STAT COE-Report-06-2013

Controlling Conditions is Key for Meaningful Operational Test Results

Authored by: Michael Harman

4 December 2013

Revised on: 13 November 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

Rev 1, 13 Nov 2018: minor editing and format changes

Executive Summary

Operational test planners struggle to balance realism and test control. Recent mandates for the use of design of experiments (DOE) in operational testing specifically call for increased rigor, requiring detailed test planning to specify all relevant performance measures, the input factors to be changed that allow for the different test conditions, quantifying the risks of incorrect conclusions, and assuring that a comprehensive set of operating conditions are covered. This rigor is intended to improve the quality of operational test results that are collected under realistic conditions. Unfortunately, the requirement for operational realism is too often seen as mutually exclusive to, or in competition with, the rigor and in-depth planning of DOE. Leveraging a systematic planning approach and exerting appropriate control can assist the planner in meeting both rigor and operational realism goals. This paper discusses the need for rigor, the keys to factor determination, and creative methods to maintain the necessary realism.

Keywords: Operational test, operational realism, design of experiments, factors, conditions

Introduction

Operational testing (OT) is a difficult undertaking for many reasons. The OT planner has to ensure both effective data collection for knowledge discovery, and operational realism. Test data is most useful for analysis and reflective of true system performance if the test is planned using a methodical and controlled approach. In control, the goal is to plan a test such that factors or input conditions are purposely set for a test event, and then systematically changed in subsequent tests in order to observe and quantify potential relationships between these inputs and system performance responses. The myriad operating locations, configurations, threats, tactics, and geometries present an overwhelming number of combinations to the planner. However, a systematic approach can be employed to break down the process into manageable steps and specific actions. The scientific method has been adapted to test planning as shown in Table 1.

Table 1: The Guidelines for Designing a Test

1. Define the test objectives
2. List the response variables
3. Identify contributing factors
4. Specify the degree of control for each factor
5. Design the test space
6. Develop the execution plan
7. Author the analysis plan
8. Report on the results

This sequence describes the design of experiments planning methodology (Coleman and Montgomery, 1993) and adds structure that can further assist the test planner. Unfortunately, the requirement for operationally realistic OT is too often seen as in conflict with the rigor provided by DOE. Testers may conclude that operational testing is best executed by letting a scenario unfold uninhibited. This position implies that any level of control detracts from the realism. However, operational realism is defined by authentic conditions and operationally representative equipment and operators. So, how does control impact operational realism and what degree of factor control should be exerted?

Background

Operational Testing Definition

Operational Testing has the following general guidelines as defined by the Defense Acquisition Guidebook (DAG, United States Department of Defense, 2013):

- Typical users operate and maintain the system under test conditions simulating combat and peacetime operations;
- Operation test and evaluation uses threat or threat representative forces, targets, and threat countermeasures;
- Testing should use production representative systems.

No assertion is made that operational testing conditions cannot be controlled. However in some minds, *realistic* conditions have been misinterpreted as *uncontrolled* conditions with the rationale that real life events unfold in that manner. Consider the following definitions.

Test is defined as:

- A critical examination, observation, or evaluation: specifically the procedure of submitting a statement to such *conditions* or operations as will lead to its proof or disproof or to its acceptance or rejection (Merriam and Webster, 2013).

Experiment is defined as:

- An operation or procedure carried out under *controlled conditions* in order to discover an unknown effect or law, to test or establish a hypothesis, or to illustrate a known law (Merriam and Webster, 2013).

Both definitions mention conditions that directly influence the outcomes. The definition of experiment adds language about controlling these conditions.

Design of experiments is defined as

- A test or series of tests in which purposeful changes are made to the input variables of a process or system so that we may observe and identify the reasons for changes that may be observed in the output response (Montgomery, 2010).

This design of experiments definition mirrors the other two but also suggests that pursuing meaningful information requires *purposeful* factor *changes*, what is termed in this paper as factor *control*. This definition is included specifically in light of the DoD mandate (Gilmore, 2010) for more test rigor through the use of design of experiments.

