assessing Risks and Readiness

1. Considering uncertainties explicitly when assessing schedule risks
2. Using the Defense Technical Risk Assessment Methodology (DTRAM)
3. Using objective criteria to assess technology readiness
4. Assessing manufacturing readiness

These topics were chosen by soliciting suggestions from staff in the Institute for Defense Analyses (IDA) who have expertise in testing of weapon programs. The list is not exhaustive---no single set of topics could be. But, it does reflect the judgment of the authors of practices DTE&A could adopt and/or expand the use of that would increase the technical rigor incorporated in the Office's assessments, thereby increasing the value of those assessments to decision-makers.

Applying these best practices to good effect will require data held by and obtained from program offices and the user community, as well as engagement with the associated staff. It will also require time and resources allocated by program offices that always face constraints on both. Program delays and cost increases are most often caused by

performance problems discovered during developmental and operational testing.[1] Therefore, to the extent these practices and associated assessments help identify and mitigate problems earlier than has often been the case, they will have been worthwhile. Overall, the use of these practices should help DTE&A support the Department's statement of the fundamental purpose of Test and Evaluation (T&E), which is to: "enable the DoD to acquire systems that support the warfighter in accomplishing their mission."[2]

[1] Freeman L., et al, "Reasons Behind Program Delays," Institute for Defense Analyses, D-5289, October 2014

[2] Department of Defense Instruction (DoDI) 5000.89, "Test and Evaluation," November 19, 2020.

(This page is intentionally blank.)

Contents

Tables

Figures

(This page is intentionally blank.)

1. General Rigor-Related Considerations

A. General Considerations for Incorporating Technical Rigor when Conducting Assessments.

1. Introduction and Background

Technical rigor has been defined as "the application of precise and exacting standards…to better understand and draw conclusions…based on careful consideration or investigation."[3] In the physical sciences, technical rigor is reflected in the application of the canonical "scientific method" in experiments and analyses. In many situations; however, rigorous, repeatable experimentation and analytical procedures are not possible (or applicable), due to uncontrollable variations in, for example, experimental and data collection procedures and/or data analysis methodologies. This is often the case with the testing and data used to support DTE&A assessments - test procedures and test execution, instrumentation and data collection methods, quality and quantity of data, are some of the factors that often vary from one test event to another. This lack of consistent, repeatable methods can contribute to the risk of so-called "shallow analysis," in which an analysis is of inadequate depth for a given situation, such as supporting a major acquisition milestone (MS) decision.

2. Improving analytical and technical rigor

An analogous problem has been encountered in intelligence analyses, where the qualitative nature of the data sources and the information being analyzed often obscure a decision-maker's ability to determine whether an analysis is sufficiently rigorous to provide, say, actionable intelligence. Consequently, methodologies have been developed for use in intelligence analysis that are applicable more generally to situations in which an exact, replicable analytical procedure is not appropriate.[4] The methodology recommends eight attributes that contribute to assuring a rigorous analysis (see Table 1).

---

[3] Military Operations Research Society. Terms of Reference in "Bringing Analytical Rigor to Joint Warfighting Experimentation, July 2006.

[4] Zelik et al., "Understanding Rigor in Information Analysis," Proceedings of the Eighth International NDM Conference, June 2007.

Table 1: Attributes of Analytical Rigor

Attribute	Description	Low rigor analysis	High rigor analysis
Hypothesis exploration	Describes the extent to which multiple hypothesis were considered in explaining data	Minimal weighing of alternatives	Incorporates multiple perspectives to identify the best, most probable explanations
Information search	Describes depth and breadth of the search process used in collecting data	Does not go beyond routine and readily available data sources	Attempts to exhaustively explore all data potentially available
Information validation	Details the level at which information sources are corroborated and cross-validated	Little effort made to use converging evidence to verify source accuracy	Includes systematic approach for verifying information
Stance analysis	Evaluation of data with the goal of identifying the perspective of the source and placing it into broader context of understanding	Noticeable bias in a source	Research into source backgrounds with intent of gaining more understanding of how their perspective might influence their stance
Sensitivity analysis	The extent to which the analyst considers and understands the assumptions and limitations of their analysis	Appropriate and valid explanations on a surface level	Consideration of the strength of explanations if individual supporting sources were to prove invalid
Specialist collaboration	The degree to which the analysis incorporates the perspective of domain experts into the assessment	Little effort to seek out expertise	Effort to incorporate experts in key content areas of analysis
Information synthesis	How far beyond simply collecting and listing data the analysis goes	Relevant information simply compiled in a unified form	Extracted and integrated information with a thorough consideration of diverse interpretations of relevant data
Explanation critique	Captures how many different perspectives were incorporated in examining the primary hypotheses	Little use of other analysts to give input on explanation quality	Peers and experts have examined chain of reasoning and identified strengths and weaknesses

Source: Zelik (2007).

3. General best practices for performing DTE&A assessments

The summary of attributes displayed in Table 1 suggests DTE&A, when judging the sufficiency of the rigor incorporated in its assessments, gauge the extent to which the following approaches have been employed:

- **Consider multiple hypotheses when explaining data, and incorporate multiple perspectives to identify the best, most probable explanations**. Multiple factors usually contribute to a particular assessment outcome. For example, poor system performance in a test may be due to hardware or software deficiencies, operator error, or adverse environmental conditions. The assessment should attempt to investigate these multiple factors and present a rationale, based on analysis, for the most likely explanations for the assessed outcome.
- **Explore all data potentially available in the relevant sample space**. Rather than rely on a single set of data, analyses should include as much relevant data from as many sources as feasible. Examples include data from previous tests and assessments, data on other similar systems, and M&S (M&S) (including digital M&S and hardware- and software-in-the-loop). As discussed in the section entitled: Using Statistical Methods to Analyze Reliability, rigorous approaches are available for combining data to generate performance estimates.
- **Incorporate a systematic approach for verifying data sources and cross-validating data from different sources**. Analyses should attempt to incorporate multiple sources of data into the assessment, which can either help strengthen the evidence supporting a particular conclusion or highlight inconsistencies in the data. For example, test data from a single event can be compared with pre-test predictions obtained through M&S, engineering assessments, and/or data from different test events. The assessment should attempt to explain any discrepancies from different data sources, if applicable.
- **Attempt to understand and explain any sources of data bias and how they may influence the assessment**. For example, data provided by the contractor or tests conducted solely by contractor representatives and staff may suggest better system performance than is possible under more realistic test circumstances; or test schedules proposed by a schedule-driven program may be overly optimistic. Assessments should describe the sources of information and data, and explain how their context could affect conclusions.
- **Explain the assumptions and limitations of the analysis, and discuss how individual explanations or conclusions may be affected by any weaknesses in supporting data sources**. An example of this might be an assessment of system performance based on data from one single test event, or data from M&S

that was based on certain modeling assumptions. The analysis should describe the characteristics of the data and the potential impact those data characteristics and assumptions may have on the assessment (e.g., "The conclusion that the system is effective at X mission may be invalidated if the modeling assumption is off by more than Y percent.")

- **Incorporate subject matter expertise in analyses where necessary and appropriate**. Analysts should solicit and include input from the appropriate subject matter experts (SMEs), which may include testers, system engineers, M&S developers and users, and operational end users. Such input can provide context to analyses, explain data inconsistencies or confirm test or analysis results. Additionally, analysts should incorporate a review process in which relevant peers and SMEs provide input on the quality of the assessment and examine the strengths and weaknesses of the chain of reasoning.

Not every assessment will incorporate or fulfill all eight attributes displayed in Table 1 in whole or even in part. Although necessarily accomplished subjectively, gauging the extent to which an assessment employs these attributes can provide DTE&A a basis for deciding whether the assessment is defensible, as well as explicit rationale for conducting that defense.

2. Considering Operational Context and Properly Characterizing Test Results

A. Incorporating Operational Context and Including Sensitivity Analyses in Assessments

1. Introduction and Background

Incorporating operational context and realism into early testing is critical for understanding system performance in its intended operational environments.[5] Similarly, operational context should be incorporated in DTE&A assessments, including developmental test assessments (DTAs) and independent technical risk assessments (ITRAs), as follows:

- When performing first-order analyses of test results and technical risks, to provide the decision-maker with an understanding of the linkage between the system's key technical parameters/specifications and its ability to fulfill its operational missions.
- When performing sensitivity analyses to determine the implications for a system's ability to fulfill its mission of uncertainties in meeting technical specifications.[6]

2. Incorporate operational context when analyzing test results

Analyses of test results should consider how the system's prescribed technical specifications and requirements relate to the system's overall mission performance, and seek to answer the following questions:

1. Do the specifications/requirements accurately reflect the system's mission requirements?
2. What are the mission impacts if the system does not meet the specification/requirement?

Technical specifications may not always capture all aspects of the system's operational mission scenarios and vignettes. Thus, evaluating system performance against

[5] See the section entitled Incorporating Operational Context and Realism in "IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy," Institute for Defense Analyses paper D-22771, October 2021.

[6] The uncertainties will have been informed by risk analyses and test results.

the technical specifications is necessary, but not sufficient for assessing the system's ability to perform its intended operational missions. In seeking to answer the first question, the assessment should describe whether meeting the specifications/requirements as written implies operational mission success or failure, preferably with a statistically significant degree of confidence, but at least using quantitative indicators and metrics. In other words, if the test data indicate the system has or has not met certain specifications and/or requirements, do those results suggest that the system can or cannot successfully accomplish its operational missions?

The assessment should first describe the relationship between the system's operational scenarios and environments, and the key technical parameters and metrics that drive mission performance in those scenarios and environments.[7] Then, compare those key technical parameters and metrics with the assessed technical specifications and requirements and discuss the implications for mission success. For example: "Sensor X met all of its technical specifications for detection and identification range against Target Y. These results indicate that the system has the capability to successfully conduct Mission Scenario 1 in an environment where Target Y is the primary threat."

To address the second question, the assessment should discuss the specifications/requirements that were not met, and the potential operational implications of those test results. Using the example above: If, on the other hand, the system's sensor does not meet the requirement for detection range, how is the system's ability to accomplish its mission(s) affected? What modifications to concepts of operations (CONOPs) and/or tactics, techniques, and procedures (TTPs) might be needed as a result?

DTE&A should include operational end users in the discussions of the operational context and mission impacts of the test results, and incorporate their input into the assessment.[8]

3. Incorporate operational context in risk assessments

Although the discussion above is focused on assessments and analyses of test results, a similar approach should be taken when assessing technical risks, such as for ITRAs.

Early in development, it is possible that data are limited because the specifics of design alternatives, including hardware, software, and other attributes of the system, are not mature. In these cases, it is advisable to focus on the required capabilities, system

---

[7] See the section entitled Incorporating Operational Context and Realism in "IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy, Institute for Defense Analyses paper D-22771, October 2021.

[8] See the section entitled: Solicit User Feedback on the Implications of Test Results and Analyses.

requirements, and mission needs of the system. Using the information available in CONOPs, TTPs, and requirements documents, the risk assessment should aim to answer:

1. What technological capabilities are required to achieve the missions envisioned in the CONOPs and otherwise satisfy mission needs and requirements?
2. Based on current technology development efforts and the demonstrated capabilities of current and previous analogous systems or development efforts, what technological capabilities are feasible?

The process for using mission needs and requirements to assess technical risks is analogous to the process for analyzing test results described in the previous section, with some key differences. As a first step, the mission needs and requirements of the system need to be decomposed into key mission tasks and functions. If the system design is finalized, the mission tasks and functions can be traced to the system components and/or subsystems required to perform those tasks and functions. The risk assessment then focuses on comparing the technical capabilities of those subsystems/components with the state of current technology and the specifications/requirements.

If the system design has not been finalized, or if multiple designs are under consideration, the assessment should consider the range of design options and technological capabilities that could be used to satisfy the mission requirements. Risks can be assessed by comparing the technological capabilities needed to those that were actually achieved in analogous programs or are judged likely to emerge from current development activities. Considering a broad range of analogous programs and associated technological capabilities will help identify the uncertainty inherent in assessing those risks. Modeling and simulation (M&S), if it is applicable, can help determine the potential effects of the uncertainties in technical risks on mission capabilities. As the program proceeds and those uncertainties are reduced, M&S-generated results should be updated.

4. Perform quantitative risk assessments

As data become available, use them to assess program risks and use quantitative metrics whenever possible. For example, system performance can be assessed based on demonstrated results from test events (including at the subsystem and component levels), hardware-in-the-loop, and M&S. Bayesian techniques are available to combine information from multiple test venues and events, as well as from engineering assessments, to generate estimates of both performance and uncertainty.[9]

Risk analysis involves evaluating both the likelihood a risk will occur and the consequence(s) if it occurs. Risk consequences should be assessed based on quantifiable

[9] See the section entitled: Use Statistical Methods to Analyze Reliability.

impacts to the program where applicable (e.g., mission impacts resulting from not meeting a key performance parameter (KPP) or a key system attribute (KSA)).[10]

The likelihood a risk will occur is often an uncertain qualitative estimate based on subject matter expertise. However, quantitative data may be available in some cases. For example, statistical modeling techniques, such as Monte Carlo simulations (MCS), can be used to quantitatively estimate the likelihood of different consequences. Data from other similar programs or historical trends can also be analyzed to provide additional insights. The use of such approaches can enable the uncertainties inherent in any assessment of risk to be presented, which should be more informative than a point estimate for determining useful risk mitigation strategies.

5. Performing sensitivity analyses

When assessing technical metrics in the context of the system's intended mission and operational environment, the system's performance should initially be examined relative to the most likely conditions the system will face. However, this initial set of conditions can be limited, and may include only a subset of the operational conditions and environments the system will encounter in its operational usage. Furthermore, due to schedule, cost, and/or safety constraints, it can be the case that even the most comprehensive test and evaluation efforts are unable to cover the full spectrum of a system's potential operational conditions and environments. Thus, assessments of test results should include a sensitivity analysis that covers the reasonable ranges of critical technical and performance parameters that could be realized, conducted as soon as feasible and updated as appropriate.

The same principles also apply to assessments of technical risk. In particular, during early risk assessments, where the system design is not yet finalized and test data are limited, it is critical for stakeholders to understand how the system's performance may vary in relation to variations and uncertainties in its operating environment, achievable technological capabilities, and system design decisions.

In essence, a sensitivity analysis seeks to answer the following questions:[11]

- How would the results change if we use other assumptions?
  - E.g., How would the system perform, and what are the mission impacts, if the conditions and environments are different from those assumed and that were explicitly tested?

---

[10] See the section entitled: Using the Defense Technical Risk Assessment Methodology (DTRAM).

[11] Saltelli, A., et al, "How to avoid a perfunctory sensitivity analysis," *Environmental Modelling and Software*, 15 May 2010.

- How sure are we of the assumptions?
  - E.g., How sure are we that the tested conditions and environments are representative of the range of conditions/environments the system will encounter throughout its operational life cycle?

A simple example of such an analysis might aim to determine how a system's demonstrated radar detection range for a particular threat would change if the threat's radar cross section were reduced relative to the tested condition.

The process for conducting the sensitivity analyses consists of the following key steps:

1. Identify potential uncertainties and variabilities in the system's threat environment and other key aspects of mission context. Consider uncertainties spanning a reasonable range of what is technically and operationally possible, not just what would be projected based on what was directly observed. Such aspects include:
   - Threat technical capabilities
   - Threat CONOPs and TTPs
   - Numbers and types of threats
   - System CONOPs and TTPs
   - System's physical environment and operational conditions (e.g., terrain, radio frequency background, weather and visibility, etc.)
2. Identify key system tasks and functions, and how those tasks and functions would change with variations in the aspects identified in 1.
3. Based on 2, identify key components and subsystems and their associated key technical parameters/specifications that drive operational mission performance in the identified mission scenarios/vignettes.
4. Relate the uncertainties identified in 1 to potential variations in the system's technical parameters identified in 3. The traceability between key aspects of system's threat and mission environment and its key technical parameters should have been completed during requirements development, development of the request for proposals and associated contract specifications, and/or during early test planning.
5. Using the relationships identified in 4., determine significant changes in mission performance that could occur due to potential variations in the system's key technical parameters. The potential variations will be informed initially by engineering analyses and subsequently by the test data that accumulates as the

program proceeds. Use all available data to conduct these analyses, as well as the available and appropriate analytical tools and M&S (more details below).

The assessment should describe in detail the information used to support each of the process steps described above, the data sources, the outcome and implications of the sensitivity analyses, including potential implications for the decision(s) the assessment is intended to support.

6. Sensitivity analysis methods

Numerous approaches exist to conduct sensitivity analyses. The specific type of sensitivity analysis and analysis method used in an assessment will depend on many factors, including: the system under test, the technical parameters and performance metrics of interest, the operational conditions/environments and their potential variations and uncertainties, and the data and M&S tools available. DTE&A should engage with system engineers, testers, and analysts to determine the appropriate type of analysis needed to support their assessment.

Although the details may vary, most sensitivity analyses adhere to the following general procedures:[12]

- Quantify the uncertainty in each input
- Identify the model output to be analyzed
- Run the model a number of times using a statistical process (e.g., design of experiments (DOE))
- Using the resulting model outputs, calculate the sensitivity measures of interest

Although rigorous, quantitative approaches to sensitivity analyses are preferred, in instances where data and/or analytical modeling tools are limited or not applicable, such analyses can also be performed qualitatively by incorporating, for example, input from SMEs.

7. Data sources

The sources of data used to support the sensitivity analyses as described above include:

- Test data from the system under test and its components and subsystems

[12] Norton, "An introduction to sensitivity assessment of simulation models," *Environmental Modelling and Software,* 15 April 2015.

- M&S, including full-scale system simulations such as data obtained from a systems integration laboratory (SIL) and hardware-in-the-loop (HWIL) facility, and analytical and/or statistical models
- Test data from other systems, including legacy platforms and other similar or closely related systems
- Input from SMEs, including system engineers, testers, and operational end users[13]

The evaluation of technical metrics should be conducted in the context of the system's intended mission and operational environment. For example, a description of measured radar detection ranges should discuss whether those demonstrated ranges are sufficient for the system to accomplish its mission against its primary threats and targets. Additional discussion of best practices for the incorporation of operational context in DTE&A assessments can be found in the following sections of this paper:[14]

- Soliciting User Feedback on the Implications of Test Results and Analyses;
- Characterizing the Implications of Test Results;
- Conducting Modeling and Simulation (M&S) using Operational Vignettes.

B. Soliciting User Feedback on the Implications of Test Results and Analyses

1. Introduction and Background

It is important for any test assessment to provide adequate operational context of the test results, as discussed in (See the section entitled Incorporating Operational Context and Realism in "IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy, Institute for Defense Analyses paper D-22771, October 2021). An important element of this is incorporating feedback from operational end users who understand the system's operational usage, and are best-suited to provide mission context. When analyzing test results and their implications for meeting specifications and requirements, user inputs should be solicited regarding the operational significance of any potential shortfalls or exceedances. In particular, one should solicit input regarding potential changes to CONOPs

[13] When quantitative data are limited, system engineering technical reviews can provide opportunities to obtain information from contractors on system performance and enabling technologies. These meetings often include technical discussions of the highest risk or technically challenging aspects of programs. They can assist in gaining a more complete understanding into risks associated with the programs, as can progress in technology demonstrations conducted by the contractors or the government. The section entitled: Using the Defense Technical Risk Assessment Methodology (DTRAM), provides additional discussion.

[14] These section titles are cross-referenced links.

that could occur to compensate for shortfalls or take advantage of better-than-expected performance.

Such feedback can provide the operational context for decision-makers to assess system performance and to select courses of action based on the information provided (e.g., choose to invest additional time and resources to address a mission-critical requirement shortfall, or to proceed as-is with a shortfall that is operationally insignificant, or proceed as-is with a shortfall that is operationally significant while acknowledging risk). This information will also be useful to operational testers in planning operational scenarios and vignettes for the operational test (OT), and to operational end users who must fully understand the capabilities and limitations of the system they are operating. The following are key pieces of information that should be included in discussions with the operational user and discussed in assessments of test results:

- Impact to mission
- Implications for CONOPs/TTPs
- Test artifacts
- Implications for operational testing

2. Impact to mission

This discussion should focus on how the assessed requirement affects the user's ability to accomplish the intended operational mission(s). For example, in a situation where a technical metric fails to demonstrate the specification value, possible outcomes include:

- No or low mission impact – e.g., Mission can be accomplished as intended without any significant performance degradation
- Moderate mission impact – e.g., Mission can be accomplished with an acceptable level of performance degradation, or with some workarounds in place
- High mission impact – e.g., Mission can be accomplished, but with severe performance degradation
- Prevents mission accomplishment – e.g., Mission cannot be accomplished

The objective of this discussion should not be solely to provide a mission impact rating (i.e., low, moderate, high), but rather to describe in detail how the mission may be impacted, and/or which specific workarounds are needed to successfully conduct the mission as intended.

