Guide to Developing an Effective STAT Test Strategy V7.0

Sarah Burke, PhD
Emily Divis
Seth Guldin
Michael Harman
Kyle Kolsti, PhD
Alex McBride
Cory Natoli
Steve Oimoen, PhD
Francisco Ortiz, PhD
Aaron Ramert
Bill Rowell, PhD
Gina Sigler
Lenny Truett, PhD
Troy Welker, PhD

31 December 2019

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

2FI: Two Factor Interactions
ACAT: Acquisition Category
AFFTC-TIH: Air Force Flight Test Center Test Information Handbook
ANCOVA: Analysis of Covariance
AOTR: Assessment of Operational Test Readiness
ATO: Authorization to Operate
ATRRS: Army Training Requirements and Resources System
A&A: Assessment & Authorization
C: Control
CDD: Capability Development Document
COE: Center of Excellence
COI: Critical Operational Issues
CSE: Cyber Survivability Endorsement
CT: Component Testing
CTI: Critical Technical Issues
CTP: Critical Technical Parameters
DoD: Department of Defense
DOE: Design of Experiments
DT&E: Developmental Test and Evaluation
DT: Developmental Testing
DTI: Developmental Test Issues
FDS: Fraction of Design Space
FOT&E: Follow-on Operational Test and Evaluation
GEM: Good Enough Missile
H: Hold Constant
IATT: Interim Authority to Test
IPO: Input-Process-Output
IT&E: Integrated test and Evaluation
IT: Information Technology
JCIDS: Joint Capabilities Integration and Development System
JCS: Joint Chiefs of Staff
KPP: Key Performance Parameter
KSA: Key System Attribute
LFT&E: Live Fire Test and Evaluation
LRIP: Low Rate of Initial Production
ME: Main Effect
MOE: Measure of Effectiveness
MOP: Measure of Performance
MOS: Measure of Suitability
MS: Microsoft
MTBF: Mean Time Between Failures
N: Noise
NHPP: Non-homogenous Poisson Process
NIST: National Institute of Standards and Technology
OA: Operational Assessment
OC: Operating Characteristic
OSD: Office of the Secretary of Defense
OT&E: Operational Test and Evaluation
OT: Operational Testing
Q: Quadratic
R: Record
RG: Reliability Growth
RMF: Risk Management Framework
SCA: Security Controls Assessment
SEMP: System Engineering Management Plan
SME: Subject Matter Expert
SNR: Signal to Noise Ratio
SPRT: Sequential Probability Ratio Test
STAT: Scientific Test & Analysis Techniques
SoS: Systems of Systems
T&E: Test and Evaluation
TEMP: Test and Evaluation Master Plan
TPM: Technical Performance Measures
US: United States
USAF: United States Air Force
VIF: Variance Inflation Factor

CHANGE SUMMARY

- V1.0: Original
- V2.0: Updated figures and terminology
- V3.0: Addition of learning resource appendix and errata
- V4.0: Full review of technical descriptions, updated images and sections, glossary of STAT terms, and new examples
- V5.0: Updated figures, expanded Section 3.3 (Setting Test Objectives) and Section 3.5 (Factors and Levels), updated Appendices D and E, updated references and added hyperlinks
- V6.0: Added subsection on Identifying STAT Candidates (Section 3.2), added descriptions of additional statistical techniques, added Appendix C, Reliability Test Planning methods, general editing and formatting updates
- V7.0: Added subsection on capabilities-based testing and evaluation, added description of sequential testing in mission decomposition section, added subsection on measuring responses and noise, updated several figures for readability, updated hyperlinks, general editing and formatting updates

Executive Summary

Scientific Test and Analysis Techniques (STAT) are deliberate, methodical processes and procedures that seek to relate requirements to analysis in order to inform better decision-making. All phases and types of testing, such as developmental, operational, integrated, and live fire testing, strive to deliver meaningful, defensible, and decision-enabling results in an efficient manner. The incorporation of STAT provides a rigorous methodology to accomplish this goal.

Although academic references and industrial examples of STAT are available in large quantities, many do not adequately capture the challenges posed by complex systems. Case studies, such as those that are used in test planning course curricula, are more relevant since they are application-oriented. However, even these rarely contain enough detail to be useful in practical settings. Another potential source of confusion arises from test documentation that not only fails to adequately define system requirements, but also omits critical conditions for the employment of these systems, such as a contested cyber environment. Such difficulties must be overcome, usually at a high cost in both time and effort, to produce efficient and highly effective, rigorous test plans (including designed experiments) that instill a sufficient level of confidence in the inferred performance of the system under test.

This guide addresses the critical planning processes and procedures that are required for the effective integration of STAT into test planning. Since many of the key aspects of such processes and procedures reside in the problem definition and mission decomposition phases of test planning, this guide provides greater emphasis on these areas. For further information concerning mathematical and statistical details relevant to the design of test, we direct the reader's attention to specific references at various locations throughout this guide.

1 Overview

Applying STAT requires implementing the scientific method throughout the planning, design, execution, and analysis of experiments or observational studies to address test objectives. Although STAT has traditionally been applied to physical systems, its use can be extended to software, software-intensive systems such as business systems, simulations, and even test objectives that address cybersecurity requirements. Using STAT identifies and quantifies the risk and cost associated with different choices of testing and thus assists practitioners in developing efficient, rigorous test strategies that will yield defensible results. Freeman, et al. (2014) and Coleman and Montgomery (1993) both outline considerations and processes for planning objective tests. The anticipated product of the sustained use of STAT is a more progressive, sequential testing approach that correctly leverages past test information while adhering to established systems engineering methodology.

Design of Experiments (DOE) is often the gold standard for testing because (as we explain throughout this guide) it is an efficient and effective methodology that yields defensible results. However, DOE is not always the appropriate implementation of STAT. While we focus on DOE in this planning guide, other rigorous methods may be acceptable if DOE is not feasible or applicable. Table 1 lists several additional statistical methods for various applications including reliability, parameter estimation, and observational studies. For more details on recommendations for test planning for reliability, see Appendix C.

Table 1: Additional Rigorous Statistical Methods

Application	Sub-Application	Method	Description
Reliability	Reliability Assessment	Sampling Plans	Sampling is the selection of a subset (a statistical sample) of members from a population to estimate characteristics of the whole population. Two advantages of sampling are that the cost is lower and data collection is faster than measuring the entire population (NIST).
		Sequential Probability Ratio Test	SPRT is a specific sequential hypothesis test that permits concurrent pass/fail analysis during testing and provides stopping/continuing criteria (MIL-STD-781D).
		Parametric Survivability Analysis	Fits the time to event Y (with or without censoring) using linear regression models that can involve both location and scale effects (JMP.com).
	Reliability Growth	Non-Homogenous Poisson Process (NHPP)	NHPP analysis permits the estimation of a variable failure rate that reflects a change in reliability, typically due to configuration changes designed to improve reliability (ReliaWiki).
		Bayesian Analysis/Inference	Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available (Gelman et al., 2004, Meeker and Escobar, 1998).
Parameter Estimation	Estimate Data Distribution Parameters or Interval that Data Occupies	Confidence Interval	A confidence interval is a type of interval estimate of a population parameter (e.g., the mean), computed from the statistics of the observed data. It provides a range of values that might contain the true value of the unknown population parameter (Kensler and Cortes, 2014).
		Prediction Interval	A prediction interval is a type of interval estimate used to estimate the value of one or more future (new) observations (Ortiz and Truett, 2015).
		Tolerance Interval	A tolerance interval is a statistical interval within which, with some confidence level, a specified proportion of a sampled population falls (Ortiz and Truett, 2015).
Observational Data Analysis	Evaluate Observational Data Sets	Survey Design	Survey design employs methods to generate data from users (Lohr, 2009).
		Inferential Statistics	These are classical statistical techniques including linear regression, logistic regression, graphical data analysis, and time series analysis.
		Data Mining Techniques	These methods include decision trees, neural networks, and Bayesian networks and are used to uncover patterns in large, unstructured datasets.

As systems become more complex, the application of STAT becomes simultaneously more necessary and more challenging. This challenge is best addressed by breaking down the requirement, system, and/or

mission into smaller pieces, which can then be readily translated into rigorously quantifiable statistical designs, a point that is expressed in the following excerpt from Montgomery (2017):

> *"One large comprehensive experiment is unlikely to answer the key questions satisfactorily. A single comprehensive experiment requires the experimenters to know the answers to lots of questions, and if they are wrong, the results will be disappointing. This leads to wasting time, materials, and other resources, and may result in never answering the original research question satisfactorily. A sequential approach employing a series of smaller experiments, each with a specific objective … is a better strategy."*

To support such a breakdown, STAT drives an iterative procedure that begins with the requirement and proceeds through the generation of test objectives, designs, and analysis plans, all of which may be directly traced back to the requirement. Critical questions at every stage help the planner keep the process on track. At the end, the test design and plan for analysis are reviewed to ensure that it supports the objective that began the process. Figure 1 provides a concise process flow diagram that summarizes the application of STAT to the test and evaluation (T&E) process.

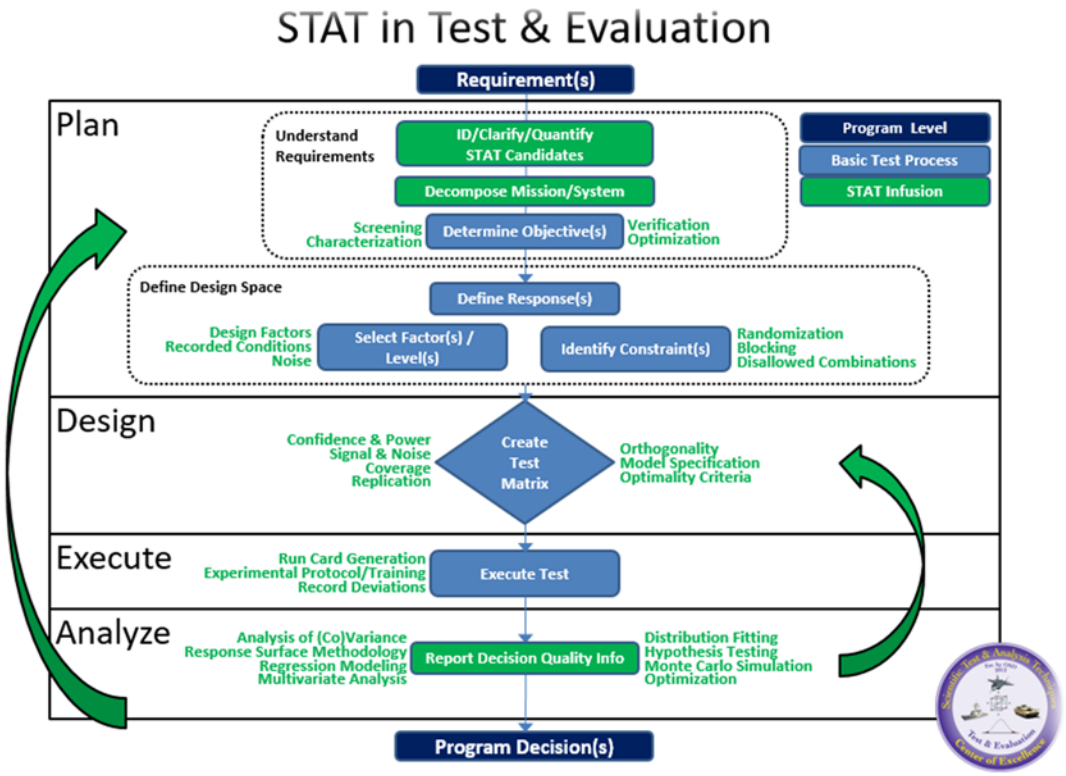

Figure 1: STAT in the Test & Evaluation Process Schematic

2 Test and Evaluation Phases

The DoD exercises three formal and statutory categories of tests administered by the Office of the Secretary of Defense (OSD): developmental test & evaluation (DT&E), operational test & evaluation (OT&E), and live fire test & evaluation (LFT&E). The simultaneous execution and independent

assessment of developmental and operational testing is called Integrated Test & Evaluation (IT&E). (Note: For complete definitions, refer to the Defense Acquisition Guidebook (DoD, 2019) and the Test and Evaluation Management Guide (DoD, 2017)). In the following sections, we provide more details of each of these phases of testing.

2.1 Developmental T&E

DT&E verifies that a system is built correctly and meets the technical requirements specified in the contract. DT&E is conducted throughout the lifecycle of a system to inform the systems engineering process and acquisition decisions, to help manage design and programmatic risks, and to evaluate the combat capability of the system and its ability to provide timely and accurate information to the warfighters.

2.2 Operational T&E

OT&E informs the decision authority by testing production-representative systems executing their intended mission in a realistic operational environment. Every operational test evaluates a system against three major categories: effectiveness, suitability, and survivability. Operational effectiveness is the extent to which a system is capable of accomplishing its intended mission(s) when used by representative personnel in the environment planned or expected for operational employment. A representative operational environment is considered to be a realistic arrangement of various elements occurring in time or static. These elements may be physical, such as operators, maintainers, weather, and terrain, or conceptual, such as organization, training, doctrine, tactics, survivability, vulnerability, and threat (including the growing cybersecurity threat). Operational suitability is the degree to which a system can be placed in field use when limitations on the system's reliability, availability, compatibility, transportability, interoperability, wartime usage rates, maintainability, safety, human factors, manpower supportability, logistics supportability, documentation, environmental effects, and training requirements are considered. The evaluation of operational suitability informs decision makers about the ability of a system to sustain operations over an extended period of time in the anticipated operating environment(s). Survivability testing is required by law for some systems, and may often be conducted through modeling and simulation that is validated by live fire testing.

2.3 Live Fire T&E

LFT&E provides an assessment of the lethality of a system with respect to its intended targets and vulnerability to adversary countermeasures as it progresses through design and development. Anticipated threats to the user of the system as a result of vulnerabilities that are induced by system design shortcomings are a major consideration with regard to LFT&E. A sound LFT&E strategy proactively incorporates design changes resulting from testing and analysis before proceeding beyond the Low Rate of Initial Production (LRIP) phase of acquisition.

2.4 Integrated T&E

IT&E is "the collaborative planning and collaborative execution of test phases and events to provide shared data in support of independent analysis, evaluation, and reporting by all stakeholders, particularly the development (both contractor and government) and operational test and evaluation communities" (OSD Memorandum *Definition of Integrated Testing*). Integrated testing focuses the test strategy on designing, developing, and implementing a comprehensive, yet economical, test design for collaborative use among the various organizations participating in the test. The goal is to improve the quality of the overall system evaluation through synergistic interactions between the organizations.

2.5 Cyber T&E

Before a DoD Information Technology (IT) system can operate, it must undergo a formal Assessment & Authorization (A&A) process to verify that it is secure to operate in the expected cybersecurity environment. DoD Instruction 8510.01 "Risk Management Framework (RMF) for DoD Information Technology (IT)" documents DoD guidance on implementing the National Institute of Standards and Technology's RMF process – the current DoD verification approach. The key part of this process is verifying that the necessary security controls are in place to allow the system both to begin testing (Interim Authorization to Test (IATT)) during the Engineering & Manufacturing Development phase and to fully operate (Authorization to Operate (ATO)) during the Production & Deployment phase. Planned cybersecurity testing during the T&E process is essential to inform the Security Controls Assessment (SCA) part of the RMF process.

