Critical STAT Questions DoD Leadership Should Ask:

Injecting Rigor into the Test and Evaluation Process through Effective Inquiry

Authored by: Jennifer Kensler, PhD
Francisco Ortiz, PhD
Lenny Truett, PhD
Jim Cordeiro, PhD
Luis Cortes
Michael Harman
18 March 2015

Revised 7 September 2018

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

Revision 1, 7 Sep 2018: Formatting and minor typographical/grammatical edits.

Executive Summary

The Deputy Assistant Secretary of Defense (Developmental Test and Evaluation) (DASD(DT&E)) and Director Operational Test and Evaluation (DOT&E) have advocated the use of more rigorous test methods to ensure risk is identified and quantified as early as possible. Scientific test and analysis techniques (STAT) are the scientific and statistical methods and processes which enable the development of efficient, rigorous test strategies that will effectively quantify and characterize system performance and provide information that reduces risk and informs decisions. This paper focuses on the leadership aspect of the planning process; playing the quality assurance role in identifying if the proposed testing will be sufficient. Critical leadership questions are proposed along with acceptable response content.

Keywords: STAT, scientific test and analysis techniques, leadership, test and evaluation, critical questions, test planning, test readiness review

Introduction

The Deputy Assistant Secretary of Defense (Developmental Test and Evaluation) (DASD(DT&E)) and Director Operational Test and Evaluation (DOT&E) have advocated the use of more rigorous test methods to ensure risk is identified and quantified as early as possible. Scientific test and analysis techniques (STAT) are the scientific and statistical methods and processes which enable the development of efficient, rigorous test strategies that will yield defensible results. STAT encompasses such techniques as design of experiments, observational studies, reliability growth, and survey design. The suitability of each method is determined by the specific objective(s) of the test. These techniques can be applied to a broad array of missions, systems, and functions. Despite this wide applicability, some testing does not lend itself to the application of STAT. This is typically due to existing standards or safety reasons which dictate how a test must be performed. Test planners should make every effort to apply STAT until a reason to deviate is justified.

This paper focuses on the leadership quality assurance role of the STAT planning process. It outlines the critical questions that those in a leadership position should be asking. These questions are appropriate at Test Plan Working Groups (TPWG), Working Integrated Product Team (WIPT) meetings, Test Readiness Reviews (TRRs), and milestone reviews. Their purpose is to help the leader understand if the proposed testing will be sufficient and ensure program risk is quantified and minimized.

Note: Practitioners requiring more technical process details should refer to the ***Guide to Developing an Effective STAT Test Strategy*** available at https://www.afit.edu/stat/statcoe_files .

Questions to Ask

The questions that follow are broken up by phase (refer to Figure 1) to make it easier to understand why they are being asked. Sub-bullets below each question outline acceptable response content. The phases should be addressed in order because effective planning at one stage sets up the next stage for good results. Note that there are some questions that pertain to planning before any testing is conducted and some that are relevant only after data has been collected. These pre- and post-testing questions are identified.

This is not an exhaustive list but a reference tool to assist in meeting preparation. More detailed technical questions and process details can be found in the ***Guide to Developing an Effective STAT Test Strategy.***