3. Implications for Concept of Operations (CONOPs) and Tactics, Techniques and Procedures (TTPs)

Similar to the discussion on mission impact, the assessment should describe how any demonstrated requirements shortfalls or exceedances may affect, or necessitate modifications to, existing CONOPs and/or TTPs, or inform the development of new CONOPs/TTPs. For example, if a sensor demonstrates a target detection range short of its specification value, how might the procedures for identifying and prosecuting the target be affected?

4. Test artifacts

It is important for the assessment to put the test results in the proper context by clearly describing the conduct of the test and any test limitations that may have affected the test execution and/or results. For example, were there safety constraints that limited the testing of certain system capabilities? Equally important are any test artifacts that may have been introduced by the test environment, test procedures, or system operators. In particular, the analysis should seek soldier input in describing any potential effects introduced by contractor personnel or field service representatives (FSR) operating the system, and how the test outcome may have been different if soldiers were operating the system.

5. Implications for operational test

Based on the mission impacts and implications for CONOPs/TTPs resulting from the demonstrated system performance, the assessment should highlight areas of emphasis to be evaluated during OT. Specifically, the impacts to mission scenarios and vignettes, and CONOPs and/or TTPs should be evaluated during OT under operationally realistic conditions and with operational users.

6. Example

The following example provides a notional template to present the information described above in an assessment. The example is illustrated using a hypothetical radar of a ground-based air defense platform.

Test results for radar detection range specification testing: Radar detection range for a fixed-wing (FW) aircraft target has a specification value of X, but the demonstrated value is $0.8X \pm 0.01X$.

Impact to mission: For the counter-fixed wing mission, the demonstrated detection range shortfall has a moderate impact on the mission. The reduced detection range will require threat targets to fly closer to the defended area before the radar can pick it up, which will decrease the amount of time the system has to generate a firing solution and intercept the threat. The shortfall of 0.2X in detection range does not prevent mission accomplishment.

As a potential workaround, an off-board radar can detect fixed wing targets at X range and can interoperate with the system under test to pass the radar information to the interceptors, with sufficient time and range to defeat the threat.

Implications for Concept of Operations (CONOPs) and Tactics, Techniques, and Procedures (TTPs): If the system's on-board radar is to be used, no changes to CONOPs or TTPs are necessary, but operators may need to be trained to proceed through the kill-chain sequence more expediently, to accommodate the shorter threat engagement timeline. If the off-board radar is needed, and operators will need to be trained in its operation.

Test artifacts: The contractor FSR operating the radar was familiar with the system. The FW target flew predictable flight patterns, so after the first radar detection, the FSR knew to look for detections around the same range, which an operational user is unlikely to be able to do. This potentially skewed the test results, and system performance under operational conditions would likely be less than 0.8X.

Implications for operational testing (OT): The operational impact of the shortfall in detection range against fixed-wing targets should be examined during OT. The FW target should fly operationally realistic mission profiles at ranges that will challenge the radar system under test, as determined by a statistical test design. Particular emphasis should be placed on evaluating the time required for the operator to proceed through the entire kill-chain, from target detection to defeat, and the ranges at which the target is defeated. The ability of the off-board radar to interoperate with the system under test, and user operation of both platforms, should be evaluated in operational scenarios as well.

C. Characterizing the Implications of Test Results

1. Introduction and Background

Properly characterizing the implications of test results in assessments can help decision-makers understand how well the data represent actual system performance, and how much confidence to place in the data reported in the assessments. The primary objective of such characterizations is to provide an objective assessment of system performance using the available data and information.

2. Data characterization

Insufficient data quantity and quality are common challenges for assessments. For example, tests involving live missile firings are often limited in number due to test resource (and safety) constraints. Or, some tests may have extensive and complex instrumentation requirements and/or complicated data collection processes that make it difficult to capture all of the relevant data throughout the entire test event. This is a common occurrence

during, for example, ballistic vulnerability tests, where some of the test instrumentation is often destroyed as an unintended consequence of the test itself.

Assessments should distinguish between metrics/requirements that were not met due to insufficient (or statistically insignificant) data, versus metrics/requirements not met due to poor system performance. This is especially important in situations where the test data are inherently limited in quantity (e.g., live missile firings). Assessments should also identify whether the data collected and made available to support the assessment are sufficient. Data gaps, and potential means to fill those gaps, should be identified as well. Details on how to assess test sufficiency and the statistical significance of test data are described in Best Practices for Using Statistical Techniques to Assess Test Sufficiency.

For situations in which the test data are found to be statistically insufficient, additional insight on the system performance can be obtained by supplementing the test data with other sources of information. These sources can include the following:

- M&S results
- Data from testing on other, comparable systems
- Data from subsystem and/or component testing of the current system under test
- SME input from system engineers, testers, and/or operational end users

3. System characterization

Prior to OT, the system under test will likely not be in its final production-representative configuration, and will often undergo hardware and software changes as system deficiencies are discovered and corrected, and the system design is refined and finalized. Since the system configuration is likely to change from one test event to another throughout contractor and developmental testing, it is important for assessments to accurately characterize the system configuration for each test event, and to discuss the potential effects of the tested system configuration on the test data reported in the assessment.

Assessments should describe the system hardware and software configuration(s) used in the test, and identify major differences between the tested configuration(s) and the expected production- representative configuration and the implications of those differences for interpreting test results. This is especially important for rapid prototype and rapid fielding programs, where the system configuration can change quickly during the course of the test period. If system performance deficiencies are identified during the test, the assessment should distinguish between deficiencies for which there exist planned upgrades and/or fixes, and those deficiencies for which fixes have not been identified.

For example, if a system using a non-production representative software version demonstrates poor performance during an early test event, the assessment should describe

the known limitations and/or deficiencies of the tested software version, and how those limitations/deficiencies contributed to the system's performance. The assessment should also describe whether and how any planned software upgrades will or will not address the identified performance shortfalls.

3. Using Statistical Methods

A. Using Statistical Techniques to Analyze Data

1. Introduction and Background

Many statistical methods are available to choose from for analyzing test data. Empirical models produce objective conclusions based on the observed data. Parametric regression models maximize information gained from test data, while non-parametric methods provide a robust assessment of the data free from model assumptions. Bayesian methods provide the means for integrating sources of information in addition to the data from a particular test. These statistical techniques for analyzing the results of testing have several benefits, including the following:

- Quantitative estimation of the uncertainty in the results. Because testing always results in a finite set of data, there is uncertainty associated with using those data to estimate system performance. Statistical methods enable that uncertainty to be quantified.
- Quantitative estimation of the risk of drawing incorrect conclusions regarding system performance. Use of statistical techniques enables quantification of the probability of concluding a system does not satisfy performance requirements when it actually does (Type 2 error or manufacturer's risk), as well as probability the system meets the requirements when it actually does not (Type 1 error or government's risk).
- Ability to rigorously combine information from multiple test events and engineering assessments, thereby reducing uncertainty in estimates of system performance.
- Ability to combine information from multiple variants or configurations of a system under development to better use collected test data.
- Ability to generate models that can be used, with the appropriate care, to predict performance under conditions not explicitly tested.
- Ability to rigorously compare predictions of M&S with test data to support verification, validation, and accreditation (VV&A).

The construction of confidence bounds quantifying uncertainty, Type 1 and Type 2 errors, and Bayesian techniques enabling information from multiple test events and other sources to be combined are discussed in the section entitled: Using Statistical Methods to Analyze Reliability, as well as in "IDA Support to DTE&A Initiative: "Shift Left" Baseline

Framework Strategy."[15] Although focused on reliability, the uses of the statistical methods described in that section and other paper are applicable to planning other types of tests and analyzing their results. This section focuses on generating statistical models and using statistical techniques for VV&A of M&S.

2. Consider using statistical models to analyze data

Fitting a statistical model, such as linear regression, is one way to make good use of the available test data. Statistical models enable inferences to be made across a wide range of conditions using information from the full set of data simultaneously.[16] The models can be used for prediction and inference about the response variable (i.e., system performance) or on parameter estimates (the conditions and other factors affecting performance significantly). The models enable conditions that have an impact on system performance to be identified, as well as the conditions under which the system's specifications and requirements are met (or not). They also enable quantitative estimates of the uncertainties in both the parameters and predictions to be generated.

Linear Models[17,18]

A linear model is of the form:

$$y = X\beta + \epsilon,$$

where:

- X is a $n\ x\ p$ matrix comprised of n rows of the values of the p factors affecting performance (such as network load, processor load, temperature, range to radiofrequency emitter, etc.) associated with each of the n test data points;
- y is the $n\ x\ 1$ vector of test data (or response variables such as time to detect, track accuracy, distance of impact from aim-point, etc.)
- β is the $p\ x\ 1$ vector of model parameters; and
- ϵ is the normally distributed error with $variance(\epsilon) = \sigma^2 I_n$, where I_n is the $n\ x\ n$ identity matrix and σ^2 is the overall variance in the responses.

---

[15] Institute for Defense Analyses paper D-22771, October 2021. See the section entitled: Reliability and Test Planning.

[16] The introductory discussion is adapted from Thomas, D. et al., "Statistical Methods for Defense Testing," Institute for Defense Analyses, NS D-8893, December 2017.

[17] Thomas, D. et al., "Statistical Methods for Defense Testing," Institute for Defense Analyses, NS D-8893, December 2017.

[18] See Searle, S. et al., *Linear Models*, John Wiley and Sons, 2016 for a comprehensive discussion of linear models.

The model is linear in the parameters, β, but not necessarily in the factors, X, which can be quadratic or higher order polynomials. Maximum likelihood estimation is used to determine the parameters:[19]

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Prediction. A linear model can be used to predict what would have been the system's performance (i.e., the test results) if the tests had been conducted under different conditions; i.e., with different values of the factors X, than associated with the actual tests. The predicted values of the responses, $\hat{y}$, for the values of the factors, $\hat{X}$, for which predicted performance is to be estimated are:

$$\hat{y} = \hat{X}\hat{\beta},$$

where $\hat{\beta}$ is a vector containing the model parameters determined using maximum likelihood estimation. The standard error, SE, in the prediction y_0 for any single $1\ x\ p$ set of factors x_0 is:

$$SE(y_0(x_0)) = \sqrt{\frac{{y_0}^2 - \hat{\beta}^T {x_0}^T y_0}{n-p} x_0 (X^T X)^{-1} {x_0}^T}.$$

Although the level of significance of the factors can be of interest (e.g., whether range to an electronic warfare threat emitter is a significant factor affecting system performance) decisions on defense programs can often involve understanding system performance across the full range conditions under which it could be used and under which the system is expected to satisfy its requirements. Such predictions can be made within the set of conditions (i.e., factors) examined during testing, in which case they are interpolations of the test data. When predictions are made outside the range of test conditions, they are extrapolations, and questions then arise as to whether the linear model (or, indeed any model) remains valid. Careful judgments should be made when performing extrapolations, informed by sensitivity analyses exploring the model's behavior outside the test conditions. Of course, given that the tested conditions often do not include the full set under which the system could be employed, extrapolations are often of interest.

Categorical test conditions. Test data can be collected under different environmental conditions or conditions of employment; i.e., under different categories of conditions. Examples of such categories include day/night, vehicle type or variant, presence or absence

[19] See, for example, "Lecture 6: The Method of Maximum Likelihood for Simple Linear Regression," available at https://www.cs.princeton.edu/courses/archive/fall18/cos324/files/mle-regression.pdf, accessed April 28, 2021.

of jamming or electronic countermeasures, operating mode, etc. Linear models can handle these categorical (vice continuous) factors via the use of dummy variables; e.g.,

$$x_1 = \begin{cases} 1, if\ wheeled \\ 0, if\ tracked \end{cases}.$$

See Thomas (2017) for additional discussion.

Lognormal data. Linear models assume test data are normally distributed. However, testing of defense systems can produce data on performance (or response variables) that improve from one design iteration to the next, and that can even improve within one test event. Examples of such data include detection times, detection ranges, miss distances, and target location errors. These data are right-skewed and therefore not normally distributed.

Nonetheless, if the variable T follows a lognormal distribution, the transformed variable Y = log(T) is normally distributed, and a linear model can still be used to analyze the transformed data. Estimates for the mean of the untransformed data, as well as the associated confidence intervals for predictions generated using a statistical model of the transformed data can also be calculated (see Thomas (2017)).

Generalized Linear Models (GLM)

GLM can be used to analyze test situations involving random effects, systematic components, and the linkage between them. The random effects allow for any test data that follow a normal, or any other, exponential distribution. The systematic component is a linear, non-random function (or predictor) of the model parameters, and the link function relates the mean of the distribution of random effects to the linear predictor. Logistic regression (also used in machine learning) is a type of GLM.[20] Generalized Linear Mixed Models (GLMM) further extend linear models to situations involving systematic variation across test events as well as random variation and systematic effects within specific test events.[21] Among many purposes, such models can be used to analyze binary pass/fail test results such as those that occur in ballistic missile testing (the incoming re-entry vehicle is either destroyed or not) or torpedo testing (the incoming torpedo, which has a point detonation fuse, either impacts its target or not).

3. Consider using statistical methods to employ test data for M&S VV&A

The VV&A conducted should account for the model's intended use; the quantities of interest for evaluation (response variables); the range of input conditions for the model; the

[20] See, for example, Hosmer, D., et al., *Applied Logistic Regression*, Second Edition, John Wiley and Sons, 2013.

[21] See Myers, R., et al., *Generalized Linear Models with Applications in Engineering and the Sciences* Second Edition, John Wiley and Sons, 2010.

range of input conditions over which test data are available; and the acceptability criteria (the allowable differences between the model and actual test data). [22] Thus, the ideal process for M&S VV&A would proceed as follows:

1. Develop the intended use statement.
2. Identify the response variables or measures.
3. Determine the factors that are expected to affect the response variable(s) or that are required for operational evaluation.
4. Determine the acceptability criteria.
5. Estimate the quantity of data required to assess the uncertainty within the acceptability criteria.
6. Iterate the Model-Test-Model loop until desired model fidelity is achieved.
7. Verify that the final instance of the simulation accurately represents the intended conceptual model (verification process).
8. Determine differences between the model and real-world data for the acceptability criteria determined for each response variable using the appropriate statistical methods.
9. Identify the acceptability of the model or simulation for the intended use.

Wojton (2019) provides detailed discussion of the statistical methods that can be employed during each of the above nine steps in the ideal VV&A process. This section will highlight the salient points regarding the application of statistical methods to conducting the last two steps. Unfortunately, in many real-world cases, testers seeking to use M&S data are presented with the model itself (i.e., the ability to request that the model owners perform certain runs generating output) and the actual test data, with little, if any, participation in the model's development and the other activities conducted during the first seven process steps.[23]

Statistical validation techniques that account for possible effects due to factors are preferred over one-sample averages or roll-ups across conditions. If the data enable creation of a statistical model containing factors such as those discussed above, that approach can be used to compare the test data and simulated output. In all cases, even if no factors are identified and a one-sample approach is taken, the uncertainty about the

---

[22] This discussion is adapted from Wojton, H. et al., "Handbook on Statistical Design &Analysis Techniques for Modelling & Simulation Validation," Institute for Defense Analyses, NS D-10455, February 2019.

[23] This is not to say that earlier involvement by the test community in the conduct of the earlier steps of the ideal VV&A process should not be attempted.

difference between live data and simulated output should be quantified. Hypothesis tests (e.g., where the hypotheses being tested is whether the distributions of the test and M&S data are the same) and confidence intervals are simple ways to quantify statistical uncertainty.

Table 2 summarizes the methods of statistical analysis that can be used depending upon the distribution of the test and M&S data, the factors available, and the amount of data available. Where there are multiple methods per cell, more than one test may be required to determine with statistical significance whether there are differences in both the mean and variance of the test and M&S data. Some tests are more sensitive to cases where the live test data are more variable than the simulation data, while others perform better in the reverse case, where the simulation data are more variable than the live data. Thus, it may be necessary to use up to three techniques. The table does not include every possible appropriate technique or prohibit the use of any method. There are entire classes of methodologies, such as statistical process control, time series techniques, and Bayesian analyses that also may be applicable.

Table 2: Statistical Methods for Analyzing Test and M&S Data

		Sample Size		
Distribution	**Factors**	**Small**	**Medium**	**Large**
Skewed (Lognormal)	Univariate / Categorical	Fisher's Combined	Log t-test Fisher's Combined Non parametric K-S	Log t-test Fisher's Combined Non parametric K-S
	Distributed / continuous	Log t-test Fisher's Combined Non parametric K-S	Non parametric K-S	Non parametric K-S
	Determined using statistical model	Log-Normal Regression Emulation and Prediction	Log-Normal Regression Emulation and Prediction	Log-Normal Regression Emulation and Prediction
Symmetric (Normal)	Univariate / Categorical	Fisher's Combined	t-test Fisher's Combined Non parametric K-S	t-test Fisher's Combined Non parametric K-S
	Distributed / continuous	t-test Fisher's Combined Non parametric K-S	Non parametric K-S	Non parametric K-S
	Determined using statistical model	Regression Emulation and Prediction	Regression Emulation and Prediction	Regression Emulation and Prediction
Binary	Univariate / Categorical	Fisher's Exact	Fisher's Exact	Fisher's Exact
	Distributed / continuous	Logistic Regression	Logistic Regression	Logistic Regression
	Determined using statistical model	Logistic Regression	Logistic Regression	Logistic Regression

Source: Wojton (2019). Note: K-S = Komolgorov-Smirnov

The interpretation of the column headings in Table 2 is as follows:

Distribution of the response variable---

Skewed - data generated from the lognormal distribution

Symmetric - data generated from the normal distribution

Binary - data generated from the binomial distribution

Structure of factors:

Univariate - no varying factors across the collected data

Distributed level effects - factors significantly affect the mean. The difference between simulation and test varies across factor levels. The amount of variation across factor levels is represented by a distribution (hence the name, distributed level effects).

Determined using a statistical model - factors that are significant have been determined using statistical modelling.

Data size:

Small- 2-5 (continuous data)/ 20 (binary data)

Moderate - 6-10 (continuous data)/ 40 (binary data)

Large- 11-20+ (continuous data)/ 100+ (binary data)

The tests cited in Table 2 are as follows:

Fisher's Combined Test---used to test for the equality of the means of two independent samples (e.g., of the test and M&S data). The test assumes the samples are randomly selected from infinite populations (equivalently the observations are independent); the samples come from normal populations; and the two populations have equal variances.[24]

t-test---used to determine if two sample means (e.g., of the test data and the M&S data) are equal. Can also be applied to log-normal transformed data (Log t-test).[25]

Fisher's Exact Test---used to analyze two samples of data that fall into one or the other of two mutually exclusive classes (e.g., test and M&S data for hit versus miss for a torpedo). The test will determine whether the two samples differ in the proportion with which they fall into the two classes.[26]

Non-parametric Kolmogorov-Smirnov test (also called the two-sample K-S test)---used to determine whether two data samples (e.g., the test and M&S data) come from the same distribution, which is not specified.[27]

Regression---fitting the test and M&S data (or the transformed log normal data) using a linear or non-linear model and comparing the resulting the models.[28]

---

[24] See, for example, https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/fishrand.htm, accessed April 29, 2021.

[25] See, for example, https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm, accessed April 29, 2021.

[26] See, for example, https://www.itl.nist.gov/div898/handbook/prc/section3/prc33.htm, accessed April 29, 2021.

[27] See, for example, https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/ks2samp.htm, accessed April 29, 2021.

[28] See, for example, https://www.nist.gov/itl/sed/statistical-reference-datasets/strd-background-information/linear-regression, and https://www.nist.gov/itl/sed/statistical-reference-datasets/strd-background-information/nonlinear-regression, accessed April 29, 2021.

Logistic Regression---used to perform regression on binary data.[29]

Emulation and prediction - data are generated using a MCS to create prediction intervals and actual data (test and/or M&S) are evaluated against those intervals.