The Cybersecurity Test and Evaluation Guidebook (DoD, 25 April 2018) outlines the following 6 Cybersecurity T&E phases:

1. Understand the Cybersecurity Requirements
2. Characterize the Attack Surface
3. Cooperative Vulnerability Identification
4. Adversarial Cybersecurity DT&E
5. Cooperative Vulnerability and Penetration Assessment
6. Adversarial Assessment

These phases occur from pre-Milestone A test planning to cybersecurity T&E after Milestone C. The first four phases primarily support DT&E while the last two phases support OT&E. Section 6.3.1 of this guidebook discusses the role of STAT in test planning.

The overarching guidelines for cybersecurity T&E are outlined below:

- Planning and executing cybersecurity DT&E should occur early in the acquisition lifecycle.
- Test activities should integrate RMF security control assessments with tests of commonly exploited and emerging vulnerabilities early in the acquisition life cycle.
- The Test and Evaluation Master Plan (TEMP) should detail how testing will provide the information needed to assess cybersecurity and inform acquisition decisions.
- The cybersecurity T&E phases should support the development and testing of mission-driven cybersecurity requirements, which may require specialized systems engineering and T&E expertise.

Recently there has been an important change in the process of developing cybersecurity requirements for warfighting systems. During 2015 and 2016, the Joint Staff led development of the Cyber Survivability Endorsement (CSE) process as part of the DoD's Joint Capabilities Integration and Development System (JCIDS) requirements process (JCS, undated). The objective of the CSE process is to ensure cyber survivability requirements are articulated sufficiently to ensure that Joint Warfighting Systems are designed to prevent, mitigate, and recover from cyber-attacks throughout their life cycle by applying a risk managed approach to building and maintaining systems. The CSE process ensures cyber survivability is considered as part of the operational risk trade space throughout a weapon system's lifecycle rather than as a simple threshold-based compliance requirement. It provides a structured, validated approach to developing cyber survivability warfighting capability requirements with the level of granularity appropriate for use in existing DoD acquisition requirements documents.

2.6 T&E Strategy

In defense acquisition, programs are designated by category and type. The acquisition category (ACAT) is a statutory set of cost thresholds maintained by the Defense Acquisition Authority. ACAT classifies acquisition programs by estimated total life cycle cost, thereby prescribing acquisition strategies according to the scale of life cycle costs. The test strategy, in turn, is developed to describe how T&E supports the acquisition strategy. Although test strategies may differ, they must all adhere to sound T&E processes and practices that promote thrift, timeliness, and system performance.

In defense testing, the TEMP is the master document for the planning, management, and execution of the T&E program for a particular system. It describes the overall test program structure, strategy, schedule, resources, objectives, and evaluation frameworks.

As indicated in Figure 2, T&E planning begins with the evaluation of the capability requirements, which are usually identified by those individuals and organizations either with intimate knowledge of, or are directly involved in, operational activities. The requirements are then sorted into technical and operational categories, where they are expressed in the form of measurable and testable technical performance measures (TPMs). The developmental and operational issues, objectives, and measures comprise the twin pillars of the T&E strategy, namely the Developmental and Operational Evaluation Frameworks. The evaluation frameworks describe how a system will be evaluated against its technical and operational requirements to inform programmatic, technical, and operational decisions.

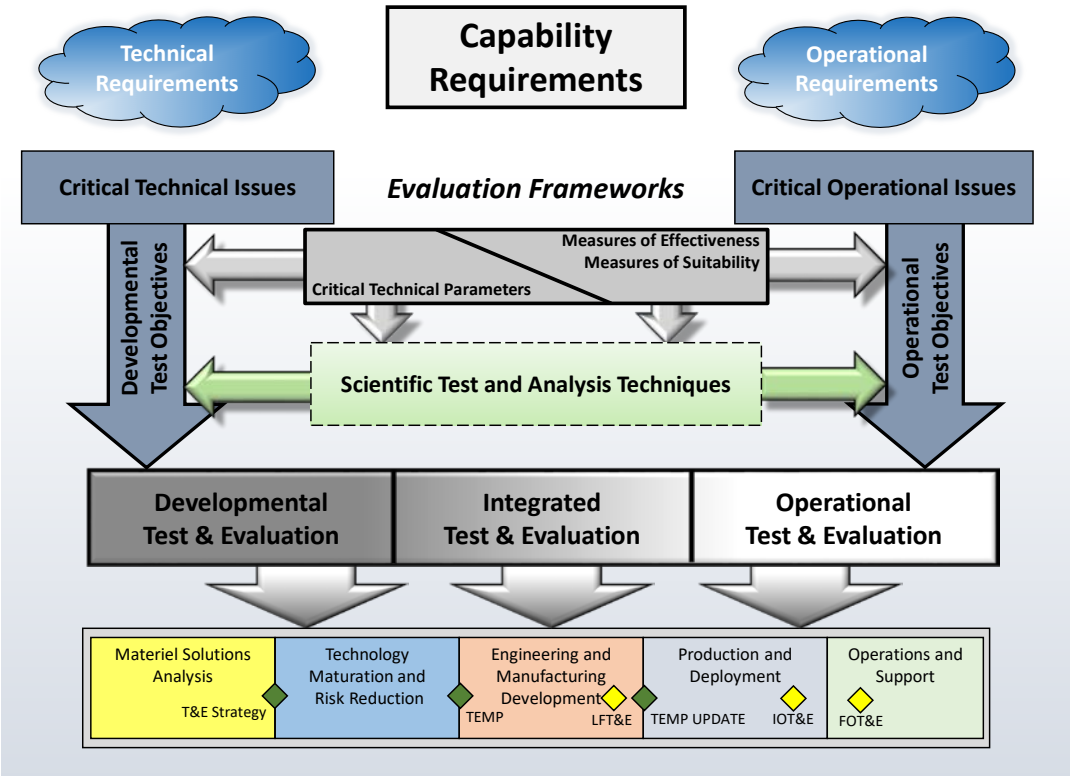

Figure 2: Test and Evaluation Strategy

The requirements are typically associated with two types of test issues: the critical technical issues (CTIs), which are associated with the technical requirements, and the critical operational issues (COIs), which are associated with the operational requirements. The CTIs and COIs guide technical and operational evaluations throughout the acquisition process in refining performance requirements, improving the system design, and enriching milestone decision reviews. They are typically formulated as questions for which multiple measures are often required to adequately determine an answer.

CTIs, also referred to as developmental test issues (DTIs), are the technical evaluation counterparts of the COIs. CTIs must be examined during DT&E to evaluate technical parameters, characteristics, and engineering specifications. They are normally resolved by demonstrating that a system has fulfilled key performance parameters (KPPs), key system attributes (KSAs), and critical technical parameters (CTPs). KPPs, KSAs, and CTPs are a subset of the TPMs that are derived from the requirements document and the System Engineering Management Plan (SEMP). They provide quantitative and qualitative information on how well a system, when performing the mission essential tasks specified in the requirements document, is designed and manufactured. The failure of a system to meet KPP and KSA thresholds may provide grounds for program reevaluation, further technical review, or even program termination.

COIs are operational effectiveness and operational suitability issues that must be examined in OT&E to evaluate the capability of a system to perform its mission. The resolution of the COI is based on the evaluation of measures of effectiveness (MOEs) or measures of suitability (MOSs) using standard criteria. MOEs and MOSs comprise the subset of TPMs that reflect the operational requirements derived from the KPPs and other requirements. MOEs indicate how well a system performs its mission under a given set of conditions while MOSs indicate how ready, supportable, survivable, and safe a system is to sustain effective performance in combat and peacetime operations. COIs address the overall system's operational capability when operated by the warfighter in realistic operational mission environments.

2.6.1 Capabilities Based Test & Evaluation (CBTE)

Both Air Force Test and Evaluation and Navy Test and Evaluation utilize capabilities-based test and evaluation (CBTE) concepts. The capabilities-based concepts that apply to various parts of the Joint Capabilities Integration and Development System (JCIDS) process, such as capabilities-based requirements and capabilities-based acquisition, are applied to test and evaluation as well. These concepts relate to Joint planning guidance that assesses the criticality of an activity to successful task completion, and the criticality of a task to the mission (CJCSI 3126.01A, 31 January 2013).

Air Force guidance on CBTE instructs testers to evaluate the capability of the system to effectively accomplish its intended mission in a realistic mission environment in addition to meeting individual technical specifications. The current emphasis on joint military operations in an information-intensive environment means that Air Force systems will seldom operate in combat as completely independent entities. Air Force systems are expected to fully integrate with systems, activities, and products from all Services and National agencies. Capabilities-based testing requires a full understanding of joint operational concepts in order to develop test scenarios that will provide meaningful results (AFI 99-103, 6 April 2017).

Naval Air Systems Command (NAVAIR) developed its CBTE test strategy to support its capability based acquisition strategy (Charles and Turner (2004)). The motivation for this T&E strategy is that warfighting increasingly involves multiple platforms/systems performing mission tasks in an integrated fashion, thereby driving the need to test across platform/system boundaries and to focus on verifying high-level

warfighting capabilities such as anti-submarine warfare. A key Navy CBTE concept is to collaborate with the Navy's Operational Test Agency (OTA) to implement Mission Based Test Design (MBTD) in DT. Other CBTE enablers include DOE, System of Systems (SOS) testing, Live-Virtual-Constructive (LVC) testing, Mission T&E during system development, Mission Analysis, Mission Training, and Human Performance. For additional details on CBTE see Senechal (2018) and Auborn, et al. (2017).

3 STAT Considerations

In this section, we provide an overview of the key STAT considerations, originally depicted in Figure 1. We review each of these topics in more detail throughout the remainder of this guide.

- Requirements
  - The information, purpose, and intent contained in the requirements drive the entire process
  - All subsequent steps support the selection of test points that will provide sufficient data to definitively address the original requirement
- ID/Clarify/Quantify STAT Candidates
  - Identify what systems or tests may benefit from STAT (not all tests require STAT)
  - Clarify the type of results the STAT candidate tests should produce
  - STAT candidates need to be associated with quantifiable requirements/metrics
- Decompose Mission/System
  - Break the system or mission down into smaller segments
  - Smaller segments make it easier to discern relevant response variables and the associated factors
- Determine Objective(s)
  - Derived from the requirements and reflect the focus and purpose of testing
  - Serve to further define the scope of testing
  - Should be unbiased, specific, measurable, and of practical consequence (Coleman and Montgomery, 1993)
- Define Response(s)
  - The measured output of a test event
  - The dependent outputs of the system that are influenced by the independent or controlled variables, otherwise known as factors
  - Used to quantify performance and address the requirement
  - Should be quantitative whenever possible
- Select Factor(s) and Level(s)
  - Design factors are the input (independent) variables for which there are inferred effects on the response(s), or dependent, variable(s)
  - Levels are the values set for a given factor throughout the experiment. Between each experimental run, the levels of each factor should be reset, even when the next run may have the same level for some factors. Failing to do so can violate the assumption of independence of each result and therefore introduce bias into the analysis.
  - Tests are designed to measure the response over different levels of a factor or factors
  - Statistical methods are then used to determine the statistical significance of any changes in response over different factor levels

    - Uncontrolled factors contribute to noise and are referred to as nuisance factors. The design should account for these nuisance factors.
- Identify Constraint(s)
    - Anything that limits the design or test space
    - May be resources, range procedures, operational limitations, and many others
    - Limitations affect the choice of design, execution planning, execution, and analysis
- Create Test Matrix (or Design)
    - Provides the tester with an exact roadmap of what and how much to test
    - Provides the framework for future statistical tests of the significance of factor effects on the measured response(s)
    - Allows you to quantify risk prior to executing the test
    - All the aforementioned considerations combine to inform the final test design
- Execute Test
    - The planning accounts for, and requires, certain aspects of the execution to be accomplished in a particular manner
    - Since variations from the test plan during execution may potentially impact the analysis, they must be approved (if the change is voluntary) and documented as data is collected during test execution
- Report Decision Quality Information/Analysis
    - Begins when all of the data is collected and you can perform statistical analysis
    - Must reflect how the experiment was actually executed (which may be different from the original test plan)
    - Conclusions as to the efficacy of the system during test are documented, organized, and then reported to decision makers
    - The information collected from the previous test phases may either influence the design of the next phase of testing or inform future program decisions
- Program Decision(s)
    - The analysis can next be used to quantitatively support program decisions
    - The program decides how effectively the user needs are satisfied by the system capabilities, if there are any areas of concern, and if any follow-on testing is required to address any untested conditions or new concerns

TIP: The most important ingredient in a recipe for test planning success is full and active test team participation. This is best accomplished through a formal STAT working group initiated early and maintained throughout the acquisition cycle.

3.1 Understanding the Requirement(s)

Requirements are the starting and ending point for T&E. Planning starts with mapping out a path to report on the requirement. Many choices must be made during T&E planning to address concerns such as operating conditions, resource constraints, and range limitations. These choices help focus the process toward the test design and execution methods so the analysis will provide the right information to make an informed decision about the requirement. If the requirement is not understood clearly at the beginning of the test-planning process, the test team may find that it cannot produce the data needed to report on the requirement. Understanding what is written, what is missing, or what needs to be clarified in the requirement is the first step in effective planning (Harman, 2014). Figure 3 depicts the

translation of requirements to performance measures which define the outcome of test events. One should construct requirements in a thoughtful manner that facilitates a meaningful analysis of the performance of the system under test.

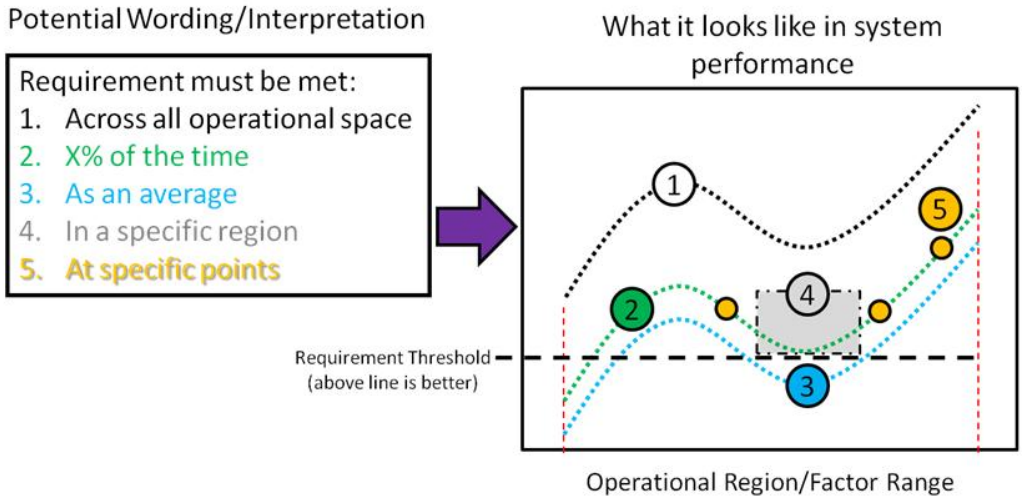

Figure 3: A Depiction of the Flow from Requirements to Performance in the Operational Envelope

Critical questions:

1. To what part of the mission does this requirement pertain?
2. Are specific factors described?
    a. Is there anything specifying performance or conditions?
    b. What are the operating conditions relevant to this requirement?
3. What analysis is implied?
    a. Is there a percentage of the time it must pass?
    b. Is the evaluation to be based on the overall average performance?
4. What remains to be clarified in the requirement?
    a. Can I effectively characterize the system?
    b. If not, where are the risk (un-testable) regions?