Why Exert Factor Control?

Ostensibly there are two decisions when it comes to factors of interest: control or do not control. For some factors a choice may not exist as in the case of weather or resource availability. As will be discussed later via example, factors can be ultimately separated into four general categories: control, covariate, constant, and noise. Control factors are purposely varied, covariate factors are allowed to take on values without intervention but are recorded, constant factors are held fixed at a particular level, and noise factors are essentially ignored (Figure 1). The more important decision rests with the factors of priority, control and observe.

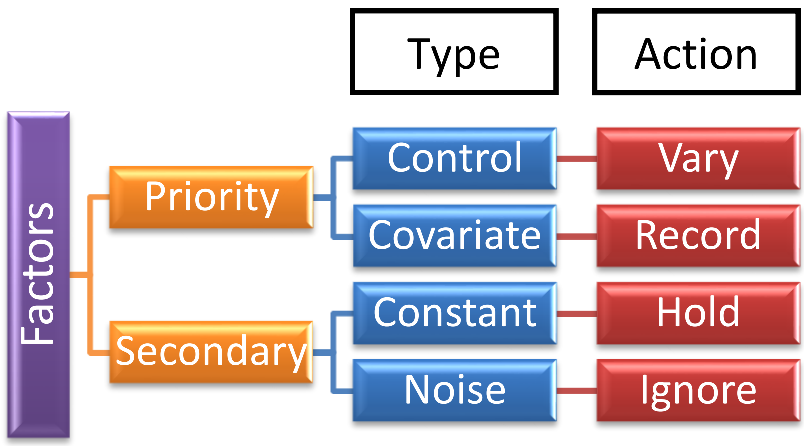

Figure 1: Factor Types and Decisions to be Made for Each Type

If at least one factor is control, the test is considered a designed experiment. The more factors controlled, the more useful the data and the more knowledge gained in terms of factors affecting responses. If no factors are controlled, the test is considered an observational study. Observational data comes from experiments where the explanatory or predictor variables of interest are not controlled. The major limitation of these studies is that they often do not provide adequate information about cause-and-effect relationships (Neter et al., 1996). Observational studies can ensure accurate, complete, and reliable data but these studies often provide very limited information about specific relationships among the data (Montgomery, Peck and Vining, 2003). Observational studies are common in the medical research community where controlled experiments are unrealistic and/or unethical. Cochran and Chambers (1965) discuss the difficulties associated with observational studies by first comparing the process to designed experiments. If practical, the preference would be to exert control over the experimental factors.

The Need for Rigor: Factor Type Choices Affect Response Clarity

The factor labeling decisions are often driven by test execution restrictions and resource limitations, but many times happen due to insufficient planning or a lack of realizing the implications of the labeling decisions. Most responses are affected by a number of factors and some factors contribute more than others. Assuming the operational environment allows for factors to be purposely varied, control factors provide the most insight into drivers of system performance. If factors are allowed to vary and are recorded (covariates), insight can also be gained, but the chances increase greatly that changes to these factors coincide with changes in other factors, preventing determination of any causal influence. Factors fixed (constant) limit the scope of the test and possible conclusions that could be drawn on any role those factors play. Those noise factors allowed to vary (known or unknown) without attention cause the most harm by increasing the variability of the response such that changes in control or covariate factors cannot be otherwise assessed. But, does controlling factors in OT come at the cost of operational realism?

What is Operational Realism?

Operational realism is a much discussed term and an important aspect of OT. It refers to the system configuration, conditions, and presentation of events (DAG, Section 9.5.8.1, 2013). United States code defines the term "operational test" as "the field test, under realistic combat conditions, of any item of weapons, equipment or munitions for the purpose of determining their effectiveness and suitability" (Title 10, Section 139, 2013). Service-specific guidance also typically requires that the test articles be as production-representative as possible. It is essential that operational realism be maintained, which is why our military operators are essential not only in execution, but in planning.