Examples of the applications of these tests are provided in the Analysis Appendix of Wojton (2019). Note that even in the absence of M&S data, the statistical tests discussed above can be used to analyze data sets collected during different test events to gain an understanding of whether the test results from those events are different, and thereby gain insight regarding, for example, whether performance is improving or degrading.

B. Using Statistical Methods to Analyze Reliability

1. Introduction and Background

This section focuses on hardware reliability.[30] Software reliability is also an important element of overall system reliability; collection and analysis of metrics for software development and testing that are indicators of the reliability of the resultant code are discussed in the section entitled: Choosing Software Development Metrics.

2. Estimate initial reliability defensibly

The estimate of initial reliability for a system provides the starting point for the reliability growth curve. Therefore, its estimate must be realistic and defensible, otherwise the reliability growth curve and its use to assess progress toward achieving reliability requirements will be at best suspect, and at worst misleading.

Data collected from previous versions of the system, from related systems, and from component and subsystem testing can be used to estimate initial reliability. However, to combine such data to produce a defensible estimate, including the uncertainty in that estimate, requires understanding of the conditions under which the testing was conducted, the details of component and subsystem interdependencies (e.g., fault tree analyses), and the use of specific statistical methods. Reliability estimates based on subsystem/component testing can often be optimistic. Estimates based on the performance of analogous systems can be either optimistic or pessimistic, depending upon many factors. Statistical methods and the considerations associated with their use to combine data to estimate reliability

---

[29] See Myers, R., et al., *Generalized Linear Models with Applications in Engineering and the Sciences* Second Edition, John Wiley and Sons, 2010. Also see http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html, accessed April 29, 2021.

[30] The primary sources for this section of the paper are: "Reliability Basics, Key Reliability Concepts for DT&E," M. Ambroso et al., Institute for Defense Analyses paper NS-P-4925, May 2013. Conlon et al., *Test and Evaluation of System Reliability, Availability, Maintainability, A Primer*, Department of Defense, 1983.

(initial or otherwise) can be found in: Rausand, M., A. Hoyland. *System Reliability Theory*, Hoboken: John Wiley & Sons, 2004.

3. Use confidence intervals to analyze test results

This discussion considers computing reliability estimates using continuous and discrete data obtained from system testing. In either case, confidence intervals associated with point estimates of reliability can and should be constructed. Those intervals provide the lower and upper bounds of the region within which the system's reliability is expected to occur with the confidence desired. For example, an 80-percent confidence interval is one in which the system's true reliability would be expected to fall for 80 percent of the measurements made/testing done on that configuration of the system. Because tests are always of finite duration and data are limited, there is uncertainty in what the system's true reliability is and point estimates alone, which do not capture that uncertainty, will not provide sufficient information to assess reliability.

Continuous systems

For a system with reliability requirements specified as a function of its duration operating, a point estimate for Mean Time Between Failure (MTBF) is typically computed as:

$$MTBF = \frac{Test\ Duration\ (D, hours)}{Number\ of\ Failures\ (F)},$$

or

$$MTBF = \frac{Test\ Duration\ (M.miles)}{Number\ of\ Failures\ (F)}$$

and a point estimate for the system's reliability is:

$$Reliability = e^{-D_0/MTBF},$$

where D_0 is the interval (hours or miles) over which the reliability is to be estimated given the MTBF estimated from the test results. These equations apply only when the system's reliability does not degrade over time.[31] Test duration and miles or hours should be conducted and measured using multiple test articles to capture random variations in

[31] If reliability does degrade over time, a non-homogeneous Poisson statistical model, which is appropriate for cases in which the failure rate, F, varies (e.g., increases) with time or miles, must be used to estimate reliability. See https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/course-notes/MIT6_262S11_chap02.pdf, accessed January 27, 2021, for a discussion of non-homogeneous Poisson processes.

performance due to variations in manufacturing processes. So, if seven tanks are tested for 10 hours each and four failures, F, occur, point estimates for MTBF and Reliability during four hours of operation are the following:

$$MTBF = \frac{7\ X\ 10}{4} = 17.5\ hours\,,$$

$$Reliability = e^{-4/17.5} = 0.796\,.$$

Thus, the system's point estimate for the probability of not experiencing a failure (i.e., that it will be reliable) over four hours is about 80 percent.

To construct a confidence interval in which the true reliability would be expected to occur, the confidence, α ($0 \leq \alpha \leq 1$), must be selected. The greater the value of α, the greater the probability the true reliability will lie within the confidence interval, and the larger the interval. Values of α of at least 0.8 (or 80 percent) are recommended. Confidence intervals can be estimated as follows:

$$Lower\ MTBF\ Bound = \frac{2D}{CHIINV(\frac{1-\alpha}{2}, 2F+2)}$$

$$Upper\ MTBF\ Bound = \frac{2D}{CHIINV(\frac{1+\alpha}{2}, 2F)}\,,$$

where $CHIINV$ is an Excel function that computes the inverse of the chi distribution.[32]

So, for the case above with F = 4, $D = 7X10 = 70\ hours$, and $\alpha = 0.8$ (i.e., an 80-percent confidence interval) and a point estimate of 17.5 hours for MTBF:

$$Lower\ 80-percent\ MTBF\ Bound = 8.8\ hours$$

$$Upper\ 80-percent\ MTBF\ Bound = 40.1\ hours$$

$$Lower\ 80-percent\ Reliability\ Bound\ for\ 4\ hours\ of\ operation = 63\ percent$$

$$Upper\ 80-percent\ Reliability\ Boundfor\ 4\ hours\ of\ opeation = 90.5\ percent.$$

[32] These expressions use the relationship between the Poisson and chi distributions. See Sahai, H., and Khursid, A., *Confidence Intervals for the Mean of a Poisson Distribution: A Review*, Biometrical Journal, January 1993. Available at https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.900.3700&rep=rep1&type=pdf, accessed January 27, 2021.

Thus, in this case the MTBF could span a substantial range and be as low as 8.8 hours or as high as 40.1 hours, with a correspondingly wide range in probability of successfully completing a four-hour mission.

Discrete, pass/fail systems

For a system with reliability requirements specified as a function of its probability of successful operation whenever it is employed, a point estimate for reliability is typically computed as:

$$Reliability = \frac{Number\ of\ Successful\ Tests}{Total\ Number\ of\ Tests}.$$

So, a system that operates successfully in nine of 10 tests has a point estimate for reliability of 0.9 or 90 percent. If continuous data are available, those data should be used to estimate reliability as they provide more information and a better estimate. In particular, continuous data should not be converted to pass/fail data to estimate reliability. And, for evaluation of both continuous and discrete test results, if data are collected using different operational conditions under some of which performance is substantially degraded or enhanced, such data cannot, in general, be combined to compute an aggregate estimate of reliability.

For N tests with F failures, an α-percent confidence interval for the reliability of this assumed-to-be binomially distributed process can be constructed as follows:

$$Lower\ Reliability\ Bound = BETAINV\ (\frac{1-\alpha}{2}, N-F, F+1)$$

$$Upper\ Reliability\ Bound = BETAINV\ \left(\frac{1+\alpha}{2}, N-F+1, F\right),$$

where $BETAINV$ is an EXCEL function that computes the inverse Beta function.[33] So, for the pass/fail case described immediately above, the 80-percent confidence interval for reliability is:

$$Lower\ Reliability\ 80-percent\ Confidence\ Bound = 66.3\ percent$$

$$Upper\ Reliability\ 80-percent\ Confidence\ Bound = 99\ percent.$$

[33] These equations provide the Clopper-Pearson confidence interval for the binomial distribution. The Beta function can be used to compute the cumulative distribution function of the binomial. See Clopper, C. and Pearson, E. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial, Biometrika, Vol. 26, No. 4, December 1934. Available at https://www.jstor.org/stable/2331986?seq=4#metadata_info_tab_contents, accessed January 27, 2021.

Therefore, in the case the available test data indicate the reliability could be as low as 66.3 percent, which would be a concern if the requirement was 90 percent.

4. Confidence intervals, growth curves, and requirements

The system's true reliability can lie anywhere within the interval generated with the probability selected for the confidence level, α. In particular, the reliability could be equal to the lower bound of the confidence interval. If that was the case, the question arises as to whether that value of reliability would be acceptable. If not, additional testing may be warranted and/or work/re-design to improve system reliability. The true reliability could also lie outside the confidence interval (above or below) with probability 1-α. To assure the range of true reliability is well bracketed, α should be chosen so that the chance is low that the true reliability lies outside the interval; typically, a 20-percent chance or less is considered low. Values greater than 20 percent would yield a narrower confidence interval with larger lower bounds, but could lead to false assurance reliability is adequate because the probability is greater that the true reliability lies outside the interval, and in particular, below the lower bound. On the other hand, broad confidence intervals with small lower bounds likely indicate the need for additional testing and data that would produce narrower confidence intervals. Broad intervals with small lower bounds would not constitute a cogent case for the need for system re-design to assure progress is consistent with the reliability growth curve or satisfies requirements. In particular:

- If the confidence interval lies entirely above the requirement, the data indicate the system has met its requirement with confidence $\frac{1+\alpha}{2}$. Calculating the largest value of α for which the interval lies entirely above the requirement provides an estimate of the greatest confidence that can be associated with being "on" the growth curve or satisfying requirements.
- If the confidence interval lies entirely below the requirement, the data indicate the system has failed its requirement with confidence $\frac{1+\alpha}{2}$. Calculating the largest value of α for which the interval still lies entirely below the requirement provides an estimate of the greatest confidence that can be associated with not being "on" the growth curve or not satisfying the requirement.

Decisions regarding the values of alpha used to determine whether reliability is sufficient or insufficient are matters of judgment. However, it seems clear that values of .5 or less would not provide a sound basis for decision-making.

5. Consider using parametric models to analyze reliability data

Parametric models that can be used to analyze reliability data include the exponential, Weibull, lognormal, and Gamma.

Exponential distribution. The simple, single-parameter form of the exponential probability density function (PDF) is:

$$f(t) = \lambda e^{-\lambda t}.$$

The reliability function is one minus the cumulative distribution function (CDF) or:

$$R(t) = e^{-\lambda t}.$$

The failure rate function or hazard function is:

$$H(t) = \frac{f(t)}{R(t)} = \lambda = \frac{1}{\alpha},$$

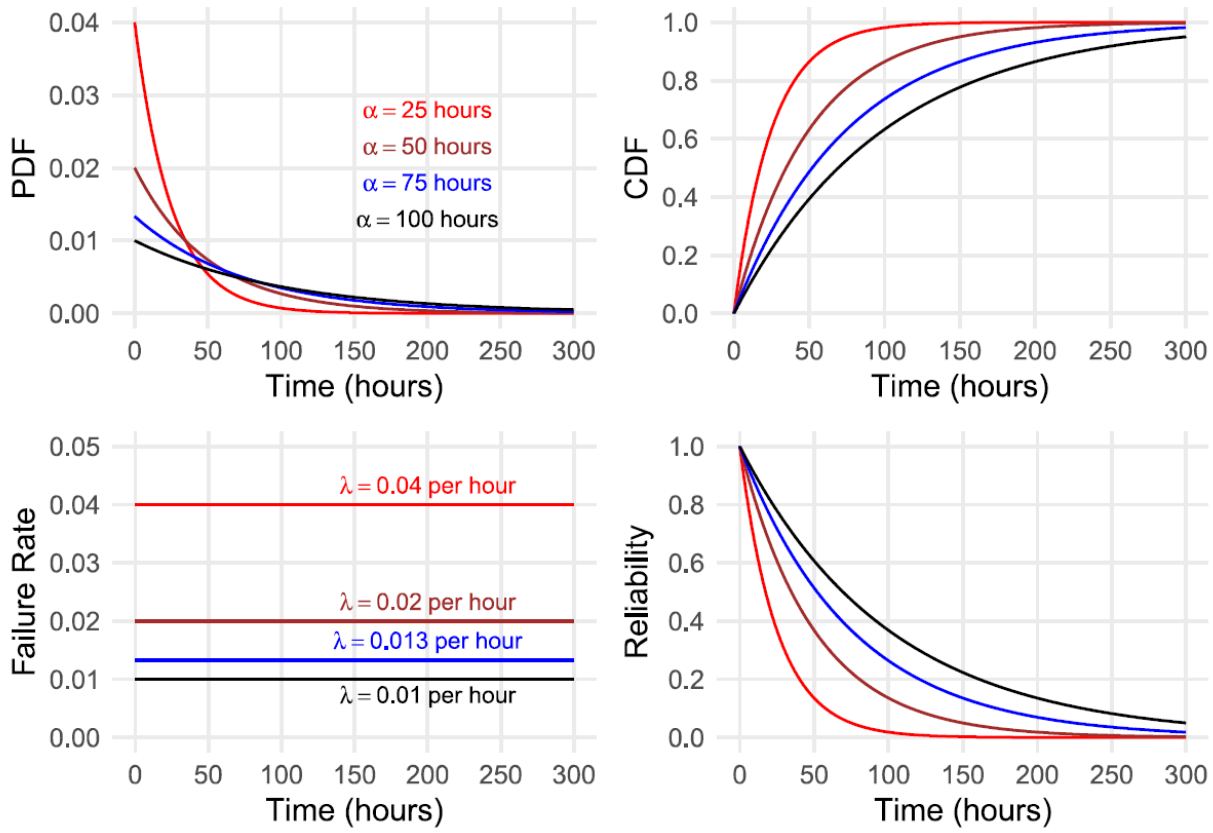
which is constant in time. So, the exponential PDF may not be the best choice for systems exhibiting degradation in reliability (increased failure rates) as they age.

The mean of the exponential distribution is:

$$\overline{T} = \frac{1}{\lambda} = MTBF.$$

Given a dataset, many packages are available that can be used to determine a value of the single parameter λ that yields the best (but perhaps still imperfect) fit, as well as upper and lower confidence bounds for λ.[34] Figure 1 provides graphical examples of the exponential PDF.

[34] JMP and Reliasoft are two of many such commonly-used packages. See https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/fit-curve.shtml, and https://www.reliasoft.com/, accessed February 4, 2021. JMP and Reliasoft are cited as examples of packages providing such capabilities, which does not constitute endorsement.



Notes: PDF = probability density function; CDF = cumulative distribution function; $\alpha = \frac{1}{\lambda}$, α = MTBF and λ = Failure Rate; Reliability = 1- CDF.

Source: Pinelis, Y. and Whitledge, W., Tutorial: Parametric Reliability Models, Institute for Defense Analyses, NS D-9171, September 2018.

Figure 1: Exponential Model Examples

Weibull distribution. The two-parameter form of the Weibull PDF is:

$$f(t) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}},$$

and its reliability function is:

$$R(t) = e^{-\left(\frac{t}{\alpha}\right)^{\beta}}.$$

The Weibull failure rate or hazard function is:

$$H(t) = \frac{f(t)}{R(t)} = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}.$$

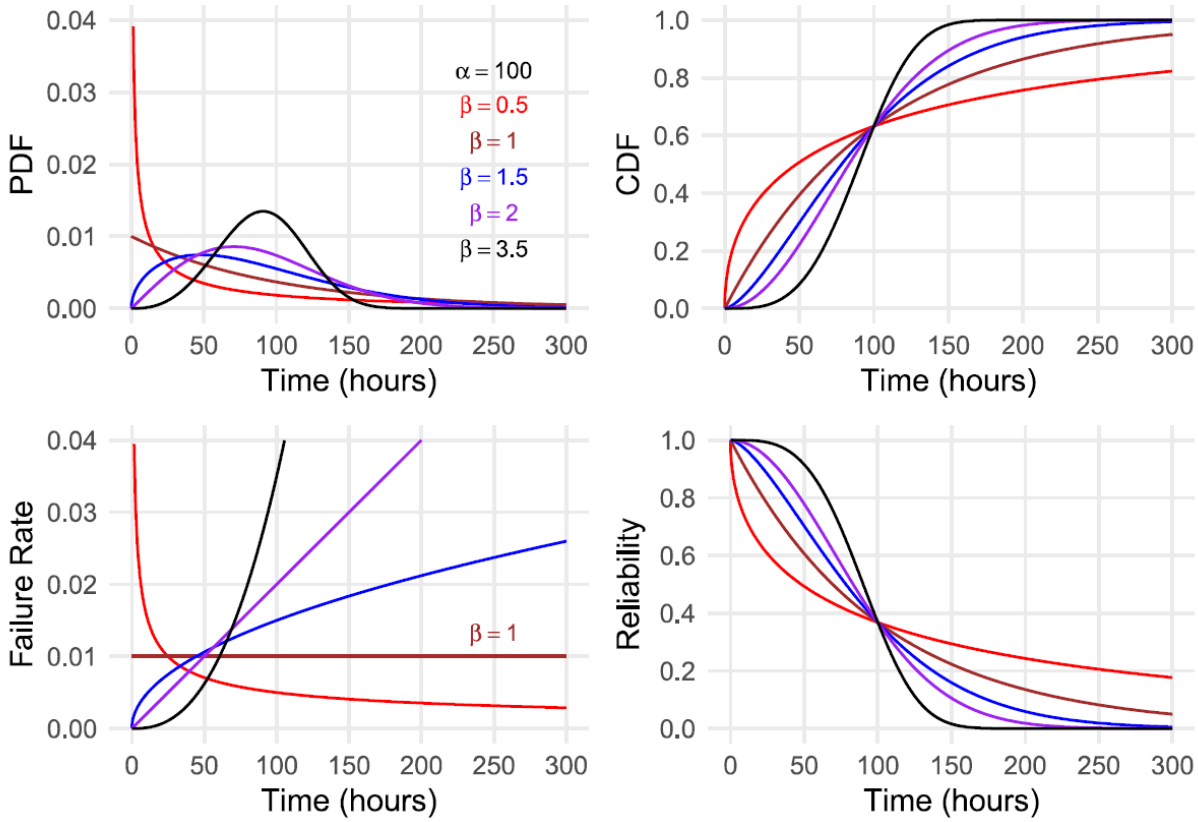
The Weibull failure rate increases with time if $\beta > 1$, decreases with time if $\beta < 1$, and is constant with time if $\beta = 1$, in which case the Weibull reduces to the exponential PDF with

$\lambda = 1/\alpha$ (see Figure 2). So, the Weibull can be used to model reliability that changes with time.

The mean of the Weibull is:

$$\overline{T} = \alpha\Gamma\left(\frac{1}{\beta} + 1\right),$$

where $\Gamma(x)$ is the gamma function.[35] Again, the choices of the two parameters governing the scale (α) and slope (β) of the distribution that best fit a set of data can be computed using many packages.



Notes: PDF = probability density function; CDF = cumulative distribution function; Failure Rate = $H(t)$; Reliability = 1- CDF.

Source: Pinelis, Y. and Whitledge, W., Tutorial: Parametric Reliability Models, Institute for Defense Analyses, NS D-9171, September 2018.

Figure 2: Weibull Model Examples for Various Values of α and β

Inverse Gamma distribution**.** The two-parameter inverse Gamma distribution PDF is of the form:

[35] $\Gamma(z) = \int_0^\infty e^{-x}x^{z-1}dx.$

$$f(t;\alpha,\beta) = \frac{\beta^{\alpha} t^{-\alpha-1} e^{-\beta/t}}{\Gamma(\alpha)}.$$

As discussed below, the Inverse Gamma can be used to model the prior and posterior PDFs of an MTBF using different sets of information and test data. The Inverse Gamma rises to its maximum from 0 exponentially and decays to 0 like $1/t$ (see Figure 3). The mean of the Inverse Gamma (MIG) is $\frac{\beta}{\alpha-1}$ and its variance (VIG) is $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$. Given an MTBF and information/assumptions about its variance, α and β can be determined.