Takeaway: Question, understand, clarify, and document details in the requirements first.

3.2 Identify STAT Candidates

Once all requirements are understood by the test team, the team can next assess which requirements will require STAT to verify and which will not. Figure 4 displays a simple flowchart to help determine the method needed to verify different types of requirements. Not all requirements will require rigorous testing methods. For example, suppose a weapon system is required to weigh a certain amount. The weight of an item does not generally change under different conditions, so verifying this requirement has been met only requires an inspection. Some requirements may be verified by industry standards or commonly accepted best practices. For example, an established military standard may outline the

recommended test approach for certain requirements. Be aware that not all standards dictate which test points should be executed and may only give basic recommendations for testing. If the standard does not provide this level of information, then the requirement should be considered a STAT Candidate. Requirements that do not fall into these two categories will require more rigorous test methods to verify. These requirements are typically related to performance, reliability, or sub-system integration.

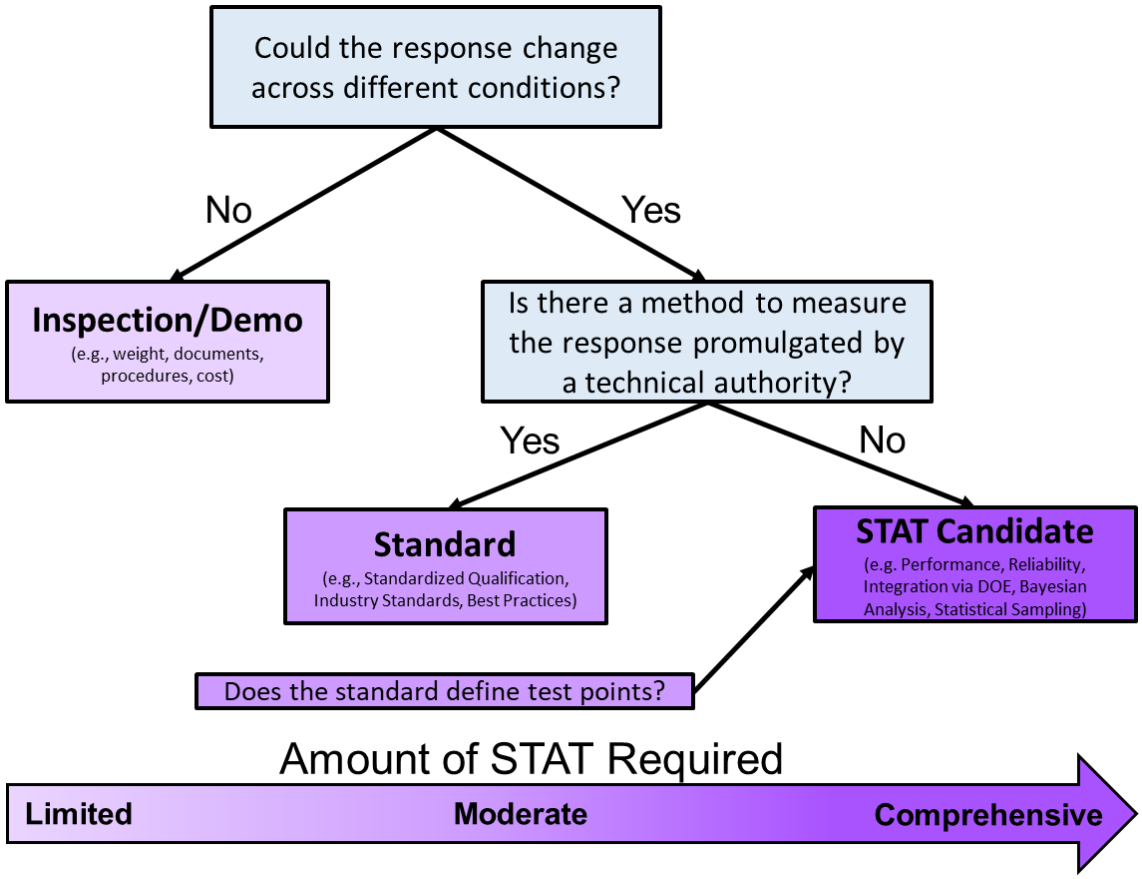

Figure 4: Steps to Determine Which Requirements Should Be Verified Using STAT

Takeaway: Not all requirements need to be verified using STAT.

3.3 Mission & Task Decomposition and Relation

Every test design must enable evaluation of top-level requirements in realistic operational conditions. Because end-to-end, systems-of-systems (SoS) testing in realistic operational conditions is very costly, it is usually impractical to test every possible mission scenario. So, how does one develop a test program that addresses all critical functions over a comprehensive set of operational conditions and provides rigorous and defensible results? In this section, the process of adapting higher-level requirements to testable criteria is discussed at length.

After determining the requirements and assessing those that require STAT, the next step of the process is to understand how the system must function to meet requirements. These functions can then be decomposed into the components that are required to accomplish them. This will enable the straightforward determination of the objectives, responses, factors, and levels needed to evaluate those critical components. Decomposing a mission from top level objectives to segments to functions might look like the example illustrated in Figure 5. At the top level, the "kill targets" objective (1) is too broad and may have numerous factors over the mission space. The lower level segments (2) are more defined and will have fewer contributing factors. At another level below the segments, the functions (3) are further refined with an even smaller factor list. Early DT should focus on this lowest level (level determined when obvious and meaningful responses and factors are revealed) to ensure performance is characterized, quantified, and verified.

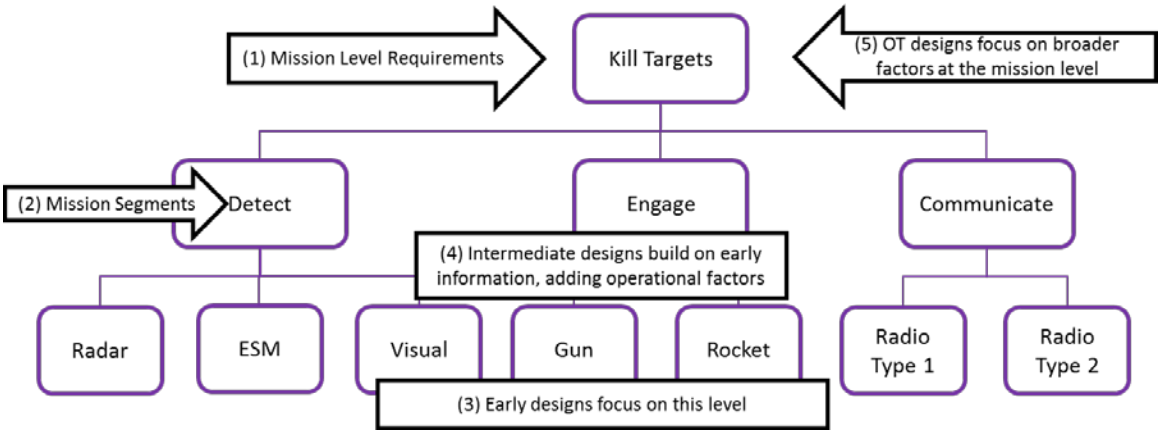

Figure 5: Mission Decomposition Example

If one begins testing at the mission level, it is often very difficult to determine how individual system components influence performance. Even if one could identify a particular component as being the specific cause of observed behavior during test, the exact mechanism of this causal relationship may not be well understood. This issue is mitigated by a sequential, progressive test strategy that starts with low level test designs and culminates in top-level operational testing. For more information on sequential test strategies, see Simpson (2014). Box and Liu (1999) and Box (1999) emphasize the importance of sequential testing and learning as follows:

- The fundamental starting point for any test should be to understand the performance of each critical component. The type and the extent of the component testing required is a function of criticality and previous knowledge.
- The next level of testing is to evaluate individual functions. To the greatest extent possible, functional characteristics should be evaluated over the entire range of operational conditions. Testing over a subset of possible conditions greatly increases the risks of delaying failure discoveries until OT&E.
- Next, the combination of all of the functions should be evaluated in the system. The goal is to discover failures caused by the integration of all of the functions.
- Finally, operational testing should be conducted in order to evaluate the highest level system-of-systems measures and to validate all previous testing.

The goal of the preceding sequential test strategy is to identify and correct failures as early as possible, while also validating each phase of testing. It is an iterative process as each phase of testing informs a new phase, and the new phase is used to validate the prior phase. Sequential testing does not necessarily draw a line between DT&E and OT&E. DT&E should be performed over the entire range of operational conditions. OT&E may then focus on the SoS mission level testing, thus augmenting or verifying the DT&E results. For example, M&S may be the first phase for testing a new system. Then it would be followed by ground testing and finally flight testing. Each phase of testing would help inform the next phase, while the new phase would assist in validating the previous phase. All phases of testing help to manage the overall risk.

TIP - Whenever a fundamental failure of a component or a function is discovered during complex system-level testing, it is more difficult to isolate, more complex to determine the root cause, and more expensive to correct than if discovered in component/function level testing.

Critical questions:

1. Are all the steps necessary to meet the requirements included in the decomposition?
2. Are responses and factors easily defined for critical components?
3. Will DT cover the entire range of operational conditions?
4. Will early testing be used to inform more complex test designs?
5. Will risk areas (for failing to meet requirements) be identified before developing OT test plans?

Takeaway: Decompose missions, systems, and functions until obvious and meaningful responses and factors are revealed; then, develop a test strategy that builds on previous testing as complexity increases

3.4 Setting Test Objectives

Planning cannot proceed without a set of clear test objectives. Objectives are derived from the requirements and serve to focus resources and designs toward addressing the requirement in a clear, quantitative, and unambiguous way. Properly identifying the test objectives is a critical step in the overall process. The test objectives will influence the design choices made; improperly identifying the test objectives can result in a meaningless test. As the objectives form a roadmap that can be referenced throughout the planning process to ensure that subsequent steps remain on course to produce a relevant design, it is important that the objectives be clearly defined. The objectives must be precise enough that there is no confusion as to how the system measurement will take place.

When writing objectives, precise and deliberate wording is critical. The action verb used in the objective is especially important as it implies (or sometimes defines) the nature of the test program and the type of answer it will provide. Confusion quickly arises when different people perceive a different specific meaning or connotation to a given verb. The problem is amplified when experts from different

communities are accustomed to applying the same word in different ways. To provide ideas and spark discussion on possible objectives, we list some useful action verbs as defined by two sources. The overlap and cross-referencing between the two lists demonstrates the risk of confusion when figuring out the objectives of a test. Therefore, regardless of the final choice of words, it is critical that the test team explicitly defines each objective to remove any ambiguity or misinterpretation.

The United States Air Force (USAF) flight test community uses the six action verbs defined in AFFTC-TIH-93-01, Air Force Flight Test Center Test Information Handbook, Feb 1999, as shown in Table 2. In this approach, compelling the test team to pick one of these distinctly defined verbs clarifies the discussion.

Table 2: Example Action Verbs for Objectives as Defined For Use in USAF Flight Test

Verb	Action	Example
Observe	To watch carefully, especially with attention to detail or behavior for the purpose of arriving at a judgement.	Observe the actions of the self-driving car on representative city streets and highways.
Compare	To examine in detail the likenesses and differences in the quality or performance of the test items.	Compare the sensitivity of the upgraded sensor versus the legacy sensor.
Demonstrate	To reveal something qualitative or quantitative which is not otherwise obvious. (It either passes the test or it does not.)	Demonstrate that the utility truck can carry a payload of 2,000 pounds on level terrain.
Determine	To discover certain measurable or observable characteristics of a test item.	Determine the maximum payload that the utility truck can carry on level terrain.
Evaluate	To establish overall worth (effectiveness, adequacy, usefulness, capability) of a test item. (Requires the development of evaluation criteria that lead to a rating of the system.)	Evaluate the radar's maximum detection range against small targets.
Verify	To confirm a suspected, hypothesized, or partly established contention.	Verify the mean time between failures (MTBF) exceeds 10,000 hours.

Montgomery (2017) also provides common reasons for running an experiment or test provided. They are listed in verb form in Table 3 to be consistent with the previous list, and the definitions are paraphrased from the book. The author points out that this list is non-exhaustive. As indicated in Figure 6, the objective(s) (there may be more than one) are ideally measurable in terms of continuous values in order to facilitate statistical analysis.

Table 3: Example Action Verbs for Objectives from Montgomery (2017)

Objective	Action	Example
Characterize	To measure the response across a design space.	Characterize the effects of solder temperature and conveyor speed on circuit board defect rates.
Screen	To learn which factors have the most influence on the response.	Screen solder temperature, solder depth, or conveyor speed to see whether they affect the defect rate.
Optimize	To find the factor levels that result in a desired response.	Optimize solder temperature and conveyor speed to achieve the lowest defect rate.
Confirm	To verify the system behavior is consistent with theory or experience.	Confirm that the defect rate during full-rate production is the same as that during LRIP.

Discover	To determine what happens when factors are added/removed or the factor ranges are increased.	Discover the effect a new soldering material has (if any) on the defect rate.
Robustness	To find the factor levels that both provide desired response, AND reduce the variance of the response.	Find the solder temperature and conveyor speed that result in the lowest defect rate AND lowest variability from batch to batch.

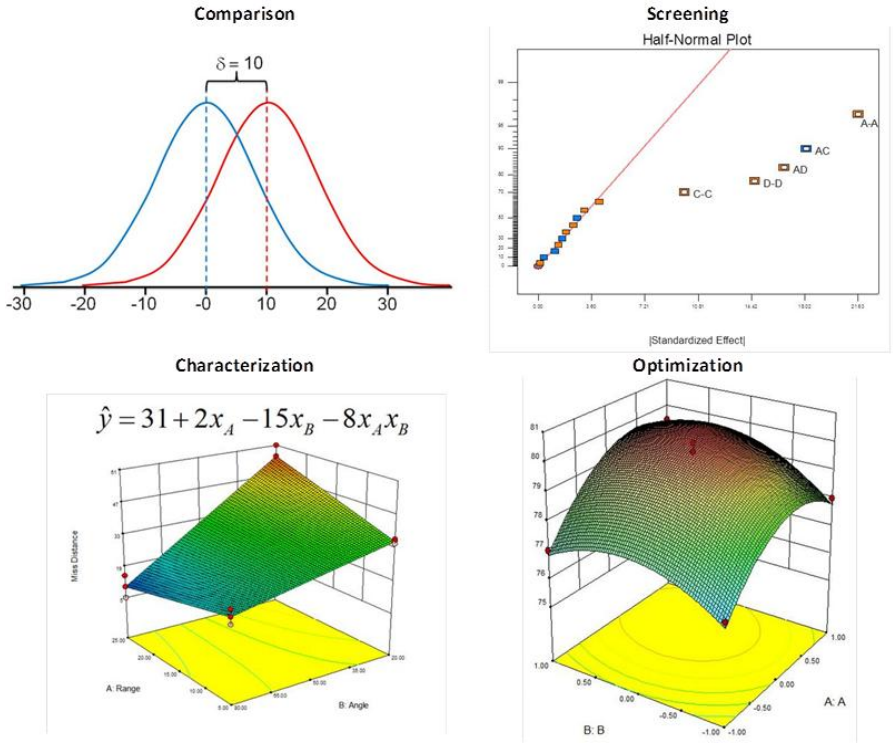

Figure 6: Assessment of Quantitative Objectives over the Operational Envelope

Finally, objectives must be achievable within cost, schedule, and safety constraints and should address the following questions at a minimum:

1. Is the objective stated clearly and unambiguously?
2. Which requirement does the objective address? (Traceability)
3. Will the objective lead to being able to answer a clear question at the conclusion of the testing?
4. Is the objective feasible and measurable?

