GAO

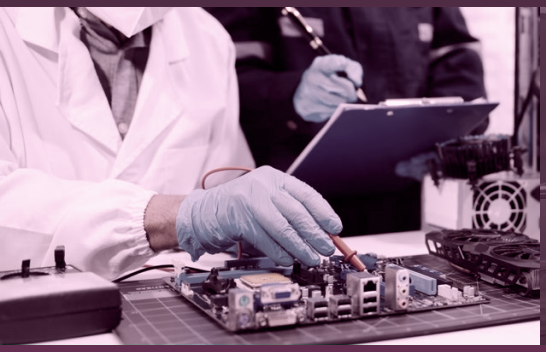
U.S. GOVERNMENT ACCOUNTABILITY OFFICE

GAO BEST PRACTICES

TESTING AND EVALUATION GUIDE

Best Practices for Commercial Off-the-Shelf Systems

(EXPOSURE DRAFT)



GAO-26-108068
September 2026

Contents

Tables

Figures

Abbreviations

COTS	commercially available off-the-shelf
DHS	Department of Homeland Security
DOD	Department of Defense
NASA	National Aeronautics and Space Administration
NIST	National Institute of Standards and Technology
QLR	quick-look report
SOP	standard operating procedure
T&E	testing and evaluation
TR	testing report
VHA	Veterans Health Administration

U.S. GOVERNMENT ACCOUNTABILITY OFFICE
441 G St. N.W.
Washington, DC 20548

Preface

Federal agencies spend billions of dollars each year buying commercially available off-the shelf (COTS) systems, including hardware and software.[1] These systems typically have a wide range of applicability across multiple users, but they also differ based on operational needs. While some COTS systems operate independently (e.g., medical diagnostics instruments), many systems are integrated into larger systems-of-systems (e.g., planes, ships, spacecraft, and ground vehicles). A COTS system might collect data that are critical to key decisions or the overall performance of a larger system. For example, a laboratory diagnostic device that detects the presence of a biological agent will drive treatment decisions, and different systems (e.g., power, communications) need to work together on a spacecraft to manage critical functions.

Effective testing and evaluation (T&E) of a COTS system before purchase allows agency officials to make sound, evidence-based procurement decisions and prevent cost overruns and schedule delays. Effective T&E entails systematic, evidence-based processes that assess the system against specified requirements and might also identify potential risks, such as performance or safety risks, that can be mitigated.

In this guide, we establish a methodology for evaluating COTS system T&E, providing best practices that can be used across the federal government. This methodology can help agencies make sound decisions when evaluating and procuring a COTS system, as well as identify and mitigate risks. Existing guidance from government agencies and industry might include similar strategies for COTS system T&E, but there is no widely accepted process or organization for sharing best practices.

This guide is intended to provide T&E practitioners, program managers, and procurement officials throughout the federal government with a framework of the best practices for conducting high-quality T&E to better understand how systems perform against stated requirements. Manufacturers can also use this methodology to evaluate system performance against agency requirements. Organizations that have their

[1]Federal Acquisition Regulation (FAR) § 2.101 defines “commercially available off-the-shelf (COTS) item” as any item of supply, other than bulk cargo, that is (i) a commercial product, (ii) sold in substantial quantities in the commercial marketplace, and (iii) offered to the government, under a contract or subcontract at any tier, without modification, in the same form in which it is sold in the commercial marketplace. Our scope excluded some of the applicable items as defined by the FAR.

own T&E guidance can use the guide to supplement their practices. Organizations that have not yet developed their own policies might use it to establish their own guidance. This guide can also help GAO and other oversight organizations evaluate federal agencies' T&E and its basis for their conclusions and decisions about a COTS system.

GAO plans to periodically update this guide based on user experience and comments.

If you have any questions concerning the guide, you may contact me at HowardK@gao.gov. Contact points for our Offices of Congressional Relations and Media Relations may be found on the last page of this report. GAO staff who made key contributions to this report are listed in appendix V.

//SIGNED//

Karen L. Howard, PhD
Director
Science, Technology Assessment, and Analytics

Introduction

Federal agencies might opt to purchase a COTS system that meets their mission requirements instead of investing in developing a custom system. To determine whether vendors' offerings meet agencies' requirements, agencies consider the system's technology, scenarios the system might encounter, and actual operational environments where users will deploy the system. One way agencies do this is by conducting testing and evaluation (T&E) of the system, using the results to inform decision-makers during its acquisition.

Many agencies have established policy and guidance documents for T&E that are generally useful as a baseline framework (see app. I). Some of these documents, however, lack the specificity needed to address the unique challenges presented by COTS systems. For example, agencies might not distinguish between mandatory testing for COTS systems and what might be flexible testing. Also, agency documents might not include threshold criteria for COTS systems that affect acquisition decisions. These gaps could be due to policies focused on developing a new system instead of buying and implementing an existing COTS system.

Prior GAO work has identified problems with how agencies test and evaluate COTS systems.[2] We have found that without effective T&E programs, agencies risk cost overruns and schedule delays by procuring and implementing ineffective technologies that fail to meet mission needs. Further, no official government T&E communities of practice currently exist to share best practices and lessons learned between agencies.

For this guide, we focus on operational T&E by federal agencies of complete, integrated, and technically sophisticated hardware systems,

[2]For example, see GAO, Weapon Systems Testing: DOD Needs to Update Policies to Better Support Modernization Efforts, GAO-26-107009 (Washington, D.C.: Dec 11, 2025), GAO, Biometric Identification Technologies: Considerations to Address Information Gaps and Other Stakeholder Concerns GAO-24-106293 (Washington, D.C.: Apr 22, 2024), GAO, Air Cargo Security: TSA Field Testing Should Ensure Screening Systems Meet Detection Standards, GAO-21-105192 (Washington, D.C.: July 29, 2021), and GAO, Biodefense: DHS Exploring New Methods to Replace BioWatch and Could Benefit from Additional Guidance GAO-21-292 (Washington, D.C.: May 20, 2021).

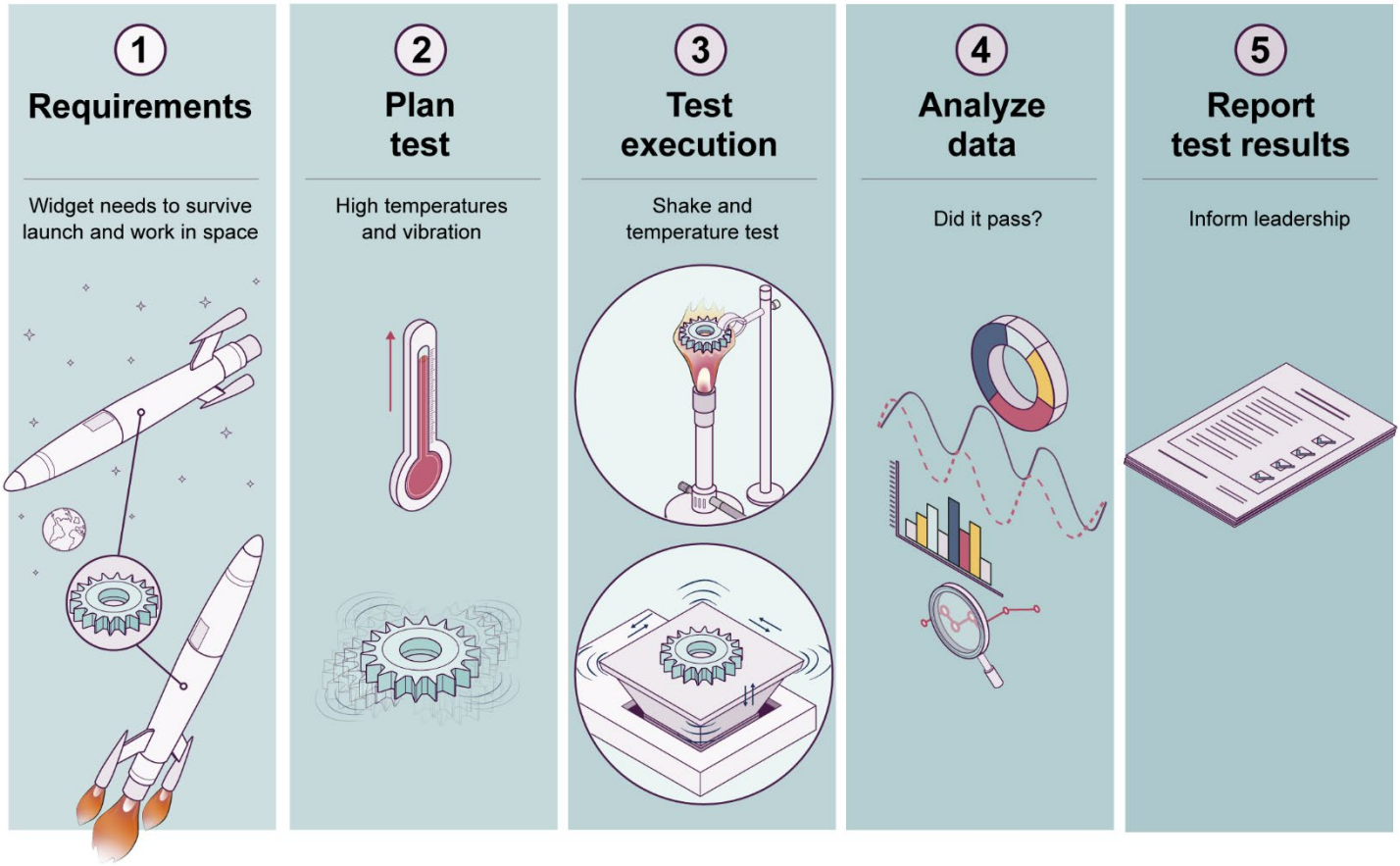
which generally include software. We exclude standalone software, as well as basic components such as nuts, bolts, and tools.[3]

This guide presents a five-step process that can be used across agencies for planning, executing, and reporting the T&E results of a COTS system. This guide also identifies best practices for designing and executing COTS system T&E to encourage agencies to standardize their approaches, manage risks, and help inform system vendors. The best practices might also serve as criteria that GAO or other independent entities can use to evaluate government T&E of a COTS system. We provide key questions that auditors can use for this purpose in appendix II. We used information from existing policies and guidance and the collective experience of T&E experts to develop this method and identify examples of agencies using the practices we present (see app. III for more details).

Agencies should use each of the five steps and the related best practices to ensure a system will enable them to fulfill their specific missions. For example, an agency procuring a system to use on a spacecraft should evaluate the system's performance on withstanding high temperatures and vibration the system would experience when launched into space (see fig. 1).

[3]Federal agencies should refer to T&E standards such as the Department of Defense, DOD Instruction 5000.87 Operation of the Software Acquisition Pathway, Washington, D.C, or the National Institute of Standards and Technology NIST SP 800-53 Rev. 5 Security and Privacy Controls for Information Systems and Organizations, Washington, D.C. for standalone software. Federal agencies should also refer to Section 508 of the Rehabilitation Act of 1973 (29 U.S.C. §794d) for electronic and information technology systems.

Figure 1: The five steps of testing and evaluation for commercially available off-the-shelf systems



Source: GAO. | GAO-26-108068

Although this guide describes the five T&E steps that each agency should follow to ensure program success, it is not intended to overly prescribe how those steps should be accomplished. The specific activities that should be followed to fulfill each step will vary between agencies, and this guide does not describe those in detail. Similarly, the best practices presented in this guide are not all-encompassing; there are likely many other supplemental practices that programs should follow on an agency-by-agency basis.

While this guide implies a linear sequence of T&E steps, in practice agencies might conduct significant evaluation prior to testing. For

example, agencies might conduct extensive background research into a COTS system's part lists, board designs, and other data. Further, agencies should consider whether a system is mission critical or noncritical, because mission-critical systems require more extensive T&E than noncritical systems. Mission-critical systems are those whose failure might have severe consequences, including loss-of-life. In contrast, noncritical systems are those whose failure might not have any significant consequences. For example, an oxygen generation system on a spacecraft would be considered mission critical while a camera might be considered noncritical.

What is COTS Systems Testing and Evaluation?

T&E determines whether a COTS system can perform the functions an agency requires by running tests. These tests use representative end users under a range of operationally realistic conditions and for stated periods of time. Agencies should test and evaluate a COTS system in environments as close as possible to the environment in which it will operate.[4] T&E also evaluates a system's critical characteristics—key performance parameters—essential for mission success.[5] While key performance parameters vary for each system, one parameter often examined during T&E is operational effectiveness—the ability of the system to perform its required function (e.g., detect a biological threat agent).

T&E also often assesses reliability—the ability of the system to operate without a failure over a period of time. T&E reliability metrics might include the number of failures over specific operating times, the average time a system successfully operates between failures, and the average time needed to restore a system to full operation after a failure. T&E can also address potential cybersecurity issues and compatibility with other systems, if part of larger systems-of-systems.

