ITEA Journal 2015; 36: 121-138

Twenty-one Parameters for Rigorous, Robust, and Realistic Operational Testing

Richard A. Kass, Ph.D.
Senior Analyst, GaN Corporation
(d/b/a Geeks and Nerds®)

Parameters for designing rigorous, robust, and realistic operational tests are proposed. These parameters unite principles of Design of Experiments (DOE) with test architecture and test execution considerations.

Keywords: design of experiments, operational test rigor, robustness, and realism

Overview

This paper proposes 21 parameters for designing and executing rigorous, robust, and realistic operational tests (OT) within the larger context of implementing Design of Experiments (DOE) into testing. These 21 parameters supplement classical seven-step design and analysis approaches to DOE (Montgomery, 2011, 2012) by providing rationale and examples for implementing basic DOE concepts such as power and confidence, systematically varying variables, randomization, control, and factor resolution. These parameters also expand test design steps to include considerations for covering the operational envelope and operational realism. The terminology and examples derive from Army T&E, but these 21 parameters should have wide applicability.

Director Operational Test and Evaluation (DOT&E, October 2010; June 2013; and July 2013) has published guidance for implementing DOE methodology into OT. A key DOT&E statement *revises the goal* of OT.

> The goal of operational testing is not solely to verify that a threshold requirement has been met in a single or static set of conditions. I advocate the use of DOE to ensure that test programs (including integrated testing where appropriate) are able to determine the effect of factors on a comprehensive set of operational mission-focused and quantitative response variables. The determination of whether requirements have been met is also a test goal, but should be viewed as a subset of this larger and much more important goal. [DOT&E June 2013 op.cit. p.1, underline added]

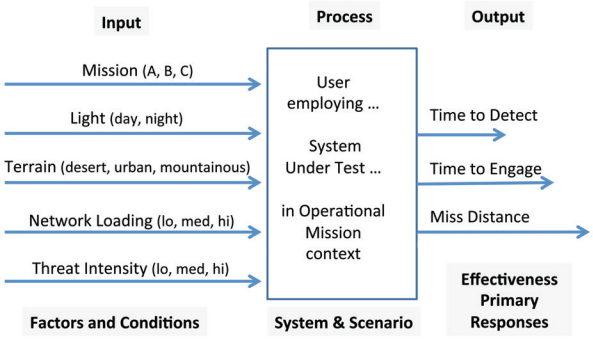

Figure 1: Process Map of Testing

This is a paradigm shift. Prior to this statement, Army OT focused on determining whether systems met requirements in an operational environment. Now, we will test to determine if and where system performance variations occur across battlefield conditions. Requirement verification in an operational environment is still essential but now a secondary goal. Characterizing performance across operational battlespaces is the superseding goal that drives test scope and provides sufficient data to verify requirements.

Figure 1 provides a process map1 of this shift in primary goals where OT is now designed to answer three questions:

- Are there variations in system output (peaks and valleys in system performance)?
- If so, what combination of battlefield factors and factor levels caused these changes?
- How confident are we in answers to these questions?