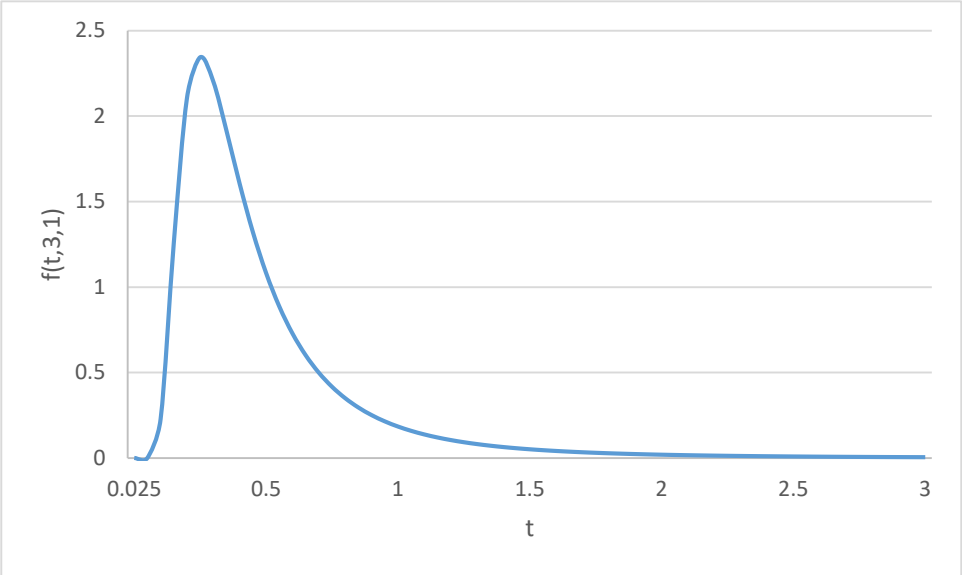


Figure 3: Inverse Gamma PDF for $\alpha = 3, \beta = 1$

6. Combine data from all analysis and test phases

Although a system's configuration will change, sometimes substantially, across developmental testing (DT), it can be advantageous to combine reliability data (and other performance data) when assessing the system's performance. Combining data from multiple test events, when done rigorously, can provide a better estimate of reliability or other aspects of performance than using a single dataset, even if those data are those obtained most recently. Frequentist and Bayesian inference techniques can be used to combine data from different test events through the use of a parametric model. These approaches have been used to combine data from developmental and operational testing; but, the approaches are also applicable to combining data obtained from different DT events in which the configuration of the system tested can vary.[36] Parametric modeling can

[36] See Stefan Steiner, Rebecca M. Dickinson, Laura J. Freeman, Bruce A. Simpson & Alyson G. Wilson (2015) *Statistical Methods for Combining Information: Stryker Family of Vehicles Reliability Case Study*, Journal of Quality Technology, 47:4, 400-415, DOI: 10.1080/00224065.2015.11918142.

yield information regarding whether reliability is significantly different (in the statistical sense) across the different test events and configurations. That information is of obvious use for evaluating whether reliability is on track to satisfy requirements and what changes in configuration, if any, have affected reliability.

A Bayesian approach can be used to combine prior knowledge with recently observed data to produce an estimate of reliability with less uncertainty than would otherwise be the case and even if the recent observations include no failures. The Bayesian approach to estimating reliability uses:

$$\pi(1/\lambda \mid x) \propto L(x \mid \lambda)\, \pi(1/\lambda),$$

where $1/\lambda$ is the MTBF,

$$\pi(1/\lambda \mid x)$$

is the posterior distribution of the MTBF incorporating prior knowledge and recent data,

$$L(x \mid \lambda)$$

is the likelihood distribution of the failures, and

$$\pi(1/\lambda)$$

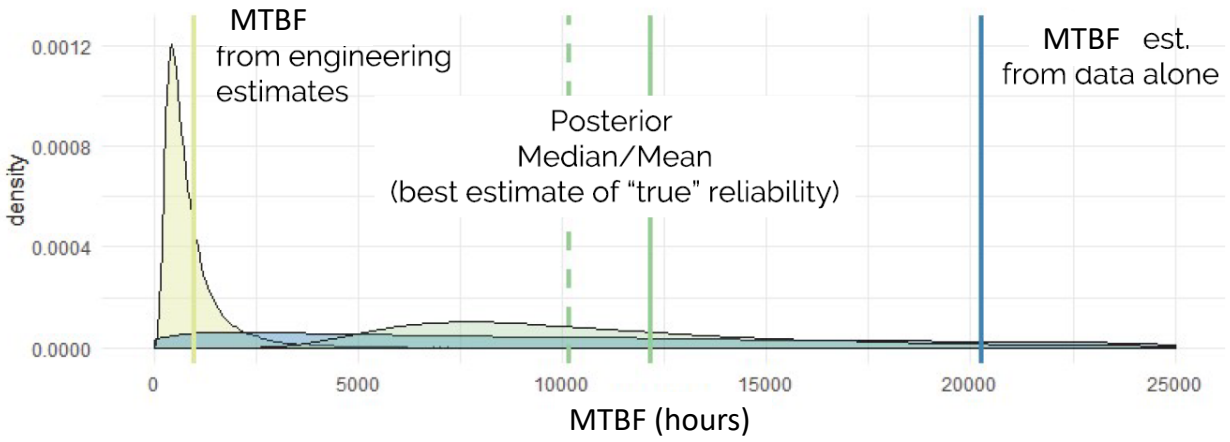
is the prior distribution of the MTBF. Assumptions sometimes used (but not always applicable) are that the likelihood distribution is exponential and the prior distribution is Inverse Gamma. The exponential and Inverse Gamma distributions are a conjugate pair; so, the posterior distribution will also be an Inverse Gamma.[37] If we set $MIG = MTBF_{prior}$, $VIG = p^2 MTBF_{prior}^2$, and p=1.5, α and β for the prior distribution are determined. The posterior distribution for this conjugate pair will be an Inverse Gamma with $\alpha^* = \alpha + N$ and $\beta^* = \beta + T$, where N is the number of failures observed during the most recent observation period and T is the time span of the recent observations.[38]

Suppose we have a prior estimate for MTBF of an aircraft component based on engineering judgment of 990 hours, but have more recently observed two failures during 40,000 hours of flight. In this case the MTBF estimated using the more recent data alone would be 20,000 hours. However, the Bayesian approach yields a lower estimate of MTBF

[37] See *Bayesian Inference and Decision Theory Unit 3*, available at http://seor.vse.gmu.edu/~klaskey/SYST664/Bayes_Unit3.pdf, accessed February 2, 2021.

[38] Ibid. Also note that N=0 is not prohibited.

of about 12,000 hours by incorporating the knowledge of the shorter engineering estimate (see the solid green vertical bar in Figure 4).[39]



Source: Lillard, A. and Medlin, R., "Bayesian Component Reliability Estimation: an F-35 Case Study," Institute for Defense Analyses, NS D-10561, March 2019

Figure 4: Prior and Posterior PDFs for an Example Bayesian Reliability Analysis

For any prior estimate of MTBF substantially less than the most recent observation period, T, as the number of failures observed, N, increases, the estimate for the posterior MTBF will approach the "traditional" result of T/N. For example, if 10 failures were observed during the 40,000-hour period, the posterior estimate for the MTBF would be about 3,600 hours. Thus, as the amount of more recent data increases, the prior estimate has less influence on the posterior estimate.

The Bayesian approach for combining data to produce estimates of reliability can also be used for cases in which analytic prior distributions are not available or appropriate given the data. See, for example, Spalding, D., "Bayesian Reliability Projection for Developmental Systems with Early Test-Fixes," Institute for Defense Analyses, D-16389, November 2020.

C. Using Statistical Techniques to Assess Test Sufficiency

1. Introduction and Background

The National Research Council (NRC) reviewed the state of defense testing in 1998 and found: "Current practices in defense testing and evaluation do not take full advantage of the benefits available from state-of-that art statistical methodology." The NRC

[39] See Lillard, A., et al, "Bayesian Component Reliability Estimation: an F-35 Case Study," Institute for Defense Analyses, NS D-10561, March 2019.

recommended such methods be used.[40] Statistical Test and Analysis Techniques, in particular, Design of Experiments (DOE), provide a rigorous way to construct tests that efficiently cover the set of factors affecting performance (e.g., presence of jamming, type of jamming, network load, illumination, etc.).[41] These methods can be used to calculate the probability the test will correctly pick up significant differences in performance when they are present; as well as the probability the test will correctly not indicate differences in performance exist when they actually do not. These probabilities are key metrics enabling understanding of whether the testing planned is sufficient; i.e., whether the risks of reaching mistaken conclusions on the basis of the test data planned for collection are acceptable.

2. Consider constructing tests using experimental design

The key steps that should be taken to apply statistical techniques to construct tests are the following:

1. Identify the questions to be answered; i.e., the goals or objectives of the testing.
2. Identify the responses of the system (e.g., time to detect, hit versus miss, accuracy of geo-location, etc.), or dependent variables, that will be measured.
3. Identify the factors affecting the system's performance; i.e., affecting the response variables. Also identify the levels of those factors (e.g., jamming to signal ratio, jamming waveform used, high, medium, or low network load, day versus night, etc.). If non-linear performance is expected, three or more levels are needed.
4. Identify and use the applicable statistical techniques to design the test. These depend on the questions, metrics, types of factors (numeric (e.g., jamming to signal ratio) or categorical (day versus night)) and resources available for the testing. Identify the combinations of factors and levels that will be tested. Use statistical measures such as power, prediction variance, and correlation among factors to determine whether testing is sufficient.
5. Conduct the test.
6. Analyze the data.
7. Draw conclusions.

---

[40] Statistics, Testing, and Defense Acquisition, New Approaches and Methodological Improvements, National Academies Press, 1998, available at https://www.nap.edu/catalog/6037/statistics-testing-and-defense-acquisition-new-approaches-and-methodological-improvements, accessed May 3, 2021.

[41] The introductory discussion, as well as later specifics are adapted from Freeman, L. et al., "Testing Defense Systems," Institute for Defense Analyses, NS D-8551, January 2017.

Examples of the use of statistical techniques to design tests and assess their sufficiency are provided in the section entitled Applying Design of Experiments to Cyber Testing in "IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy, Institute for Defense Analyses paper D-22771, October 2021. Statistical methods for data analysis (the sixth step) are the subject of a separate section of this paper: Using Statistical Techniques to Analyze Data. This section will discuss the first four steps listed above, assuming the fifth and seventh will be executed appropriately.

Table 3 provides a non-exhaustive list of common objectives and the test design techniques associated with them. Such objectives include prediction, characterization, and optimization.[42]

Table 3: Test Objectives and Statistical Design Techniques

Test Objective	Potentially Useful Design Techniques
Characterize performance across a set of conditions and determine whether a system meets requirements across those conditions	Response surface designs, optimal designs, factorial designs, fractional factorial designs
Compare two or more systems across a variety of conditions	Factorial or fractional factorial designs, matched pairs optimal designs
Screen for important factors driving performance	Factorial or fractional factorial designs
Test for problem cases that degrade performance	Combinatorial designs, orthogonal arrays, space filling designs
Optimize system performance with respect to a set of conditions and inform system design	Response surface designs, optimal designs
Predict performance, reliability, or material properties at use conditions	Response surface designs, optimal designs, accelerated life tests
Improve system reliability or performance by determining robust system configurations	Response surface designs, Taguchi designs (robust parameter designs), orthogonal arrays

Source: Freeman (2017).

The design techniques cited in Table 3 include the following (Freeman (2017) provides more detailed discussion):

- ***Full factorial*** designs include at least two factors and examine all combinations of each of the factors' levels.
- ***Fractional factorial*** designs are a variation of full factorial, but do not use all combinations of factors and levels. Therefore, they include only a fraction of the test points used in a full factorial design. Fractional designs trade-off the ability

[42] See *NIST/SEMATECH e-Handbook of Statistical Methods*, available at http://www.itl.nist.gov/div898/handbook/, accessed May 4, 2021.

to detect interactions among factors to achieve a smaller, less expensive test composed of fewer test points.

- ***Response surface designs*** spread test points across the factors and levels, collecting sufficient data to enable a model of the pattern of responses to be generated.
- ***Optimal designs*** use criteria, such as minimizing prediction variance (see below), to identify, using software algorithms, the test points satisfying the optimality criteria. Most useful when the number of test points is constrained.[43]

Using statistical models to analyze test data is discussed in the Using Statistical Techniques to Analyze Data section of this paper. These same kinds of statistical models, asserted a priori, also can be the basis for determining test sufficiency. First order models enable estimation of main effects only (no interactions among factors, linear in the factors). Second order models enable estimation of two-way interactions (e.g., quadratic in the factors). Higher order models are possible; but, the performance of most systems is dominated by a few main effects and lower-order interactions.[44] In particular, tests meant to screen for or discern large effects can use lower-order models.

Statistical measures of merit quantify the quality of the test design, enabling comparison across designs as well as understanding the risk of reaching erroneous conclusions based on the test data collected (see Table 4).

Confidence is the probability the test will not indicate a false positive; i.e., that a factor has a significant effect on performance when it actually does not, or that a system meets requirements when it actually does not. Thus, confidence is the probability of avoiding a Type 1 error. However, imposing levels of confidence that are too high can also mean that tests with many data points will be needed to detect effects that are small, but real. Thus, high confidence levels in conjunction with a limited test can result in a false negative; i.e., that a factor does not have a significant effect when it actually does, a Type 2 error.

Power is the probability the test will find a statistically significant relationship between a response and a particular factor. Statistical significance is determined by the confidence level selected for the test: the higher the confidence required, the less likely effects will be detected (unless the effects are large and/or so is the test), and the lower the power.

---

[43] D-optimal, I-optimal, and G-optimal criteria/designs are commonly implemented in statistical software packages. See Meyers, R. et al., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 4th Edition, John Wiley and Sons, February 2016.

[44] Meyers (2016).

All else being equal, large effects will be associated with higher power. However, whatever the size of the effects, if there is substantial variance in the responses (the performance data collected during the test), power will be reduced. Signal-to-noise ratio (SNR) is the effect size divided by the standard deviation of the random error in the data. Lower SNRs indicate substantial "noise' in the data and yield lower power; high SNRs indicate the converse.

The number and types of test points also influence power. Generally, larger tests yield higher power. However, test points must be balanced across the combinations of factors and levels to yield high power.

Thus power, depending upon effect size, test size, and test point coverage, is a measure of the precision that will be possible in reporting the test results. Therefore, it determines how well the test will characterize the system's performance, a key indicator of the test's sufficiency.

Collinearity indicates the degree of linear relationship among two or more factors, which the test design should attempt to minimize. Factors are collinear if they vary together linearly; i.e., if one increases, so does the other. Collinearity causes large variances in the estimates for the parameters in the statistical model underlying the test, thereby increasing the likelihood of both false negatives and false positives. The large variances will also yield large uncertainties when using the model for interpolation or extrapolation.

Variance inflation factors (VIFs) are also a measure of collinearity. They depend upon the number of test points, the number of factors and levels, and how the factors and levels vary across the test points. The VIF associated with the i^{th} factor is:

$$VIF_i = \frac{1}{1 - R_i^2},$$

where R_i^2 is the coefficient of determination of regression on the statistical model when the i^{th} factor is treated as a response variable with all other factors held constant. VIFs range from one to infinity, with values greater than one indicating dependence among factors. Generally, the square root of the VIF should be less than about five.

Scaled prediction variance (SPV) indicates the error involved when using the statistical model underlying the test to predict a system's performance. Those statistical models are generally linear in their parameters, but not necessarily in their factors; they are of the form:

$$y = X\beta + \epsilon,$$

where:

- X is a $n\ x\ p$ matrix comprised of n rows of the values of the p factors affecting performance associated with each of the n test data points;
- y is the $n\ x\ 1$ vector of test data
- β is the $p\ x\ 1$ vector of model parameters; and
- ϵ is the normally distributed error with $variance(\epsilon) = \sigma^2 I_n$, where I_n is the $n\ x\ n$ identity matrix and σ^2 is the overall variance in the responses.

The SPV is:

$$SPV = nx_0(X^T X)^{-1}{x_0}^T ,$$

Where x_0 is the point in the factor space where the prediction is being estimated. SPV should be evaluated across the factor space; the lower it is, the better the ability to use the test data and model to predict performance via interpolation or extrapolation. When considering the sufficiency of the test, values of SPV no greater than in the mid- to low-single-digits are preferable.

Table 4: Statistical Measures of Merit

Measure of Merit	Purpose	Test Design Criteria
Confidence	The true negative rate (versus the corresponding risk, the false positive rate). Quantifies the likelihood in concluding a factor has no effect on the response variable when it really has no effect.	Maximize
Power	The true positive rate (versus the corresponding risk, the false negative rate). Quantifies the likelihood in concluding a factor has an effect on the response variable when it really does.	Maximize
Correlation coefficients	The degree of linear relationship between individual factors.	Minimize
Variance inflation factor	A one number summary describing the degree of collinearity with other factors in the model (provides less detail then the individual correlation coefficients).	1.0 is ideal; otherwise less than 5.0
Scaled prediction variance	Gives the variance (i.e., precision) of the model prediction at a specified location in the design space.	Balance across factors and levels of interest (i.e., the design space)
Fraction of design space	Summarizes the scaled prediction variance across the entire design space.	Keep close to constant (horizontal line) for a large fraction of the design space

Source: Freeman (2017).

Other best practices to follow when using statistical techniques to design tests and determine their sufficiency include the following:

- Where possible, use continuous metrics for system performance and test factors, as opposed to pass/fail probability-based or categorical metrics. Using continuous metrics can reduce test resource requirements by 30 percent to 50 percent for the same power and confidence.
- Use factor-by-factor power calculations to evaluate test sufficiency rather than a single "roll-up."
- Include all relevant factors when designing the test.

There are a large number of software packages that can be used to design tests applying the techniques discussed in this section. Examples of such packages include (but are not limited to) the following:[45]

- JMP (see https://www.jmp.com/en_gb/offers/design-of-experiments.html)
- Minitab (see https://www.minitab.com/)
- Design-Expert (see https://www.statease.com/software/design-expert/)
- Robust Testing (see http://phadkeassociates.com/index_files/robusttesting.htm)

[45] The presence or absence of a particular package on this list does not constitute endorsement or non-endorsement of its use.

4. Assessing and Using Software and Modeling and Simulation

A. Choosing Software Development Metrics

1. Introduction and Background

This discussion focuses on agile software development, which Office of the Under Secretary of Defense, Research and Engineering OUSD(R&E) recommends be used.[46] Early planning for involvement with agile software development is discussed in "IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy."[47] Agile development emphasizes incremental product development and delivery; its primary measure of progress is working software delivered and in use. Software is developed and delivered in iterations and is being tested and evaluated continually for functionality, quality, and whether it meets the expectations/requirements of the customer. Appropriately chosen, metrics can provide objective and quantifiable means to assess software development progress, software reliability, as well support more insightful and accurate estimates of development schedules and costs.

2. Agile development metrics

Government Accountability Office (GAO)'s agile framework contains best practices that can be used to evaluate a program's plans when agile development is being used.[48] Those best practices include appropriate choices of metrics; GAO indicates:

- Metrics---Assuring key metrics are being used that align with and are tailored to the project's goals and objectives, are validated, are readily available and continually updated, provide good measurement of progress and quality, and have management commitment for use.

[46] See https://ac.cto.mil/swe/, accessed May 6, 2021.

[47] Institute for Defense Analyses paper D-22771, October 2021. See the section entitled: Planning for Software Testing and Collecting Metrics.

[48] *Agile Assessment Guide, Best Practices for Agile Adoption and Implementation*, Government Accountability Office, September 2020, draft.

Categories and specifics of agile development metrics suggested by Office of the Under Secretary of Defense, Acquisition and Sustainment (OUSD (A&S)) include the following:[49]

- Metrics of process performance indicating how well planning, execution, and delivery are being performed. Velocity, velocity variance, and completion rate are examples.
- Metrics for the quality of work being delivered. Defect count, test coverage, and first-time pass rate are examples.
- Metrics indicating the value of the delivered products, both quantitative and subjective. Number of features or capabilities delivered and level of user satisfaction (indicated by surveys) are examples.
- Development, Security, and Operations (DevSecOps)-related metrics enabling continuous measurement of the efficiency with which capability is being delivered. Mean time to restore, deployment frequency, lead time, and the failure rate of code changes are examples.
- Cost metrics for both government and contractor efforts, considering both near-term and longer-term work. Estimates can be updated continually as iterations are finished and their actual costs are realized. Total costs (including for sustainment), hardware, software, cloud, and licensing costs, computing costs, storage costs, and bandwidth costs are examples.