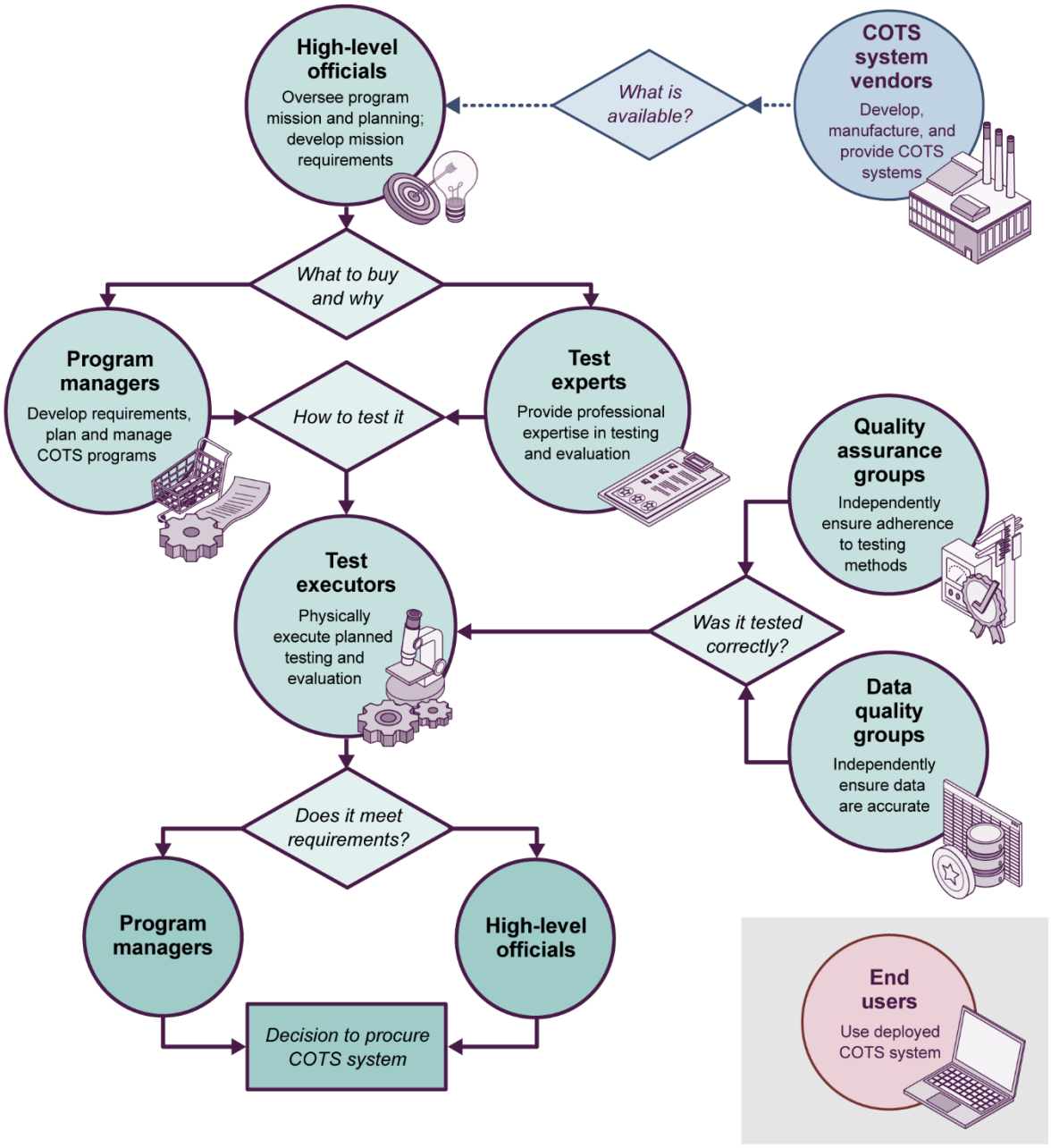
Who is Involved in COTS Systems T&E?

Many different stakeholders are involved in the COTS systems T&E process (see fig. 2). They oversee a broad range of testing-related duties such as developing requirements, planning and executing tests, analyzing data, developing reports, and procuring the system.

[4]Throughout this guide, we will use the term 'test and evaluation' or 'T&E' to refer to the operational testing and evaluation of COTS systems in realistic, representative operational environments with representative users. We will distinguish this from 'developmental' or 'functional' testing conducted in controlled, laboratory environments.

[5]Key performance parameters include operational effectiveness, reliability, operational availability, maximum performance, logistics footprint, and network readiness.

Figure 2: Stakeholders for commercially available off-the-shelf (COTS) system testing and evaluation



Source: GAO. | GAO-26-108068

Stakeholders involved in the COTS T&E process:

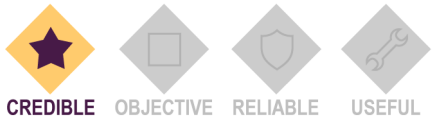
- High-level officials. Agency officials who oversee high-level mission and program planning and develop mission requirements.
- Program managers. Agency officials who develop operational and testing requirements, plan and manage COTS systems T&E, and provide input to the procurement.[6]
- Test experts. Individuals with professional expertise in testing and evaluation and relevant technical skills, for example engineers, statisticians, chemists, pilots, or explosive technicians. These individuals might also be referred to as operational test agents.
- Test executors. Individuals who physically execute planned testing and evaluation. This includes operating test equipment and the COTS systems at the test site and ensuring collection of all needed data. They should be independent from high-level officials that develop the operational and testing requirements.
- Quality assurance groups. Independent groups from the test executors that ensure testing methods follow standards, such as instrument calibration to a standard. They can be internal to the agency or outside contractors
- Data quality groups. Groups that ensure testing data are complete and accurate and independent from the test executors. They can be internal to the agency or outside contractors.
- COTS systems vendors. Companies that develop, manufacture, and provide COTS systems to government agencies.
- End users. The individuals who would be using the COTS system if it is purchased and deployed.

[6]For more information on requirements, see Step 1: Develop and Communicate Requirements, on page 14 of this report.

High-Quality Testing and Evaluation is Credible, Objective, Reliable, and Useful

We have identified four characteristics that define high-quality T&E and pervade all five steps:

T&E should be credible, objective, reliable, and useful. These characteristics are defined as:



Credible T&E has defined requirements that describe the environment in which the COTS system will operate and criteria that the system should meet. Credible testing is comprehensive and will account for all relevant external factors, such as potential interactions with other systems. The test environment will closely resemble the operational environment to ensure the results reflect real-world performance.

Objective T&E draws conclusions based on objective, relevant, and trustworthy data and data analysis. Independent teams conduct testing and report results with a level of detail consistent with the level of evidence provided by test results. Practices such as using statistical analysis improve the objectivity of testing.

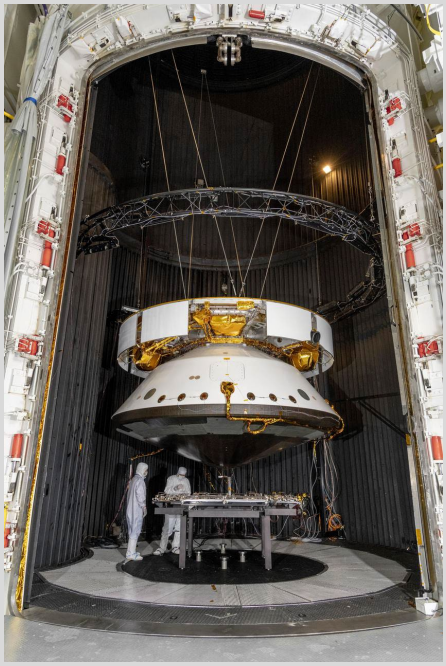
Reliable T&E follows a disciplined process that facilitates repeatability, reproducibility, consistency, and regularity. Reliable testing reports key information to stakeholders, and every test links to a requirement and mission need. Reliable testing will document all actions taken and decisions made such as following a checklist of testing procedures that includes recordkeeping.

Useful T&E provides timely information with sufficient detail to allow program managers and high-level officials to make important COTS system acquisition decisions and modifications to existing programs. For example, if a system fails testing, then decision-makers might try alternative testing, modify the system, procure a different system altogether, or accept the system as-is.

Risk Management is the Foundation of T&E

Example: Considering risk

NASA often tests large COTS systems or their components for spacecraft in a simulated space environment, such as a vacuum chamber. If NASA's vacuum chambers cannot accommodate the size of a COTS system component, testing might need to be done in a contractor's larger vacuum chamber, or the COTS system component might need to be tested in space by putting it on a satellite. Both options might significantly increase testing costs. Program stakeholders would need to weigh the performance and safety risks of not performing the more costly testing. This might include developing a different testing approach such as testing individual parts of the COTS system component that fit inside NASA's vacuum chamber.



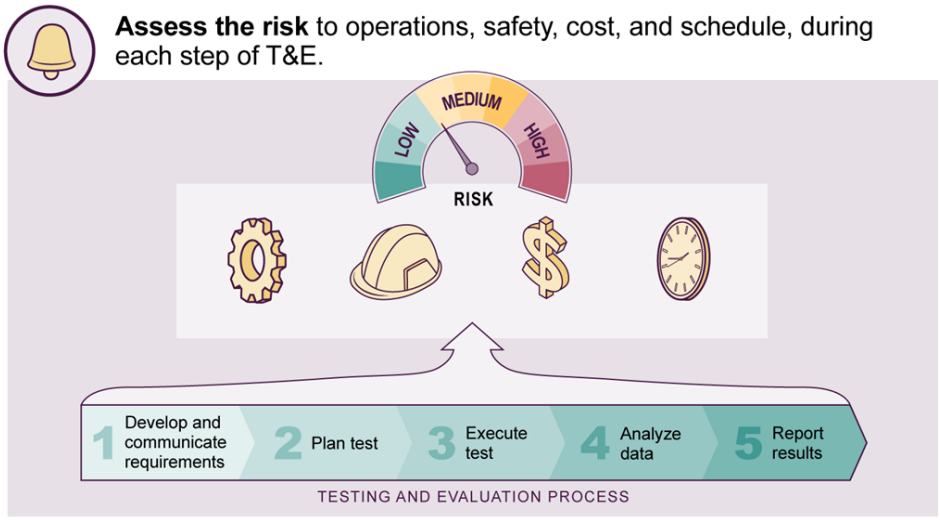
Source: NASA/JPL-Caltech. | GAO-26-108068

For T&E stakeholders, identifying, assessing, and mitigating risks—technical, operational, and financial—throughout the entire T&E process is foundational. Risk is defined as an uncertain factor or event that has an adverse or unwanted effect. Risks specific to COTS system T&E include those that affect schedule, cost, resources, operations, safety, security, or a combination of these. Risk management involves identifying potential risks, assessing their likelihood and potential consequences, and creating mitigations to reduce the risks. All T&E stakeholders have the responsibility to regularly manage risk from the beginning to the end of the entire T&E process. Not considering risk can have negative effects on many aspects of an agency's program.

Assessing the risks to operations, safety, cost, and schedule is an important part of testing a COTS system. There are two types of risk that agencies should consider throughout the T&E process: performance risks and testing risks. Performance risks include those that might result in a system's failure to meet operational and testing requirements. For example, skipping, reducing, or modifying testing of a system might save time and reduce costs, but it might also create performance risks such as increased system failures or decreased operational efficiency of the COTS system. Testing risks include those that might pose safety risks to testing personnel. Testing risks—such as malfunctions that result in a fire or explosion and endanger test executors—might occur during testing in harsh operational environments.

Risk ranges from low to high (see fig. 3). The risk will depend on the COTS system or its components' importance in the overall system. For example, a COTS system that is part of a greater critical pathway system (e.g., a thruster for spacecraft propulsion) will have a higher performance risk (i.e., a failure could seriously impact the spacecraft's mission) than a stand-alone laptop that does not connect to any critical spacecraft systems.

Figure 3: Assess risk during the testing and evaluation (T&E) of commercially available off-the-shelf systems



Source: GAO. | GAO-26-108068

Risk management includes the potential for being too risk averse or overly cautious (i.e., “better or best is the enemy of good enough”). Some testing stakeholders might overcompensate for risk during testing, such as planning for excessive numbers of tests, testing over long time periods to improve data collection, or setting overly strict testing requirements for a COTS system. For example, testing a new system on a ship could be limited by the amount of time the ship will be in port and available for testing. While ideal testing might require hundreds of individual tests over many months, the ship might only be in port for a few weeks. In reality, the test plan should consider the time the ship will actually spend in port and weigh the risks associated with conducting a reduced amount of testing within the limited time against performing the ideal increased number of tests.

Example: Involving end users to reduce risk

Infusion pumps are used throughout a Veterans Health Administration (VHA) hospital to supply medication to patients, including in high-risk areas such as anesthesia for surgeries. One facility purchased infusion pump replacements without soliciting stakeholder input from the anesthesia department. As a result, the anesthesia department could not use the pumps, and VHA had to buy appropriate replacements. VHA used this experience to improve stakeholder reviews and established a standard procurement review process used by all VA medical facilities.