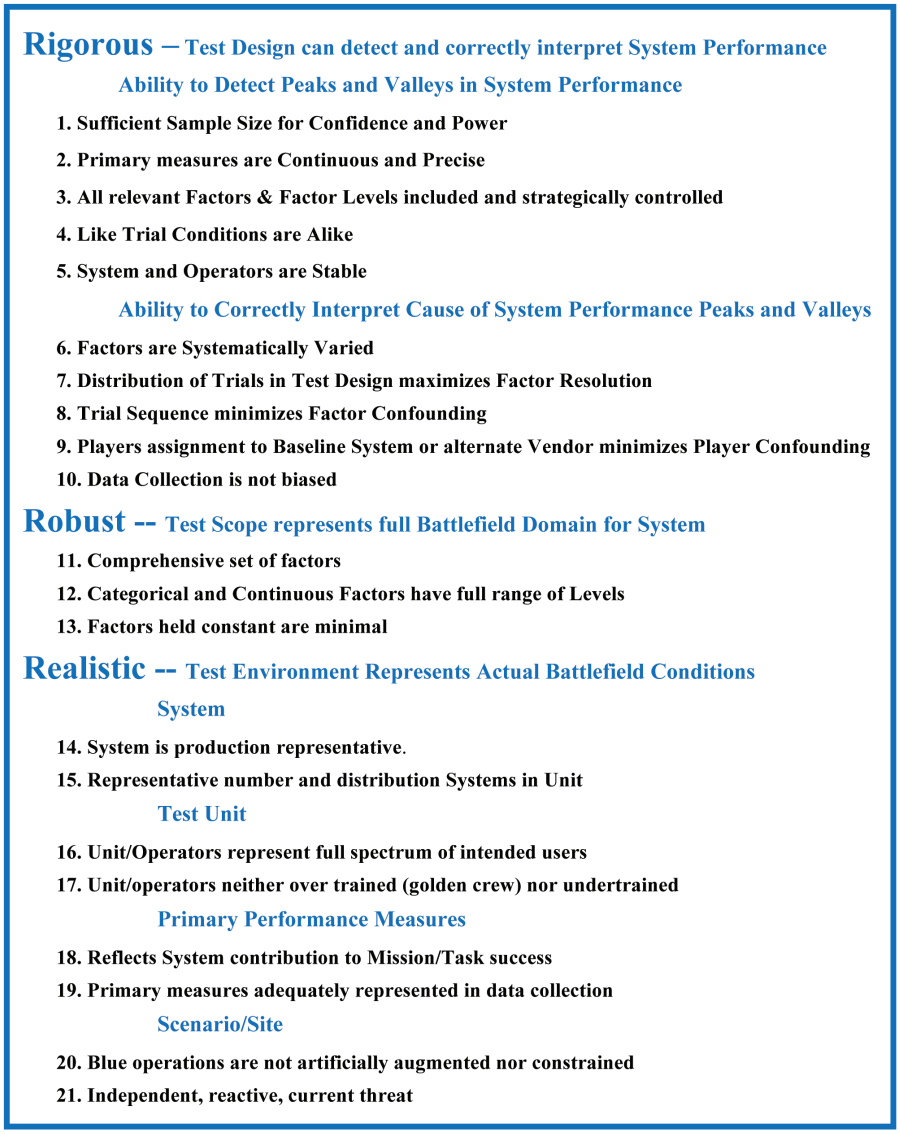

Figure 2: 21 Parameters for Rigorous, Robust, and Realistic OT

Montgomery (2012) provides statistical foundations for DOE and Hill et al. (2014) summarize key concepts and discuss misperceptions and challenges to implementing DOE into Department of Defense (DOD) testing. Additional implementation challenges involve synthesizing tenets of DOE with traditional operational test practices of data collection calibration, scenario development, and virtual augmentation. The following discussion proposes a logical framework of 21 parameters (Figure 2) to synthesize tenets of DOE with test range best practices to design and execute rigorous, robust, and realistic (R3) OT. Discussion of these parameters includes excerpts from DOT&E guidance on DOE implementation. We begin with a short, heuristic definition of each "R."

- *Rigorous* tests provide sufficient power, confidence, and resolution to detect and correctly interpret causes of peaks and valleys in system performance.
- *Robust* tests cover the full range of factors and factor levels expected in a system's operational employment envelope.
- *Realistic* tests include representative systems, operators, measures, and scenarios to reflect intended operational environments.

Our focus here is system effectiveness, not system suitability (reliability, availability, and maintainability). R3 considerations for system suitability are topics for follow-on discussions.