If you were to ride along with an operational unit for some period of time and experience a "day in the life" of that unit then you would witness (a portion of) what is operationally realistic for them. Both the conditions and the "unknown" nature of future events are parts of the operational experience.

As the observer (or tester) you may look back on a particular event to examine the outcome. You may believe that one particular condition exhibited more of an effect on the outcome than another but, as a single event, you may not be able to conclude that with any real conviction. You would have data points for the events you observed but you may have missed certain conditions of interest during your observation. Perhaps they were operating in the summer in the mountains with a certain opposing force. That may not be sufficient to answer your curiosity about their performance during the winter in a forest terrain against another force. Or, what if they had a new piece of equipment that you were interested in evaluating? If the conditions for employing that equipment never presented themselves then you would leave without any data because you were unable to create and/or control the required conditions.

Similarly, conducting OT strictly as an observational study limits the depth, breadth and quality of data collection and analysis. To the contrary, OT needs to include systematic data collection under realistic controlled conditions in order to effectively assess the significance of the factors on the response to obtain the proper information required to support acquisition decisions.

Operational Scenarios and Factor Control

Operational testing is typically accomplished through the execution of mission scenarios. These scenarios attempt to replicate complete missions from beginning to end. Mission scenarios help to ensure that information, conditions, system configuration, and the order of events most closely replicate the operational environment. Because of this, operational mission scenarios may contain many potential paths (typically called threads) and subsequently include myriad factors. Assuming the planner wants to incorporate factor control into his design, the seemingly high number of potential factors may confuse the process of defining them.

Decomposing the mission into subtasks is a good way to take complex scenarios and make them understandable. Remember that factors are associated with specific responses, not the whole mission. Also, requirements may align closely with certain tasks in a mission. Understanding which tasks these align with allows a clearer understanding of where to apply design of experiments and factor control.

As a simple example, consider a ground mission to position a laser on a target for use by another weapon delivery system with specific delivery accuracy limits. The process by which the operator plans his mission, gathers pertinent data, moves to the site, illuminates the target, and safely returns comprises the whole mission. But the requirement only applies to the laser spot. While the whole mission scenario is important, the mission segment “illuminate target” is where the responses (“time to target threat” and “laser spot error”) will be measured. It is for this response that the factor list will be generated. Elements like target data and the path to the targeting location may be dependent on the factors under control or may be completely independent. Either way, full mission scenarios can be generated around the factors and levels under control. This process of decomposition becomes more vital as the mission complexity increases and/or the number of responses increases. Figure 2 shows an example of the targeting mission after decomposition.

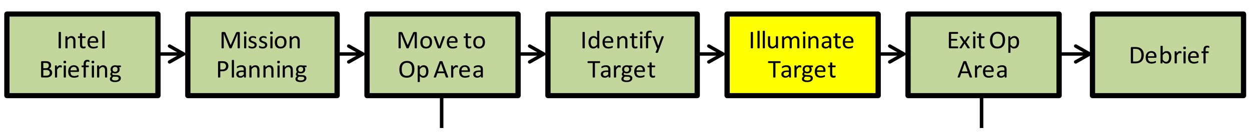

Figure 2: Sample Mission Decomposition with Highlighted Segment

After decomposing the mission and defining response variables, factor brainstorming can begin. So what is the method and process to list factors and define the level of factor control?

Method

Overview

The process of discerning the relevant factors and deciding the role each will play takes careful consideration and requires a team effort from all the test team contributors: management, engineering, operators, and analysts. Each element of the entire 8-step process for planning, executing and reporting an operational test (Table 1) is essential. For instance, both defining the objective (Step 1) and listing the response variables (Step 2) are success critical and influence the direction of the design. A well thought out and consensus objective and clearly defined responses facilitate clear factor definitions. This discussion will primarily focus on Steps 3 and 4.