Source: Agency officials. | GAO-26-108068

We identified several best practices that practitioners should use to manage risks throughout the T&E process. These include:

- **Identify and collaborate with stakeholders early and often**. Without input from stakeholders—such as end users—early in the requirements or test plan development processes, decision makers might not appropriately test COTS systems, resulting in the selection of systems that do not actually address end-user needs.
- **Consider requirements for the COTS system at each step**. As testing proceeds through each step, test executors should ensure testing will fulfill testing requirements and address mission needs by engaging high-level testing stakeholders.
- **Share lessons learned with other agencies that conduct similar testing**. Agencies that procure similar COTS systems should share testing lessons learned with each other. This will help prevent duplication of efforts and repeated mistakes, which will save time and money.

Effective risk management practices used within each step of T&E and across all steps should help reduce the probability of a COTS system performance or testing failure or reduce the consequences of the failure should it occur.

The Five Steps of Testing and Evaluation

We divide the process of testing and evaluating COTS systems into five steps, with best practices specific to each: (1) develop and communicate requirements, (2) plan test, (3) execute test, (4) analyze data, and (5) report test results (see table 1).

Table 1: The Five Steps of Testing and Evaluation (T&E) for Commercially Available Off-the-Shelf (COTS) Systems

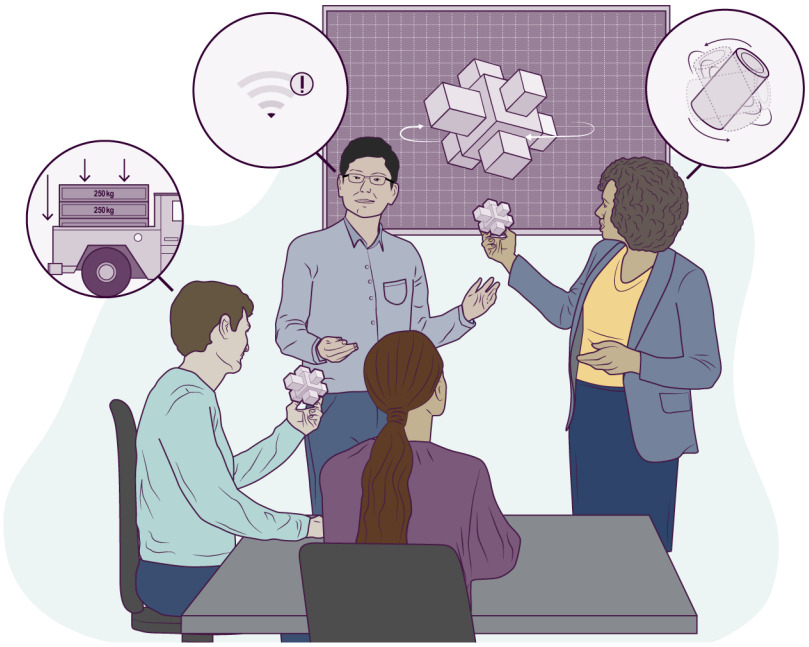
Step	Best practices
1. Develop and communicate requirements	Develop operational requirements
	Develop testing requirements
	Involve testing personnel in requirements development
	Communicate requirements to COTS system vendors
2. Plan test	Use Design of Experiments to develop test plans and procedures
	Develop and follow a test plan
	Develop and update test standard operating procedures (SOP), which are agency specific
	Conduct a test readiness review
	Identify and address resource challenges
	Tailor standard testing procedures from what is provided by agency guidance or from high-level officials, if necessary

Step	Best practices
	Use analysis, modeling, and simulation to evaluate the scope of testing and adapt to testing conditions when appropriate
	Verify any vendor-provided testing data
3. Execute test	Follow SOPs
	Calibrate testing equipment
	Perform dry run testing
	Conduct operational testing
	Appropriately record and review testing data, results, and observations
	Retain raw testing data and results
4. Analyze data	Use data authentication processes to validate testing data
	Use appropriate statistical data analysis to increase the quality of T&E results
5. Report test results	Develop and maintain T&E reporting guidance
	Use Quick Look Reports (QLR) to quickly communicate preliminary testing data
	Prepare and distribute testing report (TR)
	Review and approve the TR prior to release
	Wait for review and authentication of test results before committing to the next procurement cycle
	Archive reports
	Share lessons learned

Source: GAO. | GAO-26-108068

As previously noted, high-quality COTS T&E should be credible, objective, reliable, and useful. Each of the five T&E steps listed in table 1 includes several best practices, each of which should reflect one or more of these four characteristics as shown in the following sections of this guide. Collectively, the best practices for each T&E step should reflect all four of these characteristics.

Step 1: Develop and Communicate Requirements



Source: GAO (illustration). | GAO-26-108068

Best Practices

- Develop operational requirements
- Develop testing requirements
- Involve testing personnel in requirements development
- Communicate operational and testing requirements to commercially available off-the-shelf system vendors

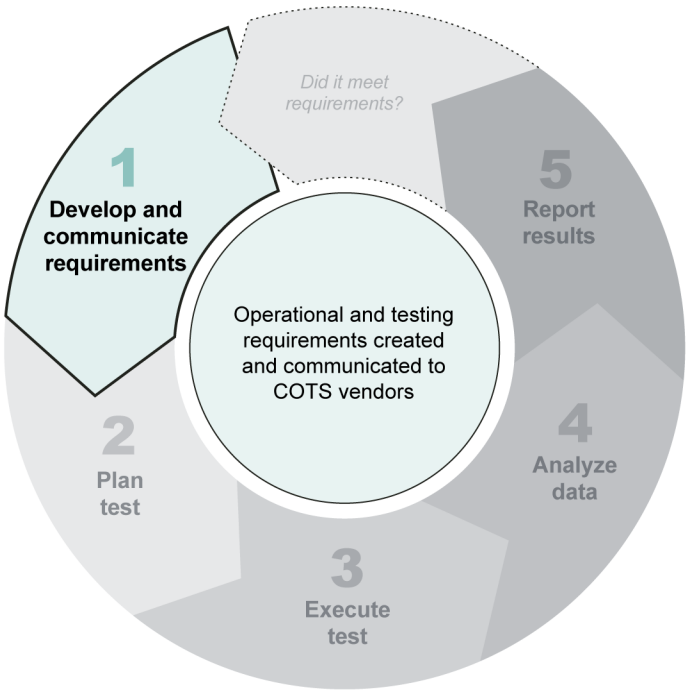
Remember to consider:

Risk

Source: GAO (icons). | GAO-26-108068

Requirements describe the metrics that a COTS system needs to meet. Requirements are the foundation of credible T&E; they define what a system should do, how well it should do it, and under what conditions it should do it. Developing and communicating requirements ensures that T&E results are reliable and useful to agency officials who can rely on these results to make informed decisions about how well a COTS system performs against established metrics (see fig. 4).

Figure 4: Steps of the commercially available off-the-shelf (COTS) systems testing and evaluation cycle
Step 1—Develop and communicate requirements



Source: GAO (image). | GAO-26-108068

Two types of requirements are relevant to COTS systems operational T&E:

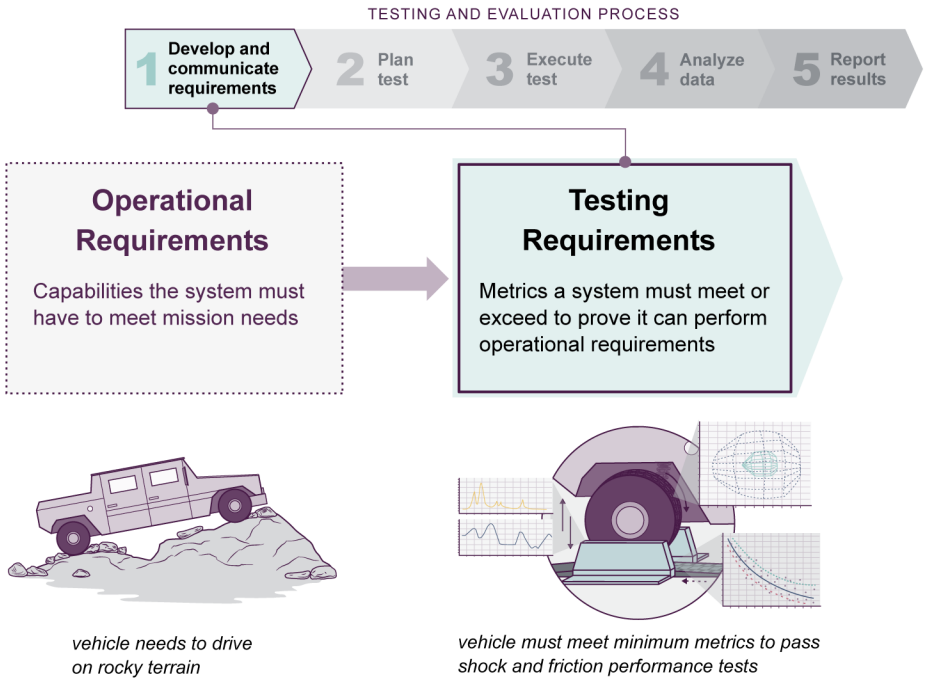
- Operational requirements describe the specific conditions, capabilities, and processes needed for a COTS system to operate effectively in a specific environment and meet an agency's mission needs.[7]

[7]Functional requirements are those used to verify that individual COTS system components perform basic actions in highly controlled, developer-led laboratory environments and are outside the scope of operational testing and evaluation for the purposes of this guide.

- Testing requirements describe the metrics a COTS system must meet or exceed during a test to demonstrate that it meets operational requirements.

Operational requirements drive testing requirements (see fig. 5).

Figure 5: Types of commercially available off-the-shelf system testing and evaluation requirements



Source: GAO (illustrations). | GAO-26-108068

Best Practices for Developing and Communicating Requirements

Develop operational requirements. Operational requirements should use clear and concise language and be measurable, testable, and achievable. Comprehensive operational requirements describe and incorporate the environment in which the COTS system will operate. They also assess and incorporate the unique features of cutting-edge systems and identify the expertise needed to conduct testing. Human factors, ergonomics, and environmental constraints should also be included as necessary. For example, operational requirements should include the systems' ability to perform under high shock, vibration, or extreme temperatures, or account for the complex, high-pressure, decision-making tasks in high-stress environments, such as those found on battlefields or

in space. Good operational requirements will include needs in terms of system function, inputs, processes, outputs, external interfaces, performance, system availability, and maintainability.

High-level officials and program managers should begin developing operational requirements by eliciting input from all appropriate stakeholders, including end users. COTS systems manufacturers might also be engaged to determine recent technical developments and identify any significant capability gaps (see How do agencies know what is out there? text box below). Use cases and use studies might also be helpful for this process. The preliminary operational requirements should be validated with all stakeholders, refined (if necessary), prioritized, and finalized.

This makes testing credible and reliable.

Thorough and thoughtful requirements serve as a foundation for later steps in the testing process and connect the performance of the COTS system during T&E to the actual operational mission need. Further, operational requirements should describe key information to convey to stakeholders and link to a mission need.

Develop testing requirements. Testing requirements describe the criteria a COTS system needs to meet during testing. These requirements should directly reflect the operational requirements to ensure the system can meet the overall operational mission, and they should be measurable, testable, and achievable.

As with operational requirements, high-level officials and program managers should develop testing requirements by eliciting input from appropriate stakeholders, including end users. Testing requirements should define the operational conditions that the COTS system will be tested in, the characteristics of the general test(s) to be conducted under those conditions, and the test metrics the system must meet to pass. The testing requirements might also include the degree of variability in the testing results that will be accepted for the COTS system. For example, a communications microprocessor for a spacecraft might need to show optimal functionality with less than 1 percent deterioration in performance during testing over the course of 240 hours in a vacuum chamber that mimics a space environment.

This makes testing credible and reliable.

Comprehensive testing requirements use the operational performance environment as a basis for selecting test methods, reflect potential adverse interactions with other systems, and describe the criteria the product will need to meet. Thorough and thoughtful requirements make testing credible by serving as a foundation for later steps in the testing process. Additionally, testing requirements should describe tests that are repeatable, reproducible, and link to operational requirements and mission needs.

Involve testing personnel and end users in requirements development. High-level officials developing testing requirements should solicit input from relevant stakeholders, including end users and test executors. Test executors have experience that can help program officials set realistic testing goals using metrics that are testable within a reasonable budget and schedule. For example, a program manager might ask for testing a device up to 99 percent reliability, when a test operator knows that 90 percent reliability would be sufficient. In this hypothetical scenario, improving reliability from 90 to 99 percent would require additional testing that would exceed the operational requirements mandated by high-level officials, wasting time and money. Experienced test operators can also help high-level officials generate better requirements that will lead to improved testability. End user input helps ensure requirements measure how the COTS system will be used in an operational environment.

This makes testing credible, objective, and useful.

Including testing personnel in requirements development will help ensure the correct experts are available to conduct the testing, and the correct operational performance environments are used when selecting testing methods. Additionally, involving test executors and end users should help create realistic testing methods that are not overly prescriptive. This should allow for the generation of relevant and trustworthy testing data in sufficient detail to assist in acquisition decisions.

For some new, unique COTS systems, identifying specific testing requirements might not yet be possible, and flexibility in the testing process may be required for these systems.

Communicate operational and testing requirements to COTS system vendors. Agencies should relay operational and testing requirements to vendors via contracting language in solicitations, including requests for information, requests for proposals, and statements of work that specify testing metrics.