Complex test programs may require a further breakdown of objectives which resolves big-picture program objectives into specific data requirements using some combination of MOEs, MOSs, and Measures of Performance (MOPs). These design concepts ensure traceability so that every requirement is addressed by data that are acquired, and conversely that data are only obtained if they pertain to a requirement. Details can be found in the Test and Evaluation Management Guide available online from

the Defense Acquisition University (link provided in the references section), or you can check the equivalent guidance for your organization.

Takeaway: Objectives should be unbiased, specific, measurable, and of practical consequence (Coleman and Montgomery, 1993).

3.5 Define Responses

As discussed in the previous section, the key to a good test design is to fully understand the goals of the test. A critical characteristic of a good objective is that it is measurable. Responses are the measured outputs of a test event and are used to assess the objective of the test. There may be several responses measured for a given test and/or in support of a requirement. One tool to help in determining these measures is a process flow diagram, a figure that details each step of functionality. To illustrate, consider the process flow map of an armed escort mission in Figure 7.

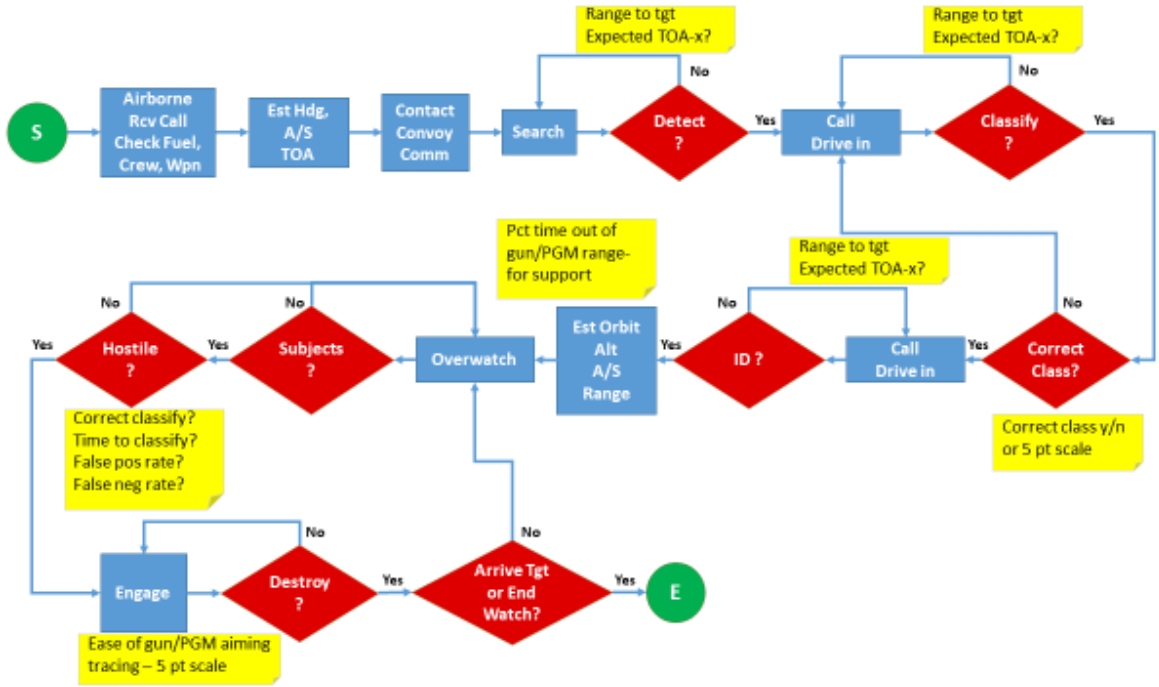

Figure 7: Process Flow Diagram, Armed Escort Mission (Adapted from Hutto, 2011)

Steps and decisions are coded in blue and red while possible responses are in yellow. Within such a flow diagram, stated system performance measures should be traceable back to KPPs, MOPs, MOEs, and MOSs. Moreover, measures that already exist in capability documents, particularly those metrics that were used in similar test programs or were developed by subject matter experts (SMEs), should be documented in the flow diagram. The process flow diagram also helps to understand how the system works. This is a chance for SMEs on the system to share a step by step walkthrough. The process flow

diagram can show any limitations that may exist when executing the experiment. These limitations must be discussed prior to designing a test as they can have an impact on the possible designs.

Critical questions:

1. Does the response relate to top level requirements (e.g., through KPP, MOE)?
    a. Is the response expressed as a continuous variable? (more detail on this in Section 3.6.1)
    b. If not, can it be converted to a continuous variable?
2. Is the response directly measurable?
3. How is the response measured?
4. Is the response defined in a clear and unambiguous manner?

Takeaway: The response must be clear, concise, measureable, preferably continuous, and directly related to the requirement.

3.6 Measuring Responses

The goal in response measurement is to determine if there is a cause-and-effect relationship between a factor and the response. By measuring a smaller change in response, a more precise relationship can be defined. However, many of the systems we test are stochastic, and the response has an inherent level of variation because of noise (σ) in the system. In a stochastic system, the response will change even when all controllable factors are held constant. The way to overcome this is to measure changes in the response that exceed the expected variation.

The desired level of change to be measured is the difference-to-detect, or signal (δ). The optimum value for δ is large enough to indicate that system performance has been meaningfully changed and small enough to define a precise relationship between the factor and response. Setting δ is not a trivial task and will nearly always involve discussions with program leadership, where decision makers can decide what a *meaningful* change is, and test engineers, who understand the system and the test equipment and how precise the response can be measured (Ramert, 2019).

It is common to compare the difference-to-detect to the noise with the signal-to-noise ratio (SNR). A higher SNR will require fewer runs to create a test with adequate power. Conversely, a lower SNR will require more test runs to achieve the desired power, but will yield a more precise model. The trade-offs between SNR and power and test size are a large component of test design comparison.

3.7 Factors and Levels

Factors are inputs to, or conditions for, a test event that potentially influence the value of and variability in the response. Factors can be derived from prior testing, system knowledge, or insight into the underlying physics of the problem. Factors may include configurations, physical and ambient conditions, and operator considerations (e.g., training, skills, limitations, shifts). Factors should not be ruled out without proper screening or technical analysis. Levels, which are the distinct values that are set for each factor in a design, may number anywhere from two (a low and high value, for example) to many values. Factors should be made continuous whenever possible to ensure the most information is contained in

the design and to maximize the level of detail contained in the analysis. Categorical (non-continuous) factors with more than two levels also add additional runs to the test matrix and should thus be avoided if possible. We discuss the benefits of using continuous factors (and responses) in section 3.6.1 on data type consideration.

Figure 8 depicts a convenient brainstorming tool, a fishbone diagram, whose purpose is to facilitate extracting causal factors. There are six broad categories often referred to as the "6 Ms": Machines, Manpower, Materials, Measurements, Methods, and Mother Nature. These categories help to further identify potential factors that influence the response variable. The fishbone diagram is often used in a collaborative environment in which all potential reasonable sources are categorized into one of the "6 Ms." When creating this diagram, write down all potential factors in the test. The test team will further classify each factor as:

- Vary (V) – a factor we can systematically vary during test and that could impact the response as we go from a low level to a high level. Appropriate levels must be agreed upon that adequately stress the system, but do not go outside the expected operating envelope (unless the test objective is to do so).
- Hold constant (H) – a factor that likely influences the response that we can set constant throughout all of the tests as it may not be one of primary interest or too difficult to vary. The test team must accomplish due diligence in setting these at the appropriate level. Realize that even if these factors are not "in the test," the levels set may have a very large impact on system performance. Unknown interaction effects between constant and control factors could also be possible. Most importantly, if you never change the levels of these variables, you will never know how they influence the responses.
- Record (R) – a factor that will not be controlled during test, but is recorded. If measured, they can be included in the analysis with advanced statistical methods such as analysis of covariance (ANCOVA).
- Noise (N) – a factor that likely influences the response but that we cannot control, though we may be able to observe and record it during test (often an environmental factor). These are sometimes referred to as nuisance factors. The test must be protected against these as much as possible. They are either going to be measured, randomized against, or blocked.

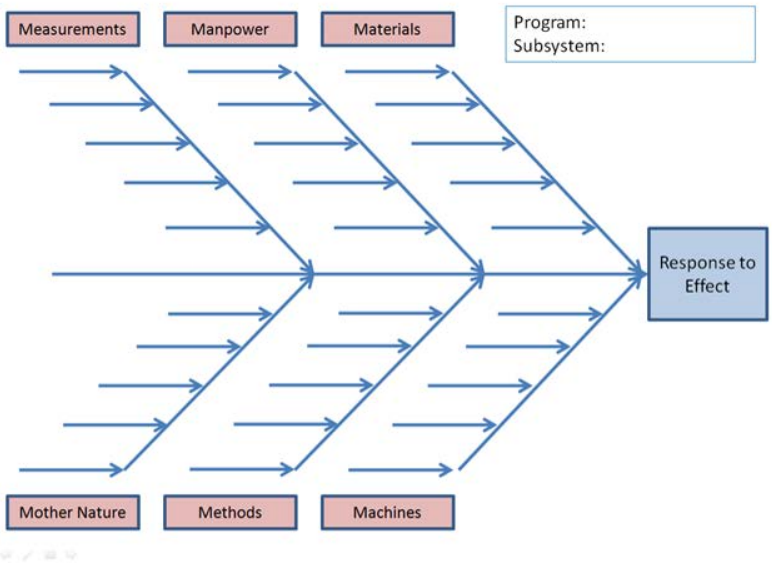

Figure 8: Fishbone Diagram

Other methods for brainstorming potential factors include affinity diagrams and inter-relationship diagrams, which can be used in conjunction with the fishbone diagram. An affinity diagram, which has been used successfully in industry, can be an effective tool to organize many items into natural groupings or when a group needs to come to a consensus. Creating the affinity diagram is a group exercise. Each member of the group writes down ideas (e.g., potential factors that may affect the response) on separate sticky notes or cards. All of the cards are then placed before the entire group. Then, with no communication between group members, each person looks and groups similar items together. Some of these cards may be moved more than once and some may not belong in a group with others. A generic example of the results of this process is shown in Figure 9. One advantage of the affinity diagram process is that it encourages creative thinking from the group and can generate new ideas that may have been previously missed (Tague, 2004, pp. 96-99).

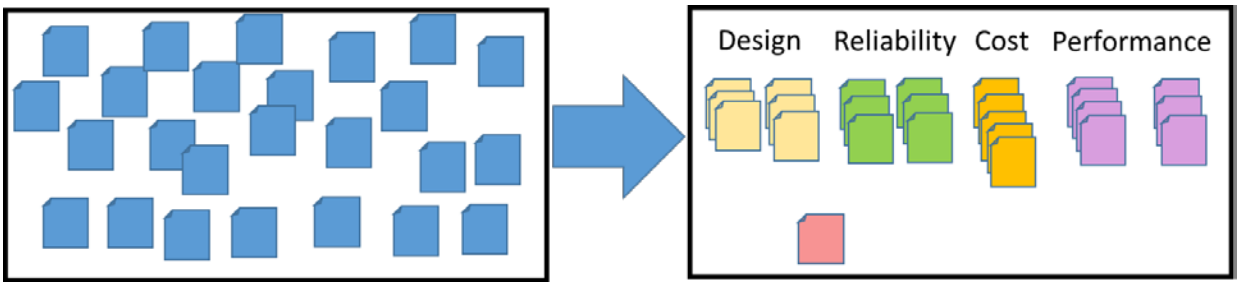

Figure 9: Affinity Diagram (General Example)

An inter-relationship diagram (Figure 10) can be used after generating a cause-and-effect or affinity diagram in order to further understand and explore links between items (or steps) in a complex solution (or process). For every item generated from the fishbone or affinity diagram, ask “does this item cause or influence any other item?” Draw arrows from each item to the items it influences. After completing

this process for all of the items, analyze which items have the most arrows going in and out of them. In general, those that have mostly outgoing arrows are basic causes to investigate; those that have mostly incoming arrows are critical effects to address (Tague, 2004, pp. 444-446).

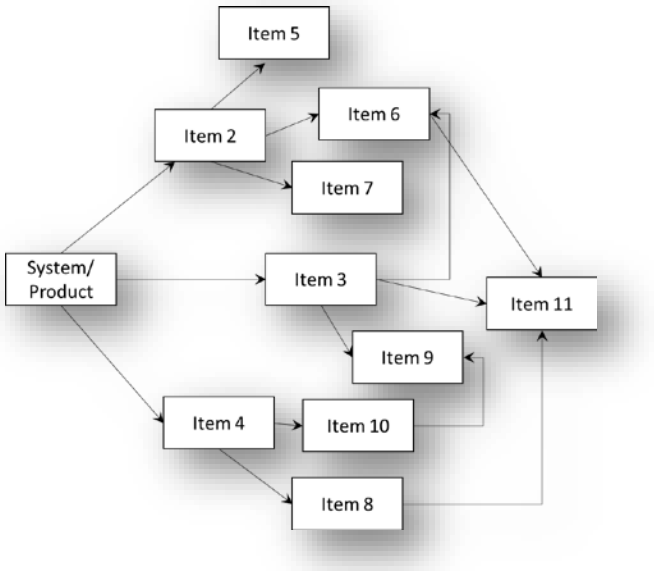

Figure 10: Inter-Relationship Diagram (General Example)

Again, the test team must identify these factors and then agree on how to treat each one for the test program. This is an iterative process where continuous reassessment is required.

The Input-Process-Output (IPO) diagram is an effective way to summarize the results of a planning effort. The inputs shown flow from the control factors identified in the affinity diagram and/or fishbone diagram, the process is the test program objective, and the outputs are the MOPs identified through requirements documents and process mapping. It is also useful to display the noise factors for both the input factors and output responses. A good IPO diagram is the foundation for a good test design to allow the test team to efficiently quantify and develop insights on system performance. A generic example of an IPO is shown in Figure 11.