Defining contributing factors (Step 3) is the process of listing everything that influences the response. This list should include potentially uncontrollable factors like weather and natural events (e.g. sunlight). Specifying the degree of control for each factor (Step 4) is an often iterative process that considers realities like testability, complexity, and cost in order to arrive at the final factor list.

Listing Factors

Factors may be part of the operating environment (e.g. weather), system design (e.g. speed), or configuration (e.g. choices of tactical links), among others. The first step in factor determination is for the test team to brainstorm a list using historical knowledge, technical sources and/or subject matter expertise. One of several methods is to employ a cause-and-effect, or fishbone diagram to elicit ideas on how factors influence the response. The categories are:

- Measurement
- Manpower
- Materials
- Mother Nature
- Methods
- Machines

List as many factors as possible, and keep in mind that this is not the time to judge whether the factor will be controlled. Keep the factors clear and defined to enable assignment of levels later in the process. Ambiguous factors or levels may result in noisy responses. Also, every effort should be made to ensure factors are continuous (e.g. all values between 1 and 10) instead of categorical (red, blue, green). Continuous factors often reduce the number of runs, provide the capability to mathematically model response values between points, and support a more informed and insightful analysis.

A Short Commentary on Continuous vs. Categorical Factors

Consider the factor "speed" with continuous levels (10, 30 mph) or categorical levels (slow: <15, fast: >15mph). Conducting the test in categorical speed bands may produce similar magnitude responses in the case that the two speeds are close in value (e.g. 14 and 16) or dramatically different responses when the levels are far apart (1 and 40). In both cases, the speed bands are "slow" and "fast" and the analysis will attempt to characterize the effect of speed on the response. In the case where the actual speed values differed very little (14 and 16) one can imagine that little difference will be sensed in the response. If speed truly impacted the response then it may not be sensed in this test and incorrectly deemed insignificant. Whereas, the wider levels (1 and 40) may result in large differences in the response magnitude and correctly lead to the determination that speed is a significant factor. So, allowing the speed values to vary uncontrolled inside "slow" and "fast" bands may produce inconclusive results. Therefore, using and controlling continuous factors with discreet levels (e.g., 10 and 30) is a best practice, where applicable and practical.

Figure 3 shows an example of a completed fishbone diagram. In this case, the response is "time to target threat." Assume this is a ground unit tasked with laser targeting of an opposing vehicle for attack by another friendly unit. Note that “laser spot error” (or other responses) may have many of the same factors. Similarities like this may reduce the time to generate a factor list for other responses.