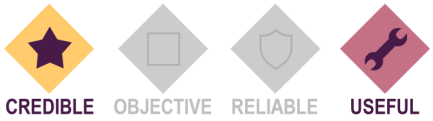
Example: Communicate operational and testing requirements to vendors

Officials from DOD's Director, Operational Test and Evaluation (DOT&E) office said that testing COTS components while integrated with complex systems (such as submarines and ships) necessitates specialized testing requirements that need to be communicated to vendors. For example, testing a COTS component installed in a system on a ship resulted in a reading above the reading produced when the component was tested in isolation. This caused a shutdown of the ship's system. The officials noted that if the specific requirements for testing the COTS component in the ship's system had been communicated to the vendor, the COTS component might have performed better in the operational testing on the ship.

Source: agency officials. | GAO-26-108068

Communicating operational and testing requirements via solicitations can also help vendors when developing new COTS systems that agencies might need in the future. For example, Veterans Health Administration (VHA) officials described a process where they communicated operational requirements to vendors through standardized forms and instructions that helped guide potential vendors on what was specifically needed to deploy systems on its network. As a result, vendors were able to deploy their systems on VHA's network faster and with less confusion and rework than in previous similar situations.

This makes testing credible and useful.



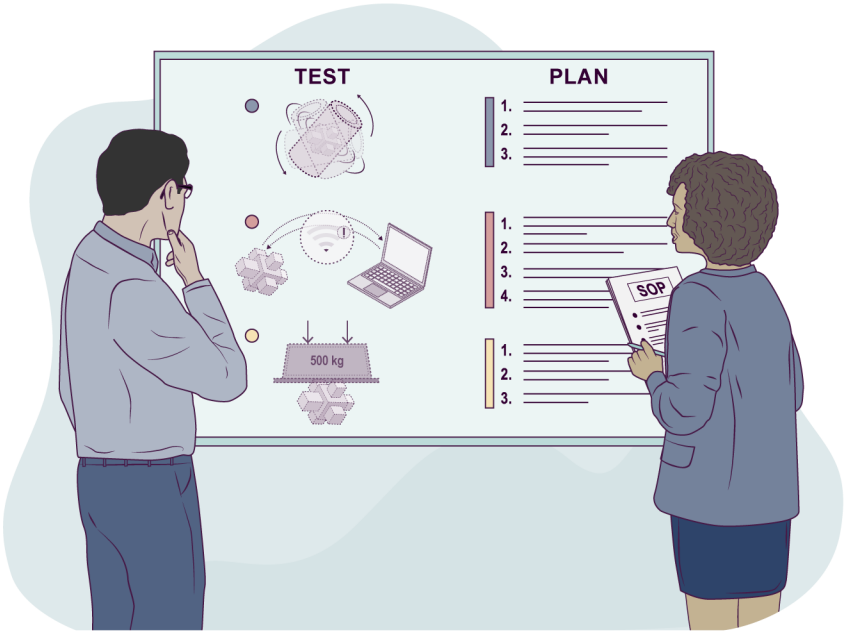
Communicating both operational and testing requirements to vendors allows them to aim to provide systems that will operate as needed for mission needs. This will allow vendors to include testing data in their application packages as support. When agencies receive testing data with vendor applications, the data might also provide them with a starting point for test planning and data analysis, making testing more timely and thus more useful.

How do agencies know what is out there? Agencies use market research to determine what COTS systems are available that might meet operational requirements. Market research might include internet searches, direct outreach to potential vendors, and attending trade shows. Industry days or conferences geared toward certain types of manufacturers are a good opportunity for agencies to advertise upcoming system needs and answer questions from vendors, particularly if attended by agency T&E personnel.

Market research can also save agencies time and money by uncovering more mature or less expensive versions of COTS systems than originally mandated by operational requirements. Additionally, if market research does not reveal the existence of a system that already meets requirements, then an agency will learn that it might need to establish a program to develop a new system instead.

Source: GAO. | GAO-26-108068

Step 2: Plan Test



Source: GAO (illustration). | GAO-26-108068

Best Practices

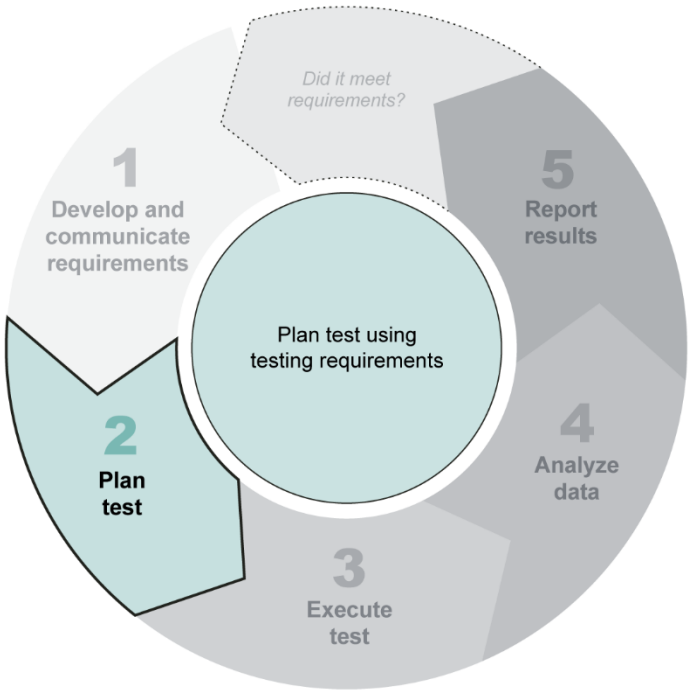
- Use Design of Experiments to develop test plans and procedures
- Develop and follow a test plan
- Develop and update standard operating procedures (SOP), which are agency specific
- Conduct a test readiness review
- Identify and address resource challenges
- Tailor standard testing procedures from what is provided by agency guidance or from senior officials, if necessary
- Use analysis, modeling, and simulation to evaluate the scope of testing and adapt to testing conditions when appropriate
- Verify any vendor-provided testing data

Source: GAO (icons). | GAO-26-108068

After developing the requirements for a COTS system, agencies should plan testing activities that will demonstrate the COTS system meets those

requirements (see fig. 6). Further, every test should link to a requirement and operational use case.

Figure 6: Steps of the commercially available off-the-shelf systems testing and evaluation cycle
Step 2—Plan test



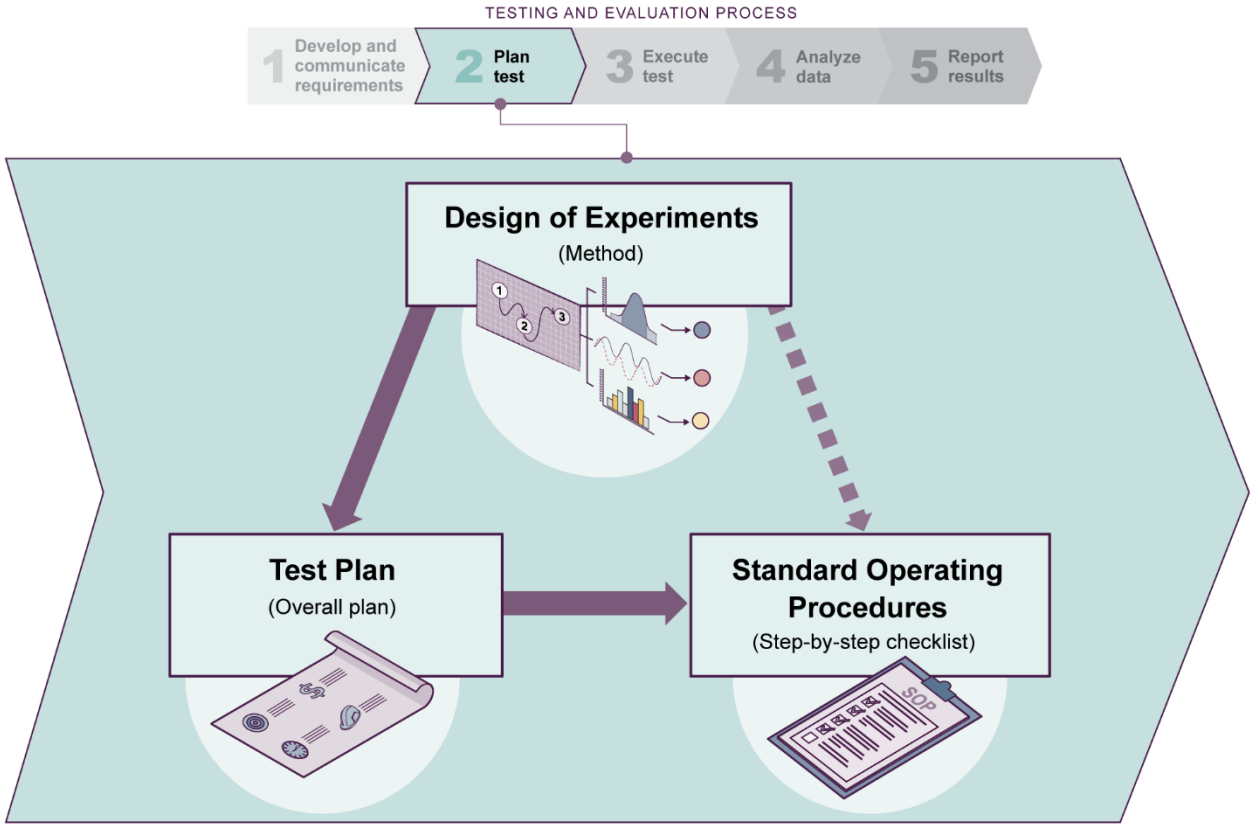
Source: GAO. | GAO-26-108068

The purpose of planning a test is to determine the specific kinds of tests that will be used to evaluate a COTS system, the data needed to assess the system against the requirements metrics, and the resources those tests will require. Resources include staff, schedules, funding, and necessary equipment. Test experts should oversee test planning. Test planning should also identify test deficiencies (i.e., what the testing cannot address) and incorporate mitigation strategies that reduce risk to acceptable levels.

There are three important components of T&E planning: Design of Experiments, the test plan, and SOPs. Design of Experiments is the methodological approach used to design and prepare for the test. It

should be used to develop the test plan, which addresses the overall test strategy and aspects such as safety and resources. Then the test plan should be further refined into SOPs, which contain step-by-step instructions on how the test plan will be completed. While Design of Experiments is used primarily to develop the test plan, its principles also influence the SOPs (see fig. 7).

Figure 7: Components of commercially available off-the-shelf systems test planning



Source: GAO. | GAO-26-108068

Best Practices for Planning a Test

Use Design of Experiments approaches to develop test plans and procedures. Many factors (e.g., temperature, pressure, vibration, radiation) in an operational environment can have an impact on a COTS system's performance. When multiple factors need to be examined during

testing, agencies should use Design of Experiments.[8] Design of Experiments is a methodology that combines multiple factors in a single study, reducing the number of tests needed to evaluate a COTS system while maintaining statistical robustness. Since multiple factors are tested at once rather than a single factor in a single study, overall testing efficiency increases. Additionally, Design of Experiments defines the minimum number of tests needed to achieve desired testing precision. This helps to limit potential over-testing that would result in increased cost and time. The method also provides solutions to the problems caused by "nuisance variables" or "noise," which are factors that cause random variability in test results.

Design of Experiments use should begin early in the test planning process. Program managers and test experts should use subject matter experts—for example, those familiar with technical methods such as ballistics testing —to identify the test measures that will be used to characterize the performance of the COTS system. Program managers and test experts should also identify environmental and operational factors that the COTS system will be expected to encounter, such as variable temperatures or pressures.

This makes testing objective and reliable.

Using Design of Experiments helps ensure T&E is based on objective, relevant, and reliable data and data analysis. It also helps ensure that T&E follows a reliable, disciplined, repeatable, and traceable process.

Develop a test plan. Test planning should include standard methods that describe how the testing should be done, how the data should be collected and analyzed, and how the results should be reported. Agencies have different methods of developing and documenting their test plans, such as through Scientific Test and Analysis Techniques.[9] Agencies sometimes refer to the test plan as a Test and Evaluation Master Plan or an Operational Test Plan. While the scope and type of testing will depend on the COTS system being tested and the operational environment being tested, most test plans should include the following at a minimum:

- Purpose, goal(s), and objective(s) of each test—why testing is being done and what it is supposed to achieve,

[8]For example, see Jacqueline K. Telford, *A Brief Introduction to Design of Experiments,* Johns Hopkins APL Technical Digest, Volume 27 No. 3 (2007), *Guidance on the use of Design of Experiments in Operational Test and Evaluation*, DOT&E, October 19, 2010.

[9]See, for example, Department of Defense, *Test Resources Guide 3.0*, Washington, D.C.

- Test resources needed—personnel, testing materials, data collection instruments, COTS system(s) and responsibilities,
- Safety and security requirements—how testing will be performed to avoid harm to personnel or the test environment as well as to protect sensitive data and operational objectives,
- Test execution specifics—detailed testing execution instructions (methods and techniques) usually described in SOPs (see Develop and update SOPs below) and work papers including test schedule, instrument calibration instructions, what data to record, and evaluation criteria,
- Data analysis plan—describe how data will be analyzed and reviewed and any statistical measure(s) to be used (important for understanding how much testing is needed and instilling confidence in results; included in Design of Experiments),
- Test reporting—results of testing and whether the COTS system met test requirements.