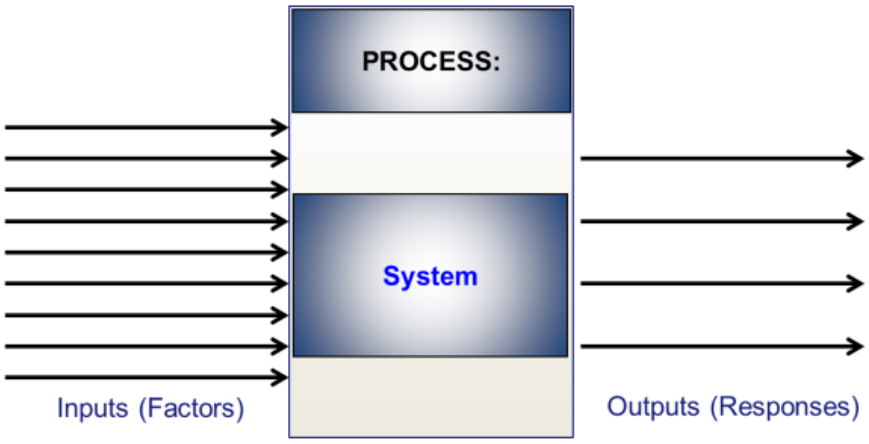

Figure 11: Input-Process-Output Diagram (General Example)

Critical questions:

1. Have all factors been considered?
    a. Do all the listed factors contribute to the response?
    b. How were potential factors determined?
    c. Were factors determined via a collaborative brainstorming session?
    d. Has input from all SMEs been collected?
2. Are the factors clear and unambiguous?
    a. How many levels can (and should) be allocated for each factor?
    b. Will the number of levels support the objective of the test?
    c. Will the number of levels adequately sample the design space?
3. How is each factor expressed?
    a. Are the factors expressed as continuous or categorical variables?
    b. Can categorical variables be converted to continuous variables?
4. Has each factor been classified as vary, hold constant, record, or noise?
    a. Can the factors be varied?
    b. Is there any difficulty in varying or setting the factor levels?
5. What are the factor priorities for testing?
    a. Which are to be varied, held constant, or recorded?
    b. Which will contribute to noise in the response?

Takeaway: Factors must be clear, concise, controllable, preferably continuous, and must relate directly to the response.

3.7.1 Data Type Consideration

The data type chosen to represent factors (inputs) and responses (outputs) in a test may affect the resources needed to conduct the test itself and, consequently, the quality of the statistical analysis. Categorical data types are too often used when describing factor levels and responses. This may be due to vaguely-defined requirements and test objectives. It could also be that planners find it easier to conceptualize and plan experiments using generic settings and measures. Perhaps there is difficultly in measuring the inputs and outputs in great detail. For example, live fire testing could be destructive and thus it is impossible to precisely measure the impact point. As shown in Figure 12, it may prove easier to use a count of hit and misses instead of measured missed distance from the target.

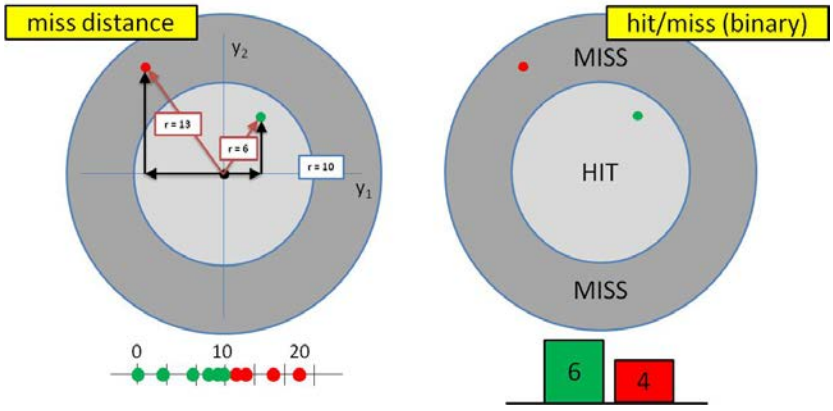

Figure 12: Continuous Versus Categorical Variables

Another example would be that we are testing the system during different times of the day and list the factor settings as “day” and “night” instead of using some measure of illumination. Whatever the reason, any perceived savings in favoring categorical over continuous data types may be paid for in wasted resources and complications during analysis (Ortiz, 2018).

As mentioned previously, categorical factors add additional runs to the test and may also have an effect on the quality of the analysis (see Figure 13). In addition, categorical data types do not allow for prediction between levels. Suppose temperature is a factor of interest. Using low, medium, high as the levels for this factor, you can estimate the response for each of these levels. However, if you associate degrees to these levels and treat temperature as a continuous factor, you are now able to estimate the response at temperatures within the entire range of these levels (e.g., a one unit increase in temperature leads to an x-unit increase in the response).

In the case of responses, categorical data types contain a relatively poor amount of information in comparison to continuous data types (38% to 60% less in some cases (Cohen, 1981)). This state of reduced information results in an increased difficulty for tests to detect changes in the presence of noise. More specifically, these tests will have a poor SNR, which will in turn require a greater number of replications/runs to achieve sufficient power.

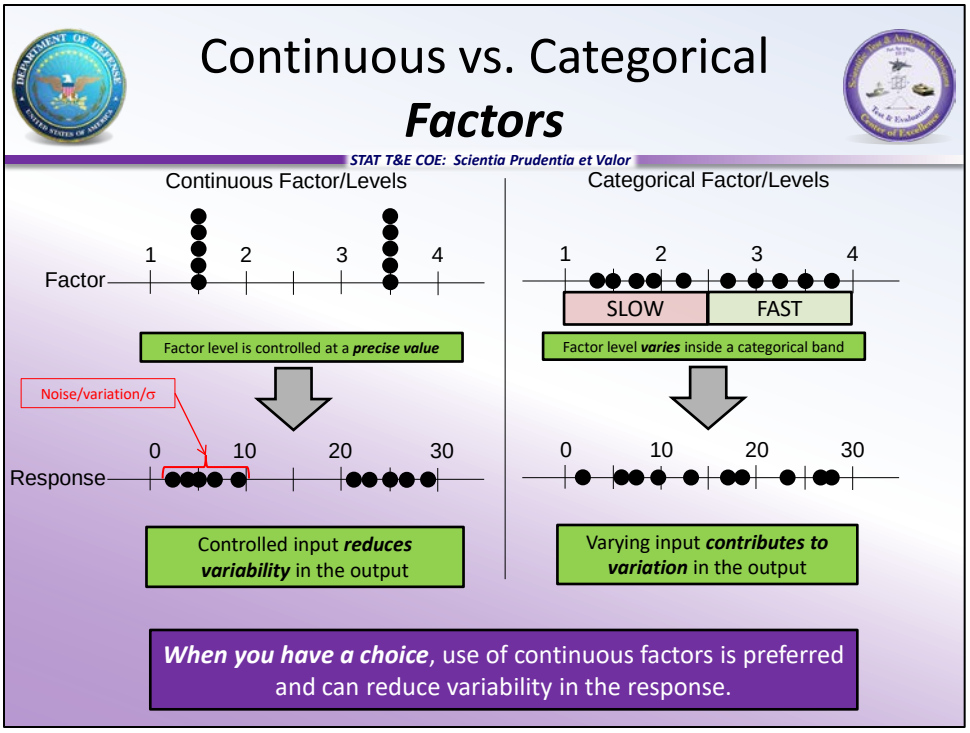

Figure 13: Impact of Factor Type on Response

TIP – Just because the requirement is stated as a probability should not dictate how the data are collected! Continuous responses can be converted to probabilities in the analysis.

Critical questions:

1. Is the factor or response a continuous variable? If not, can it be converted to a continuous variable?
2. What is the estimated signal-to-noise ratio for each response?
3. Do I need to have an understanding of what is happening between categorical levels?
4. How many fewer runs are needed if a response were continuous rather than categorical?

3.8 Constraints

An important step in the test planning process is to identify any restrictions on the test design or execution. Test design and execution restrictions can influence the design choice and analysis. Constraints affect which combinations of levels among different factors may appear in the design simultaneously, thus influencing the geometrical arrangement of the design points. Common constraints include the budget, the experimental region, and restrictions on randomization (which includes difficulty changing factor levels).

3.8.1 Budget Constraints

The test budget, in terms of both time and money, is an important consideration when planning an experiment. Planning a "should execute" design and then factoring in cost constraints to create the actual design enables the team to assess the risk imposed by budget constraints. This risk can be expressed as the loss of statistical power (refer to section 3.8 for more information on power).

3.8.2 Restrictions on the Experimental Design Region

The test planning team must consider any restrictions on the test region. The team must decide on the test region of interest, which may be a subset of the region of operability. This test region need not be rectangular; a characteristic which indicates that disallowed factor combinations may exist. For example, the high levels of two factors (say pressure and temperature) considered separately may be within the region of operability; however, the combination of both factors at high levels may be outside the region of operability. Prohibited combinations may stem from factor interactions on the response, safety considerations, undesired results occurring under certain combinations, and known scientific principles. It is important that all prohibited combinations be identified early in the planning process.

3.8.3 Restrictions on Randomization

Randomization is one of the three principles of designed experiments and refers to both the randomization of the order of test runs and the random allocation of test materials. Failure to randomize could result in factors confounded with nuisance variables, meaning that the analysis will be unable to discriminate effects due to the factors and nuisance variables. Therefore, any restrictions on randomization must be identified in the planning stages, as these may affect the design and analysis.

The ease with which factor levels can be varied must be considered as well. Are there any factors that are difficult or expensive to change many times? A decision to vary levels of such hard-to-change factors less often will represent a restriction on randomization, which may subsequently lead to a split plot design (Anderson-Cook, 2007). Another common restriction on randomization is blocking, which is a variance reduction technique used for dealing with nuisance variables, factors that could influence the response, but which are not factors of interest. Blocking is discussed in more detail in section 3.8.

In section 3.8, we discuss one of the other principles of DOE: replication, independent test runs of each factor level combination. Distinct from replication, repeated observations, also called sub-sampling or pseudo-replicates, are multiple measurements made at once for a given factor level combination. For example, suppose you have two factors with a and b levels, respectively, and n replicates. In a completely randomized design, all abn observations would be taken in a random order. Suppose instead that the n replicates are taken all at once for one of the levels of factor A and one of the levels of factor B. This may be done because both factors are difficult to change (or it is impractical to do so). There are analysis methods to account for this restriction on randomization, often done with a nested model. See Kutner, et al. (2005, pp. 1106) for details.

Critical questions:

1. What is the budget for this test?
2. What is the total time allotted for this test?
3. How much time does each run take to complete?
4. How much does each run cost?
5. Are there restrictions on the operational space or design region?
6. What range restrictions are imposed?
7. How many disallowed combinations of factor levels are there?
8. What factors are hard or costly to change?
    a. Does the system configuration remain set for multiple runs before being changed?
    b. Are there factor levels that can only be changed in a linear manner (e.g., small to large)?
9. Is blocking necessary?
    a. Does the test span multiple days?

b. Will different operators perform different test runs?

Takeaway: Detail, define, and document how constraints will limit or impact the design or execution.

3.9 Test Design

Many considerations impact the choice of design, some of which include the following and which have appeared in previous sections of this guide:

- What is the test objective (e.g., screening, characterization, and optimization)?
- Are the factors quantitative or categorical?
- How many factors are there?
- How many levels does each factor have?
- Are there any constraints?
- Are there any restrictions on randomization?

Table 4 lists some sample designs for various test objectives. Details of many of the designs listed in Table 4 can be found in Box, Hunter, and Hunter (2005), Montgomery (2017), and the National Institute of Standards and Technology (NIST) website.

Table 4: Design Types

Test Objectives	Sample Designs*
Screening for Important Factors	Factorial Designs, Fractional Factorial Designs, Definitive Screening Designs, Optimal Designs
Characterize a System or Process over a Region of Interest	Factorial Designs, Fractional Factorial Designs, Response Surface Designs, Optimal Designs
Process Optimization	Response Surface Designs, Optimal Designs
Test for Problems (Errors, Faults, Software bugs, Cybersecurity vulnerabilities)	Combinatorial Designs, Orthogonal Arrays
Analyze a deterministic response (e.g., from a computer experiment)	Space Filling Designs, Optimal Designs
Reliability Assessment	Sampling Plans, Sequential Probability Ratio Test, Design of Experiments

* The design choices listed in this table are general guidelines. The design should be chosen to match your goals as well as account for any restrictions in the test execution as discussed in section 3.7. For example, your goal may be to screen factors; but if there are restrictions on the design region, a factorial or fractional factorial design is not appropriate.

There are many considerations to weigh when choosing between candidate designs. Table 5 provides a list of criteria that can be used to evaluate and compare designs, thereby minimizing the risk and cost for a given design.

Table 5: Design Metrics

Criteria	Definition	Application	Additional Notes
Statistical Model Supported	The effects that can be estimated by the design (e.g., main effects, 2-factor interactions, quadratic effects, or higher order terms)	Corresponds to test objective	
Confidence	The probability of concluding a factor has no effect on the response, when it does not have an effect (True Negative Rate and equal to 1 – type I error rate)	Maximize	It must be determined before the experiment is executed, and it is used to help determine power of a design.
Power	The probability of concluding a factor has an effect on the response when in fact it does. (True Positive Rate and equal to 1 – type II error rate)	Maximize	A design with power greater than 80% for model terms is ideal.
Correlation Coefficient between Model Parameter Estimates	Correlation coefficients measure the strength and direction of the linear relationship between two model parameters	Minimize correlation between model parameters	If model parameters are *orthogonal*, the estimated parameter for one model term is the same value whether the other model term is included in the model or not.
Variance Inflation Factor (VIF)	The VIF for a factor measures the degree of multicollinearity with the other factors. Multicollinearity occurs when the factors are correlated among one another.	If a factor is not linearly related to the other factors, the VIF is 1; otherwise, it is greater than 1. VIFs greater than 10 indicate serious problems with multicollinearity.	If there's multicollinearity, the values of the estimated model parameters change depending on which terms are included in the model.

Design Resolution	Design resolution is a metric of a screening design that indicates the degree of confounding in the design. In general, a design is resolution R if no p-factor effect is aliased with another effect containing less than R-p factors.	The higher the better, but also more expensive	The higher the resolution, the less restrictive the assumptions are on which higher-order interactions are negligible.
Prediction Variance	Variance of the predicted response	Balance over regions of interest	This is important when the goal is prediction, but may be less of a priority for screening.
Fraction of Design Space (FDS) Plot	Summarizes the prediction variance across the design region	Ideally, the curve is relatively flat with small values so that the prediction variance does not change drastically across the design space.	The plot is useful to identify minimum, median, and maximum prediction variance across the design region.
Design Efficiency	Evaluation measure related to computer-generated optimal designs (Could be related to parameter estimation or response prediction)	Maximize	There are several criteria available (D-optimal, I-optimal, etc.). Higher efficiency is better, but does not give full picture of design properties.
Strength (Applies to combinatorial and orthogonal array designs)	Indicates the level of interactions that are fully covered by the design	Sets the design size and is used to scope or reduce risk for finding errors	

The first decision that one must make is that of choosing between a classical design and an optimal design. The STAT COE recommends the "classical first" approach to DOE due to the fact that classical designs possess certain desirable properties such as capturing the entire design space, orthogonality, and optimality with respect to several optimality criteria such as D-optimality, Myers et al. (2016, pp. 467). Optimal designs are most appropriate for tests in which the sample size is unusual, the design region is highly constrained, or there are several categorical factors at more than two levels.

To pick a final design, the design metrics can be compared to reach the final choice. Table 6 shows an example of how design options can be evaluated. The number of points, the supported empirical regression model, the confidence and power levels, and other critical design metrics can be varied to balance cost and risk for the test program. In addition, Myers et al. (2016, pp. 370) discuss 11 desirable

properties of test designs which are repeated below. These properties should be balanced to achieve the priorities of your experiment.