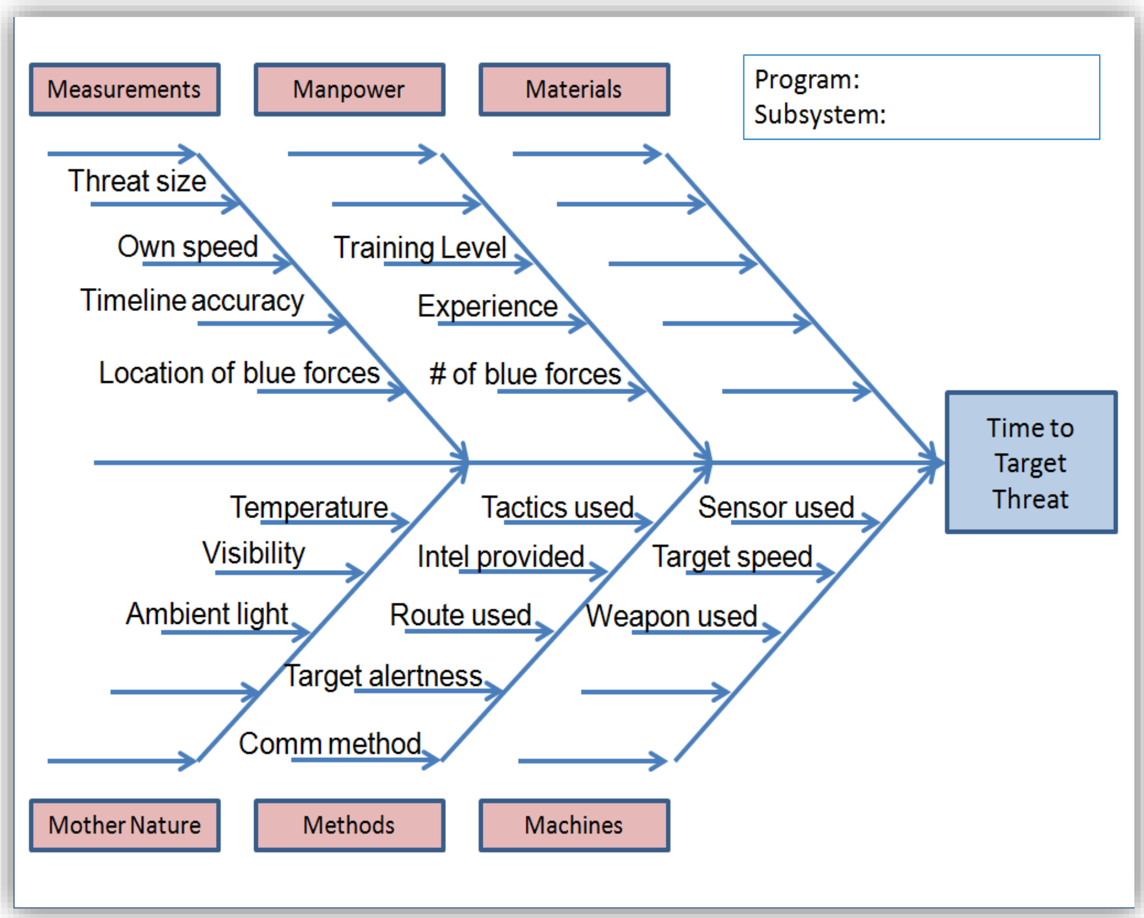

Figure 3: Fishbone Diagram for Laser Targeting an Opposing Vehicle.

Further Factor Evaluation: The Preliminary Factor List

The fishbone diagram is a method to generate a list, but these factors must now be further evaluated for interest, ease of control, and level range.

Factor interest is simply identifying which factors the test team is interested in learning about through the design. This interest may be due to changes in installed technology, particular tactics involved, or threat attributes, among others. Factors of interest should be controlled. Comments captured during the evaluation process can help document the level of interest. For the laser targeting example, a preliminary factor list (

Table 2) is provided that addresses each fishbone factor, the factor level ranges, estimates of ease of control, and relevant comments.

Table 2: Example Preliminary Factor List for Laser Targeting

Factor	Range	Ease of Control	Comment
Threat size (sq ft)	1-8	Easy (E)	Primary concern for response
Own speed (kts)	80-150	E	Must vary with tactics used
Timeline accuracy (min)	1-30	E	May be tied to intel/other factors
Location of blue forces (yds)	50-150	E	Key for timeliness and decisions
Training level	1, 2, 3	Hard (H)	Hard to scope and control
Experience (yrs)	1-10	Very Hard (VH)	Test team is already defined
# of blue forces	0-3	H	Critical for accuracy and decisions
Tactics used	Overt, covert	E	Contribute to time and accuracy, not defined easily
Intel provided	Specific, vague	E	Contains many details which may vary widely
Route used	Direct, indirect	H	Contributes to time, varies with tactics used
Target alertness	Alert, passive	E	May be linked to target size and speed
Communication method	Active, passive	E	Linked to tactics and intel
Sensor used	Visual, radar, EO, IR	E	Weapon and tactics specific
Target speed (kts)	0-25	E	Key contributor to time
Weapon used	A, B, C	H	Impacts crew workload
Ambient light (lux)	0-120K	VH	Conduct day and night ops
Visibility (sm)	0.5-10	VH	Key for certain sensors
Temperature (F)	0-120	VH	Key for certain weapons, sensors

Ease-of-control indicates how much facility the team will have to exert control during testing. Choices in labeling are easy (E), hard (H) meaning limited ability/opportunities to control, or very hard (VH), which will drive test design, cost, and/or scheduling. Ease-of-control values will be used to determine the factor type in a subsequent iteration.