Some test plans might include a description of modeling and simulation that agencies might use to verify instrument setup, testing schedules, proposed testing scenarios, and predict results.[10] For example, modeling and simulation might help to predict a COTS system performance under operationally representative conditions that might be impractical to achieve during field tests due to resource constraints or safety issues. While data from modeling and simulation might be used to inform T&E planning, the data from T&E might also be used to inform the development and refinement of the modeling and simulation.

This makes testing objective and reliable.

Creating test plans that follow established testing methods such as Design of Experiments facilitates consistency in testing, helps ensure all steps are followed, and mandates that appropriate testing data are being collected and measured.

[10]Modeling and simulation refers to the use of computer models and virtual simulations that approximate, represent, or idealize selected aspects of the structure, behavior, operation, or other characteristics of a real-world process, concept, or system. See Institute of Electrical and Electronics Engineers, *IEEE Standard Glossary of Modeling and Simulation Terminology*, IEEE Std 610.3-1989 (New York: IEEE, 1989), 1–0, doi.org.

Example: Positive and negative controls in clinical diagnostics

In clinical diagnostics, a positive control uses a sample spiked with a known material such as virus nucleic acid that will yield a positive test result. A negative control uses a sample of pure water or other material that is known not to yield a positive result.

Source: Agency officials, DC Studio/stock.adobe.com. | GAO-26-108068

Develop and update SOPs. SOPs are derived from the test plan and provide detailed instructions on how to conduct T&E to meet testing requirements. For example, if the test plan mandates conducting a heat test on a COTS system, the SOP breaks down that mandate into actionable steps. These steps would include detailed instructions on how to turn on the heat source (e.g., oven) and set the thermostat to a desired temperature, calibrate a thermometer to measure the temperature, place the COTS system into the oven, remove the system at a certain time, and measure the effect of the heat on the system. SOPs should include the purpose and scope of the T&E, responsibilities, work instructions, checklists, lists of materials and equipment, and potential hazards. SOPs should also include instructions for calibrating testing equipment to ensure it is accurately capturing testing data. SOPs might also include a description of testing controls (negative and positive) as appropriate. Positive controls such as spiked samples demonstrate that the testing procedures, materials, and equipment all work and produce known, expected results. Negative controls such as blank samples demonstrate that the testing procedures, materials, and equipment all work and produce a zero response or a negative result.

SOPs should be regularly reviewed and updated, with all revisions tracked. Typically, SOPs are supported by work papers documenting that each outlined step before, during, and after testing has been followed.

This makes testing credible, reliable, and useful.

Documenting and following SOPs facilitates consistency in testing and helps ensure all steps are followed, making testing credible, reliable, and useful. Outlining details such as controls might add confidence in testing results. Positive controls ensure reliability and help to confirm that a negative result is not due to experimental error. Negative controls also ensure reliability and help to confirm that a positive result is caused by a test variable. They can also reveal test system contamination, improper instrument calibration, or the presence of other factors that cause unintended reactions.

Conduct a test readiness review. A test readiness review demonstrates that the test plan for the COTS system is sound, and testing is ready to initiate.[11] In a typical test readiness review, the testing team (e.g., test experts, test executors, quality assurance group) review the test plan with program management to identify and address any concerns or issues

[11]For operational testing, the test readiness review might also be referred to as the operational test readiness review.

prior to beginning testing. For example, during the review the testing team might note that the test plan specifies using pressures higher than the pressure rating of a COTS system. Since these pressures might pose a safety risk to the test executors and destroy the COTS system, the testing team would need to modify the test plan to use lower pressures or only test systems with higher pressure ratings. This review also includes confirmation that all resources are available to conduct testing according to the planned schedule.

This makes testing reliable and useful.

Test readiness reviews implement a disciplined process for test plan review and informing decision-makers about potential areas of concern or risk. If issues are identified during the test readiness review, management might choose to cancel or postpone the testing. Management might also choose to accept the risks and proceed with the testing as is or mitigate the risks to an acceptable level.

Identify and address resource challenges. Choosing what tests will be appropriate for assessing a COTS system includes determining what resources will be needed to perform those tests. Resources and their challenges include:

- Schedule - how long testing will take. The urgency of mission needs for a system can affect what kind of testing is possible in the time available. Time limits might require test method substitutions or adjustments to testing requirements.
- Cost - how much funding is available for testing. Overall program budget limits will determine how much funding is available for testing. Limited funding availability might require test method substitutions or adjustments to testing requirements.
- Personnel - who will work on testing. Testing requires that enough test executors with the correct skillsets are available. If personnel lack the appropriate skillsets, training might be needed.
- Materials and equipment - what will be used to test the COTS system and collect data. The types and amount of materials and equipment will vary based on the complexity of the COTS system and the testing.
- Facilities - where the testing will be performed. For example, testing might be conducted in laboratories, specialized facilities, test ranges, or operational sites.
- Data rights - access to underlying software data. Sometimes vendors will maintain ownership of data within a COTS system, which means

that agencies must rely on vendor-provided testing information on certain aspects such as software code.

This makes testing useful.

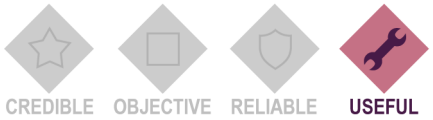
Identifying resource challenges early in the testing process can help mitigate problems or prevent them altogether, avoiding delays and making testing useful for decision-makers. For example, a Defense Innovation Unit official said that for a surface ship program, identifying testing resources, including testing location availability, and any applicable regulations was necessary for efficient testing. The program's test team did not plan for their vendor testing and evaluation requirements, which contributed to schedule and cost increases for the surface ship program. The team learned to identify requirements early on and designed testing around limited resources.

Agencies might address limited resources by removing COTS components from testing, replacing them with similar components, isolating or testing individual components instead of an entire system, or reducing the extent of testing.

Tailor standard testing procedures from what is provided by agency guidance or from high-level officials, if necessary. COTS systems procurement requires streamlined, agile approaches, and testing teams should have the flexibility to adjust tests as necessary. This allows for robust T&E information to be quickly provided to acquisition decision-makers so that systems can be fielded as rapidly as possible. Experts often note the need to deviate from or tailor standard testing procedures when necessary to expedite the collection of testing information. This ability is especially important with cutting-edge technology that might not have been tested before or quickly evolves, and rapid fielding is necessary for mission requirements.

For example, NASA officials noted that they tailored procedures when the agency needed to test COTS pumps in a flight-like environment. Vendor-provided internal testing information indicated the testing environment could damage the pumps. This information allowed NASA to tailor its testing requirements during test design to prevent pump damage. NASA also modified the pumps to prevent damage in a flight-like environment and still meet testing requirements. To ensure tailoring is only used as necessary, testers should obtain permission from high-level program officials to deviate from or tailor standard testing procedures and document any testing modifications.

This makes testing useful.



Tailoring can speed up testing and help reduce costs by reducing the amount or length of testing, or by eliminating the need for specialized testing facilities.

Use analysis, modeling, and simulation to evaluate the scope of testing and adapt to testing conditions when appropriate. Analysis, modeling, and simulation approaches might vary between agencies, and can be used to evaluate the need for facilities, amount of testing needed, requirements, length, and scope of testing for COTS components.

Agencies might consider conducting analysis, modeling, and simulation to predict COTS performance, assess risk and safety issues, and optimize testing conditions prior to conducting testing in specialized facilities (e.g., NASA vacuum chambers) with schedule demands.

This makes testing credible and reliable.

Modeling and simulation help testers use the operational performance environment as a basis for selecting testing methods and creating a testing environment as close to the operational environment as possible.

Verify any vendor-provided testing data. In most instances, any vendor-provided testing data should be verified by additional agency testing. Vendors will often test their own COTS systems to demonstrate that their systems will meet requirements and are therefore eligible for procurement. They then present this information as a qualification data package to agency officials, but agencies should still verify those data by conducting their own in-house testing. Further, vendors might not have conducted all necessary testing themselves, such as radiation testing, which requires specialized equipment.

The verification of testing data might not need to be extensive if the vendor designed, developed, and tested the COTS system against an operational requirement previously communicated by an agency during the solicitation process. Extensive verification might also not be required when acquiring a COTS system that is readily available in a standard configuration, such as laptop computers.

This makes testing objective and reliable.

Confirming that vendor testing data are accurate helps ensure testing is based on objective, relevant, and reliable data and data analysis. Inaccurate data from vendors without verification by agency testing might increase the risk of a COTS system's performance or test failure.

Step 3: Execute Test

Source: GAO (illustration). | GAO-26-108068

Best Practices

- Follow standard operating procedures (SOP)
- Calibrate testing equipment
- Perform dry run testing
- Conduct operational testing
- Appropriately record and review testing data, results, and observations
- Retain raw data and results

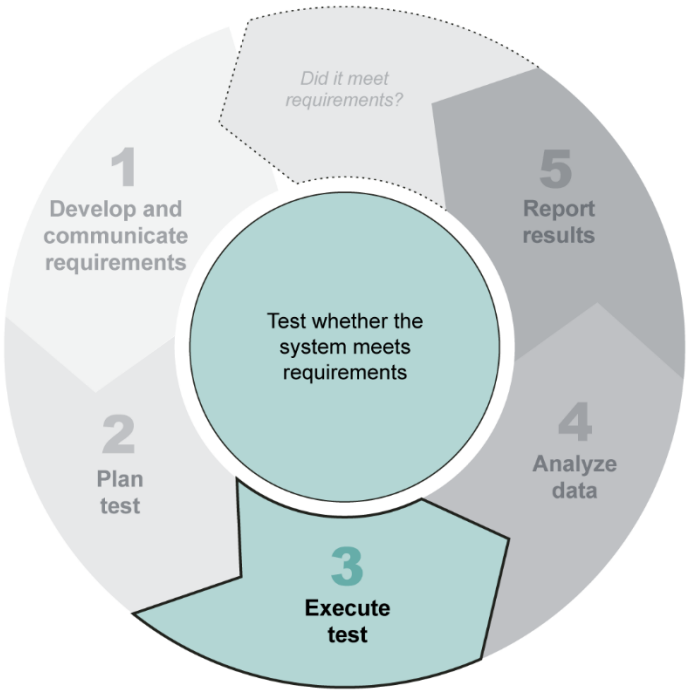
Remember to consider:

Source: GAO (icons). | GAO-26-108068

The purpose of executing a test is to collect data that officials will use to determine whether a COTS system meets defined testing requirements. While actual testing methods will vary greatly by agency and type of COTS system, there are best practices that can apply to all types of testing. Depending on the test plan, test officials might execute testing in-house, externally, or by a third party, and best practices will vary depending on who executes the testing. The application of the best practices in this step will help ensure testing is credible, objective, reliable, and useful.

Executing a test implements the test plan and allows agency officials to determine whether a COTS system meets testing requirements and, in turn, operational requirements (see fig. 8).

Figure 8: Steps of the commercially available off-the-shelf systems testing and evaluation cycle
Step 3—Execute test



Source: GAO. | GAO-26-108068

Best Practices for Executing a Test

Follow SOPs. SOPs should be followed as closely as possible. Any deviations from the SOPs during any phase of testing should be approved by management before the change in procedure is used. Any change should also be appropriately documented.

This makes testing reliable.

Following SOPs for executing a test ensures a disciplined, repeatable, and traceable process (i.e., every test links to a requirement and operational use case).

Calibrate testing equipment. Before conducting operational testing, testing equipment should be appropriately calibrated. Calibration is the process of adjusting or checking a measuring instrument against a known

standard or reference material to ensure data accuracy.[12] For example, any pressure gauges, thermometers, radiation detectors, or analytical balances should be calibrated prior to use in T&E. Calibration should use standards traceable to the National Institute of Standards and Technology (NIST) or specified by other authorized national or international standards as appropriate.[13]

This makes testing objective and reliable.

The calibration of testing equipment ensures that data are accurate, repeatable, and traceable.

Perform dry run testing. Dry run testing (sometimes referred to as burn-in, system, or integration testing) involves performing practice runs after installing the COTS system but before conducting full operational testing. This provides the testing team with the opportunity to become familiar with the COTS system under operational conditions and ensure the SOPs needed for operational testing are correct, complete, and produce repeatable results. This testing also allows the team to work out any problems and refine SOPs as necessary in advance of operational testing. If the system experiences significant issues during dry run testing, the team—with approval from program officials—might opt not to proceed with operational testing until the issues have been addressed.

This makes testing credible and reliable.

Dry run testing helps ensure tests have accounted for all relevant external factors and the test environment closely resembles the operational environment. Additionally, it enables the test team to determine whether the test plan and SOPs have all the essential information needed to correctly conduct the testing. Dry run testing also allows for modifications to the test plan and SOPs if deficiencies are identified.