Table 1: Sample Size impact on Confidence and Power

Test Design Matrix Conditions			Sample Size required for different levels of confidence/power.		
			70%/70%	80%/80%	90%/90%
Attack	Low Threat	Day	1	2	4
		Night	2	3	5
	High Threat	Day	2	3	5
		Night	1	2	4
Defend	Low Threat	Day	1	2	4
		Night	2	3	5
	High Threat	Day	2	3	5
		Night	1	2	4
Total Test Trials = Total Sample Size =			12	20	36
Sample size calculation is approximate and based on S/N =1 with main effects and two-factor interactions in model.					

Test Rigor

Test rigor focuses on detecting systematic changes in system performance and correctly interpreting causes for these changes. Tested systems may exhibit performance changes, peaks and valleys[ii] in performance, due to changes in battlefield conditions. Ability to detect performance peaks and valleys is a signal-to-noise problem and cause-and-effect problem combined. True peaks and valleys are *signal effects* to be detected against various sources of random performance variations (*noise*). Test factors and conditions are potential *causes* of performance peak and valley *effects*. Rigor is concerned first with detecting performance differences, followed by concern for correctly identifying causes of these variations. Signal detection problems are discussed first.

Ability to Detect Peaks and Valleys in System Performance

Parameters 1-5 increase detection of performance signals. Parameter 1, sufficient sample size, is the most recognized technique. Parameters 2-5 serve to reduce test noise which indirectly enhances detection of performance signals, thereby reducing dependence on sample size. *Confidence* and *power* are test design metrics that quantify ability to detect performance signals from background noise.

1. Sufficient sample size is necessary to ensure confidence and power in test conclusions.

How much testing is enough; how many test trials are needed? Answers to these questions are a two-step process.

Step one is estimating power for a given confidence value and sample size. DOE software programs make this calculation easy. Power calculation is based on choosing a desired level of confidence and assumptions about tests' ability to detect desired performance differences (signal) from anticipated random performance variation (noise).

Table 1 provides sample size calculations for a notional test design with three factors with an anticipated

Table 2: Confidence and Power impact on Ability to Detect System Performance Variations

Test Objective Determine impact of factors …	**Conclusions from Test Data**	
	Confidence (typically 70% - 95%)	**Power** (typically 70% - 95%)
… on System performance	**Low:** *incorrectly detect false performance difference as battlefield conditions change*	**Low:** *inability to detect real performance differences as battlefield conditions change*
	High: *correctly detect no performance differences*	**High:** *correctly detects real performance differences*
… on performance differences between New System and legacy or baseline (BL) system	**Low:** *incorrectly detect false differences between System and BL*	**Low:** *inability to detect real differences between System and BL*
	High: *correctly detect no differences from BL*	**High:** *correctly detects real differences between System and BL*
… on performance differences between alternative Vendors	**Low:** *incorrectly detect false differences between Vendors*	**Low:** *inability to detect real differences between Vendors*
	High: *correctly detect no differences*	**High:** *correctly detects real differences*
Supplemental Test Objective … on whether System meets KPP or COIC requirement	**Low:** *incorrectly detects poor System meets requirement*	**Low:** *inability to detect good System meets requirement*
	High: *correctly detects poor System does not meet requirement*	**High:** *correctly detects good System* meets *requirement*

signal-to-noise (S/N) ratio of 1.0. S/N ratios compare the magnitude of factor effects on system performance to the magnitude of unaccounted remaining system variation (error). This comparison is represented in the numerator and denominator of the t-or F-statistics.

In this example, S/N=1.0 indicates our tester anticipates that differences between performance peaks and valleys will be about the same magnitude as random system variation. Tests with more anticipated random variation might require a smaller S/N, such as S/N=0.5. Smaller S/Ns reflecting more relative noise require larger sample sizes to achieve equivalent power and confidence.

Examination of the "80%/80%" column indicates 20 test trials will provide 80% confidence and 80% power in test conclusions. Individual rows indicate how 20 test trials might be distributed across eight combinations of test conditions.