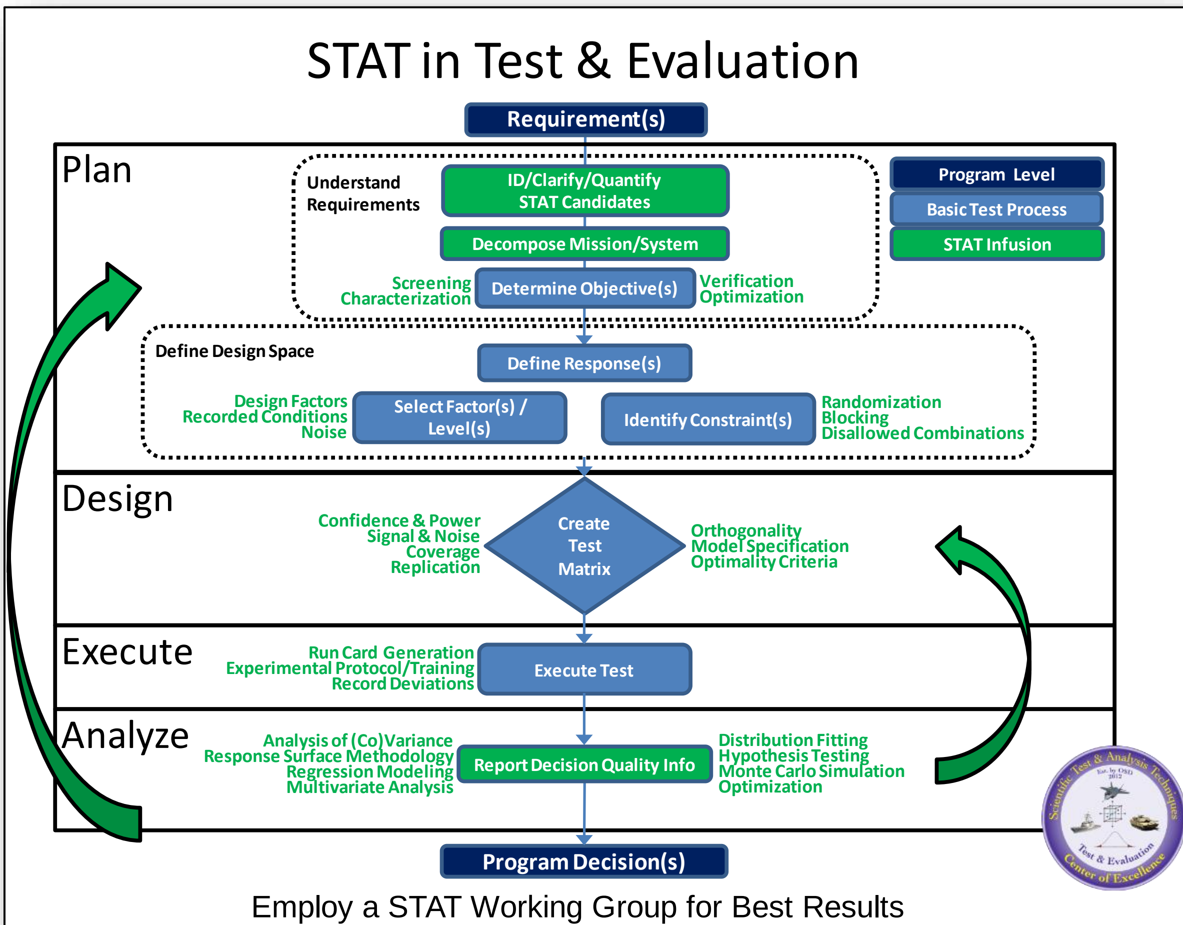

Figure 1: STAT Process for Test Planning

STAT Differences in Developmental (DT) and Operational Testing (OT)

The application of STAT will differ in DT and OT for various reasons. In early DT the subsystems may be tested individually before executing fully integrated systems testing and the test designs will reflect this. Later in DT the full-up system testing may resemble more operationally realistic conditions as should the test design. In general, OT is the most difficult realm to apply STAT due to the potential lack of control of many factors. Despite these differences, the STAT planning principles apply in all test phases: determine appropriate objectives, define the response and factors, create a design for the required data, execute the testing in the proper manner, and analyze the results to address the requirement. Moreover, a well integrated test team (ITT) will ensure that early testing will support later testing. Test planners should always strive to collect data that relates to operational goals so results can be rigorously compared, risk can be assessed, and the test readiness determination is well informed.

Strategy/Planning

- **What STAT resources have you consulted to plan your strategy?**
    - Leverage support from the STAT COE.
    - Utilize local STAT/DOE Subject Matter Expert (SME).
    - STAT/DOE class/trainer consultation or notes.
    - STAT/DOE knowledge base/website.
    - Lessons learned from previous versions or similar systems

- **How has STAT planning been incorporated into your test team?**
    - STAT is a component of the Integrated Test Team and Test Integrated Production Team.
    - The STAT section was developed through an integrated process, not a separate working group.
    - Both DT & OT organizations address STAT considerations and the ITT approves the approach and documents this in the TEMP test strategy.
    - TEMP STAT plans are translated into the system level test planning activities by the responsible test organization.

- **Are the test objectives clearly articulated?**
    - Note: Clear objectives are fundamental to the planning process and facilitate good design. Objectives should be specific and define desired outcomes.
    - Example objectives:
        - Screening: a low resolution design focused on identifying main effects
        - Characterization: a high resolution design with low prediction variance
        - Software factor coverage: a combinatorial design covering all triplets

- **What is your strategy to ensure early testing informs later testing?**
    - Operational testers are part of the planning team from the beginning.
    - Follow-on testing will employ the same designs and/or responses and be used to validate previous findings.
    - The empirical predictive models will allow later testing to validate earlier results.

- Earlier testing with simulations, labs, and hardware in the loop (HWIL) will cover and control factors unavailable in later real-world testing.
- A data sharing process flow map shows what data is being collected throughout testing and how it is being used.

- **What process was used to determine which test requirements need STAT for assessment?**
  - Working group meetings employ SMEs from all relevant disciplines.
  - All test requirements are reviewed by the group and STAT requirements are identified based on the test objective at each particular phase.

- **How are constraints (inability to randomize, lack of factor control, resource or range limitations, system limitations, etc) limiting our ability to gather a statistically relevant size data set?**
  - Early testing (lab, HWIL, simulation) will cover the full operating region and factors.
  - Present a comparison of the cost ($ and number of points) required for a relevant data set against the proposed plan.

- **Can the test labs/ranges facilitate the fundamentals of STAT/DOE (randomization, replication, and blocking)?**
  - Ranges allow randomization without negatively impacting resources or time to test.
  - Range personnel understand DOE concepts including why points are tested in a particular order.
  - Data collection, reduction, and analysis can be done in a timely manner to facilitate screening designs and sequential testing.