3. Metrics for continuous iterative development (CID)

CID is a kind of agile development. It is defined as "A method of managing development, testing, and release of software, or systems, to continually, or iteratively, provide working functional systems of increasing capability to internal and external customers."[50] Specific measures, consistent with those discussed above, were developed for CID using surveys of SMEs; the measures include the following: [51]

- Automated test coverage
- Burndown

---

[49] *Agile Software Acquisition Guidebook Version 1.0,* Office of the Under Secretary of Defense for Acquisition and Sustainment (OUSD (A&S)), February 2020. The Guidebook defines the metrics cited in the bullets.

[50] See *Practical Software and Systems Measurement [PSM] Continuous Iterative Development Measurement Framework Version 2.1 Parts 1, 2, and 3*, Practical Software and Systems Measurement (Product No. PSM-2021-03-001), National Defense Industrial Association, International Council on Systems Engineering (Product No. INCOSE-TP-2020-001-06), April 15, 2021.

[51] Ibid.

- Committed versus completed
- Cumulative flow
- Cycle time/lead time
- Defect detection
- Defect resolution
- Mean time to restore / mean time to detect
- Release (or deployment) frequency
- Team velocity

Definitions for the multiple individual metrics associated with each of the measures were updated in April 2021.[52] The remainder of this section provides a synopsis of the explanatory discussion contained in Practical Software and Systems Measurement (PSM) (2021) for those definitions. Readers interested in additional detail and explanation can refer to the PSM website provided in the footnote.

Automated test coverage. The amount of automated testing that should be used depends upon numerous factors associated with the specifics of the project and the environment in which the project is being conducted. Typically, organizations plan to cover about 70 percent to 80 percent of testing using automation. Such tests can be integrated within and executed upon each build of the code or accomplished during nightly regression testing. Results can be distributed automatically (e.g., via E-Mail or deposit in a central database) so that errors can be identified and corrected quickly. Table 5 provides specific metrics that can be collected or calculated and monitored.

[52] See http://www.psmsc.com/CIDMeasurement.asp, accessed May 6, 2021. The material includes a copyright notice granting permission for use with attribution to PSM, NDIA, and INCOSE.

Table 5: Specific Metrics for Test Automation

Total Requirements
Requirements Tested
Requirements Tested through Automation
Requirements Tested Manually
Code Constructs (e.g., classes, conditions, files, lines packages)
Code Constructs Tested by Automated Test
Automated Test Cases Passed
Automated Test Cases Failed
Requirements not Tested
Percentage Requirements Tested through Automation
Percentage Requirements Tested Manually
Percentage Requirements not Tested
Percentage Code Constructs Tested

Source: PSM, National Defense Industrial Association (NDIA), International Council on Systems Engineering (INCOSE) (2021)

Regarding requirements with agile development, an initial set of requirements is defined that subsequently changes over time (with additions, modifications, and deletions) as the project proceeds.

The extent of code coverage using automated testing is an indicator of the risks remaining in the code's quality including, in particular, the risk there are undiscovered defects. Nonetheless, automation will likely not suffice for all testing needs. Manual testing may be needed to evaluate some code functionality, as well as to assure that automated testing is exercising a representative set of code functions.

Burndown. Completed work items versus planned work items (e.g., story points, features, capabilities) for an iteration or release are measured using burndown.[53] The

[53] A story is a capability or behavior of the system being developed that can be implemented and demonstrated in a single iteration of agile development. Story points are a subjective value assigned to each story indicating its complexity and the level of effort needed to complete the story. For definitions of other terms, see the separate best practices paper discussing software development, as well as Agile 101—An Agile Primer, Office of the Under Secretary of Defense for Acquisition and Sustainment, November 18, 2019. Available at https://www.dau.edu/cop/it/_layouts/15/WopiFrame.aspx?sourcedoc=%2Fcop%2Fit%2FDAU%20Sponsored%20Documents%2FAgilePilotsGuidebook%20V1%2E0%2027Feb20%2Epdf&action=view, accessed April 21, 2021; also see the definitions provided in PSM (2021).

activities associated with these items include design, code, and test; i.e., everything involving requirements, development, configuration management, and quality engineering. Table 6 provides specific metrics that can be collected or calculated and monitored.

Table 6: Specific Metrics for Burndown

Planned Work (e.g., story points/features/capabilities)
Completed Work (e.g., story points/features/capabilities)
Open Work = Planned Work – Completed Work (e.g., story points/features/capabilities)

Source: PSM, NDIA, INCOSE (2021)

Variances from the plan of more than 5 to 10 percent should be reviewed to determine causes and the need for re-planning and/or changing priorities.

Committed versus completed. Measurement of progress completing planned and expected features and capabilities for iterations and releases is done using committed versus completed. This metric is also an indicator of quality---the capability and functionality ready versus that planned and expected. Table 7 provides specific metrics that can be collected or calculated and monitored.

Table 7: Specific Metrics for Committed versus Completed

Work Items Committed Each Iteration (e.g., stories, story points)
Work Items Completed Each Iteration (e.g., stories, story points)
Work Items Committed Each Release (e.g., features, capabilities)
Work Items Completed Each Release (e.g., features, capabilities)
Percent Work Items Completed = (Sum of All Work Items Completed) * 100 / (Sum of All Work Items Committed) for a desired iteration, release, or program (e.g., stories, story points, features, capabilities)

Source: PSM, NDIA, INCOSE (2021)

Cumulative flow. CID, as well as all forms of agile development, must manage the flow and throughput of work on the product. Measuring flow enables understanding of whether development is proceeding stably and efficiently. Work in progress (WIP) is the number of work units in progress between steps in the development process. Table 8 provides specific metrics that can be collected or calculated and monitored.

Table 8: Specific Metrics for Cumulative Flow

Work items in each workflow state--- • To do, items accepted for development, but not started • In progress, items that have started development • Done, items that have completed all development in the iteration • Deployed, items on which work is finished (including integration and test) and are in use via a release
Average cycle time = average duration for all completed work items
Throughput = average number of work items Done per unit time

Source: PSM, NDIA, INCOSE (2021)

The values taken by cumulative flow metrics will vary, but should do so within a fairly stable range or band of values. Variations of greater than 10 percent relative to the near-term average values can indicate performance issues and delays in completing work that should be investigated.

Cycle time/lead time. Efficiency of development is measured by cycle time and lead time, which can also be used to estimate the work that can be accomplished in the future. Cycle time is the time elapsed from when a work item is started until completion. Lead time is measured from an earlier point in time; i.e., from when work is identified and a request for it to be done is submitted to the development team. Thus, lead time includes backlog, planning, assigning priority to work, analysis, and design. Table 9 provides specific metrics that can be collected or calculated and monitored.

Table 9: Specific Metrics for Cycle Time/Lead Time

Start time for a process activity
End time for a process activity
Elapsed time = End Time – Start Time + 1 [time unit]; units (e.g., hours, days, week, months) will vary

Source: PSM, NDIA, INCOSE (2021)

When collected under consistent conditions, data for cycle time and lead time are indicators of the development team's capability and throughput. They can be used instead of measures such as lines of code per hour. Outliers in the data collected (i.e., data points deviating more than 10 percent from a longer-time average), should be investigated.

Defect Detection. Quality of the delivered product is directly related to the number of defects that have escaped detection; i.e., those that accompanied the delivery. There are trade-offs between defect detection and speed and cost. Nonetheless, development teams

generally strive to minimize the number of defects escaping detection. Peer review, automated testing conducted throughout development, and other testing are means to minimize escaped defects. Table 10 provides specific metrics that can be collected or calculated and monitored.

Table 10: Specific Metrics for Defect Detection

Contained Defects
Internally Escaped Defects
Externally Escaped Defects
Total Defects = Contained Defects + Internally Escaped Defects + Externally Escaped Defects
Internal Defect Escape Ratio = Internally Escaped Defects / Total Defects
External Defect Escape Ratio = Externally Escaped Defects / Total Defects
Total Defect Escape Ratio = (Internally Escaped Defects + Externally Escaped Defects) / Total Defects

Source: PSM, NDIA, INCOSE (2021)

Defect Resolution. The efficiency of the resolution of critical problems and, therefore, the quality of the delivered product is measured by defect resolution used in conjunction with defect detection. Table 11 provides specific metrics that can be collected or calculated and monitored.

Table 11: Specific Metrics for Defect Resolution

Defects detected per iteration
Defects resolved per iteration
Iterations to Resolve (number of iterations between detection and resolution)
Resolved 0...n Iteration = the number of defects that are resolved 0…n iterations after being detected; defects resolved in iteration 0, are contained defects.
Defect Resolution Lag Time = iteration the defect was resolved – iteration the defect was planned to be resolved
Open Defect Lag Time = current iteration – iteration the defect was detected

Source: PSM, NDIA, INCOSE (2021)

Defects should preferably be resolved during the same iteration they are discovered. Defects not resolved after multiple iterations may indicate an inherent problem with quality, but could also be lower priority problems that have not been assigned priority for

correction. If the number of defects detected is greater than the number of defects resolved, the backlog is growing, and conversely. Defects whose resolution crosses multiple iterations and growing backlogs merit investigation. Nonetheless, in agile development deferring resolution of lower priority defects to later iterations is to be expected. Thus, investigations should focus on the metrics associated with the resolution of high priority defects.

Mean time to detect (MTTD) and Mean time to restore (MTTR). These are metrics directly related to continuity of services, which is highly valued by any user, and are a direct indicator of quality.

- MTTD measures the amount of time taken by system operators to detect an incident has occurred affecting delivery of services to users.
- MTTR measures the amount of time taken by system operators to restore services to a previously established, good operational state.

Table 12 provides specific metrics that can be collected or calculated and monitored.

Table 12: Specific Metrics for MTTD and MTTR

Failure Occurrence Time
Failure Detection Time
Service Restoration Time
Time to Detect = (Failure Detection Time) – (Failure Occurrence Time) (hours, minutes, days, etc., as appropriate)
MTTD = ∑ (Time to Detect) / N (rolling average Time to Detect, based on N previous failures)
Time to Restore = (Service Restoration Time) – (Failure Occurrence Time) (hours, minutes, days, etc., as appropriate)
MTTR = ∑(Time to Restore) / N (rolling average Time to Restore, based on N previous failures)

Source: PSM, NDIA, INCOSE (2021)

Although speed of delivery is important for any agile development, quality should be maintained. In particular, it is important to be able to restore services quickly if a new deployment introduces a failure in the operational environment. Comparison across outages indicates general trends, including severity and other effects on service delivery. Rolling averages should be calculated, and simple statistical measures, such as mean, median, and standard deviation are also useful. Users should provide feedback regarding the lengths of outages causing severe impacts. Trends in the data collected and their potential implications are displayed in Table 13.

Table 13: Potential Implications of Trends in MTTD and MTTR

Trend	MTTD	MTTR
Increasing	Ineffective monitoring, detection processes, tools, training Incomplete knowledge of failure modes	Increasing complexity of system, software, or architecture Lack of rollback capability or strategy Lack of effective redundancy Developer changes / inexperience
Steady	Established MTTD met and satisfied - no further improvement needed Predictable capability Lack of continuous improvement	Established MTTR met and satisfied - no further improvement needed Predictable capability Lack of continuous improvement
Declining	Improved monitoring effectiveness Effective defect prevention initiatives	Effective improvements through automation, tools Added capability or capacity (redundancy, etc.) are effective

Source: PSM, NDIA, INCOSE (2021)

During development, values of MTTD or MTTR exceeding the objective or mean by more than 10 percent should be investigated and root causes determined. Generally, if trends over time are not improving, the system may not be sufficiently mature for operational deployment.

Release or deployment frequency. Agile development features continual, rapid delivery of useful capability. Nonetheless, the schedule (including frequency and intervals between releases) for delivering capabilities can vary substantially or be predictable and can vary between those extremes as well. Whatever the pattern, the time and effort needed to deploy useful products are direct measures of the efficiency of the development process. Release failures will erode confidence in the development effort. Key terms include:[54]

- Minimum viable product (MVP)---An early or initial version of the software providing basic capabilities to users for evaluation and feedback.
- Minimum viable capability release (MVCR)---A set of capabilities suitable for deployment on a rapid timeline in the operation environment that provides value to users.
- Next viable product (NVP)---the next set of features in the planned sequence of product delivery.

Table 14 provides specific metrics that can be collected or calculated and monitored.

[54] See PSM (2021), Part 1, for a comprehensive set of definitions associated with these and other metrics.

Table 14: Specific Metrics for Release or Deployment Frequency

Release Start Date (release, candidate release, or operational release)
Release End Date (release, candidate release, or operational release)
Effort Hours to generate a release (release, candidate release, or operational release)
Number of releases during a specified period of interest
Release Duration = (Release End Date) – (Release Start Date) May be tracked for capabilities at various stages of maturity; i.e., •Time to Minimal Viable Product (MVP) •Time to Minimal Viable Capability Release (MVCR) •Time to Next Viable Product (NVP)
Release Frequency = (# of Releases) / date range (e.g., days, weeks, months, quarters, years)
Average Release Duration = ∑ (Release Duration) / (# of Releases)
Average Release Transition Time = ∑ (Release Transition Time) / (# of Releases)

Source: PSM, NDIA, INCOSE (2021)

Deployment for military programs must be certified and coordinated with many stakeholders including groups outside the immediate program office and not subject to the program's control or authority. Thus, the times needed to prepare for that coordination and successfully conduct it can substantially affect release frequency. Automation of the build, test, and release process can help increase frequency. The inherent tension between speed of release and quality must be managed.

Velocity. The amount of work completed during a development iteration is measured using velocity, typically by a count of story points completed during specific amounts of time.[55] Acceleration is the change in velocity across iterations. Velocity and acceleration can be used to plan and predict performance of the development team. For example, the current velocity and acceleration can be used to predict whether the development team can complete its commitments by the end of the project. Table 15 provides specific metrics that can be collected or calculated and monitored.

[55] A story is a capability or behavior of the system being developed that can be implemented and demonstrated in a single iteration of agile development. Story points are a subjective value assigned to each story indicating its complexity and the level of effort needed to complete the story.

Table 15: Specific Metrics for Velocity

Story Points Completed (during a specified time interval)
Iterations Completed (during a specified time interval)
Average Velocity = Story Points Completed / Iterations Completed Team Acceleration = (Current iteration Velocity – Reference Comparison iteration Velocity) / Reference Comparison iteration Velocity
Average Acceleration = ∑ (Team Acceleration 1 …Team Acceleration N) / N, where N is the number of iterations

Source: PSM, NDIA, INCOSE (2021)

Changes in velocity of more than 10 percent should be analyzed. Note that because stories and story points will vary across development teams, each team's performance must be assessed separately.

B. Incorporating Technical Rigor in Assessments of Modeling and Simulation

1. Introduction and Background

M&S activities to support decision-making are comprised of three key phases: M&S planning, M&S execution, and the evaluation and reporting of M&S results. These phases may not always be distinct, and there may be overlap among the activities throughout a system's test and evaluation program. To ensure a technically rigorous M&S assessment can be generated, DTE&A should be involved in all phases of M&S activities consistently throughout the test program. Best practices for DTE&A in each of these phases are described in greater detail in the following sections.

2. M&S planning

During M&S planning activities, DTE&A should work with the M&S stakeholders (model developers, testers, and decision-makers) to ensure that the appropriate models providing the desired computational capabilities, fidelity and transparency necessary to satisfy their intended use are selected and available when needed. Details on the types of models and considerations for model selection are discussed in "IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy."[56]

When assessing the sufficiency of plans and activities for using M&S to support decision-making, consider the extent to which the program has developed and is executing

[56] Institute for Defense Analyses paper D-22771, October 2021. See the section entitled: Using Modelling and Simulation.

a comprehensive, systematic approach to conducting M&S validation and is updating the associated plans as needed (i.e., at least once per year) given real-world considerations. These activities should include all "major" M&S capabilities and for each of those should do the following:

- Develop the intended use statement
- Identify the response variables or measures
- Determine the factors expected to affect the response variable(s) or that are required for evaluation purposes
- Determine the acceptability criteria
- Estimate the quantity of data needed to assess the uncertainty within the acceptability criteria
- Determine differences between the model and real-world data for acceptability criteria of each response variable using appropriate statistical methods[57]
- Based on the above, identify shortfalls in the acceptability of the model that exist given available data and their implications for evaluation
- Define the content of test events needed to collect data addressing the shortfalls and include that content in upcoming test events.

3. M&S execution

The execution of the M&S is generally left to the system engineers and M&S SMEs. However, it is important that DTE&A remain engaged throughout the process and stay abreast of any developments that may occur during the M&S "test events." This ongoing engagement is similar to DTE&A's typical involvement throughout a program's "live" test activities. During the M&S execution, as with "live" test events, DTE&A should:

- Ensure M&S is being conducted in accordance with the approved M&S plans
- Ensure M&S data are of sufficient quantity and quality to support DTE&A's evaluation of the M&S results and, ultimately, the decision those results are intended to support. It is important that any data gaps are identified early (ideally, before the completion of M&S "test events"), because some M&S, especially detailed physics-based models, can require extended computational run times. If there are delays in gathering sufficient data, the advantages and benefits provided by the M&S (e.g., enabling early design and requirements

[57] See, for example, *Handbook on Statistical Design & Analysis Techniques for Modeling and Simulation Validation*, IDA Document NS D-10455, 2019.

trade-off analyses, reducing uncertainty in system performance, or optimizing the use of live test resources) may not be fully realized.

4. Assessment of M&S results

An assessment of M&S results should describe the M&S background, context and data such that the decision-maker can judge the adequacy and accuracy of the M&S results to support the intended decision. To that end, assessments should include the following information on each of the models and/or simulations that are included in the evaluation:

1. M&S requirements and use cases
   - What is the purpose of the M&S and what decision(s) is it supporting? For example, is the M&S used to make design decisions (e.g., down-selecting to a final design from several initial options), or is the M&S used to augment live test events by extrapolating system performance to not-yet-tested conditions?
2. Description of M&S tools
   - What type(s) of models/simulations (e.g., physics-based or empirical) are used in the evaluation, and what are their pedigrees?
   - For empirical models, it is especially important to describe that data sources were used to develop the model.
3. Capabilities and limitations of the M&S
   - What are the model assumptions, approximations and uncertainties? The discussion should also address how these capabilities and limitations affect the M&S's suitability for the prescribed use cases, and whether the M&S meets its requirements (see 1. above).
4. VV&A activities
   - If the M&S has undergone VV&A, the assessment should describe the data sources used to support the VV&A (e.g., description of the test data against which the M&S has been compared).
   - If the M&S has not gone through a formal VV&A process, the assessment should describe the processes and data used to establish confidence in the M&S for the use cases described in section 1. Additionally, any plans for formal VV&A, and data requirements or test activities needed to support the VV&A should be included in the discussion.

Assessments should use systematic approaches employing statistical techniques to validate and evaluate M&S results, following the best practices described for rigorous analyses and evaluation of all test data, which are summarized below. (Further details on

related best practices can be found in the sections entitled: Using Statistical Techniques to Analyze Data and Using Statistical Techniques to Assess Test Sufficiency.)

- Employ quantitative analysis and use statistical techniques to analyze all available data: Assessments should use quantitative metrics to evaluate compliance with specifications and requirements where applicable, and include statistical confidence of results and confidence intervals; point estimates should be avoided.
- Use statistical techniques such as DOE to generate quantitative assessments of test sufficiency when possible. Assessments should describe how the M&S data are generated and discuss whether the M&S data and results are sufficient to support the intended decision. For example, can statistically significant conclusions be drawn from the M&S data?
- Compare M&S results with live test data when possible. If there are discrepancies between the M&S and live test results, explain the possible sources of the discrepancies (e.g., test execution procedures, data collection methods, or modeling approximations and simplifications).
- Highlight areas of high uncertainty in the M&S results, and data and/or knowledge gaps that require further investigation.

Using these practices will not constitute a panacea; nonetheless, it will provide visibility to all regarding what shortfalls in a program's use of M&S exist and their significance for evaluation. It will also enable progress (or lack thereof) to be tracked in addressing the shortfalls, and can help provide rationale to the test and resource communities to plan and resource the testing by other means to mitigate the shortfalls in M&S with significant implications for evaluation.