Conduct operational testing. Once dry run testing has been successfully completed and the testing team is comfortable with the COTS system and the testing SOPs, operational testing begins. Operational testing assesses effectiveness and suitability of the COTS

[12]Accuracy describes the closeness of agreement between a measured value and a known, true value. Standards are reference instruments or materials that are used to verify and adjust the accuracy of other instruments.

[13]Examples of calibration and measurement standards include MIL-STD-1839D (DoD Calibration and Measurement Requirements), ANSI/NCSL Z540.3 (Requirements for the Calibration of Measuring and Test Equipment), ISO 10012 (Measurement Management Systems), and specific NIST calibration guidelines, among others.

system in accordance with the Design of Experiments, the test plan, and all SOPs.

This makes testing reliable.

Operational testing is conducted in an environment that is representative of the operational environment, is conducted by an independent and objective testing team, and determines whether requirements were met.

Appropriately record and review testing data, results, and observations. The testing team should record all data, results, and any observations in real time as testing proceeds. Data should be recorded using official written documentation or approved electronic processes as described in the SOPs. Any procedural deviations, data anomalies, test failures, safety issues, or other testing observations should also be recorded.

T&E managers and program managers should conduct a review of the testing results with the testing team at appropriate times throughout the process depending on the COTS system and test plan. Interim reviews might be completed at the end of each test or at the end of multiple tests to clarify and address in real time any issues that are not fully documented and identify how the system might be performing against the testing requirements. The overall performance of the system, however, is unlikely to be determined from cursory review of test data alone. Testing teams will need to conduct further in-depth analysis, which includes statistical analysis, to determine whether the COTS system meets operational and testing requirements (see step 4).

Upon completion of all testing, a final testing review should be conducted by testing teams to ensure SOPs were followed and data recorded appropriately and address any significant issues in the data or observed during the testing that might require repeat testing of the COTS system.

This makes testing objective, reliable, and useful.

Recording all key information in an appropriate amount of detail helps ensure testing is objective and reliable. Providing information on how well a test functioned is useful, giving decision-makers the ability to take action, such as by considering alternative testing or an alternative COTS system.

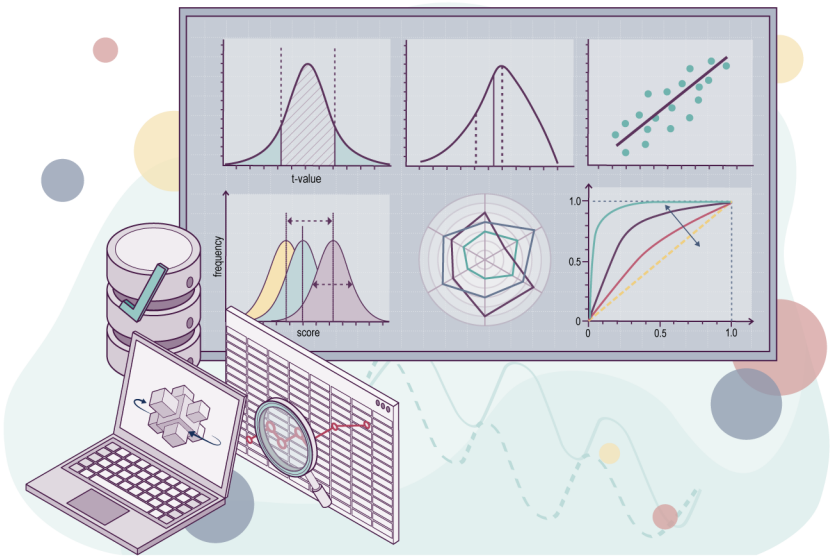
Retain raw testing data and results. A common data repository should be established to retain testing raw data and results. The repository should also include records of testing irregularities such as unexpected data, COTS systems failures, and safety issues, and should be readily accessible for review by stakeholders. Stakeholders can then address

irregularities, failures, and other issues prior to repeating T&E. It also provides important information for audits of T&E. Record retention should be effective and readily retrievable and follow agency record retention policies and governance to ensure high-quality data as well as all rights, reproducibility, and chain of custody of the data.

This makes testing useful.

Retaining testing data helps stakeholders to identify data trends and mitigate issues that may have been observed during previous tests.

Step 4: Analyze Data



Source: GAO (illustration). | GAO-26-108068

Best Practices

- Use data authentication processes to verify the quality of test data or to validate system performance
- Use appropriate statistical analysis to verify and validate testing

Source: GAO (icons). | GAO-26-108068

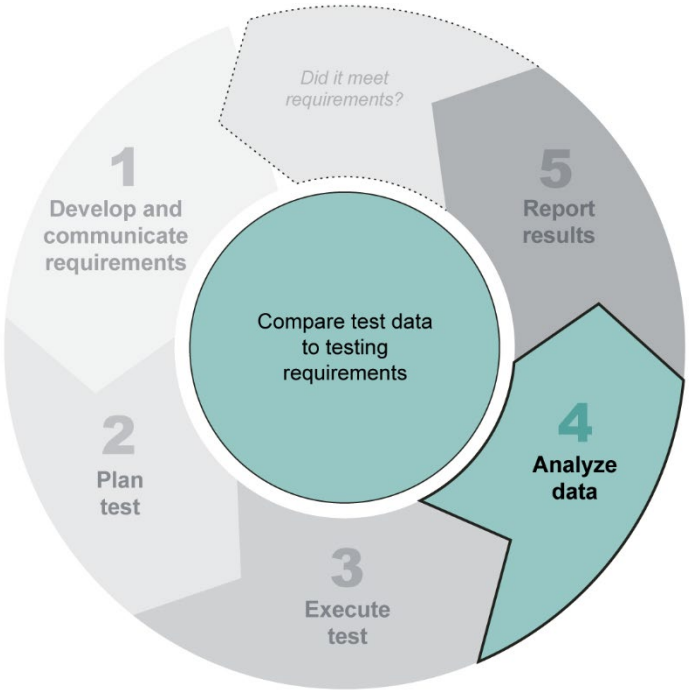
The purpose of analyzing test data is to determine whether a COTS system met the testing requirements, which in turn demonstrates the system meets the operational requirements and can support the operational mission. For example, a COTS explosive detection device that meets requirements for fast and accurate identification; false positive and false negative rates; and resistance to shock, vibration, and dust would therefore demonstrate it meets the agency's needs for accurately and reliably detecting explosives.

Analysis involves several steps, including preparing and validating the data. Analysis often begins by reducing, translating, and analyzing raw, often electronic, signals collected from testing instruments into organized and meaningful data packages. Additionally, data validation involves systematic review of data to identify uncertainties and establish measured confidence in testing results. Validation provides an overall appraisal of the quality of a dataset that the testing team can use to accept or reject data. Validation also allows for the identification of gaps and variances in raw test data to identify voids or outliers that might affect the testing

results. The application of best practices in this step will help ensure testing is credible and objective.

Data should be provided timely and regularly to support efficient and effective operational testing objectives and acquisition decisions, as well as to support timely assessments of technical and operational performance and risks to the T&E. Data analysis takes the information produced by the execution of a test and determines whether a COTS system met or did not meet test requirements (see fig. 9).

Figure 9: Steps of the commercially available off-the-shelf systems testing and evaluation cycle
Step 4—Analyze data



Source: GAO. | GAO-26-108068

Best Practices for Analyzing Data

Use data authentication processes to validate testing data. The test execution team should put together a data package for a comprehensive quality assurance check by an independent quality assurance or data authentication group. These groups will review the testing data to confirm that SOPs were followed, equipment was appropriately calibrated, and

data were accurately recorded. The data authentication review is usually done on a continuous basis as data become available to make sure that the information is consistent, complete, accurate, and representative of the operational environment. For example, a test result in the field might report a review time of 22 seconds, but the timestamps of the data indicate the review time was 1 minute and 22 seconds. This discrepancy would be reviewed by the test execution team to determine the cause (e.g., a data entry error or a mis-calibrated data recorder) and correct the issue. If the cause cannot be determined and corrected, a decision should be made to determine whether the data are still usable or if testing needs to be repeated.

Conducting data authentication reviews on a continuous basis enables testing and data errors to be identified and corrected during T&E rather than at completion. This will increase the overall quality of the data and might help to prevent schedule delays.

Only after the testing data have been validated should test execution teams conduct data analysis to demonstrate that the COTS system met operational and testing requirements.

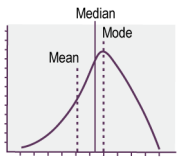
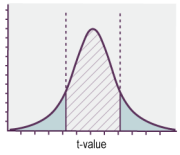
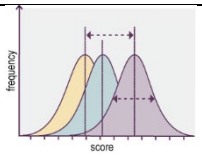
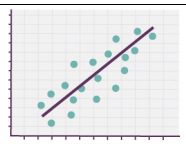
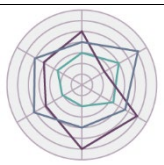
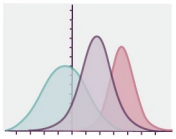
This makes testing credible, objective, and reliable.

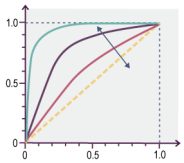
Data authentication processes validate that testing data are representative of the operational environment, high quality, relevant, and trustworthy. Data authentication also helps ensure data are consistent and reproducible. The data authentication step assures that subsequent analysis and evaluation of the COTS system is based on data that accurately reflects the real performance of the system.

Use appropriate statistical data analysis to increase the quality of T&E results. Leveraging the use of statistical data analysis can be beneficial to COTS system stakeholders and acquisition decision-makers by providing defensible and informative test results. While statistical analysis might not be necessary in all cases (e.g., small sample sizes or simple “pass/fail” tests), it might be used to confirm that a system meets or exceeds the required level of reproducible performance (e.g., 95 percent accurate, 99 percent specific, available 99.9 percent of the time). Statistical analysis helps to verify that systems are within threshold tolerances, establishes confidence intervals and uncertainty levels, and helps stakeholders understand whether test failures are within the realm of an acceptable variation to meet requirements. Statistical analysis might also allow stakeholders to identify patterns and relationships in the testing data or predict future outcomes.

Agencies can use a variety of statistical analysis methods for T&E. These can include simple statistical analyses such as descriptive statistics, or complex methods such as Bayesian analyses (see table 2).

Table 2: Selected statistical analysis methods for testing and evaluation analysis

Statistical method	Description	Benefits	Limitations	Example
Descriptive	Summarizes and depicts the average (mean), middle (median), or most frequent value of a dataset (mode) and the spread, range, and variation of the data	Simple—depicts main data features	Cannot infer causal relationships or predict future trends	
T-test	Compares the averages (means) of two datasets	Simple—determines significant differences between the averages (means) of two datasets	Cannot compare more than two datasets; relies on strict assumptions; sensitive to outliers	
Analysis of variance	Compares the averages (means) of three or more datasets	Simple—determines significant difference between the averages (means) of three or more datasets	Only indicates that at least one dataset average is different; does not specify which group differs	
Simple linear regression	Finds the best fitting straight line through multiple data points	Simple—models relationships between data and allows for predictions of an outcome or value	Assumes linear data relationship; cannot be used for complex, nonlinear data; sensitive to outliers	
Multivariate analysis	Simultaneously analyzes relationships among three or more variables	Shows complex relationships, patterns, and dependencies among variables	Requires large datasets; very sensitive to outliers	
Bayesian	Calculates the probability of a hypothetical event (posterior) based on combining current data (likelihood) with prior data (prior)	Complex—ability to calculate probabilities of present or future events based on prior results; can analyze complex systems; allows for analysis of uncertainty	Requires the identification of which prior data to use and how to incorporate those data into the analysis	

Statistical method	Description	Benefits	Limitations	Example
Receiver operating characteristic curve	Evaluates how well a model (e.g., COTS system) performs by comparing sensitivity (true positive rate) against specificity (false positive rate)	Complex—allows for performance estimates and threshold setting	Might provide misleading high- performance estimates if one data class dominates	

Source: GAO analysis of government and academic documents; GAO (icons). | GAO-26-108068

Example: statistical analysis to improve reliability estimates

The Institute for Defense Analyses used a Bayesian analysis to combine engineering reliability estimates with actual failure data to estimate component reliability for the F-35. The analysis technique allowed the Institute to overcome the problems of cases where few or no failures were observed. Combining engineering knowledge of reliability with observed operational reliability resulted in a more informed estimate of each individual component's reliability and a more informed estimate of the overall F-35 maintenance costs.

Source: Agency officials. | GAO-26-108068

Statistical analysis of test data should align with the test designs developed using Design of Experiments, which lays out the number of tests to be conducted, and accounts for environmental and operational factors that are expected to drive the performance of the COTS system. Statistical analysis should clearly describe the number of samples to collect and modify the number accordingly based on whether comparing against an established baseline or an established requirement. Specific assumptions should also be included for field-based testing.