Levels are the factor values or settings to be controlled during testing. At this point, defining a range helps the team decide where their specific interest lies in the operating envelope. It also can help identify poorly described factors. When a factor level cannot be described it may indicate an ambiguity requiring further refinement. Factor levels can be set at the extremes of the operating envelope but more realistically should reflect the edges of the expected operating region for the given response. Also, the range should not be so small that the natural variation in the response would overwhelm any effect caused by the factor. Arriving at the number of levels (typically 2-5) and values is an iterative process as the design type (Step 5) and analysis requirements (Step 7) are reviewed throughout the 8-step design process. Less levels, particularly for categorical factors, results in more efficient testing, so deliberation here is warranted.

The Final Factor List

Considering interest and ease-of-control assessments, the team must then determine which factors should be controlled and purposely varied. The goal should be to control as many factors as practical, without influencing operational realism or test execution efficiencies. Some factors are not practically controlled (e.g. winds), but can be recorded. The remaining uncontrolled, unrecorded factors will contribute to variation (noise) in the response.

Control Factors

Control means that the factor level will be set at a prescribed value on a given event. Across the complete design, these levels will be systematically varied to enable a determination of their effect on the response. The number of controlled factors/levels is a major contributor to determining the size of the design and the number of runs. Controlled factors are labeled “V” for vary.

Constant Factors

Holding a factor constant is a level of control however the value will not be varied between events. This is used to reduce noise without adding additional runs into the design. These are labeled “H” for hold.

Covariate Factors

Recorded factors are uncontrolled but of some interest to the team. Weather data is typical of this category. These factors contribute to noise in the response although they will be available for some level of analysis. These factor effects will be confounded with other factors, but sometimes can be effectively statistically modeled. They are labeled “R” for record.

Noise Factors

Noise factors are those that are not controlled, regardless of whether or not they are recorded. Once all controlled and recorded factors have been identified, any remaining factors are considered noise factors and are not included in the design. No factor of interest should be a noise factor. Noise factors are labeled “N.”

Table 3 is an example final factor list. The bold green factors are those that will become part of the test design. In this example there are 7 noise factors of the 18 total.

Table 3: Example Final Factor List

Factor	Range	Ease of Control	Vary, Hold, Record, Noise
Threat size (sq ft)	**1-8**	**Easy (E)**	**Vary (V)**
Own Speed (kts)	**80-150**	**E**	**V**
Timeline accuracy (min)	1-30	E	Noise (N)
Location of blue forces (yds)	**50-150**	**E**	**V**
Training level	1, 2, 3	Hard (H)	R (Record)
Experience (yrs)	1-10	Very Hard (VH)	N
# of blue forces	0-3	H	Hold (H) at 2
Tactics used	Overt, covert	E	N (allow crew to decide)
Intel provided	Yes, no	E	N
Route used	Direct, indirect	H	H at indirect
Target alertness	Alert, passive	E	H at alert
Communication method	**Active, passive**	**E**	**V**
Sensor used	**Visual, radar, EO, IR**	**E**	**V**
Target speed (kts)	**0-25**	**E**	**V**
Weapon used	**A, B, C**	**H**	**V**
Ambient light (lux)	**0-120K**	**VH**	**V**
Visibility (sm)	0.5-10	VH	R
Temperature (F)	0-120	VH	R

This final factor list becomes the foundation for the rest of the design process. Determining the type of design, the number of runs, and what the analysis will entail and support is derived from this list.