C. Conducting Modeling and Simulation (M&S) using Operational Vignettes.

1. Introduction and Background

The importance of incorporating operational realism early and throughout a test program and the key processes for doing so are discussed in detail in the report "IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy."[58] Following the same rationale, a similar approach should be taken when conducting M&S. DTE&A should work with the operational community to develop operational vignettes and use them in conjunction with appropriate M&S, including but not limited to whatever capabilities are

[58] Institute for Defense Analyses paper D-22771, October 2021. See the section entitled: Incorporating Operational Realism.

available using Model-Based Systems Engineering (MBSE) linking requirements and specifications to system performance.[59]

2. Incorporate realism into both the conduct of M&S and assessment of its results

Incorporating operational realism into both the conduct and assessment of the M&S allows DTE&A to use the M&S to identify the system specifications and characteristics having potentially significant and substantial effects on operational performance/success. This type of analysis is especially key during early assessments, when test assets or system prototypes may not yet be available, and "live" (e.g., open-air range) test data are likely to be limited. As test data are obtained indicating likely system performance and compliance (or not) with specifications, models can be used to quantify the potential effects on operational performance of the emerging test results. The key best practices and processes for incorporating operational realism into early testing are also applicable to the conduct and assessment of M&S, and are summarized below.[60]

- Based on the system's intended operational missions and environments, identify operational mission scenarios, warfighting vignettes and threats.
  - Consult the appropriate resource documents, including CONOPs, TTPs, as well as threat and requirements documents.
  - Engage operational end users and incorporate their insights in the development of the scenarios and vignettes.
- Identify key system tasks and functions (i.e., those which are required for mission success) for the mission scenarios and vignettes.
- Based on the two steps above, conduct M&S in the context of the mission scenarios and operational environment.

The manner in which the operational realism is incorporated into the M&S will depend on many factors, including: the type of M&S (and its associated limitations and modeling approximations), the purpose of the M&S and/or the decision the M&S is supporting, and the intended operational usage and environment of the system being modeled. For example, a full-scale SIL comprised of actual system software and hardware will have the capability to incorporate operational realism to a much greater extent than a simplified sensor model. However, to the extent possible, all assessments of M&S results should present the data with the appropriate operational context. In the case of the simplified sensor model, for example, the modeled sensor detection ranges can be

[59] Institute for Defense Analyses (2021). See the sections entitled: Using Modelling and Simulation (M&S), and Using Model Based Systems Engineering (MBSE).

[60] Institute for Defense Analyses (2021). See the section entitled: Incorporating Operational Realism.

discussed in the context of the threat targets, operational environments, and potential mission scenarios in which the installed sensor is expected to operate.

The M&S resources needed to conduct an assessment following the above-described best practices need to be included as part of the test planning process. Early planning is critical to ensure the required models and simulations are available when needed. If the models/simulations are new, or being used in a new or different application, VV&A of the models/simulations may be required, and plans should account for the additional time and resources, including the collection of test data, needed to complete the VV&A process.

5. Assessing Risks and Readiness

A. Considering Uncertainties Explicitly when Assessing Schedule Risks

1. Introduction and Background

Particularly at its outset, any project schedule necessarily "represents the project plan under a specific set of assumptions, often that it will avoid new risks or even those that have occurred on previous occasions." [61] Typical approaches used by Department of Defense (DoD) program managers to develop schedules, such as the Critical Path Method (CPM), use subject matter expertise to estimate the duration of project activities based, in part, on past experience with analogous programs. [62] These kinds of approaches implicitly, if not explicitly, assume the most likely durations are known a priori, often in detail (e.g., using engineering-based build-ups) and with precision. Unfortunately, subsequent experience often demonstrates that precise foreknowledge was inaccurate, often by many months, if not, in some instances, years. [63]

As a project proceeds, information accumulates regarding the work actually accomplished versus that planned, including both costs and schedules. This information can be used to generate new schedule estimates using the methods originally applied. An Earned Value Management System (EVMS) is also often used "to assess cost, schedule, and technical progress on programs to support joint situational awareness and informed decision-making."[64] Extensions of EVMS have been developed that employ statistical techniques to estimate/predict Earned Schedule (ES) using the typical data collected for Earned Value Management (EVM), without considering in detail deviations from the

[61] Hulett, D., Practical Schedule Risk Analysis, Routledge, April 2016, https://doi.org/10.4324/9781315601885.

[62] Kelley, J. et al., "Critical-Path Planning and Scheduling," Proceedings of the Eastern Joint Computer Conference, 1959.

[63] See "Weapon Systems Annual Assessment," Government Accountability Office, GAO-21-222, June 2021 and its many predecessor reports providing quantitative data on schedule growth in DoD programs. Also see "Defense Acquisitions, Decisions Needed to Shape the Army's Combat Systems for the Future," Government Accountability Office, GAO-09-288.

[64] Macgregor, J., and Bliss, G., Department of Defense Earned Value Management System Interpretation Guide, February 2018, available at https://acqnotes.com/wp-content/uploads/2014/09/DoD-Earned-Value-Management-Interpretation-Guide-Jan-2018.pdf, accessed July 28, 2021.

originally planned schedule.[65] However, EVMS is not applicable to projects using fixed-price contracting, (e.g., the Space Development Agency's Transport Layer) because the information needed to use EVMS will not be available or collected for such projects.

2. Assessing schedule realism at the outset

Methods to develop schedules that incorporate uncertainty in duration, such as the Program Evaluation and Review Technique (PERT),[66] the Probabilistic Network Evaluation Techniques (PNET),[67] and MCS[68] have CPM as their basis, but attempt to improve upon it. Those improvements include using point estimates for potential schedule increases or probability distributions to characterize the potential variance in the duration of project activities. When they are employed, the choices made for the point estimates or those distributions, including in the latter case their shapes, means, and variances, should be informed by data, such as, but by no means limited to, those compiled by the Government Accountability Office.63 Decisions on which programs and data are applicable will involve judgment; but, if data are not used to determine the choices made, the estimates generated will, in effect, be guesses.

Another limitation is that MCS generally assumes the probability distributions it employs are independent, thereby not capturing correlations among risks that could simultaneously affect the durations of multiple activities and, therefore, potentially have substantial effects on an estimate of the project's overall duration. Bayesian networks may provide an approach to capturing the effects of correlations; but, as has been the case with many of the techniques that attempt to more rigorously incorporate uncertainty in schedule estimates, Bayesian methods do not appear to be have been used widely.[69]

Nonetheless, evaluators should consider inquiring whether uncertainties have been included in a program's schedule estimate, and if so, how. In particular, if techniques such as those cited above have been employed, evaluators should consider requesting the data upon which the employment of those techniques was based. If no or little data exist, then

---

[65] Lipke, W., et al., Prediction of project outcome: The application of statistical methods to earned value management and earned schedule performance indexes, International Journal of Project Management, 27 (2009) 400-407.

[66] Malcolm, D., et al., "Application of a Technique for Research and Development Program Evaluation," Operations Research, Vol. 7, No. 5, pp. 646-669, October 1959. Miller, R., "Program Cost Uncertainty: Prediction and Control Using PERT Techniques," Industrial Management Review, Vol. 4, Issue 2, 1963

[67] Ang, A., "Analysis of Activity Networks under Uncertainty," Journal of the Engineering Mechanics Division, Vol. 101, Issue 4, August 1975.

[68] Van Slyke, R. Monte Carlo Methods and the PERT Problem, Operations Research, 11 (5):839-860, 1963.

[69] Khodakarami, et al., "Project Scheduling: Improved Approach to Incorporate Uncertainty Using Bayesian Networks," Project Management Journal, June 2007.

the basis of the incorporation of uncertainty in the schedule estimate is largely subjective, notwithstanding the use of MCS or some other avowedly quantitative technique.

In the absence of the use of quantitative means to explicitly incorporate uncertainty in schedule estimates and the ability to review them, the realism of a program's schedule is often assessed via rough comparison with analogous past programs, which is inherently subjective.

Whether evaluating quantitative assessments of schedule uncertainty (e.g., generated using MCS) or conducting largely qualitative/subjective assessments, evaluators can consider the following approaches:[70]

- Choose a set of past programs reflecting a sufficiently comprehensive interpretation of the meaning of the term "analogous."
  - For example, when assessing schedule realism, consider programs with analogous (1) initial technological maturity: (2) extent of incorporation of modular open systems approach (MOSA),[71] (3) extent of software development, (4) extent of planned use of M&S, (5) extent of VV&A of M&S, (6) cybersecurity requirements, (7) leadership-assigned priority, (8) funding environment/stability, (9) need for or lack of test facility/range improvements (for both hardware and software), and (10) experience scheduling and actually conducting testing of analogous complexity.[72]
- Judgments made regarding what a program's key features will likely be imperfect for a variety of reasons. For example, any set of past programs chosen could be such that none of its members contain a key analogous feature, or its members, while possessing analogous features, could also contain features not common to the program at issue. Therefore, consider performing a sensitivity assessment exploring the effects on schedule of variations in the composition of the analogous set of past programs.

---

[70] These approaches can be used by an evaluator to either perform an independent assessment of schedule uncertainty, or to assess whether a program office has used methods likely to have resulted in a reasonable estimate.

[71] See, for example, https://ac.cto.mil/mosa/, accessed July 27, 2021. Also see https://breakingdefense.com/2020/11/its-open-architectures-for-all-new-weapons-as-jroc-sets-jadc2-requirements/, accessed July 27, 2021, reporting on a mandate for the use of open systems architectures issued in late 2020 by the Joint Requirements Oversight Council.

[72] Here, "extent," although in some cases capable of being measured quantitatively, will likely often be based on subjective judgment. For example, the extent of software development can be measured (or at least indicated) in terms of source lines of code to be developed. However, the extent of the use of M&S will be judged subjectively based on the descriptions available of a program's intended uses of M&S.

  - In particular, consider the potential effects on the schedule of including or omitting programs from the analogous set or including other programs consistent with different characteristics than those planned/expected.
- Consider the potential effects on schedule of delays in maturing the least mature technologies identified as being critical to the program's success.
- Consider the potential effects on schedule of delays or outright failures in the VV&A of key M&S.
- Consider the potential effects on schedule of delays in obtaining needed improvements to test ranges, as well as delays in scheduling and conducting major test events.
  - Consider the extent to which shortfalls in both hardware test capabilities and software development/test labs and capabilities (including for cybersecurity testing) could be realized and affect schedules.
  - Consider the potential for delay in processing test data and/or due to inadequate provisions for processing and readily sharing test data.
- Consider the potential for challenges satisfying and testing cybersecurity requirements and how such challenges could affect schedules.

3. Assessing schedules and uncertainty as a project proceeds

As a project proceeds, information accumulates, including dates at which activities have completed (or not) versus dates originally planned. Such information can be used in obvious ways to assess the realism of the original schedule and to make judgments regarding the likelihood of completing the project at the originally-planned date. Such judgments will, of course, be informed only by the activities currently underway and completed. Thus, at any given time prior to near the end of a project, simple comparisons of actual versus planned finish dates for specific activities, or re-generation of the original schedule estimate using updates to the previously "known" activity durations, will incorporate substantial, and potentially unquantifiable uncertainty.

As programs proceed and information on work accomplished accumulates, the methods discussed previously can be re-applied to generate new schedule estimates; including revised uncertainties. Such estimates, although useful, will have the limitations associated with the methods originally applied, as well as some of the limitations discussed immediately above. Moreover, evaluators may lack access to the tools and data required to re-generate detailed schedule estimates, having to rely on the program to do so, if it chooses.

EVM data, if collected and readily available, can be applied to calculate ES. Estimates of ES can be generated without the detailed schedule analysis that can be "a burdensome

activity and if performed often can have disrupting effects on the project team."[73] ES measures when the amount of earned value (EV) accrued should have occurred; it is the point on the performance measurement baseline (PMB) where planned value (PV) equals the EV accrued. Statistical methods can be applied to estimate ES and its variance.[74] These methods yield "time-based indicators, unlike the cost-based indicators for schedule performance offered by EVM," and can potentially be employed by evaluators outside the program office.[75]

B. Using the Defense Technical Risk Assessment Methodology (DTRAM)

1. Introduction and Background

The DTRAM provides evaluation criteria for assessing the maturity of planning and execution of defense acquisition programs.[76] These comprehensive evaluation criteria are applicable across the Department to conduct risk assessments (including for both technology and manufacturing maturity) of programs regardless of their pathway under the Agile Acquisition Framework.

The DTRAM framework and criteria can be used to conduct 1) Independent Technical Risk Assessments (ITRAs) as directed by statute and policy, 2) Test and Evaluation Sufficiency Assessments, 3) assessments supporting MS decisions, 4) assessments of the adequacy of Systems Engineering Technical Reviews (SETRs), and 5) reviews of technical planning documentation such as Systems Engineering Plans. (The 2011 Technology Readiness Assessment Guidance is also applicable specifically to ITRAs.[77])

---

[73] Lipke (2009).

[74] Ibid.

[75] Ibid. A calculator is available at https://www.earnedschedule.com/Calculator.shtml, accessed July 27, 2021.

[76] DTRAM Criteria Volume (Version 6.3, September 30, 2020), available at https://ac.cto.mil/wp-content/uploads/2021/01/DTRAM-0-1.pdf, accessed June 10, 2021.

[77] Department of Defense Technology Readiness Assessment (TRA) Guidance, Assistant Secretary of Defense for Research and Engineering, April 2011. Available at https://www.afacpo.com/AQDocs/TRA2011.pdf, accessed August 12, 2021.

2. Organization of the DTRAM

The DTRAM is organized into eight technical risk areas across seven factors (see Figure 5). It includes specific evaluation criteria for each area shown and does this for all seven factors within each area. Using DTRAM provides a standardized framework to present the results of assessments. The findings can be mapped not only by technical risk area, e.g., Technology (Area 2.), and to a specific factor, e.g., Performance & Quality (factor 2.7), but also to a specific criterion listed in the DTRAM Criteria Volume, e.g., 2.7.C2 "Results are sufficient to evaluate performance of matured technology to support program decisions."

OVERALL		
PERFORMANCE	SCHEDULE	RESOURCES

Factors ➢ Areas ⮟	Performance & Quality	Scope & Requirements	Design & Architecture	Evaluation	Schedule	Decision & Control	Resources
MISSION CAPABILITY	F. Mission effectiveness demonstration	N. Identification of needed data				U. Radio frequency waiver	
TECHNOLOGY		Aa. Technologies identified		Ab. Technologies demonstrations ahead of schedule			
SYSTEM DEVELOP. & INTEGRAT.	Ad. Contractor leveraging internal funding	R. GFE delivery requirements	S. Low maturity of XXXX T. External system data delivery	C. Low thrust (TPM) results	A. Aggressive program schedule I. Scheduling of integration assets	E. System weight growth	W. Availability of SIL X. Availability of SIL
MOSA		Ac. Use of open system standards					
SOFTWARE	Q. Software and SoS integration	D. Identification of software req.					B. Software staffing lagging development
SECURITY & CYBERSECURITY			G. PPP implementation plan for system				
MANUFACURING		O. Production gap during LRIP	Ae. Manufacturing process demonstrated	J. Manufacturing readiness for MS C	K. Accelerated schedule		M. FRP production capacity
RAM & SUSTAINMENT	Z. Impact of failures on mission essential functions	L. Realism of RAM allocations Y. Untraced allocated requirements	V. Reliability growth planning unrealistic	Aa. XXX sub-system performance		P. Sustainment planning is lacking	H. Unfilled RAM billets

Low risk | Moderate risk | High risk | Positive | Assessed - No Significant Findings | ☒ Not Assessed

Figure 5: Notional DTRAM Assessment Scorecard

3. Performing risk assessments using the DTRAM

Use of the following activities and practices in conjunction with the DTRAM should assist in assuring a successful risk assessment:

- **Plan for assessment.** Prepare/read the written plan for the assessment. [78] The plan outlines the objective and approach for the risk assessment. It includes the

[78] OUSD (R&E) DTE&A will normally prepare the written plan.

name of the program to be assessed, the date of the event at which the assessment will be briefed to the Milestone Decision Authority (MDA) (e.g., Milestone B (MS B)), and specifies the organization that will conduct the assessment. The plan should include a short system description. It should describe any system or program aspects that may impact the conduct of the assessment (e.g., new technologies, mission changes, low manufacturing rates, problems identified through testing, etc.)

Form the assessment team. Assessment team members are selected based on their expertise in areas to be assessed. A list of team members should be included in the written plan along with the assessment areas for which they will be primarily responsible, their organization and contact information. All team members should be provided an electronic copy of the full DTRAM (Criteria and Questions).

- **Scope of Assessment**. The assessment will, at a minimum, examine the program's technical, engineering, and integration risk, to include technology and manufacturing risk. Some areas or factors in the DTRAM may be excluded from the risk assessment due to specific program characteristics that make the areas or factors not applicable.

  There may be a number of DTRAM areas and factors that are key technical drivers. These should be identified and prioritized for increased focus. Explain the reasons for selecting them and how they may affect the upcoming MS review or decision. Focus areas should be program specific and not generic.

- **Assessment Schedule**. The assessment team lead will coordinate the schedule with the Program Management Office (PMO). The schedule will take advantage to the extent possible of on-going activities to reduce the burden on the program. Scheduled events may include: coordination meeting with the PMO; document collection and review; team training and planning meetings; briefings by external offices; meeting with requirements sponsors; etc.

  Scheduled engagements with the program could include: SETRs; technical working group meetings; PMO site visits; contractor site visits; etc. The schedule should also include dates for completing a preliminary assessment report and the final report, as well as the date of the program MS review.

- **Review Program Documents, Artifacts and Data.** The assessment team will review program plans, documents, artifacts, and other data to gain a full understanding of the program.

The risk assessment team lead will coordinate with the program office to obtain needed artifacts. Table 16, below, presents a list of documents typically requested for review by the team members with the appropriate expertise.

- **Conduct Site Visits.** Site visits provide the opportunity to receive briefings from the program office and contractors and to ask questions and engage in discussions on important questions that affect the risk assessment. The discussions should be conducted on a non-attribution basis.

  It is advisable to have a Defense Contract Management Agency (DCMA) representative accompany visits to contractor sites. DCMA can provide insight and visibility into day-to-day contractor activities and processes. DCMA can often identify a contractor point of contact for technical risks judged to be critical and of high priority. Side meetings on key technical questions or topics can often be arranged to permit discussions between team technical/technology experts and contractor personal responsible for the technologies of interest.

Table 16: Assessment Documentation and Artifacts

Document Name	Date of Document MM/DD/YYYY
Analysis of Alternatives	
Appropriate Joint Capabilities Integration and Decision System (JCIDS) document for phase	
Concept of Operations	
Validated Online Lifecycle Threat (VOLT) Report	
Request for Proposal	
Acquisition Program Baseline [if MS B or later]	
Cost Analysis Requirements Description (CARD)	
Acquisition Strategy	
Systems Engineering Plan (SEP)	
Risk Management Plan	
Software Development Plan or Test Plan	
Test and Evaluation Master Plan (TEMP)	
Integrated Master Plan	
Integrated Master Schedule (electronic version, native format)	
Information Support Plan	
Life Cycle Sustainment Plan	
Program overview briefing with organization charts	
Risk Register and Risk Management Board Minutes	
Technical Performance Measures, including software	
Software Test Reports, Software Measurement Plan	
Software Data: (e.g., schedule, effort hours, planned duration, Defects, Defect backlog, Planned size software (SW) test reports, etc.,)	
Reliability Data	
Presentations from SETR events [e.g., System Requirements Review (SRR), Preliminary Design Review (PDR), Critical Design Review (CDR)]	
Assessments: Technology Readiness Assessment, Independent Reviews, Non-Advocate Reviews, etc.	
DT, Operational Assessment, or Director Operational Test and Evaluation (DOT&E) Report	
Manufacturing Plan / Assessments / Manufacturing Readiness Review artifacts	
Manufacturing Data	

- **Developing Risk Assessment Reports**. After the documentation review and site visit(s), the assessment team will synthesize data, develop findings, assess risks, and provide recommendations in each area. Among a number of approaches, the team will use comparisons with the experience of analogous programs as a

means to assess the realism of program objectives, as well as planned resources, and schedule.