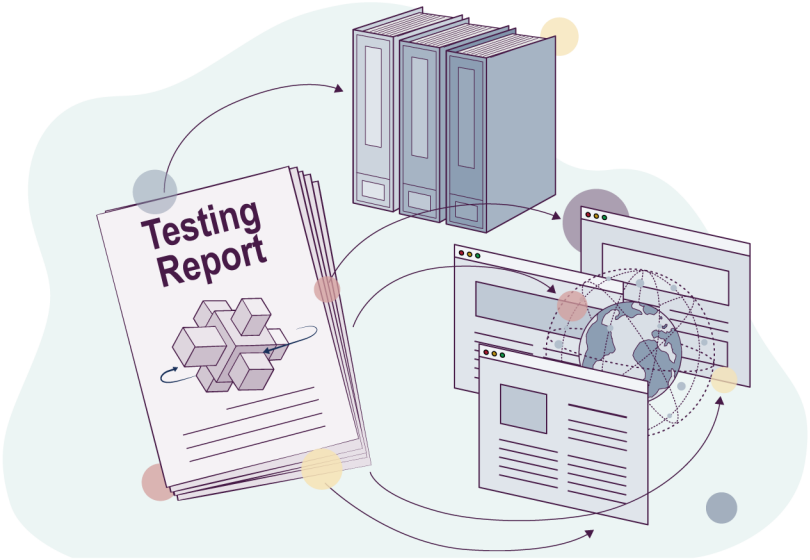
Statistical analysis methods should align with the Design of Experiments metrics and testing criteria defined for the given T&E effort or as defined by other standards. For example, some types of COTS systems used in forensic analysis or for disease diagnostics might require test teams to leverage statistical methods standards, as defined by the International Organization for Standardization or the Clinical and Laboratory Standards Institute, to the maximum extent practical.

Data analysis calculations and results should be clearly described, appropriately recorded, and reviewed in a timely fashion by the quality assurance group for accuracy. Any discrepancies should be corrected by the test execution team and resubmitted for review.

This makes testing objective, reliable, and useful.

Statistical analysis is foundational for solid, trustworthy T&E analysis. It also allows for reliable, disciplined, repeatable, and consistent analysis that enables high-level officials to make important acquisition decisions.

Step 5: Report Test Results



Source: GAO (illustration). | GAO-26-108068

Best Practices

- Develop and maintain T&E reporting guidance
- Use Quick Look Reports (QLR) to quickly communicate preliminary testing data
- Prepare and distribute testing report (TR)
- Stakeholders should review and approve the TR prior to release
- Wait for review and authentication of test results before committing to the next procurement cycle
- Archive reports
- Share lessons learned

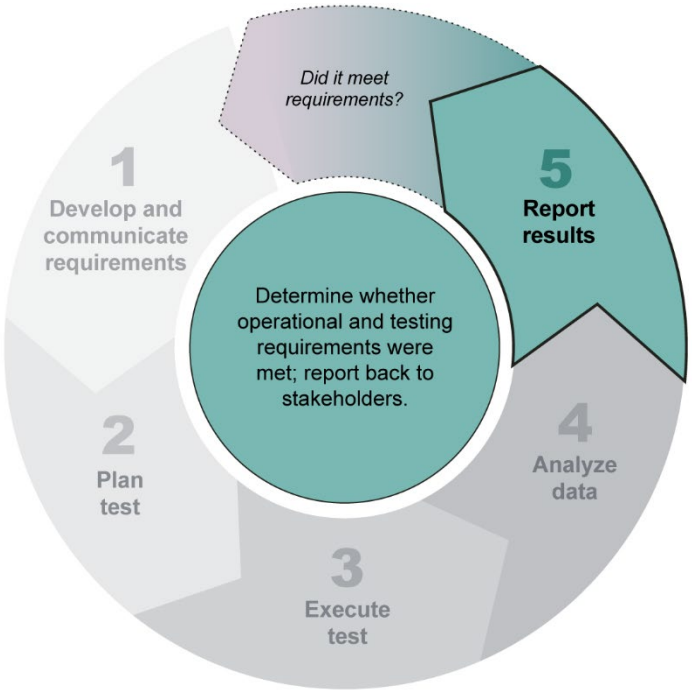
Remember to consider:

Source: GAO (icons). | GAO-26-108068

The purpose of T&E reporting is to provide timely information to all technical and program stakeholders (e.g., T&E managers, program managers, quality control officials, high-level program office officials, acquisition officials) so they can make risk-informed decisions on COTS systems acquisitions. Reporting should also describe lessons learned to constantly improve future T&E processes. Best practices in this step help ensure T&E is objective, reliable, and useful.

Reporting should thoroughly assess and document the T&E process, the results, and conclude whether the COTS system met the established operational and testing requirements. This information should be reported back to the stakeholders involved in establishing the original requirements, who will then decide whether they should procure the system, perform additional testing, or adjust requirements (see fig. 10).

Figure 10: Steps of the commercially available off-the-shelf systems testing and evaluation cycle
Step 5—Report results



Source: GAO (image). | GAO-26-108068

Best Practices for Reporting Test Results

Testing teams use reporting to communicate the results of testing efforts and provide impartial conclusions based on those results. The report permanently records all aspects of test execution, data collection, analysis, and authentication (e.g., quality assurance review), and conclusions such as whether the COTS system met operational and testing requirements. The report also documents any testing deficiencies or anomalies to determine their impact on the results and so that they might be addressed in future testing. Agencies should use a two-stage

reporting process that includes quick look reports (QLR), which summarize the testing with minimal analysis, and testing reports (TR), which contain more comprehensive analysis and describe, in detail, the COTS system performance results against each testing and operational metric.

Develop and maintain T&E reporting guidance. Agencies should have clear, written policies or guidance on how T&E reports should be prepared, including the processes and steps to create the report, such as the following:

- Reporting elements (e.g., executive summary or abstract, introduction, methodology, results, conclusions or recommendations)
- Processes for report submittal, review, and approval
- How the results are communicated within and external to the agency
- How the data and report should be stored and for how long
- What role different stakeholders have in the process

This makes testing reliable and useful.

Reporting guidance, if followed, helps ensure that testers leave a record that future testers can reference or reproduce. Following the guidance will also help testers ensure any data they report are useful for decision-makers, including enough detail but also excluding unnecessary information so that timely action can be taken toward the COTS system.

Use QLRs to quickly communicate preliminary testing data. A QLR allows for quickly communicating preliminary test results to high-level officials to keep them informed of technical performance or problems that might affect program schedules or cost. Test teams should generate QLRs within a short period of time after completion of the COTS T&E. This report should provide a very basic summary of the testing and any problems observed, including deviations from the original test plan. Typically, a QLR will not have much in-depth data analysis, but it should allow for a determination of whether the COTS system preliminarily met or did not meet operational and testing requirements. The QLR should contain appropriate disclaimers so that stakeholders do not make key decisions based on preliminary results.

Example: DOD reporting

In addition to quick-look reports, the DOD sometimes also generates Capabilities and Limitations (C&L) reports. They provide stakeholders with basic observations from testing such as results of watching people use the COTS system. C&L reports allow for capabilities to be quickly assessed in shorter timelines and help inform operational acceptance decision-makers' risk-based decisions. For example, the DOD sent a team to a space system's vendor to learn how to use the system and set up, break down, and palletize it to make it more mobile for transport. This testing did not have extensive rigor, but offered a quick glance at the system's capabilities, limitations, and recommendations that were reported in a C&L report in a 28-day cycle.

Source: Agency officials. | GAO-26-108068

QLRs vary, but they provide insight into key findings or test deviations that do not necessarily need additional analysis to determine whether the COTS system met a baseline requirement or did not. QLRs might also help support modeling or simulation through a comparison of the preliminary testing data with simulation data. Ideally, testing datasets can be analyzed using the same tools used for simulations to provide an understanding of how things are working. This might also allow testing personnel to identify and understand testing anomalies or generate a plan to identify and mitigate the problem.

The use of QLRs might also depend on who needs the COTS system and how quickly they need it. They provide enough information to generate confidence that the data can be shared and might provide details on what the next steps are in terms of expectation of final results.

This makes testing objective, reliable, and useful.

QLRs provide key testing information to stakeholders in a timely manner, which allows testing anomalies to be identified and mitigated quickly.

Prepare and distribute a TR. Testing teams should generate a more comprehensive COTS system TR that describes, in detail, all of the testing methods, data collected, analysis performed—including any statistical analysis—and the results of how the system performed against each operational and testing requirement. The TR should also report challenges and limitations in the test results. The TR should be completed after all data authentication reviews and data analysis are completed. If modeling or simulation were used, those results should also be presented, including model validation, a current performance baseline, the predicted impact, and other associated findings of value to program stakeholders.

Communication of testing results

Testing results might be communicated in presentation format when used for internal purposes only or published internally as a formal report. Results might also be published externally (e.g., in journals) as a formal report, particularly when external organizations such as NIST or academia are leveraged to conduct T&E. Various accreditation, procurement, and agency-wide policies also inform the reporting process. For example, validation records for instrumentation used (e.g., forensic casework, health care diagnostics) might be part of accreditation requirements.

Source; Agency officials. | GAO-26-108068

The TR should clearly describe whether the COTS system met all operational and testing requirements, partially met the requirements, or did not meet the requirements. In justifying those determinations, the report should present any uncertainties, confidence levels, and other calculated statistical measures. The report should explain not only whether a system met the requirements, but to what extent it met them (e.g., a system met a requirement threshold of 99 percent operational availability with 90 percent confidence).

For COTS systems that partially met or did not meet testing requirements, any testing deficiencies or reasons for unsatisfactory performance should

be identified and mitigated if possible. This might require working with the vendor to modify the system so it can be retested. If the vendor cannot address the deficiencies, the agency might need to evaluate a new system.

Reports should provide a sufficient description of the practices used to conduct high-quality T&E and unbiased results of that testing. The information in the reports should be presented clearly and logically so all stakeholders that review the reports draw the same conclusions about the COTS systems regardless of their different perspectives and goals.

This makes testing credible, objective, reliable, and useful.

High-quality TRs allow high-level program and procurement officials to compare test results to defined requirements and make sound technical and financial decisions regarding the procurement of a COTS system. TRs include all important testing evidence, document lessons learned, and identify actions to take after a system passes or fails a test.

Review and approve the TR prior to release. Program stakeholders should review and approve the TR to ensure completeness and accuracy of the results and assess whether the COTS system met all operational and testing requirements. This review provides stakeholders with a chance to identify anything in the testing or results that might be questionable or that might need further explanation. This review also provides the program stakeholders with information to address any questions that an acquisition official might ask. Program stakeholders should review whether an approved test plan and SOPs were used, the testing was conducted in accordance with that plan and SOPs, and testing deviations or other items that required changes were appropriately documented. Stakeholders should indicate their approval of the TR via signature and date stamp.

This makes testing reliable and useful.

Once the TR has been reviewed and approved by all stakeholders, high-level program and acquisition officials should be able to make an informed decision on the fate of the COTS system that was tested. Including key information and identifying actions to take for systems that fail testing will help make testing more reliable and useful.

Wait for review and authentication of test results before committing to the next procurement cycle. Waiting to ensure testing data are accurate before procuring more systems can help stakeholders avoid costly investment mistakes. Stakeholders should ensure the COTS system meets testing requirements before procurement. For example, one agency noted it might need to launch 100 satellites every 2 years to

supplement or replace existing damaged satellites. If the agency decided to buy 100 additional satellites before the reliability, availability, and maintainability of the first 100 satellites had been tested, the agency would have committed to an expensive project that might not have worked.

This makes testing useful.

High-level decision-makers can review test results before committing to procurement, saving time and money if a system fails its initial testing.

Archive reports. Agencies should maintain and store all T&E reports within a database accessible to all stakeholders. Reports in the database should be readily retrievable to allow for stakeholder and auditor reviews. Report retention should follow agency record retention policies and governance to ensure all reports, rights, and reproducibility are maintained.

This makes testing reliable and useful.

Archiving reports helps ensure that future testing can replicate the testing process using the report as reference. Including information such as the testing rationale and lessons learned will help make testing more reliable. Archived reports will be useful to decision-makers in future testing efforts.

Share lessons learned. Agencies should share lessons learned from COTS systems T&E with other agencies. Agencies could share lessons learned by providing TRs to other agencies, through informal briefings, or through structured meetings or conferences.

This makes testing reliable and useful.

Sharing lessons learned might help prevent duplication of effort and repeated mistakes, which will save time and money. This may be particularly beneficial for agencies that procure similar COTS systems.

Image Sources

Unless specified, all graphics in this report were created by GAO.

Front cover:

Source: Denis (United States Capitol building)/ Nuttapong punna (technicians testing circuit board)/ Framestock (drone testing)/ Gorodenkoff (woman at computer)/ stock.adobe.com.

Appendix I: Agency Policies and Guidance for Testing and Evaluation

Below are selected agency policies and guidance outlining testing and evaluation (T&E) standard operating procedures (SOP).

Department of Defense, Design of Experiments Guidance 3.0, Washington, D.C.

Department of Defense, Quantitative Mission-focused Measures – Guidance, Washington, D.C.