1. Result in good fit of the empirical model to the data.
2. Provide good model parameter estimates.
3. Provide a good distribution of prediction variance of the response throughout the region of interest.
4. Provide an estimate of "pure" experimental error.
5. Give sufficient information to allow for a test for curvature or lack of fit.
6. Provide a check on the homogeneous variance assumption.
7. Be insensitive (robust) to the presence of outliers in the data.
8. Be robust to errors in the control of design levels.
9. Allow empirical models of increasing order to be constructed sequentially.
10. Allow for experiments to be done in blocks.
11. Be cost effective.

There are many considerations to keep in mind when selecting a design. However, it is worth the time and effort to choose the right design for your experiment. As Montgomery (2017) says, "A well-designed experiment is important because the results and conclusions that can be drawn from the experiment depend to a large extent on the manner in which the data were collected." Choosing the right experiment allows you to leverage data to answer your objectives and verify requirements.

Table 6: Alternative Test Design Choices for a Notional Program

Design #	1	2	3
# Factors	4	4	4
Levels	2	2	2
Model Supported	ME	ME, 2FI	ME, 2FI, Q
Signal to Noise Ratio	1.0	1.0	1.0
Alpha	0.05	0.05	0.05
# Center Points	0	4	5
# Repetitions	7	7	14
Total Runs	20	24	36
Power for ME @ SNR	0.63	0.54	0.95
Power for 2FI/Q @ SNR	0.54	0.53	0.87
FDS Pred Err @50%	0.63	0.58	0.36
FDS Pred Err @95%	0.90	0.75	0.45
VIF Avg	4.09	3.73	3.45
VIF Max	11.00	11.30	11.48
Confounding/Aliasing	med	med	low

ME: Main effect	**2FI**: Two-factor interaction	**Q**: Quadratic
SNR: Signal-to-noise ratio	**FDS**: Fraction of design space	**Pred Err**: Prediction error
VIF: Variance inflation factor	**Avg**: Average	**Max**: Maximum

Takeaway: A design must be tailored to meet the test objectives and unique test circumstances.

3.10 Test Execution Planning

Some issues to consider when planning the execution of a design are the three principles of DOE: randomization, replication, and blocking.

Most runs in a test matrix design are executed in random run order in order to protect against known (and unknown) uncontrollable factors as well as nuisance factors (noise). If a design has factors of interest that are hard-to-change, or the situation otherwise prevents a complete randomization of the run order, a *split-plot* design can be used. The analysis of a split-plot design correctly accounts for restricted randomization. *Replicating* the design, or some portion of runs within the design, not only provides an estimate of pure error (the component of the error sum of squares attributed to replicates), but also mitigates the potential increase in variance induced by the presence of outliers that may result as a consequence of executing the design.

Blocking is a planning and analysis technique that sets aside (blocks out) both undesirable and known sources of nuisance variability so they do not hinder the estimation of other effects. Blocking is done when there is an expectation that the nuisance factor may introduce noise into the test data. The excessive presence of such noise may decrease the power to detect system performance responses of legitimate factors of interest. Figure 14 depicts a notional example in which test points are grouped together, or blocked, based upon a single effect; such an effect may be the test schedule, the type of equipment used, test locations, or operators. See the example in Appendix B for an example of a blocked experiment analyzed (correctly) with the block effect and (incorrectly) without the block effect. If the effect of the blocking variable *is* of interest, the design ***should not be blocked*** with respect to this factor. Because of the restricted randomization of blocks, there is not a valid statistical test to analyze the effect of a blocking variable.

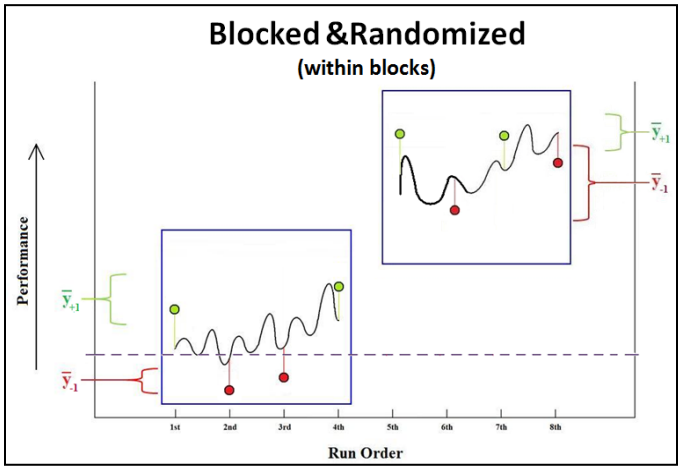

Figure 14: Illustrative Representation of the Randomized Blocking Procedure

We have limited the discussion here to block effects treated as a fixed factor. There are many cases where the block effect could (or should) be treated as a random effect (for example, day of the week or a batch of raw material). Conclusions from the experiment with random blocks are valid across the population of blocks (e.g., across all days or all batches of material). Refer to Montgomery (2017, pp. 147) for details on the analysis of an experiment with random blocks.

After the data has been collected, it must be analyzed. Any deviations from the test protocol must be recorded at the time of execution and accounted for in the analysis.

Critical questions:

1. Will the design be executed in random run order?
    a. If not, why?
    b. If randomization is restricted, must the design be analyzed as a split-plot design?
2. Are you interested in the blocking variable? If so, then the design ***should not be blocked*** according to this factor. Consider a split-plot design instead.
3. How will deviations from the test design plan be recorded and reported?

Takeaway: Try to execute the planned design, record any deviations, and use the appropriate analysis.

3.11 Analysis

It is critical to have a good plan for the analysis of the data. The requirement will guide us to the type of analysis we must perform. Is the requirement that the new system be better than the old system? Is the requirement set to a specific bench mark? If you have followed the guidelines for the planning and design of tests documented in this manual, the analysis of your test data will be straightforward and hence easily traced back to the defined objectives.

Software will play a key role in the analysis of the test data. There are various software packages for classical design analysis (JMP, Design Expert, SPSS, R, and Minitab to name a few). Since capabilities may vary depending upon the software package, you may end up using several simultaneously. It is best to experiment with a variety of software packages in order to find those that best meet your needs.

In any analysis, it is important to know how to read and interpret the software's output. For example, in a classical design, we observe the value of the F-statistic for the model to determine overall model significance. Next, we would observe the individual effect's t-statistic or F-statistic (and p-value) to determine the magnitude of its contribution to the response that is defined for a particular model. One would also require the mean squared error and the degrees of freedom. Refer to Montgomery (2017) for complete details on analyzing data from a designed experiment.

It is imperative to analyze the data as it was actually collected during the test, not the way it was planned to be collected. It is also imperative that if the test deviates from the planned protocol, then the analysis of the data must be adjusted to reflect the actual (not planned) test execution. For example, if a completely randomized design was planned but the runs were not randomized, the analysis must be that of a split plot or blocked design.

Figure 15 shows how an empirical response surface developed during DT can be used to predict responses in OT. The relative scarcity of controllable factors in OT might induce a response in the interior of the region (depicted by yellow dots). Even if these OT points were never actually tested, the response surface permits prediction for later validation with the OT points. Futhermore, failing combinations, the rate at which performance degrades across the test space, optimum operating conditions, and other analysis can be derived from the surface.

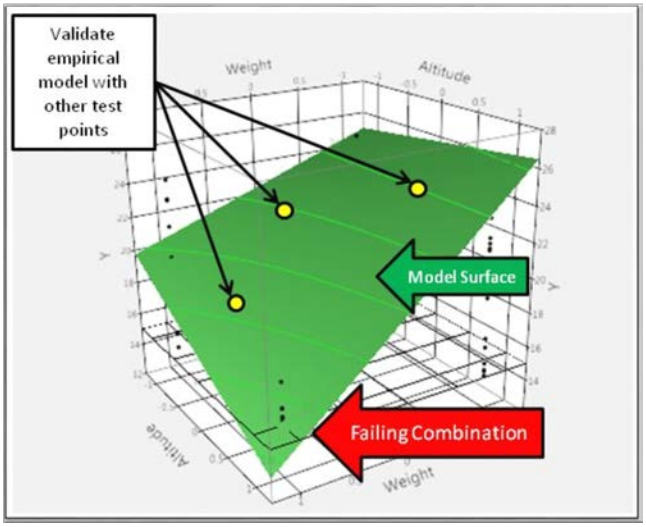

Figure 15: An Example of Test Data Analysis and Interpretation

Will the analysis address the requirement? Early in the planning process, the team defined the objectives, responses, and factors. If these were clearly and thoroughly understood, then the resulting designs should provide the right information.

Critical questions:

1. Will the analysis address the objective?
    a. Is there sufficient power to screen factors?
    b. Are the points sufficient to minimize prediction error?
    c. Are sufficient points included for assessing and fitting for curvature?
2. Will the data inform the details in the requirement?
    a. Are a sufficient number of factors being controlled?
    b. Are the factors clear, concise, and controllable?
3. Were the test points selected as part of a formal design or larger test strategy?
4. Do the test points exhibit orthogonality or space filling properties that will support the objective?
5. How is the requirement defined? Will the data answer the questions posed by it? (refer back to Figure 3)
    a. Is the requirement defined across the full range of operational conditions?
    b. Is the requirement stated as an overall probability of success (pass/fail percentage)?
    c. Is the requirement an average of all performance across the operational space?

d. Is the requirement only defined for specific conditions or a limited region?

Takeaway: Check that you are answering the right question with the analysis.

4 Conclusion

Thorough planning is fundamental to ensuring that sufficient rigor and traceability from requirements to analysis are incorporated into the test design. STAT is a collection of deliberate, methodical processes and techniques that embodies this design philosophy. STAT introduces organization and rigor into the testing process by means of a structured, iterative approach to test design, beginning with requirements and mission profiles and culminating in a realistic and executable design, the execution of which enables informed decisions on system acquisition and future planning to meet tomorrow's threats.

5 References

Anderson-Cook, Christine M. “When Should You Consider A Split-Plot Design?” Quality Progress, Oct. 2007, pp. 57–59., asq.org/quality-progress/2007/10/laboratory/when-should-you-consider-a-split-plot-design.html.

Anderson-Cook, C. M., et al. “Bayesian stockpile reliability methodology for complex systems.” Military Operations Research, 2007, 25-37.

Anderson-Cook, Christine M., et al. “Reliability Modeling using Both System Test and Quality Assurance Data.” Military Operations Research Journal, vol. 13 No. 3, 2008.

Auborn, John and Paola Pringle. ”Test and Evaluation in Acquisition of Capabilities”, http://itea.org/images/pdf/conferences/2017_Symposium/Proceedings/Auborn_Pringle%20TE%20in%20Acquisition%20of%20Capabilities.pdf, 34thAnnual International Test and Evaluation Symposium, 2017.

Box, George E. P., and Patrick Y. T. Liu. “Statistics as a Catalyst to Learning by Scientific Method Part I—An Example.” Journal of Quality Technology, vol. 31, no. 1, 1999, pp. 1–15., doi:10.1080/00224065.1999.11979889.

Box, George E. P., et al. Statistics for Experimenters: Design, Innovation, and Discovery. 2nd ed., Wiley-Interscience, 2005.

Box, George E. P., and Patrick Y. T. Liu. “Statistics as a Catalyst to Learning by Scientific Method Part I—An Example.” Journal of Quality Technology, vol. 31, no. 1, 1999, pp. 1–15., doi:10.1080/00224065.1999.11979889.

Burke, Sarah, “Reliability Test Plans for Binary Responses - Excel Tool Guide,” Scientific Test and Analysis Techniques Center of Excellence (STAT COE), www.AFIT.edu/STAT, October 2018.

Charles, Philipp and Phil Turner. “Capabilities-Based Acquisition...From Theory to Reality.” CHIPS The Department of The Navy’s Information Technology Magazine, July-September 2004.

Cohen, Jacob. “The Cost of Dichotomization.” Applied Psychological Measurement, vol. 7, no. 3, 1983, pp. 249–253., doi:10.1177/014662168300700301.

Coleman, David E., and Douglas C. Montgomery. “A Systematic Approach to Planning for a Designed Industrial Experiment.” Technometrics, vol. 35, no. 1, 1993, pp. 1–27., doi:10.1080/00401706.1993.10484984.

Freeman, Laura J., et al. “A Tutorial on the Planning of Experiments.” Quality Engineering, vol. 25, no. 4, 2013, pp. 315–332., doi:10.1080/08982112.2013.817013.

Gelman, Andrew et al. Bayesian Data Analysis. 2nd ed. New York, New York, Chapman & Hall/CRC, 2004.

Gilmore, Michael J. “A Statistically Rigorous Approach to Test and Evaluation (T&E).” International Test & Evaluation Association Journal, vol. 34, Sept. 2013.

Hamada, Michael. "The Advantages of Continuous Measurements over Pass/Fail Data." Quality Engineering, vol. 15, no. 2, 2002, pp. 253–258., doi:10.1081/qen-120015857.

Harman, Michael, "Practical Bayesian Analysis for Failure Time Data," Scientific Test and Analysis Techniques Center of Excellence (STAT COE), www.AFIT.edu/STAT, Sept. 2018.

Harman, Michael, "Understanding Requirements for Effective Test Planning, Best Practice," Scientific Test and Analysis Techniques Center of Excellence (STAT COE), www.AFIT.edu/STAT, Jan. 2014.

Hutto, Greg, Dragon Spear Test Planning Materials, 96th Test Wing, Eglin AFB, FL, 2011.

Joint Chiefs of Staff (JCS), "Cyber Survivability Endorsement Implementation Guide", v1.01a.

Kensler, Jennifer, "Reliability Test Planning for Mean Time Between Failures," Scientific Test and Analysis Techniques Center of Excellence (STAT COE), www.AFIT.edu/STAT, Oct. 2018.

Kensler, Jennifer and Luis A. Cortes. "Interpreting Confidence Intervals," Scientific Test and Analysis Techniques Center of Excellence (STAT COE), www.AFIT.edu/STAT Dec. 2014.

Kutner, Michael H., et al. Applied Linear Statistical Models. McGraw-Hill Education, 2005.

McQueary, Charles E. Definition of Integrated Testing, OSD Memorandum, 25 April 2008.

Meeker, William Q., and Luis A. Escobar. Statistical Methods for Reliability Data. 9th ed., Wiley, 1998.

Montgomery, Douglas C. Design and Analysis of Experiments. 9th ed., John Wiley, 2017.

Myers, Raymond H., et al. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley-Blackwell, 2016.

NIST/SEMATECH e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook, 19 Aug. 2014.

Ortiz, Francisco, "Categorical Data in a Designed Experiment Part 1: Avoiding Categorical Data." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), https://www.afit.edu/stat, 2018.

Ortiz, Francisco and Lenny Truett. "Using Statistical Intervals to Assess System Performance," Scientific Test and Analysis Techniques Center of Excellence (STAT COE), https://www.afit.edu/stat, Apr. 2015.

Senechal, Kenneth. Capabilities Based Test & Evaluation (CBTE), https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2018/test/Senechal.pdf, 2018.

Simpson, James, "Testing via Sequential Experiments," Scientific Test and Analysis Techniques Center of Excellence (STAT COE), www.AFIT.edu/STAT, Jan. 2014.