- ***Preliminary Report.*** A preliminary report that summarizes the technical risks, provides actionable recommendations, and is supported by appropriate documentation and analysis will be presented to the Program Manager (PM) and shared with the MDA. The preliminary report provides the PM with an early opportunity to review the risk assessment team's results and recommendations, correct any factual inaccuracies, and initiate any risk mitigation activities the PM deems appropriate. The preliminary report also provides the approval authority and the MDA with early notification of any risks that may require outside support or elevation before the MS or production decision.
- ***Final Report.*** The final report will provide the MDA, Congress, and other stakeholders with an independent analysis of the program's risk posture and provides the MDA data to support statutory reporting responsibilities. The report will consist of an executive summary and a main body containing and describing the data and analysis supporting the team's finding and recommendations.
    - The executive summary will provide an overview of the program's technical risk posture, to include critical technologies and manufacturing processes. It will identify risks to be brought to the MDA's attention and provide recommended mitigation strategies for high-risk areas.
    - The main body of the report will provide greater detail, expanding on the discussion in the executive summary. It will include the data and analyses substantiating the team's assessment of the program's risks and should be self-contained, with minimal external references. The detailed report will include a DTRAM scorecard similar to the one shown in Figure 5, above. This scorecard, coupled with the standard risk cube, provides leadership a cogent summary of the program's risks and progress mitigating them.

4. Content of risk assessment reports when using the DTRAM

Risk assessment reports normally include the following sections:[79]

- **Purpose** (one paragraph). Provide a short paragraph that identifies the program name and the MS or production decision the assessment supports.
- **Executive Summary** (one page).
- **Program Objective, Program Description, and System Description** (about six pages). State what the program is trying to achieve (e.g., new capability, improved capability, low procurement cost, reduced maintenance or manning). Briefly describe the program or program approach (not the system) as it relates to cost, schedule or performance impacts. Describe whether the program is providing a new system or is to replace or modify an existing operational system. Discuss if it is a new design, a major system modification, or a modification or repurposing of existing government or commercial off-the-shelf (COTS) equipment. Address the acquisition approach the program is using and what phase of that approach the assessment has examined. Include the program schedule and funding profile if available.
- **Summary of Technical Risks and Readiness Assessment (**pages as needed**).** Discuss the process/approach used to perform the assessment. State the team's assessment of the overall risk posture. For example, "The program shows high technical risk in meeting planned threshold performance goals. The program will likely require a schedule slip of X months due to technology and system development risks." Also discuss the overall schedule risk and any other DTRAM risk factors such as resources. For example: "Recent budget cuts have impacted the standup of new system integration labs, potentially delaying the integration of new technologies by Y months." All of the eight risk areas in the DTRAM should be addressed.
  - **Mission Capability** (Select Low/Moderate/High Risk). Identify any aspects of the mission, requirements, CONOPs, or mission profile that may not be met. Discuss significant interoperability or interdependency risks that have an impact on the program's ability to accomplish its intended mission, or meet the initial operational capability date.

    Describe whether the established KPPs, KSAs, and additional performance attributes are achievable. Provide a current status and assessed risk to achieving proposed or established requirements. Consider the program's

[79] DoD TRA Assessment guidance (2011), which provides a template for a TRA. This discussion borrows from that template, but modifies it to include details associated with the use of the DTRAM.

Technical Performance Measures (TPMs) when assessing the risk to meeting requirements.

- **Technology** (Select Low/Moderate/High Risk). MANDATORY: Include a statement that addresses the statutory reporting requirements for technology. For Milestone A (MS A), assess if there are any critical technologies that need to be matured. For MS B and subsequent production decisions assess if critical technologies have been successfully demonstrated in a relevant environment. For example-- "The program has demonstrated all critical technologies in a relevant environment." or "The program has three critical technologies, two of which still need to be matured before Milestone B, technology xx and technology yy."

  Discuss any technical risks or issues related to critical technologies, status of technology maturity, any problems with reaching needed maturity, or with demonstrating technologies in a relevant environment.

- **Manufacturing** (Low/Moderate/High Risk).[80] MANDATORY: Include a statement that addresses the statutory reporting requirements for manufacturing. For MS A, assess if there are any manufacturing processes that need to be matured. For MS B and subsequent production decisions, assess if there are any manufacturing processes that have not been successfully demonstrated in a relevant environment.

- **System Development / Integration** (Low/Moderate/High Risk). Discuss key risks associated with design considerations, technical processes, management processes, and engineering products not addressed in other areas in this report.

  Summarize any design trades made that relate to cost, schedule, or performance risks and drivers. If appropriate, comment on technical trade-off analyses conducted.

  Consider integration risks among components within the system (internal integration) as well as with external systems (external integration). Assess whether external systems having critical interdependencies and interfaces with the system of interest are on track to support planned integration, test, and production.

  Discuss significant risks related to achieving test objectives and safety certifications, as well as whether test resources have been properly identified, coordinated, and resourced. Consider risks related to training for

[80] See the section entitled: Assessing Manufacturing Readiness.

test, timing to successfully proceed with tests, and risks to successfully meeting the program's verification requirements.

Assess the risks in other areas such as spectrum supportability and Electromagnetic Environmental Effects.[81]

- **Modular Open Systems Approach** (Low/Moderate/High Risk). Assess if the system has been appropriately designed to allow evolution of capability. Discuss any MOSA risks that may hinder an evolution or opportunity for technical upgrades, reduce interoperability, or inhibit significant cost savings in the future.
- **Software** (Low/Moderate/High Risk). Assess the software development plan and the program's progress to plan.

  The program should establish a Software Development Plan (SDP) to manage the software development effort. Underpinning any successful software program is an effective process for estimating size, effort, and duration. A software estimation process should be used by the program to define the initial scope of the effort and to track progress over time. Metrics applicable to software development should be used as the basis for the estimation process. The program should identify appropriate metrics. (Software metrics are discussed in the section entitled: Choosing Software Development Metrics.)

  When conducting technical risk assessments to support a MS A decision, assess the realism of the program's SDP and software estimation. Consider using analogous programs as a baseline for comparison.

  Identify recent analogous programs that reflect the assessed system's scope, complexity, staffing, and productivity. Analogous programs can provide bounds against which the program's proposed and realized software development duration, effort, and productivity can be compared.

  When conducting technical risk assessments to support a MS B and subsequent decisions, use actual software development data to refine assessments done using the baseline data. In particular, assess the following:

  - If software development has been executed on schedule. Consider any changes to the content (scope) of the software builds/releases.

[81] See: https://www.dau.edu/cop/e3/Pages/Topics/Electromagnetic%20Environmental%20Effects%20E3.aspx, accessed August 19, 2021.

- The software quality to include defects, defect aging, and defect backlog.
- The software baseline and changes to overall software effort size; addressing new software code, software code reuse, and modified software. Address any changes to the content (scope) of the software releases and impact to software size.
- If planned software staffing and facilities are sufficient to execute the remaining software development schedule and if the metrics tracked by the program are sufficient to manage the software development and test program.

– **Security/Cybersecurity** (Low/Moderate/High Risk). Discuss any significant security or cybersecurity risks to include risks related to information assurance and system security. Assess the protection of critical program information, exposure to vulnerabilities, or any other design attributes that may impact the mission.

– **Reliability, Availability, and Maintainability (RAM)/Sustainment** (Low/Moderate/High Risk). Assess the Reliability Growth Plan (RGP) and the program's progress to plan.[82] Assess any risks or issues with meeting the RAM requirements and goals.[83] Include data tables and graphics to support the assessment team's RAM assessment such as the program's reliability growth curve, annotated with the current system performance.

– **Schedule** (Low/Moderate/High Risk).[84] Is a key factor associated with each of the eight risk areas discussed immediately above, as well as one of the three overall areas of risk (including performance and resources) that should be assessed. Schedule risk should be described including the probability of program delays based on the potential impact of all identified program risks, singly and in aggregate.

  When conducting risk assessments to support MS A and B decisions, analyze the realism of the government roadmap/schedule, to include external program dependencies. Assess contractor schedules for realism

[82] See the section entitled Planning Early for Reliability Growth and Associated Testing in IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy, Institute for Defense Analyses paper D-22771, October 2021.

[83] See the section entitled Early Consideration of Reliability, Availability, and Maintainability (RAM) in IDA Support to DTE&A Initiative: "Shift Left" Baseline Framework Strategy, Institute for Defense Analyses paper D-22771, October 2021.

[84] See the section entitled: Considering Uncertainties Explicitly when Assessing Schedules.

(e.g., DCMA 14-point assessment, Schedule Risk Assessment (SRA)) as applicable.[85]

When conducting risk assessments to support post-MS B reviews and Milestone C (MS C) and subsequent production decisions, assess contractor schedules (e.g., DCMA 14-point assessment, SRA) to include external program dependencies. Assess risks to meet upcoming MSs and technical reviews. Review the program's critical path and near critical path(s). Consider depicting the program's schedule with planned dates and the review team's assessed likely dates if significantly different from the current program estimates. Also consider using analogous programs as a baseline for comparison. Identify recent analogous programs that reflect the assessed system's scope, complexity, staffing, and productivity.

– **Risk Matrix** Identify and briefly summarize the key technical risks in a risk matrix (see Figure 6). Consider depicting only HIGH and MODERATE risks to prevent over-loading the graphic.

[85] See Defense Contract Management Agency (DCMA) Earned Value Management System (EVMS) Program Analysis Pamphlet (PAP), DCMA-EA PAM 200.1, October 2012, Section 4.0, available at https://www.dcma.mil/Portals/31/Documents/Policy/DCMA-PAM-200-1.pdf?ver=2016-12-28-125801-627, accessed August 12, 2021. Also see Department of Defense Risk Management Guide for Defense Acquisition Programs, 7th Edition, December 2014, available at https://acqnotes.com/wp-content/uploads/2014/09/DoD-Risk-Mgt-Guide-v7-interim-Dec2014.pdf, accessed August 12, 2021. The latter discusses the conduct and content of risk assessments in less detail than the DTRAM.

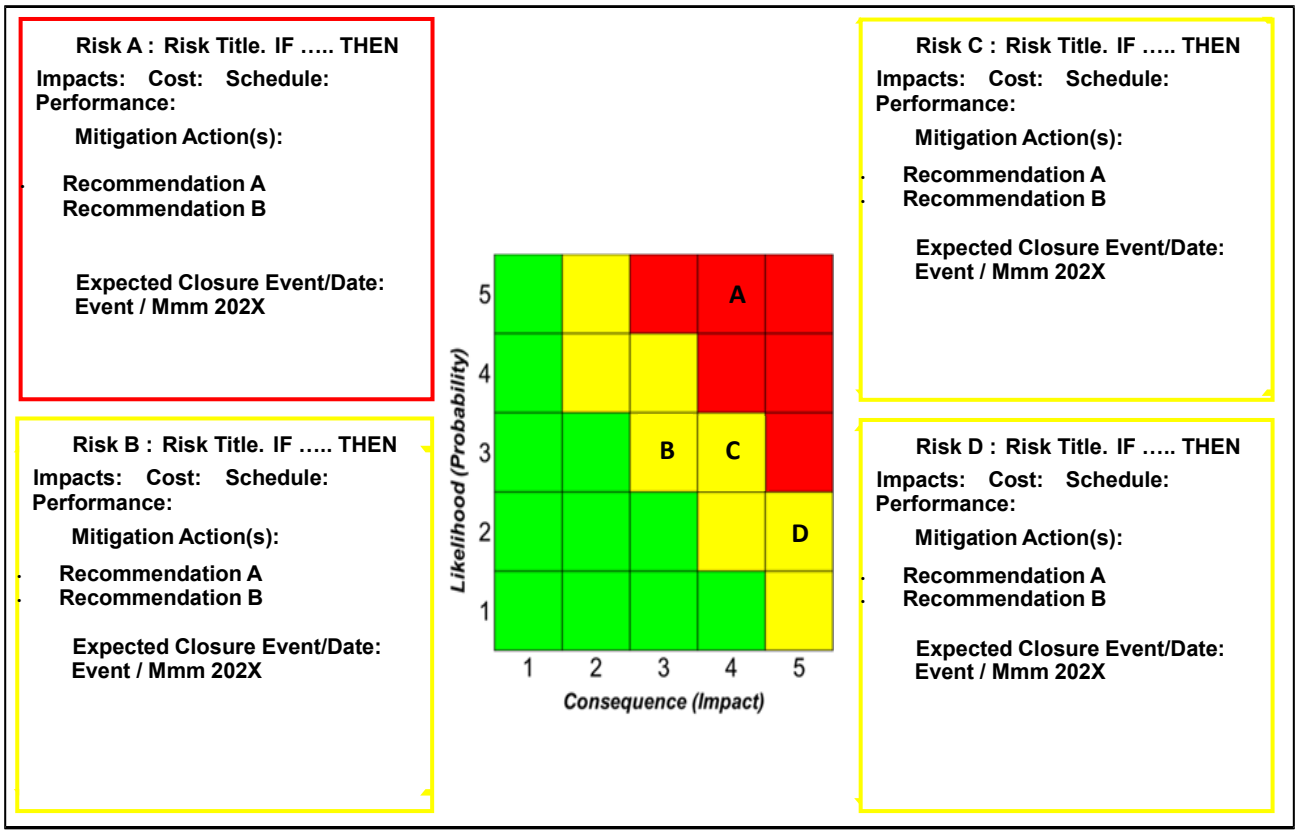


Figure 6: Notional Risk Matrix

- **Summary and Conclusions** (about one page)**.** Provide a concise overview identifying the top items for the MS Decision Authority's consideration.
- **Recommendations** (about one page)**.** Identify assessment team recommendations to reduce risk, accelerate schedule, or reduce cost. These recommendations should be clearly linked to risks, issues, or opportunities discussed in the previous sections of the report.
- **Attachments** (pages as needed). Include the following attachments to substantiate the team's assessment of program risks and ensure the report can be understood without referencing external documents excessively.

  A. Top-Level Schedule, annotated with key assessment team findings and risks

  B. DTRAM Scorecard

  C. Detailed analyses underlying key findings and assessed risks

C. Using Objective Criteria to Assess Technology Readiness

1. Introduction and Background

Technology readiness assessments (TRAs) often make use of technology readiness levels (TRLs) (see Table 17).[86]

Table 17: Technology Readiness Levels (TRLs)

TRL	Description
1	Basic principles observed and reported
2	Technology concept and/or application formulated
3	Analytical and experimental critical function and/or characteristic proof-of-concept
4	Component and/or breadboard validation in laboratory environment
5	Component and/or breadboard validation in relevant environment
6	System/subsystem model or prototype demonstration in a relevant environment
7	System prototype demonstration in an operational environment
8	Actual system completed and qualified through test and demonstration
9	Actual system proven through successful mission operations

Source:https://acqnotes.com/acqnote/tasks/technology-readiness-level, accessed August 17, 2021.

2. TRL Assessment using factors and completion criteria

The use of TRLs arguably requires assessments incorporating subjectivity (e.g., arising from ambiguity in interpretation), sometimes substantially.[87] In an effort to produce more consistent, objective assessments of technology maturity, the TRL definitions can be decomposed into five factors, with specific completion criteria for each TRL (see Table 18

[86] See *Technology Readiness Assessment (TRA) Deskbook*, Department of Defense, July 2009. Available at https://www.skatelescope.org/public/2011-11-18_WBS-SOW_Development_Reference_Documents/DoD_TRA_July_2009_Read_Version.pdf, accessed September 13, 2021.

[87] See J. Mankins, *Technology Readiness Level: A White Paper*, NASA, Office of Space Access and Technology, 1995. https://aiaa.kavi.com/apps/group_public/download.php/2212/TRLs_MankinsPaper_1995.pdf, accessed February 11, 2021; and the Department of Defense analog at https://acqnotes.com/acqnote/tasks/technology-readiness-level, accessed February 11, 2021.

and Table 19).[88] The TRL assessment is based on the extent to which the technology has met the completion criteria for each of the five factors, which are the following:

1. Performance/Function – Extent to which the technology has demonstrated key metrics for performance and/or functionality against the mission requirements
2. Fidelity of Analysis – Reflects the quality of the data and understanding used to support the TRL assessment
3. Fidelity of Build – Fidelity of the physical realization of the technology (e.g., prototype versus flight-qualified hardware)
4. Level of Integration – Extent to which the technology is integrated into its intended full-scale system assembly
5. Environment Verification – Extent to which the technology has been tested in conditions that are representative of the technology's intended operational environment

[88] M. A. Frerking and P. M. Beauchamp, "JPL technology readiness assessment guideline," *2016 IEEE Aerospace Conference*, 2016, pp. 1-10, doi: 10.1109/AERO.2016.7500924.

Table 18: TRL 1-6 Decomposition by Factor

TRL	Definition from NPR 7123.1e [1]	Completion Criteria from NPR 7123.1e [1]	Mission Reqt.	Performance/ Function	Fidelity of Analysis	Fidelity of Build	Level of Integration	Environment Verification
1	Basic principles observed and reported	Peer reviewed documented principles	Generic class of missions	Knowledge underpinning technology concept/ applications	Physical principles identified	NA	NA	NA
2	Technology concept and/or application formulated	Documented description that addresses feasibility and benefit	Generic class of missions	Concept formulated	Feasibility presented	NA	NA	NA
3	Analytical and/or experimen-tal proof-of-concept of critical function.	Documented analytical/ experimental results validating predictions of key parameters	Generic class of missions	Proof-of-Concept demonstrated analytically and/or by experiment	Low fidelity: to predict key performance parameters	NA, but could be low-fidelity bread-board	NA	NA
4	Component and/or breadboard validated in laboratory environment	Documented test performance demonstrating agreement with analytical predictions. Documented definition of relevant environment.	Generic class of missions	Basic functionality/ performance demonstrated	Medium fidelity: to predict key performance parameters and life-limiting factors as a function of relevant environments	Low fidelity: bread-board	Component/ Assembly	Tested in laboratory for critical environments. Relevant environments identified. Life-limiting mechanisms identified.
5	Component and/or brassboard validated in relevant environment	Documented test performance demonstrating agreement with analytical predictions. Documented definition of scaling requirements.	Generic or specific class of missions	Basic functionality/ performance maintained	Medium fidelity: to predict key performance parameters and life-limiting factors as a function of relevant environments	Medium fidelity: brass-board with realistic support elements	Component/ Assembly	Tested in relevant environments. Characterize physics of life-limiting mechanisms and failure modes.
6	System/ subsystem model or prototype demon-strated in a relevant environment	Documented test performance demonstrating agreement with analytical predictions	Specific mission	Required functionality/ performance demonstrated	Medium fidelity: to predict key performance parameters and life-limiting factors as a function of operational environments	High fidelity: prototype that addresses all critical scaling issues	Subsystem/ System	Tested in relevant environments. Verify by test that the technology is resilient to the effects of life-limiting mechanisms

Table 19: TRL 7-9 Decomposition by Factor

TRL	Definition from NPR 7123.1e [1]	Completion Criteria from NPR 7123.1e [1]	Mission Reqt.	Performance/ Function	Fidelity of Analysis	Fidelity of Build	Level of Integration	Environment Verification
7	System prototype demonstration in an operational environment	Documented test performance demonstrating agreement with analytical predictions	Technology demonstration mission	Required functionality/ performance demonstrated	High fidelity: to predict key performance parameters and life-limiting factors as a function of operational environments	High fidelity: prototype or engineering unit that addresses all critical scaling issues	Subsystem/ System	Tested in actual operational environment
8	Actual system completed and "flight-qualified" through test and demonstration	Documented test performance verifying requirements and analytical predictions	Specific mission	Required functionality/ performance demonstrated	High fidelity: to predict key performance parameters and life-limiting factors as a function of operational environments	Final product: flight unit; life-test unit for life-limited items	System	Tested in project environmental verification program. Completed life-tests
9	Actual system flight-proven through successful mission operations	Documented mission operational results verifying requirements	Specific mission	Required functionality/ performance demonstrated	High fidelity: to predict key performance parameters and life-limiting factors as a function of operational environments	Final product: flight unit	System	Operated in actual operational environment

Note that the "Fidelity of Analysis" is a key criterion, because the completion criteria for each of the TRLs depend on the analyses that predict or confirm technology performance. In general, the higher the TRL, the more rigorous the analyses required to support the demonstration of technology maturity – i.e., the scientific knowledge, detailed understanding of physical processes, and data available to support the assessment should increase as the technology matures. A "low fidelity" analysis can be based on "rules of thumb" and qualitative relationships without validation. On the other end of the spectrum, a "high fidelity" analysis is based on analytical physical principles and equations, statistical methods, and/or high-fidelity modeling tools, and must be validated against test results to a low level of uncertainty.