Department of Defense, Test and Evaluation Enterprise Guidebook, Washington, D.C.: August 2022

Department of Defense, DOD Manual 500.100: Test and Evaluation Master Plans and Test and Evaluation Strategies, Washington, D.C.: December 2024

Department of Defense, Software Development Test and Evaluation in DevSecOps Guidebook, Washington, D.C.: January 2025

Department of Defense, DOD Manual 5000.99: Realistic Full Spectrum Survivability and Lethality Testing, Washington, D.C.: December 2024

Department of Defense, DOD Manual 5000.101: Operational Test and Evaluation and Live Fire Test and Evaluation of Artificial Intelligence-Enabled and Autonomous Systems, Washington, D.C.: December 2024

Department of Defense, DOD Instruction 5000.98: Operational Test and Evaluation and Live Fire Test and Evaluation, Washington, D.C.: December 2024

Department of Defense, DOD Manual 5000.96: Operational and Live Fire Test and Evaluation Software, Washington, D.C.: December 2024

Department of Defense, DOD Manual 5000.102: Modeling and Simulation Verification, Validation, and Accreditation for Operational Test and Evaluation and Live Fire Test and Evaluation, Washington, D.C.: December 2024

Department of Defense, Modeling and Simulation for Test and Evaluation Guidebook, Washington, D.C.: April 2025

Department of Defense, Development Test and Evaluation of Artificial Intelligence-Enabled Systems Guidebook, Washington, D.C.: February 2025

Department of Defense, Cybersecurity Test and Evaluation Guidebook: Version 2.0, Change 1, Washington, D.C.: February 2020

Department of Defense, DOD Instruction 5000.89: Test and Evaluation, Washington, D.C.: November 2020

National Aeronautics and Space Administration, Systems Engineering Handbook, Washington, D.C.: December 2007

Transportation Security Administration, Test and Evaluation Guidebook Revision 4, Washington, D.C.: December 2019

U.S. Department of Homeland Security, Supplemental Guidance: Test and Evaluation Management, Version 4.0, Washington, D.C.: March 2025

Appendix II: Auditor's Key Questions

Following best practices helps ensure that commercially available off-the-shelf (COTS) systems testing and evaluation (T&E) processes reflect all of the high-quality characteristics of credibility, objectivity, reliability, and usefulness. We have identified five steps that, when followed, should result in T&E processes that governance bodies, program or technical managers, or COTS technology vendors can use for making informed decisions.

The following questions correlate to the best practice checklists from the guide's risk management approach and the five steps of conducting a high-quality T&E process. These questions may be used by auditors or other independent entities, internal or external to an organization, as a roadmap to assess the quality of agencies' T&E.

Risk – Assess Risks to Operations, Safety, Cost, and Schedule

1. Does the agency have a process for identifying and assessing the risks to operations, safety, cost, and schedule for each of the T&E steps?
2. How has the agency mitigated any identified risks?

Step 1 - Develop and Communicate Operational and Testing Requirements

1.1 Does the agency have a process for developing and communicating operational and testing requirements that specify operational needs and how to test whether a COTS system meets them?

1. Has the agency developed operational requirements describing what the COTS system needs to be able to do?
2. Has the agency developed testing requirements that describe the metrics a COTS system must meet or exceed to demonstrate it satisfies operational requirements?
3. Has the agency involved testing personnel in the requirements development?
4. Has the agency communicated the operational and testing requirements to COTS systems vendors?
5. How has the agency communicated these operational and testing requirements to COTS systems vendors?
    a. Are detailed operational and testing requirements in solicitations such as requests for proposals (RFP) or other contractual language provided? Do RFPs detail the need for a third party to complete testing if necessary?

b. Are internal testing data (e.g., qualification data packages) or third-party testing data submitted by COTS systems vendors used to identify whether a COTS system might meet operational and testing requirements?

Step 2 - Plan Testing and Evaluation

2.1 Does a COTS systems test plan exist?

1. Does the plan include the specific kinds of tests that will be used to evaluate a COTS system?
2. Does the plan identify the appropriate resources and challenges (staff, time, funding, equipment, data rights, security, and risk)?
   a. Does the plan describe how resource challenges should be addressed?
   b. Does the plan include challenge mitigations such as removing COTS systems components from testing, replacing them with similar components, isolating or testing individual components instead of an entire system, or reducing the extent of testing?
3. Does the plan address the availability of appropriately trained program managers, engineers, and technical personnel on the specific testing requirements and procedures? Does the plan address mitigations if these personnel are not available?
4. Does the plan identify and include the appropriate stakeholders (e.g., end users, vendors, operators, operational staff, and program offices) throughout the process?
5. Is the plan reviewed by program management prior to testing initiation?
   a. Is risk appropriately mitigated prior to testing initiation?
6. Does the plan describe a process to deviate from or tailor standard testing procedures from what is provided by agency guidance or from high-level officials if necessary?
   a. Does the plan include the need to document testing modifications?
7. Does the plan describe the use of modeling and simulation to evaluate the scope of testing when appropriate?
   a. Does the plan describe how modeling and simulation is intended to be used? For example, to evaluate the need for

facilities, predict COTS systems performance, assess risk and safety issues, or optimize testing conditions prior to conducting testing in specialized facilities?

8. Are data submitted by the COTS systems vendors verified? If so, how is the verification done?

2.2 Do standard operating procedures (SOP) for tests exist?

2.3 Did the agency use Design of Experiments to develop the test plan and SOPs?

2.4 Did the agency conduct a test readiness review?

Step 3 - Execute Testing and Evaluation

3.1 Was the ability to conduct T&E in house weighed against the ability to conduct T&E externally or by a third party? What criteria were used?

3.2 Are SOPs that describe the purpose and scope of T&E, responsibilities, detailed instructions (work instructions and checklists), list of materials and equipment, and potential hazards followed?

1. Are SOPs regularly reviewed and updated, with revisions tracked?
2. Do SOPs include a requirement to calibrate testing equipment to ensure it accurately captures testing data?
3. Do SOPs describe the use of proper controls (negative and positive) in testing when appropriate?

3.3 Is testing conducted in an environment that is as close to an operational environment as possible?

3.4 If necessary, how are COTS system vendor or third-party T&E data used?

1. Are COTS systems vendors that test their own components audited to ensure the integrity of their data?

3.5 Was testing equipment calibrated?

3.6 Was a testing dry run performed?

3.7 Are T&E irregularities such as unexpected data, COTS failures, and safety issues monitored?

1. Is T&E repeated after irregularities have been addressed?

	3.8 Are testing data, results, and observations recorded? How? 3.9 Are raw test data and results retained? How?
Step 4 - Analyze Data	4.1 Does data analysis lead to a conclusion about whether a COTS system passed testing requirements? 1. Is statistical analysis used to improve confidence in testing results? 2. Is statistical analysis used to compare testing performance to requirements? 4.2 Does data validation confirm that data used are sound and demonstrate that the COTS system meets operational requirements?
Step 5 - Report Test Results	5.1 Is reporting guidance developed and maintained? 5.2 Are preliminary test results quickly and accurately communicated to high-level officials using one-pagers, quick-looks, or other means? 5.3 Is a detailed testing report provided after more in-depth analysis? 5.4 Are challenges and limitations in test results recognized and communicated to decision-makers? 5.5 Is the testing report reviewed to determine whether all operational and testing requirements have been met? Who reviews the report? 5.6 Are test results reviewed prior to committing to the next procurement cycle to avoid cost overruns? 5.7 Were all reports appropriately archived? 5.8 Were lessons learned shared? How?

Appendix III: Objectives, Scope and Methodology

Objectives

This guide reflects generally accepted best practices government agencies use for conducting high-quality testing and evaluation (T&E) of commercially available off-the-shelf (COTS) systems, including:

1. the best practices agencies should use for designing, executing, and evaluating tests to provide decision-makers with accurate and timely information on COTS systems,
2. examples of ways agencies have successfully implemented best practices, and
3. a framework for auditors to assess the quality of agency T&E against these best practices.

Scope and Methodology

This guide focuses on federal testing that assesses the extent to which a complete, integrated, COTS product or system meets performance requirements in its operational or field environment (i.e., an environment similar to where the equipment is expected to perform). This guide focuses on COTS products or systems that are commercially available, technically sophisticated, presumed to be at technology readiness level 9,[1] and subject to operational T&E phases and protocols. Such tests typically focus on system effectiveness, reliability, and useability. This guide does not address nontechnically sophisticated COTS items such as staplers or bolts, nor software-only systems, but it does apply to integrated hardware and software systems.

Other T&E phases and requirements, such as development phases or unique requirements for mission critical needs, are not addressed in this guide. This guide does not include private sector entities that develop and internally test COTS systems, nor does it address broader parts of the

[1]Technology readiness levels are a scale of nine levels used to measure a technology's progress where level 1 generally represents paper studies of the basic concept, moving to laboratory demonstrations around level 4, and ending at level 9, where the technology is tested and proven, integrated into a product, and successfully operated in its intended environment. For more information, see Technology Readiness Assessment Guide: Best Practices for Evaluating the Readiness of Technology for Use in Acquisition Programs and Projects, GAO-20-48G (Washington, D.C.: Jan 7, 2020).

acquisition process, acquisition strategy, or developmental testing, as those have been addressed in prior GAO work.[2]

To identify generally accepted best practices government agencies use for conducting high-quality T&E of COTS systems, we reviewed past GAO work to identify agencies with experience executing T&E programs, or setting T&E policy, guidance, or standards. We then reviewed existing agency T&E guidance and documentation to identify potential best practices. Further, these reviews provided a structure for discussions with a panel of agency T&E experts.

We also conducted meetings with T&E experts from eight federal agencies and the private sector, including nonprofits, Federally Funded Research and Development Centers, and other nongovernmental organizations. These experts were identified through agency document review or through discussions with agency officials. These interviews allowed us to obtain perspectives outside of the federal government.

We conducted a 2-day panel of T&E experts from across the federal government. We asked agencies to provide a list of officials with expertise in T&E along with relevant biographical sketches. From this list of potential experts, we selected experts for our panel in coordination with a GAO methodologist. To ensure the panel reflected a broad range of knowledge and viewpoints, we assessed balance across the panel to ensure it was inclusive and as broad as possible without overrepresenting any one agency or area of expertise (e.g., DOD or acquisition professionals). The final panel consisted of experts and experienced practitioners from eight different agencies, representing 17 different offices with diverse T&E expertise in disciplines such as science and technology, defense, homeland security, medical, diagnostics, engineering, software and computer sciences, acquisition policy, and program management.

We solicited written responses from experts in advance of the meeting to help form the meeting's agenda. We also provided them with a preliminary list of best practices culled from agency guidance. During the

[2]For example, see GAO, Homeland Security Acquisitions: Opportunities Exist to Further Improve DHS's Oversight of Test and Evaluation Activities, GAO-20-20 (Washington, D.C.: October 24, 2019), GAO, Combating Nuclear Smuggling: DHS Research and Development on Radiation Detection Technology Could Be Strengthened, GAO-15-263 (Washington, D.C.: March 6, 2015), and GAO, Best Practices: A More Constructive Test Approach is Key to Better Weapon System Outcomes, GAO-00-199 (Washington, D.C.: July 31, 2000).

meeting, we led a discussion to revise and refine the best practices and gather examples of implementing these practices. We covered topics such as general principles for T&E, usefulness of agency policies and guidance, and key challenges agencies face in testing and evaluating COTS systems.

We analyzed the transcripts from this meeting to identify T&E best practices to incorporate into this draft guide. We also reached out to agencies for specific COTS T&E examples describing best practices that addressed deficiencies in T&E or examples describing a best practice and implementation approach that resulted in desired or improved T&E outcomes.

We provided the draft outline and draft sections of the guide to the experts for continued iteration and comment with a goal of reaching consensus on best practices. We adjudicated expert comments by having two analysts perform an independent review and then consolidate their edits. These iterations resulted in the development of this exposure draft, which is being made available for agency and public comment for 3 months.

We conducted our work from January 2025 to September 2026 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to our objectives. The framework requires that we plan and perform the engagement to obtain sufficient and appropriate evidence to meet our stated objectives and to discuss any limitations in our work. We believe that the information and data obtained, and the synthesis of information conducted, provide a reasonable basis for the guidance in this product.

Appendix IV: Agencies Who Helped Develop This Guide

The Department of Homeland Security, National Aeronautics and Space Administration, Department of Commerce, Department of Energy, Department of Veterans Affairs, Department of Defense, Department of Health and Human Services, and Department of Justice provided experts in commercially available off-the shelf (COTS) systems testing and evaluation who helped us develop this guide. These experts attended and participated in our expert meeting, provided text or graphics, and submitted substantial comments.

Appendix V: GAO Contact and Staff Acknowledgments

GAO Contact	Karen L. Howard, PhD at HowardK@gao.gov
Other Leadership on this Project	Will Bauder, PhD, Assistant Director, Science, Technology Assessment, and Analytics (STAA) Michael Dickens, PhD, Analyst-in-Charge and Senior Biological Scientist, STAA Kerry Burgott, MPP, Senior Analyst, STAA
Key Contributors	Victoria Aysola, MPH, Senior Communications Analyst Ryan Han, Lead Visual Communications Analyst Scott Henderson, MEA, General Engineer Hayden Huang, PhD, Assistant Director Rachael Johnson, JD, Senior Staff Attorney Jeff Larson, PhD, Senior Research Design Methodologist Jennifer Leotta, MA, Assistant Director Shenandoah Sedio, Intern