Step two is determining what levels of confidence and power are acceptable for reporting credible results. To determine this, testers and decision makers consider consequences of low or high levels of confidence and power on possible test conclusions. Table 2 summarizes these consequences for four different test objectives[iii].

How high should confidence and power be? Table 1 illustrates that higher confidence and power require more test trials. There is no "one size fits all" answer. A case can be made that acquisition systems providing body and vehicle protection should be in the high 90s while other systems might be in the low 90s or 80s. Testers, working with decision makers, need to weigh risks of providing misinformation against incremental test costs of adding additional trials.

Should confidence and power be the same level? Academic journals focus on high statistical confidence (95% or 99%) in order to publish with statistical power is often at 60-70%. However in acquisition decisions, consequences of low power can be equally onerous as low confidence, as summarized in Table 2. Selection of power and confidence levels should include discussions of S/N assumptions, incremental test costs, and risks to interpreting test results for decisions emanating from testing.

2. Primary measures of performance are continuous and precise.

Primary Performance Measures are Continuous

Choice of collecting continuous or binary response measures greatly impacts test efficiency to achieve desired levels of confidence and power. Use of binary response measures, such as those in Table 3, measure performance signals more coarsely, thereby reducing S/N and increasing sample size requirements. Sample size values provided previously in Table 1 are based on a continuous response variable. If a binary variable had been considered instead, sample requirements to obtain 80% confidence and 80% power would have been closer to 110 instead of 20! Use of comparable continuous measures in place of binary measures yields substantial test size-cost reductions.

> **Metrics must be mission oriented, relevant, informative, and not rigidly adhere to the narrowest possible interpretation of definitions in requirements documents.**
> Another pitfall to avoid is relying on binary metrics as the primary response variable on which to base test design. I expect that even if the requirements defines a probability based metric, that great effort is extended to find a related continuous measure on which to base test design. We cannot afford tests based solely on the evaluation of probability-based metrics. [DOT&E, June 2013, op.cit. p.2]

But can we substitute continuous measures for binary ones? Aren't response measures determined by key performance parameters (KPP) and critical operational issue and criteria (COIC)? To reduce test costs, DOT&E guidance instructs Operational Test Agencies (OTAs) to substitute appropriate continuous measures for binary measures *even when system requirements are written as binary measures*.

Table 3: Some Binary and Related Continuous Measures

Binary-based Responses	Related Continuous-based Responses
Percent or Probability of target or IED detection (P_{det})	Range at detection, time to detect
Probability of hit (P_h) or kill (P_k)	Miss distance; Circular Error Probable (CEP)
Mission success (yes, no); task completed (completed, not completed; met, not met)	Rating (1-5) of how well mission or task met objectives
Message completion rate (MCR)	Latency, Throughput

Table 4: Calibration of Performance Measurement for Rigor and Realism

	Calibration Goal	Reliable and Valid Measurement of System Performance			
		Data Collection Test Instrumentation (DCTI)	Test Player Surveys (workload & usability ratings)	Data Collectors (DC)	Subject Matter Experts (SMEs)
Rigor ***detect differences***	*Reliability*	DCTI returns consistent values when monitoring similar activity at different time	Individual provides consistent ratings for similar activity at different times	DC provides consistent recordings for similar activity at different times	SME provides consistent rating for similar activity at different times
	Discriminant Validity	DCTI provides different output when monitoring high and low levels of activity	Individual respondent rates high and low activities differently	DC provides different recordings for unlike occurrences	SMEs rate successful and unsuccessful activities differently
Rigor ***identify cause***	*Convergent Validity*	Different DCTI provide similar output when monitoring same activity	Different respondents provide similar ratings when undergoing same activity	Independent DCs provide similar recordings when viewing same activity	Independent SMEs provide similar ratings when viewing same activity
Realism	*Non-Intrusive Validity*	DCTI does not impact system performance	Surveys during test execution do not impact operator performance	DC presence does not impact operator performance	SME presence does not impact operator performance
	Face Validity	DCTI output provides interpretable information	Survey content adequately covers all aspects of workload or usability	DC has sufficient training	SME has sufficient knowledge and experience
	Concurrent Validity	DCTI output agrees with other indicators of output	Respondent rating agrees with other indicators of workload or usability	DC recordings agree with other data	SME rating agrees with standards of task or mission success