Maintaining Operational Realism: Knowledge of the Operator

We now return to an introductory comment that the requirement for operational realism is too often seen as mutually exclusive to, or in competition with, the rigor and in-depth planning of DOE. We have already explained why factor control is important and how to define the level of control. It is now necessary to discuss some methods to ensure or enhance realism. Beyond simply operating in realistic conditions, certain controlled factor details should be withheld from the operator so they are only revealed as the mission progresses. The fact that the conditions are controlled or manipulated is not an issue as long as they vary and unfold just as the operation typically does. This permits the operator to ***experience*** the events as if they unfolded naturally before him.

Certainly the operator must know some things like the planned mission start time, planned route, size of the opposing force, weapon loadouts, and current system status. But he will experience these things (and their changes) in the context of an operational mission experience. So how do we combine the requirement for experimental control with the realism of an operational mission?

The planner has already decided on his objectives, responses, and factors. However, when it comes time to execute the events, randomizing them and keeping the run order information away from the operator is a key element to realism. This will allow some control over the operator's choices. The initial tactical situation briefing is one opportunity to direct the operator. As per the test event conditions, the operator is instructed to plan a mission against a certain force (a factor), in a certain location (a factor). He does not need to know that you intend to have the opposing force ambush him earlier (a factor) than planned or that the force is different in size (a factor) than expected.

Also, if you want the operator to follow a certain route (without dictating that beforehand) the planner can change the conditions as they unfold. For instance, the best route might be on a paved road but you want to observe performance through the woods. The planner needs to force the operator through the woods without telling him to do so. This can be accomplished (and therefore controlled) by inserting "orders from higher authority" or telling them that their preferred route is fouled by other vehicles. Either way, the operator performs under the desired conditions in a realistic manner.

Summary and Conclusion

Operational testing requires both rigor and realism to be effective and meaningful. Controlling every factor in an operational test would generate precise results and yield the most cause-and-effect information available but this is impractical for many reasons. However, conducting operational testing without any factor control removes rigor, increases noise in the responses, and decreases the useful information gleaned from testing. More realistically, following the systematic approach outlined in this paper helps the tester to rigorously document the planning process and design tests to derive the most information from the available resources. Leveraging technical and operational knowledge of the system under test, the team can effectively decompose the mission into manageable segments that allow clear assignment of the desired responses. The fishbone diagram is an effective method to identify factors that contribute to these responses. The team must then review and revise their factor list to identify and prioritize the factors that are to be controlled in testing. Factor identification is fundamental to the creation of a solid and rigorous test design and helps ensure that test objectives are met. Lastly, during execution, creative methods of manipulating the factors can ensure that maximum realism is maintained for the operator.

References

Cochran, W. G., and Chambers, S. P. "*The Planning of Observational Studies of Human Populations*." Journal of the Royal Statistical Society, 1965, Blackwell Publishing.

Coleman, D. and Montgomery, D. *A Systematic Approach to Planning for a Designed Industrial Experiment*. American Statistical Association and ASQC, 1993.

Defense Acquisition Guidebook (2013), https://dag.dau.mil.

Gilmore, J. Michael. "*Guidance on the Use of Design of Experiments (DOE) in Operational Test and Evaluation*." DOT&E Memo, 2010.

Merriam, G. and C., and Webster N. "*The Merriam-Webster Dictionary*." www.merriam-webster.com, Encyclopedia Britannica, (2013).

Montgomery, D. C. *Design and Analysis of Experiments. 8th edition*, John Wiley &Sons, New York, 2013.

Montgomery, D. C., Peck E. A., Vining G. G. *Introduction to Linear Regression Analysis*, *3d ed*., John Wiley and Sons, New York, 2003.

Neter J., Kutner M., Nachtsheim C., Wasserman W. *Applied Linear Statistical Models, 4th ed*., McGraw-Hill, Chicago, IL., 1996.

Title 10, United States Code (2013), Role of US Armed Forces.