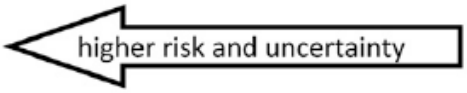
3. TRL assessment using unarguably demonstrated accomplishments

Other objective approaches for conducting TRAs are available.[89] One approach uses unarguably demonstrated accomplishments, as opposed to informed judgments, as bases

[89] Don Clausing & Maurice Holmes, "Technology Readiness," *Research-Technology Management*, 53:4, 52–59, DOI: 10.1080/08956308.2010.11657640, 2010. https://doi.org/10.1080/08956308.2010.11657640, accessed February 11, 2021.

for assessing technology readiness. The approach incorporates six levels of risk and five criteria; the latter are (1) failure modes, (2) critical parameters, (3) latitudes (robustness to perform as needed across the operational environment), (4) design and manufacturability, and (5) integrated technology models. Risks range from very high to very low. Demonstrated (or not) specific accomplishments are associated with each of the five criteria and each of the six levels of risk, composing a 30-element TRA Matrix, an example of which is shown in Table 20.

Table 20: Example TRA Matrix



Criteria	Very High Risk	High Risk	Moderate Risk	Moderate Risk	Low Risk	Very low Risk
Failure Modes	Failure mode analysis completed. Test plan approved.	Failure modes observed under short-run conditions.	Failure modes identified under stress conditions.	Failure modes confirmed in ITR test.	Failure modes updated to include CTP drift and life data.	All failure modes found and addressed.
Critical Parameters	Candidate critical parameters identified.	Parameter sensitivity studies completed.	Critical parameter first optimization completed.	Critical parameter second optimization completed.	All critical parameters verified in ITR tests.	All CTP nominals verified in stress tests.
Latitudes	Latitudes projected from analytical data.	Performance demonstrated at nominal set points.	Control parameter operating windows defined.	Control parameter latitude demonstrated under stress.	80% control parameter latitude verified in ITR.	All CTP latitudes verified.
Design & Manufactur - ability	Concept design produced. Preliminary parts/materials selected.	CTP analysis complete. Critical parts manufactur - ability confirmed with vendor involvement.	CTP specification s identified. Process capability projected.	CTP analysis and variance reduction of design intent is complete.	Supplier verification of manufactur - ability completed.	CTP audit verified CTP specification.
Integrated Technology Model	Integrated technology rig (ITR) and design layout completed.	ITR hardware built. Test plans approved.	Initial ITR stress test complete; architecture stable.	ITR upgraded and stress test completed.	Life tests and assessment completed.	Performance projected to meet goals with variations accounted for.

Note: CTP = Critical Technical Parameter.

Because not every Defense Department program will plan to accomplish all of the activities cited in the TRA Matrix, its comprehensive completion will not always be possible. Nonetheless, using sub-elements of the matrix consistent with a program's plans as well as the rationale for why activities cited in the matrix are absent from program planning could still be useful bases for assessment.

4. Other approaches to assessing technology readiness

There are a number of other approaches to assessing technology readiness than those specifically described in this section. For example, the Government Accountability Office has developed a TRA Guide that maps the characteristics of a high-quality TRA (credibility, objectivity, reliability, and usefulness) to best practices within each of five steps for preparing a TRA.[90]

And, as another example, Naval Air Systems Command has developed an approach that it describes as a "systematic metrics based process used to assess the maturity of Critical Technology Elements (CTEs)."[91] A technology element is critical provided:

1. "The system being acquired depends on this technology element to meet operational requirements (within cost and schedule limits), and
2. The technology element or its application is either new or novel or in an area that poses major technological risk during detailed design or demonstration."

D. Assessing Manufacturing Readiness

1. Introduction and Background

Manufacturing status and risk evaluations have been performed as part of defense acquisition programs for years. These reviews did not always use a uniform set of metrics to assess manufacturing risk and readiness.[92] Studies by the GAO cite a lack of manufacturing knowledge and maturity at key decision points as a cause of cost growth and schedule slippage in major DoD acquisition programs.[93]

Both Congress and GAO have placed additional focus on manufacturing readiness. Specifically, Congress has required that "the Secretary of Defense shall issue comprehensive guidance on the management of manufacturing risk in major defense acquisition programs" and "identify critical technologies and manufacturing processes that

---

[90] *Technology Readiness Assessment Guide*, Government Accountability Office, GAO-20-48G, January 2020. Available at https://www.gao.gov/products/gao-20-48g, accessed October 1, 2021.

[91] Copeland, E. "Technology Maturity, Introduction to the TRA/TMA Process," Naval Air Systems Command, April 2016. Available at https://www.dau.edu/cop/stm/DAU%20Sponsored%20Documents/Copeland%20NAVAIR%20TRA%20TMA%20Process%20Training%20Brief%20Apr%202016.pdf, accessed October 1, 2021.

[92] *Manufacturing Readiness Level (MRL) Deskbook*, Version 2020, available at https://www.dodmrl.com/MRL_Deskbook_V2.pdf, accessed August 23, 2021.

[93] *Weapons Systems Annual Assessment, Limited Use of Knowledge-Based Practices Continues to Undercut DOD's Investments*, Government Accountability Office (GAO -19-336SP), May 2019, available at https://www.gao.gov/assets/gao-19-336sp.pdf, accessed August 23, 2021. Similar conclusions were made in prior GAO reports issued annually since 2004.

need to matured by Milestone A and that have not been successfully demonstrated in a relevant environment by Milestone B." [94,95]

2. Reasons why manufacturing issues persist

The GAO has found that substantial cost growth has occurred as programs transition from development to production, and unit cost increases occur after production begins. Contributing factors to these problems include the following:[96]

- Inattention to manufacturing during planning and design, poor supplier management.
- A deficit in manufacturing knowledge among the acquisition workforce.
- Programs did not identify and resolve manufacturing risks early in development, but carried risks into production where they emerged as significant problems.

The GAO recommended DoD adopt the use of Manufacturing Readiness Levels (MRLs) to help manage manufacturing risk.

3. Policy regarding quality and manufacturing in DoD

Assessment and mitigation of manufacturing risk should begin as early as possible in a program's acquisition life cycle -- including potentially conducting a manufacturing feasibility assessment as part of the Analysis of Alternatives (AoA).[97] According to law, the PM and Systems Engineer should "consider the manufacturing readiness and manufacturing-readiness processes of potential contractors and subcontractors as a part of the source selection for major defense acquisition programs."[98] DoD policy states the following:[99]

---

[94] P.L. 112-81, 31 Dec 2011: § 834.

[95] P.L. 114-328, 23 Dec 2016: § 807.

[96] *Best Practices, DoD Can Achieve Better Outcomes by Standardizing the Way Manufacturing Risks are Managed*, GAO 10-439, Apr 2010.

[97] Interactive MRL User Guide 2020 Version, available at http://www.dodmrl.com/Interactive_MRL_Users_Guide_2020_Version.xlsm, accessed August 23, 2021. This kind of assessment is not explicitly required by DoDI 5000.84 "Analysis of Alternatives," August 4, 2021 (see https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500084p.pdf?ver=2020-08-04-131436-260) but could be included in the associated guidance.

[98] See DFARS (Subpart 215.304) available at https://www.acq.osd.mil/dpap/dars/dfars/html/current/215_3.htm#215.304 , which notes this requirement in section 812 of the National Defense Authorization Act for Fiscal Year 2011.

[99] DoDI 5000.88 "Engineering of Defense Systems," November 18, 2020, Paragraph 3.6.c. Quality and Manufacturing.

"The production, quality, and manufacturing (PQM) lead, working for the Program Manager (PM), will ensure manufacturing, producibility, and quality risks are identified and managed throughout the program's life cycle.

1. Beginning in the materiel solution analysis [MSA] phase, manufacturing readiness and risk will be assessed and documented in the SEP [Systems Engineering Plan].
2. By the end of the TMRR [Technology Maturation and Risk Reduction] Phase, manufacturing and quality processes will be assessed and demonstrated to the extent needed to verify that risk has been reduced to an acceptable level.
3. During the EMD [Engineering and Manufacturing Development] Phase, the PQM lead will advise the PM on the maturity of critical manufacturing and quality processes to ensure they are affordable and executable.
4. Before a production decision, the PQM lead, working for the PM, will ensure that:
   - Manufacturing, producibility, and quality risks are acceptable.
   - Supplier qualifications are completed.
   - Any applicable manufacturing processes are or will be under statistical process control."

4. Documented DoD methods for assessing manufacturing readiness

Achieving low-risk manufacturing readiness includes early planning for and investments in producibility, manufacturing process capabilities, and quality management to ensure effective and efficient manufacturing and transition to production. It also includes objective assessments of the capabilities of the industrial base. Manufacturing risk is evaluated through manufacturing readiness assessments, which should be included in program assessments, in particular in ITRAs, conducted throughout the acquisition life cycle.

Successful manufacturing has many dimensions. The PM and Systems Engineer can assess manufacturing readiness using the considerations described in the Defense Acquisition Guidebook to support decisions made prior to and during the MSA, TMRR, EMD, and production phases of the program.[100] Industry and DoD have also developed MRLs and associated evaluation categories for assessing manufacturing risks to support

[100] Defense Acquisition Guidebook (DAG) Chapter 3, Table 47, available at https://www.dau.edu/pdfviewer?Guidebooks/DAG/DAG-CH-3-Systems-Engineering.pdf, accessed August 23, 2021. The DAG is being updated and the specifics displayed in Table 47 are subject to change.

technical reviews and acquisition milestones. Use of the MRLs should be tailored according to product domains, complexity and maturity of critical technologies, manufacturing processes, as well as is indicated by specific risks that have been identified throughout the assessment process. The MRLs can be summarized as follows:[101]

- "MRLs 1-4: Criteria address manufacturing maturity and risks beginning with pre-systems acquisition (MRLs 1 to 3); continue through the selection of a solution (MRL 4).
- MRLs 5-6: Manufacturing maturation of the needed technologies through early prototypes of components or subsystems/systems, culminating in a preliminary design.
- MRL 7: The criteria continue by providing metrics for an increased capability to produce systems, subsystems, or components in a production representative environment leading to a critical design review.
- MRL 8: The next level of criteria encompass proving manufacturing process, procedure, and techniques on the designated "pilot line."
- MRL 9: Once a decision is made to begin initial production (LRIP), the focus is on meeting both quality, throughput, and rate to enable transition to [full] rate production (FRP)."
- MRL 10: The final MRL measures aspects of lean practices and continuous improvement for systems in production.

The nine categories to be evaluated for each of the 10 MRLs are the following:

- Technology and the Industrial Base---assess the capability of the national technology and industrial base to support the design, development, production, operation, uninterrupted maintenance support and eventual disposal (environmental impacts) of the system.
- Design---assess the maturity and stability of the evolving system design and evaluate any related impact on manufacturing readiness.
- Cost and Funding---examine the risk associated with reaching manufacturing cost targets.
- Materials---assess the risks associated with materials (including basic/raw materials, components, semi-finished parts and subassemblies).

[101] *MRL Deskbook version 2020*. An Interactive MRL Users Guide is also available for use conducting assessments; see http://www.dodmrl.com/Interactive_MRL_Users_Guide_2020_Version.xlsm.

- Process Capability and Control---assess the risks that the manufacturing processes are able to reflect the design intent (repeatability and affordability) of key characteristics.
- Quality Management---assess the risks and management efforts to control quality and foster continuous improvement.
- Manufacturing Workforce (Engineering and Production)---assess the required skills, certification requirements, availability and required number of personnel to support the manufacturing effort.
- Facilities---assess the capabilities and capacity of key manufacturing facilities (prime, subcontractor, supplier, vendor and maintenance/repair)
- Manufacturing Management---assess the orchestration of all elements needed to translate the design into an integrated and fielded system (meeting program goals for affordability and availability).

The 10 MRLs and nine categories form the MRL matrix. The criteria for evaluating each category for each MRL and for a given phase of development are also provided in the matrix. If the questions written in the MRL matrix can be answered in the affirmative, then the MRL for that phase has been met.

5. Consider using maturity models to assess manufacturing

Digital engineering and manufacturing are being employed in DoD acquisition programs in addition to MRLs and maturity models; some of which are on the Capability Maturity Model Integration (CMMI) framework and exist for assessing the capability of producers to employ digital manufacturing. [102, 103]

For example, the approach proposed by De Carolis uses five maturity levels (consistent with the CMMI) as follows:

- Initial
- Managed
- Defined
- Integrated and interoperable
- Digital-oriented.

---

[102] For information about CMMI see https://cmmiinstitute.com/, accessed February 18, 2021.

[103] See, for example, Schumacher A., et al., "A maturity model for assessing Industry 4.0 readiness and maturity of manufacturing enterprises," *Procedia CIRP* (53) 151–166, 2016. www.elsevier.com/locate/procedia; and De Carolis A. et al., "A Maturity Model for Assessing the Digital Readiness of Manufacturing Companies," APMS 2017, Part I, IFIP AICT 513, pp. 13–20, 2017.

These five levels are then used to assess maturity in four "dimensions:"

- Process
- Monitoring and control
- Technology
- Organization

This is just one example of many similar approaches available for consideration. Given the emphasis on innovation in the recently revised Department of Defense Instruction (DoDI) 5000 series of acquisition instructions, specifically analyzing the capability of defense producers to employ existing and emerging digital and cloud-based capabilities for manufacturing should add valuable insights to assessments of manufacturing readiness.

6. Discussions with contractors and using subject matter experts

Whatever assessment methods and techniques are employed, discussions with the contractors regarding the manufacturing technologies and methods they will employ and their approach to assessing manufacturing risk and readiness are ways to gain knowledge on the state of the practice and gain insight into manufacturing risks to the program. Manufacturing SMEs should participate in all assessments and technical reviews. The DCMA has qualified personal in PQM and may be a source of experienced personnel to assist in conducting assessments of manufacturing readiness.

(This page is intentionally blank.)

6. Conclusion

This document discusses 14 best practices for increasing the technical rigor incorporated in DT evaluations and assessments grouped using five categories: (1) General Rigor-Related Considerations; (2) Considering Operational Context and Properly Characterizing Test Results; (3) Using Statistical Methods; (4) Assessing and Using Software Modeling and Simulation; and, (5) Assessing Risks and Readiness. Topics discussed span a wide range and include (but are not limited to): incorporating operational context and sensitivity analyses in DTE&A assessments enabling decision-makers to more fully understand the implications of test results and technical risks for a system's ability to fulfill its operational missions; characterizing the implications of test results so that both the importance and limitations of the available data are understood; using statistical methods to rigorously analyze test sufficiency, test data, and system reliability; choosing software development metrics appropriate for rigorously assessing progress in agile development programs; and conducting rigorous, objective assessments of technical risk.

The topics included were developed based on the authors' experience across a wide variety of aspects of defense test and evaluation. Any finite set of best practices is necessarily not exhaustive. Nonetheless, this set reflects the authors' collective judgment of practices the DTE&A community can adopt or expand the use of that would improve the quality and usefulness to decision-makers of the organization's assessments by increasing the rigor used to generate them.

(This page is intentionally blank.)

7. Abbreviations

A&S	Acquisition and Sustainment
AOA	Analysis of Alternatives
CARD	Cost Analysis Requirements Description
CDR	Critical Design Review
CID	Continuous Iterative Development
CONOPs	Concept of Operations
COTS	Commercial Off-the-Shelf
CPM	Critical Path Method
DCMA	Defense Contract Management Agency
DevSecOps	Development, Security, and Operations
DoDI	Department of Defense Instruction
DOE	Design of Experiments
DOT&E	Director Operational Test and Evaluation
DT	Developmental Test
DTE&A	Developmental Test, Evaluation, and Assessment
DTRAM	Defense Technical Risk Assessment Methodology
EMD	Engineering and Manufacturing Development
ES	Earned Schedule
EV	Earned Value
EVM	Earned Value Management
EVMS	Earned Value Management System
FRP	Full Rate Production
FSR	Field Service Representatives
FW	Fixed-Wing
GAO	Government Accountability Office
HWIL	Hardware-in-the-Loop
IDA	Institute for Defense Analyses
INCOSE	International Council on Systems Engineering
ITRA	Independent Technical Risk Assessment
JCIDS	Joint Capabilities Integration and Decision System
KPP	Key Performance Parameter
KSA	Key System Attribute
LRIP	Low Rate initial Production
M&S	Modeling and Simulation
MBSE	Model Based Systems Engineering
MCS	Monte Carlo Simulation

MDA	Milestone Decision Authority
MOSA	Modular Open Systems Approach
MRL	Manufacturing Readiness Level
MS	Milestone
MSA	Material Solution Analysis
MTBF	Mean Time Between Failure
MTTD	Mean Time to Detect [Software]
MTTR	Meant Time to Restore [Software]
MVCR	Minimum Viable Capability Release
MVP	Minimum Viable Product
NASA	National Aeronautics and Space Administration
NDIA	National Defense Industrial Association
NVP	Next Viable Product
OUSD	Office of the Under Secretary of Defense
PDF	Probability Density Function
PDR	Preliminary Design Review
PERT	Program Evaluation and Review Technique
PM	Program Manager
PMB	Performance Measurement Baseline
PMO	Program Management Office
PNET	Probabilistic Network Evaluation Techniques
PQM	Production, Quality, and Manufacturing
PSM	Practical Software and Systems Measurement
R&E	Research and Engineering
RAM	Reliability, Availability, and Maintainability
RGP	Reliability Growth Plan
SDP	Software Development Plan
SEP	Systems Engineering Plan
SETR	Systems Engineering Technical Review
SIL	Systems Integration Laboratory
SME	Subject Matter Expert
SRA	Schedule Risk Assessment
SRR	System Requirements Review
SW	Software
T&E	Test and Evaluation
TEMP	Test and Evaluation Master Plan
TMRR	Technology Maturation and Risk Reduction
TPM	Technical Performance Measure
TRA	Technology Readiness Assessments
TRL	Technical Readiness Level
TTPs	Tactics, Techniques, and Procedures
VOLT	Validated Online Lifecycle Threat
VV&A	Verification, Validation, and Accreditation

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
10-2021	IDA Publication	

4. TITLE ANDSUBTITLE

IDA Support to DTE&A Initiative: Improving the Technical Rigor in DTE&A Assessments

5a. CONTRACT NUMBER
HQ0034-19-D-0001

5b. GRANT NUMBER
____ ____ ____

5c. PROGRAM ELEMENT NUMBER
____ ____ ____

6. AUTHOR(S)

John S. Hong (SED); James M. Gilmore (SED); Lance E. Hancock (SED); Olivia S. Sun (STD);

5d. PROJECT NUMBER
AX-01-3100

5e. TASK NUMBER
____ ____ ____

5f. WORK UNIT NUMBER
____ ____ ____

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
D-22772
H 2021-000293

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr. Christopher C. Collins

Director, Developmental Test, Evaluation, and Assessments (DTE&A)

10. SPONSOR/MONITOR'S ACRONYM(S)
AX / Dir, DTE&A

11. SPONSOR/MONITOR'S REPORT NUMBER
____ ____ ____

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
____ ____ ____

14. ABSTRACT

At the beginning of FY 2021, the Director of Developmental Test, Evaluation, and Assessment (DTE&A) decided to pursue 18 initiatives to improve the effectiveness of the office. The initiatives span a broad array of topics including both policy and practice. IDA has led the initiative entitled "Improving the Technical Rigor in DTE&A Assessments".

The objective of this Initiative is to identify key practices and recommend specific action and process modifications (1) To help increase the technical rigor in DTE&A assessments to better influence and inform senior leaders making critical acquisition decisions and (2) To develop specific best practices for conducting independent quantitative analyses of information collected during tests and assessing system technical performance and integration maturity in a mission context.

This paper provides IDA's suggestions for 14 best practices DTE&A could employ consistent with achieving the goals DTE&A has stated for the initiative.

15. SUBJECT TERMS

Developmental Test Assessment (DTA); Independent Technical Readiness Assessment (ITRA); Developmental Test and Evaluation; Integrated Decision Support Key (IDSK); Defense Technical Risk Assessment Methodology (DTRAM); Modeling and Simulation (M&S); Software development Metrics; Manufacturing Readiness; Reliability, Availability, and Maintainability; Software Development Plan; Schedule Risk Assessment; Technical Readiness Assessments (TRA); Technical Readiness Level (TRL); Model-Based Systems Engineering (MBSE); Design of Experiments (DOE)

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			John Hong (SED)
Unclassified	Unclassified	Unclassified	Unlimited		**19b. TELEPHONE NUMBER** *(include area code)* (703) 845-2564