Tague, Nancy R. *Quality Toolbox*. 2nd ed., ASQ Quality Press, 2005.

United States, Department of Defense (DoD), "Cybersecurity Test and Evaluation Guidebook," 2018. Version 2.0. https://www.acq.osd.mil/dte-trmc/docs/CSTE%20Guidebook%202.0_FINAL%20(25APR2018).pdf.

United States, Department of Defense (DoD), "Risk Management Framework (RMF) for DoD Information Technology (IT)", 2016. https://www.dau.edu/acquipedia/pages/articledetails.aspx#!245

United States, Department of Defense (DoD), "Defense Acquisition Guidebook," 2019. https://www.dau.edu/tools/dag

United States Air Force, "AIR FORCE INSTRUCTION 99-103, CAPABILITIES-BASED TEST AND EVALUATION," 6 April 2017.

United States, Department of Defense (DoD), "CHAIRMAN OF THE JOINT CHIEFS OF STAFF INSTRUCTION J-1 CJCSI 3126.01A, LANGUAGE, REGIONAL EXPERTISE, AND CULTURE (LREC) CAPABILITY IDENTIFICATION, PLANNING, AND SOURCING," 31 January 2013.

United States, Department of Defense (DoD), "Test and Evaluation Management Guide," 6th ed., Dec. 2012. https://www.dau.edu/guidebooks/Shared%20Documents%20HTML/Test_and_Evaluation_Mgmt_Guidebook.aspx

Appendix A Example Application of Sequential Testing

Consider the acquisition of a new missile called the Good Enough Missile (GEM) which is based on a legacy system. GEM has almost all of the same components as its legacy system with the exception of an upgraded targeting system. In order for GEM to replace the old missile, it needs to perform as good as or better than the old system in terms of the following requirements:

- The missile shall have a range of R meters.
- The missile shall be usable on stationary targets of varying size.
- The missile shall be usable during different times of day and under various states of cloud cover.
- The missile shall have a maximum impact time of T seconds from time of launch.
- The missile weight shall not exceed W pounds.

The weight requirement can be shown using only a demonstration (recall Section 3.2); however, the other requirements need to be fully tested across the operating space. Due to the use of legacy sub-systems, testing focuses on the new targeting system.

Objectives were stated in the form of two testable questions: 1) Does the missile hit the target at a range of R meters or less? 2) Does the missile spend T seconds or less in the air from time of launch to time of impact? It is possible to evaluate both of these questions and satisfy the remaining requirements using sequential testing across the operational test space. The test team determined that a suitable response for the first question would be radial miss distance from the target to the point of impact. Any damage done to the target is considered to be "hit" with a recorded miss distance of 0. For the second question, the response is measured as seconds from missile launch to impact.

The chosen factors were developed using both the current requirements and previous experience with the legacy missile. Factors were defined to be continuous wherever possible to allow for interpolation between values. Defined factors include distance from launch point to the target in meters, launch angle in degrees, launch point height in meters, target size, time of day, cloud cover, operator experience, and terrain and can be seen in Table A-1. For target size, it was determined that it would be impractical to measure the size of each target because of differing shapes. Thus, the targets were grouped into size categories of small, medium, and large. It was not possible to measure lumens for time of day, so day or night are the only recorded levels. However, a high resolution sensor was procured that will allow cloud cover to be measured in oktas, a unit of measurement in meteorology ranging from 0 (completely clear sky) to 8 (completely overcast). Due to a wide range of operator backgrounds, operator experience is measured as new or experienced according to previous use of the legacy system that GEM is replacing. Subject matter experts suggest that GEM could be used in terrains such as mountains, desert, jungle/forest, or coastal areas. Values of -1 and 1 mark the placeholders for the actual minimum and maximum numbers, respectively, used during testing.

Table A-1: Factors and Levels for GEM Testing

Factors	Levels			
Distance to target	-1	1	-	-
Launch angle	-1	1	-	-
Launch point height	-1	1	-	-
Target type	Small	Medium	Large	-

Time of day	Night	Day	-	-
Cloud cover	-1	1	-	-
Operator experience	New	Experienced	-	-
Terrain	Mountains	Desert	Jungle/forest	Coastal

Terrain was deemed a hard to change factor because of the limitations of the physical location during each run. Cloud cover was also marked as a potential problem since weather patterns must be predicted in advance. Since the budget will not allow for location changes between each run, the design must instead accommodate groups of runs at a single location being performed. Several designs were created for consideration for GEM testing.

Table A-2 details the potential designs for comparison. Since terrain cannot be randomized during testing, the full factorial design (Design #1) is infeasible. Instead, a split plot design was chosen using terrain as the whole plots. Budget restrictions require that the screening design have a maximum run size of 80 to leave some budget for follow-on testing for the remaining factors. The split plot designs allow for reasonable power for all main effects except terrain. Each design in the table was constructed using JMP (V. 13) and the 8 required factors. The team ideally wanted to be able to support both main effects and two factor interactions with the design. After discussions with subject matter experts, it was determined that the legacy system had large effect sizes relative to noise, making a signal-to-noise ratio of 3.0 reasonable instead of the default 2.0. In the end, the team chose Design #5, accepting a higher Type I error risk (alpha = 0.1) in order to achieve higher power. The slightly high prediction error was also deemed acceptable since the goal of this design was to screen factors.

Table A-2: Potential Screening Designs for GEM

Design #	**1**	**2**	**3**	**4**	**5**	**6**
Software Package	JMP	JMP	JMP	JMP	JMP	JMP
Name/Design Type	Full Factorial	Split Plot 1	Split Plot 2	Split Plot 3	Split Plot 4	Split Plot 5
Factors	8	8	8	8	8	8
Model Supported	ME, 2FI	ME, 2FI	ME, 2FI	ME, 2FI	ME, 2FI	ME
Signal-to-noise Ratio	2.0	2.0	2.0	3.0	3.0	3.0
Alpha	0.05	0.05	0.05	0.05	0.1	0.05
Total Runs	768	80	72	72	72	36
Power for ME @ SNR	1	0.16	0.15	0.29	0.46	0.15
Power for 2FI/Q @ SNR	1	0.81	0.67	0.88	0.95	
FDS Pred Err @50%	0.05	1.12	1.31	1.31	1.31	0.77
FDS Pred Err @95%	0.06	1.50	1.92	1.92	1.92	0.81
Aliasing	none	low	low	low	low	medium

ME: Main effect	2FI: Two-factor interaction	Q: Quadratic
SNR: Signal-to-noise Ratio	FDS: Fractional design space	Pred Err: Prediction error

Data was collected by executing the design matrix generated by JMP for the specifications of design #5. Figure A-1 shows the JMP effect summary for the response of impact distance. As expected, distance to target has the largest influence on impact distance. Both launch point height and cloud cover have been removed from the empirical model after screening. These two factors will thus be removed from any future testing.

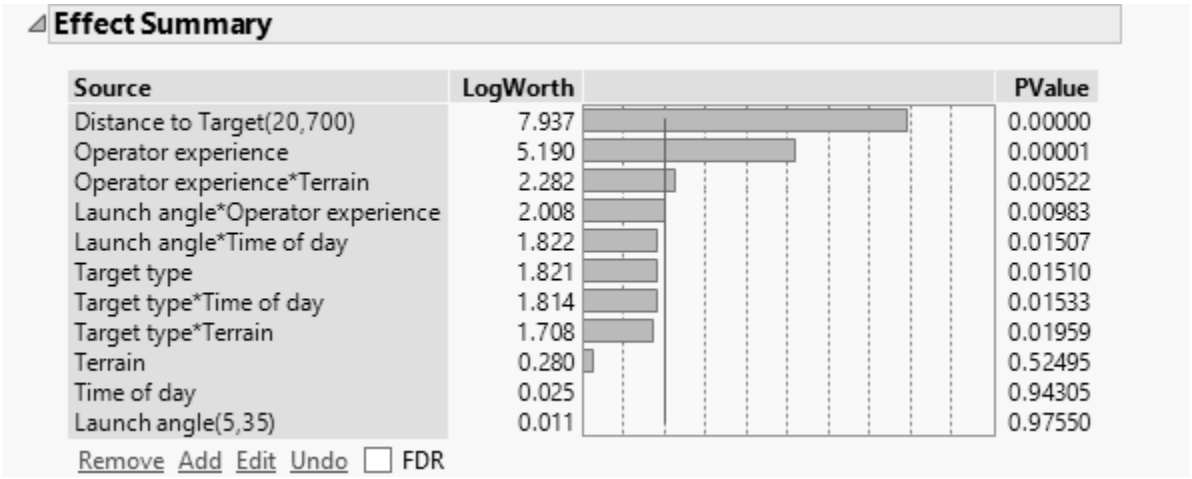

Figure A-1: Model effect summary for response of distance to target

In Figure A-2, the model effect summary for the response of time from launch to impact has a very limited number of remaining factors. Only distance to target, launch angle, and their interaction remain in the empirical model. While this information may be helpful for future testing, the follow-on testing for GEM must still include all of the factors that were significant to the other response variable.

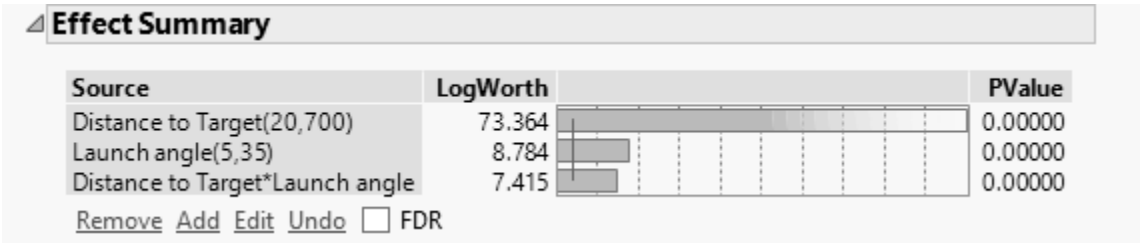

Figure A-2: Model effect summary for response of time from launch to impact

The next stage of testing will remove launch point height and cloud cover from the factor list. In addition, the testers found that external temperature might be causing some variation in the response. So, temperature will be added as factor during the next round of testing. The follow-on design will also be a split plot since terrain was not one of the eliminated factors.

Without sequential testing, the entire budget might have been used on this first round of testing. Instead, a second design can be created with a refined factor list that can be focus instead on building confidence intervals around the response variables.

Appendix B Example of Analysis of Randomized Block Design

Consider the following example as provided in Montgomery (2017). An experiment is performed to determine the effect of three washing solutions on bacteria growth. Only three trials can be performed each day. Because day could be a source of variability, a randomized block design is used. Consider the results of this experiment by including or ignoring the nuisance factor day in the analysis in Table B-1.

Table B-1: Bacteria Growth Data

	Days			
Solution	1	2	3	4
1	13	22	18	39
2	16	24	17	44
3	5	4	1	22

The correct analysis for this experiment includes day as a block effect in order to determine the effect of the washing solution on the response. The JMP output in Figure B-1 shows that there is a significant difference in growth rate for the three washing solutions. The large F ratio for the block effect indicates day did introduce variability in the response, so it was important to include this in the analysis.

Summary of Fit

Rsquare	0.972166
Adj Rsquare	0.948972
Root Mean Square Error	2.939199
Mean of Response	18.75
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Solution	2	703.5000	351.750	40.7170	0.0003*
Block	3	1106.9167	368.972	42.7106	0.0002*
Error	6	51.8333	8.639		
C. Total	11	1862.2500			

Figure B-1: JMP Output for Blocked Experiment

Now suppose we had ignored the nuisance factor day in the analysis. The results of this new analysis is shown in Figure B-2. Because we have ignored day, the block sum of squares has been folded into the error sum of squares. The error sum of squares is now 1158.75 (compared to 51.83 previously when including block effect in the analysis). Because the variability due to the nuisance factor day has not been partitioned from the error sum of squares, the effect of solution has not been identified as being significant (p-value = 0.1182). By ignoring the block effect, the analysis fails to identify the significant effect of solution on the response.

Summary of Fit

Rsquare	0.377769
Adj Rsquare	0.239495
Root Mean Square Error	11.34681
Mean of Response	18.75
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Solution	2	703.5000	351.750	2.7320	0.1182
Error	9	1158.7500	128.750		
C. Total	11	1862.2500			

Figure B-2: Analysis Ignoring Block Effect

If a blocked experiment is executed (whether it was planned that way or not), the analysis of the data should reflect that. By ignoring a nuisance factor in the analysis, you may reach incorrect conclusions and not identify important relationships.

Appendix C Reliability Test Planning

This appendix provides references and further details related to the topics previously presented in Table 1 in Section 1. Reliability testing is intent on determining the distribution of failure times, a top level metric like mean time between failures (MTBF), or a probability of failure (value between 0 and 1). Reliability growth analysis tracks recurring reliability assessments through time and seeks to determine when and if the system will achieve its threshold (minimum) reliability. Reliability testing is required to perform reliability growth tracking and several methods are summarized in Table C-1.

Table C-1: Statistical Methods for Reliability

Reliability Assessment	Sampling Plans	Sampling is the selection of a subset (a statistical sample) of members from within a statistical population to estimate characteristics of the whole population. Two advantages of sampling are that the cost is lower and data collection is faster than measuring the entire population. (NIST)
	Sequential Probability Ratio Test	SPRT is a specific sequential hypothesis test that permits concurrent pass/fail analysis during testing and provides stopping/continuing criteria. (Wikipedia)
	Parametric Survivability Analysis	Fits the time to event Y (with censoring) using linear regression models that can involve both location and scale effects (JMP.com)
Reliability Growth	Non-Homogenous Poisson Process (NHPP)	NHPP analysis permits the estimation of a variable failure rate that reflects a change in reliability, typically due to configuration changes designed to improve reliability.
	Bayesian Analysis/Inference	Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. (Gelman et al., 2004, Meeker and Escobar, 1998)

C.1 Sampling Plans

Reliability test plans may employ DOE, but the resources required for this in DoD are typically too costly. Instead, sampling plans prescribe a minimum required time, distance, or number of test units and a maximum allowable number of failures to pass the test. The sampling plans are sized based on the reliability requirement and the desired statistical metrics. Unlike DOE, these data are sampled across all conditions encountered over the duration of testing and not individually prescribed per event. Tools to create continuous sampling plans (e.g., MTBF, mean miles between failures) can be found in Kensler (2018) and binary response plans can be generated using Burke (2018).

C.2 Sequential Probability Ratio Test (SPRT)

SPRT is a modified sampling plan that enables testing to stop early if certain conditions are met. A minimum test time (or number of samples) is determined and testing begins. At every failure, a ratio is calculated and the method determines if the system is passing (or failing) with sufficient margin that additional testing would not reverse the decision to pass (or fail) the system. If stopping criteria is not met, testing continues until the stopping criteria is eventually met or until the pre-determined test duration is achieved. An MS Excel-based SPRT analysis tool can be acquired by emailing the STAT COE at COE@AFIT.edu. This tool supports both continuous and discrete reliability measures.

C.3 Parametric Survivability Analysis

This method can be used to analyze reliability data by including the test conditions associated with the failures. This is similar to the regression analysis performed with DOE but it uses more generalized models which can better represent the true failure distributions. Commercial software packages provide this capability (e.g., JMP.com or Reliasoft.com) and facilitate complex data sets to include many factors and censored data, among others.