- **For the non-STAT designs what determined the test point selection?**
  - Note: Previous designs are not necessarily a good reference (What has changed? Was it a robust design?).
  - Accepted standards are referenced (e.g. Industry, Mil-Spec, SysCom procedures).

- **How was reliability test time determined?**
  - Statistical software should be used; `rules of thumb' should be avoided (e.g., "three-times" rule is insufficient).
  - Operational characteristic (OC) curves should show a low probability of acceptance (Pa) at the metric (e.g. MTBF) threshold value and high Pa at the objective value.
  - Reliability growth curve starting point should be based on data from a similar system.
  - Reliability will not grow unless improvement insertion points are planned during development.
  - Software tool assumptions should be realistic (e.g. fix effectiveness should not be >70%).

- **Has the system (or functions) been decomposed to facilitate lower level designs?**
  - Note: Single designs that cover the entire mission are typically too broad.
  - Many smaller designs (instead of one large design) can more effectively address requirements throughout a mission.
  - Designs should focus on smaller components or functions.

  - The designs are usually the right size when the responses and factors are clear, preferably continuous, and operationally meaningful.

Design Development

- **Will the designs adequately explore the expected region of operability? If not, what regions will we not be able to assess performance? Why?**
  - M&S should cover the entire region of operability, boundary to boundary.
  - Test range constraints could limit certain operating regions. The program needs to identify, quantify, and mitigate the risk of not examining these scenarios.

- **Did you minimize the use of categorical factors and responses?**
  - NOTE: The data type chosen to represent factors and responses can have a major effect on the resources needed to conduct an experiment and the quality of its respective analysis. Continuous data types (e.g. mile per hour vs. slow/fast) are preferable when possible.
  - The reason for any categorical data types left in the design should be discussed.
  - Responses must be objective and measureable and not require a value judgment like “satisfaction” or “success.”

- **How is test risk quantified?**
  - Lot sampling plans: Number of trials and acceptable failures can be plotted as operational characteristic (OC) curves. The accepted curve should pass through a minimum 80% probability of acceptance at objective and 20% (or less) probability of acceptance at threshold.
  - Characterization testing: Minimum 80% power and confidence goals should be used to size the design of experiments.
  - Simulations: Space filling designs should report how the space was covered (design type and density).
  - Combinatorial (software) testing: Designs should have at least 2-way (all pairs) coverage. Three way (triplet) or higher coverage will reduce risk further.

- **How would the design need to change to reduce this risk?**
  - Note: This risk may be driven by test constraints (e.g. range limitations).
  - Provide a comparison of the existing design with a “perfect” design and analyze the difference in the number of points (and associated cost).
  - For reliability: estimate the additional operating time (after testing completes) needed to generate a sufficiently accurate estimate.

Pre-Execution Planning

- **What STAT-related training has the test team received?**
  - All testers should receive STAT familiarization training. This can be provided by the test designers or STAT subject matter experts. This provides awareness as to why events should be run in the planned (randomized) order.
  - Testers should be aware that any deviations from the test protocol must be recorded.

  - Testers must understand that the design recorded in the TEMP (especially if it is approved) must not be changed without test team concurrence.

Analysis Planning (before any data is collected)

- **What analysis do you have planned and how will it address the requirement?**
  - Traceability from requirements to design must be clear.
  - Selected analytical methods should directly inform acquisition decisions.
  - Multiple analytical methods (e.g., simulation, logistic regression, parametric analysis, etc.) may be required to fully address the requirement and quantify performance and risk across the space. Analysis of variance and an empirical model alone may not suffice to address the requirement as stated.

- **How will the analysis method help predict performance for points not actually tested?**
  - If a prediction is required the design should be capable of producing a predictive model with acceptable accuracy across the factor space.
  - Use of continuous factors will enable prediction between tested points (important in OT when conditions are less controllable).

Post-Execution Review

- **Were there restrictions in running design points randomly? How was it handled?**
  - Analysis should be based on how the data was <u>actually</u> collected, not the way it <u>should</u> have been collected.
  - Reconfiguring the run order to be “most efficient” may create statistical analysis issues. If this is the case then the actual run order and the collected data should be thoroughly analyzed by a highly competent DOE practitioner to ensure the statistical data validity before concluding testing.

Analysis Review (after data has been collected)

- **Can you describe the analysis methods used to come to your conclusions?**
  - The analysis should address the requirement.
  - Multiple analytical methods should complement each other or act to uncover risk or validate other conclusions.
  - Any changes from the original design and analysis plan should be detailed as to how they impacted the assessment and conclusions.
  - Validation of the raw data should be accomplished whenever possible.

Conclusion

DoD leadership have advocated the use of more rigorous test methods to ensure risk is identified and quantified as early as possible as determined by the specific objective(s) of the test. Test planners should make every effort to apply STAT until a reason to deviate is justified. STAT must be an integral part of the test planning process and the active involvement of program leadership enforces effective implementation.