C.4 Reliability Growth

Reliability growth (RG) analysis employs a non-homogenous Poisson process to continually update the failure rate between correction periods. This results in a plot of varying reliability values throughout the test period. RG does not create the test time or sampling plan but rather provides the method for analyzing the data. The US Army freely provides the Reliability Growth Tracking Model-Continuous (MS Excel) tool available at https://www.amsaa.army.mil/CRG_Tools.html. Commercial software is also available from sources like JMP.com and Reliasoft.com

C.5 Bayesian Analysis

The previously described methods are “frequentist” in nature as they rely on the frequency of events to estimate a statistic. Bayesian inference is a method of statistical inference that combines extensive past testing and physical theory with limited additional test data (Meeker and Escobar, 1998). Bayes can be used to assess the reliability metric for a single test period or for growth tracking over multiple test phases and configurations. The application of Bayes can range from fairly simple top level (e.g., just failure times) analyses to complex constructs involving reliability block diagrams and subsystems and components. Simple Bayes MTBF methods and software code can be found in Harman (2018).

More complex methods require support from a STAT Expert and custom code for your actual system application (using R or similar statistical applications). Bayesian analysis is especially effective when full system testing is limited. Combining subsystem or other ground testing data with the full system test data enables a more robust reliability assessment. Bayesian analysis can address multiple data sources, changing configurations, and the need for accurate, recurring estimates.

Bayesian analysis uses priors, statistical distributions of expected performance (i.e., our belief on the performance before the data is observed) and mathematically combines it with new test results (likelihood) to generate an output distribution (posterior) of the performance parameter. The posterior distribution represents updated knowledge from additional testing about the parameter. Figure C-1 shows the components required to perform Bayesian analysis.

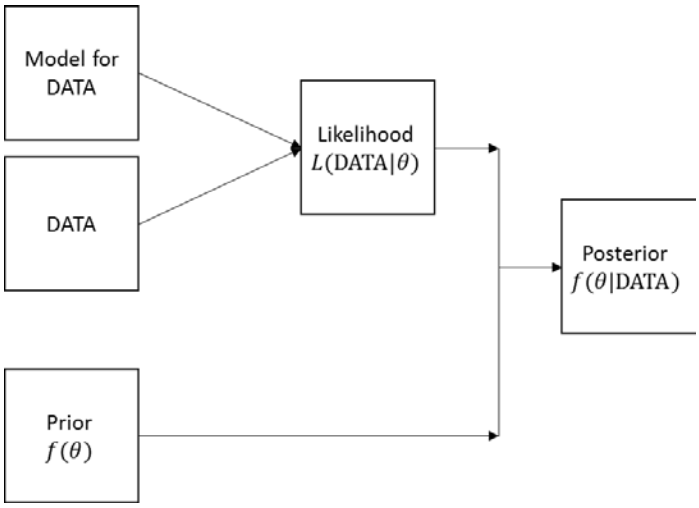

Figure C-1: Bayesian Analysis (Adapted from Meeker and Escobar, 1998)

Determination of a prior can come from two primary sources: 1) past data or 2) expert opinion. Past test data may come from (in preferential order): current system test data, previous test data from a similar system, previous component test data, current system simulation, and analysis. Expert opinion can be used when no other prior test data is available. Eliciting prior information from experts is ideally done by eliciting the general shape of the parameter distribution of interest and typical values for given quantiles (e.g., the 10^{th} and 90^{th} quantiles). If there is no prior information available from either past data or expert opinion, a vague prior, an approximately constant prior over the range of the parameter, can be used. When using expert elicitation, "wishful thinking" must not be used to determine the priors; vague priors are preferred when there is limited prior information (Meeker and Escobar, 1998).

Prior information may come in the form of continuous data (e.g., failure times or a mean time between failures) or binary data (e.g., pass/fail). Ultimately, all reliability data in the Bayesian model must reflect the way the requirement is stated (continuous or binary).

Translating continuous data into a distribution that exists only between 0 and 1 is described in Anderson-Cook et al. (2008). Anderson Cook et al. (2007) also describes how to combine information from multiple experts into one prior. The method for mapping data between domains should follow established methods and be accomplished through mutual discussion and agreement between the government program office and contractor. Sensitivity analysis should be performed to assess the reliance of the reliability estimates on the choice of the prior assumptions.

The overall process entails the use of initial prior information (expectations) regarding the components and the gathering of actual reliability data through testing. As testing proceeds, the last posterior becomes the prior for the next step. This process naturally includes any configuration changes into the posterior estimates. The program office and contractor must come to an agreement on how and which test article configurations will be utilized as priors in the reliability assessment.

Bayesian analysis for systems of this complexity requires the use of statistical software packages and custom created code. In addition to a point estimate of the system reliability, interval estimates will also be reported which would not be readily available using a frequentist approach. This code should be

developed and validated as part of a STAT working group for the system reliability block diagram. The details of the coding methodology are not provided herein.

Appendix D Glossary of STAT Terms

Affinity diagram	A tool and process to organize many items into natural groupings or to facilitate when a group needs to come to a consensus.
Aliasing	The degree to which a term in the fitted model is biased by active term(s) not in the fitted model.
Blocking	A planning and analysis technique that sets aside (partitions) both undesirable and known sources of nuisance variability so they do not hinder the estimation of other effects. A variance reduction technique used in design of experiments to account for nuisance variables.
Confidence	The probability of concluding a factor has no effect on the response, when it does not have an effect (True Negative Rate and equal to 1 – type I error rate).
Confounding	Please see aliasing for definition.
Control factor	An experimental factor that can be controlled and varied during test that could impact the response as factor levels change.
Constraint	Anything that limits the design of the test space (resources, range procedures, operational limitations, etc.).
D-optimal design	A popular computer-generated designed experiment used for screening experiments due to favorable properties to identify active effects. Mathematically, an optimal design such that the determinant of the information matrix is maximized. Equivalently, a design such that the volume of the confidence region of the parameter estimates is minimized. See Montgomery (2017) or Myers et al. (2016) for details.
Efficiency, design	An evaluation measure related to computer-generated optimal designs. The value depends on the optimality criteria used (e.g., D-optimality, I-optimality). Efficiency allows for comparing designs that have different sample sizes.
Factor	Input (independent) variable for which there are inferred effects on the response.
Fishbone diagram	A brainstorming tool whose purpose is to facilitate extracting causal factors.
Hard-to-change factor	A factor whose levels are difficult, time-consuming, or costly to change after every test run.
Hold constant factor	A factor that likely influences the response that is set to a constant value throughout all of the tests. May not be of primary interest, may be too difficult to vary, or may have been previously shown to not be significant.
I-optimal design	A popular computer-generated designed experiment often used when the goal is to optimize or predict values due to favorable properties of prediction variance. Mathematically, it is an optimal design generated such that the average prediction variance is minimized. See Montgomery (2017) or Myers et al. (2016) for details.

Inter-relationship diagram	A brainstorming tool that maps items (or steps) in a complex solution (or process). Often used after a fishbone diagram or affinity diagram to further understand or explore links between items.
Level (of a factor)	The values set for a given factor throughout the experiment.
Multicollinearity	An issue that occurs in data sets when the factors or predictor variables are correlated among themselves. Multicollinearity is very common in observational studies. Contrast with orthogonal design.
Noise factor	A factor that likely influences the response, but may not be possible (or desirable) to control in an experiment in the field.
Nuisance factor	A factor that could influence the response, but which is not a factor of interest.
Optimal design	A computer-generated design found by optimizing a specified characteristic or property of the design. Optimal designs require many inputs from the user including: test size, number of factors, factor levels, desired empirical model, design region constraints, and design criterion. See also D-optimal design, I-optimal design.
Orthogonal design	A design in which the model terms are linearly independent of each other.
Random factor	A factor that has a large number of possible levels and whose levels are randomly selected from the population of factor levels. Analysis on a random factor provides inference on the entire population of factor levels (not just those observed in the experiment).
Power	The probability of concluding a factor has an effect on the response when in fact it does (True Positive Rate and equal to 1 – type II error rate).
Pure error	The component of the error sum of squares attributed to replicates.
Repeated observation	Multiple measurements made at once for a given factor level combination.
Replicate	An independent test run of each factor level combination.
Resolution	A metric of a screening design that indicates the degree of confounding in the design. In general, a design is resolution R if no p-factor effect is aliased with another effect containing less than R-p factors.
Resolution III design	A design such that main effects are not aliased with other main effects, but are aliased with two-factor interactions and some two-factor interactions may be aliased with each other.
Resolution IV design	A design such that main effects are not aliased with other main effects or with any two-factor interaction, but two-factor interactions are aliased with each other.
Resolution V design	A design such that no main effect or two-factor interactions is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions.
Response	The measured, dependent output of a test event.
Screening	Common design objective to identify which factors are active (important/statistically significant) in the model of the response and eliminating those that are unimportant.
Sequential design strategy	A method of beginning with lower level designs and building up to more complicated and complex systems.

Signal-to-noise ratio	The ratio of δ (the desired difference in the response to detect) over σ (the magnitude of the process noise variability).
Split-plot design	A type of designed experiment used when there are restrictions on randomization due to hard-to-change factors.
Strength	A metric used for combinatorial design that indicates the level of interactions that are fully covered by the design.
Test and Evaluation Master Plan	The master document for the planning, management, and execution of the T&E program for a particular system. The TEMP describes the overall test program structure, strategy, schedule, resources, objectives, and evaluation frameworks.
Type I error	The probability of concluding a factor has an effect on the response when it does not have an effect (False positive rate and equal to 1 – confidence).
Type II error	The probability of concluding a factor does not have an effect on the response when it does have an effect (False negative rate and equal to 1 – power).

Appendix E Learning Resources

Order of Learning: We recommend you take a formal design of experiments course first so you gain knowledge in a forum in which you can ask questions and learn the fundamentals. After that, access texts and online resources to broaden your knowledge, learn additional rigorous techniques, and seek ways to solve your specific problems. Remember that the key is to first understand the problem and then apply the methods, techniques, and designs that will address the stated objectives. The following sections detail resources for learning or researching STAT related topics. These are merely a few references and this list barely scratches the surface. When in doubt, ask your STAT expert.

1. Formal Courses
    a. US Air Force (sign up using your Service-specific course portal (e.g. AcqNow or ATRRS)
        i. Design of Experiments
            1. Science of Test (SOT) 210: 2 days (for management)
            2. SOT 310: 5 days (for practitioners)
            3. SOT 410: 5 days (for practitioners)
        ii. Reliability
            1. Reliability (REL) 210: 5 days (for management)
            2. REL 310: 5 days (for practitioners)
            3. REL 410: 5 days (for practitioners)
    b. Defense Acquisition University
        i. Statistics: CLE-035 (online)
        ii. Reliability: CLE-301 (Reliability and Maintainability)
2. General websites
    a. STAT COE Website https://www.afit.edu/stat
    b. NIST Engineering Statistics Handbook https://www.itl.nist.gov/div898/handbook
    c. Test Science Knowledge Center https://testscience.org
3. Statistics:
    a. Stat Trek: https://stattrek.com
    b. Online statistics book: http://onlinestatbook.com/2
    c. Statistics textbook: http://www.statsoft.com/Textbook
    d. Which statistical test to use: https://stats.idre.ucla.edu/other/mult-pkg/whatstat
    e. Common mistakes: https://www.ma.utexas.edu/users/mks/statmistakes/TOC.html
4. Confidence Intervals
    a. 10 things to know: https://measuringu.com/ci-10things
    b. Video: https://www.youtube.com/playlist?list=PLvxOuBpazmsMdPBRxBTvwLv5Lhuk0tuXh
5. Sampling Plans/OC Curves:
    a. NIST Engineering Statistics Handbook chapter 6.2.3.2 https://www.itl.nist.gov/div898/handbook
    b. SamplingPlans.com: Modern Sampling Plans http://www.samplingplans.com/modern3.htm
6. Bayesian techniques

   a. Textbook: Gelman, Andrew et al. (2004) Bayesian Data Analysis (2nd ed), New York, New York: Chapman & Hall/CRC.
   b. Analyticsvidhya.com: Overview https://www.analyticsvidhya.com/blog/2016/06/bayesian-statistics-beginners-simple-english/
   c. JohnCook.com: Conjugate priors https://www.johndcook.com/CompendiumOfConjugatePriors.pdf
   d. Duke University: Course Notes http://www2.stat.duke.edu/~rcs46/lectures_2015/14-bayes1/14-bayes1.pdf
7. Design of Experiments
   a. Textbook: Montgomery, D.C. (2017) Design and Analysis of Experiments (9th ed.), Hoboken, New Jersey: John Wiley & Sons.
   b. DOE Overview (ASQ) https://asq.org/learn-about-quality/data-collection-analysis-tools/overview/design-of-experiments-tutorial.html
   c. DOE Overview https://www.moresteam.com/toolbox/design-of-experiments.cfm
   d. DOE Primer https://www.isixsigma.com/tools-templates/design-of-experiments-doe/design-experiments-%e2%90%93-primer
   e. DOE Basics https://www.youtube.com/watch?v=tZWAYbKYVjM
   f. Stu Hunter DOE videos https://www.youtube.com/watch_videos?video_ids=NoVlRAq0Uxs,hTviHGsl5ag,LvPWKyLTJZY,33B_fIUQJe8,6l1mZrxPUtc,jFrtzMlKsnk,AVUAt0Qly60,4hSQLqVAXT0,lyKVsd1Rda8,ttkAlcSdmuQ,O-q4af9jXR0,yQ2ONor-jdM,erEcsTE_rbs,i9ea5kawiM0,U4EhjbRbWSw,62ixqGad80o,eC0oP9zH8V8,qFdsEYRGb6Y,NKgUPxb9-iw,etIJutEwgoo,ic8wuPu6t18,cM-nlO1-tvQ,k3n9iSB6Cns,3fwoU16MHJM,F05zZL3uyRo,pAx5_uLcANA,lmMUaaHQD7U,VUw0gGK05I0,snqXKVU1ug8,MsKa59SqOrs
8. Covering arrays/Combinatorics https://csrc.nist.gov/projects/automated-combinatorial-testing-for-software
9. Reliability/Reliability Growth
   a. Textbook: Meeker, William and Escobar, Luis. (1998) Statistical Methods for Reliability Data, John Wiley & Sons, Inc: New York, New York.
   b. ReliaWiki (basic reliability growth overview) http://reliawiki.org/index.php/Reliability_Growth_Planning
   c. Tools and calculators
      i. Army Combat Capabilities Development Command (CCDC) Data & Analysis Center: Reliability Tools https://www.dac.ccdc.army.mil/CRG_Tools.html
      ii. Reliability Analytics Toolkit (tools for a variety of methods) https://reliabilityanalyticstoolkit.appspot.com

Appendix F STAT COE Best Practices

The STAT COE regularly produces best practices on different statistical procedures. The goal of each best practice is to demonstrate a practical application of a procedure that allows for repeatability. They serve as useful learning tools for increasing rigor and improving STAT procedures throughout various testing stages. The most current best practices can be found on our website: https://www.afit.edu/STAT/statdocs.cfm