OT testers can and should substitute continuous measures for binary measures to right size tests. Typically, requirement binary measures can be derived from collected continuous measures; for example computing hit probabilities from miss distance. Additionally, analytic methods can estimate missing continuous values from censored time data. In practice, both measures can be analyzed from collected data; but sample size requirements to achieve specific level of confidence and power would be based on continuous measures – reducing test trials.

Primary performance measures are Precise (Reliable)

Whether performance responses are continuous or binary, ability to detect performance differences depends on response measurement precision, here defined as reliability and discriminant validity (Table 4). Measurement reliability requires that identical performance outputs result in identical collected results (consistency). Measure reliability ensures trial-to-trial differences in system performance are not veiled due to inconsistencies in collection system performance.

Discriminant validity requires that different system outputs yield different collection results. For example, collected data should accurately reflect the full range of peaks and valleys in system detect times under different trial conditions. There are two threats to discriminant validity. Collection systems can incorrectly produce consistent output even when tested systems actually vary in output. Secondly, collection devices may produce differentiated output but only for moderate outputs, failing to yield lowest or highest values. This artificial capping is called *ceiling or floor* effects, respectively. Measurement discrimination ensures large performance differences are not artificially constrained by collection devices. Services have standardized procedures for assessing reliability and discrimination in test instrumentation. These two attributes can also be assessed and calibrated for player surveys, data collectors, and subject matter experts (SMEs).

The consistency and discrimination of data collectors (DCs) with entries on paper or electronic notebooks can be assessed during data collector training by having DCs record similar and dissimilar incidents several times. SME measurements are more susceptible to inconsistencies since SMEs judge the *value* of activities. During SME training and pilot testing, SME consistency

and discrimination can be assessed when individual SMEs rate similar and dissimilar events at different times. SME calibration occurs by having SMEs discuss why similar events were rated differently or dissimilar events were not rated differently. If needed, testers can develop more observable anchors on rating forms to better focus SME ratings.

Training does not apply to building player survey reliability and discrimination. Most player surveys provide system-use context and diagnostic information; they do not measure overall system effectiveness. These diagnostic surveys are not a concern for test rigor. In some OT however, player rating scales may provide primary quantitative measures for "effectiveness" where new systems are specifically designed to reduce workload or increase usability. In these cases, DOT&E (June 2014) encourages use of industry established surveys that have academically pre-calibrated reliability and discriminate validity. Remaining calibration goals in Table 4 are discussed later.

3. All relevant factors and factor levels are included and strategically controlled.

The next rigor challenge is to identify all relevant factors and factor levels in a test environment that can potentially cause variations in system or mission performance. If any of the five factors in Figure 1 were not identified as a test-design factor, but still occur during test execution, then the impacts of these unidentified conditions on system performance would be regulated to "noise," making it difficult to detect performance differences due to remaining factors.

> **Important factors must be strategically controlled in test design and execution.** Significant factors must be controlled whenever possible; this will ensure coverage of the full operational envelope. As an additional benefit, controlling significant factors helps us to better explain the variance in the data, leading to more meaningful results. [DOT&E June 2013, op.cit. page 3]

A comprehensive list of test factors and factor levels impacting tested performance includes test scenarios conditions, test range characteristics, and artifacts of test architecture such as different types of users, weather, inter-visibility, and so forth. Once identified, all relevant factors are accounted for in one of three control categories:

Strategically controlled: This includes factors that are systematically varied or held constant during test execution. Systematically varying or holding factors constant are central to identifying causes of performance variations and are discussed as parameter #6.

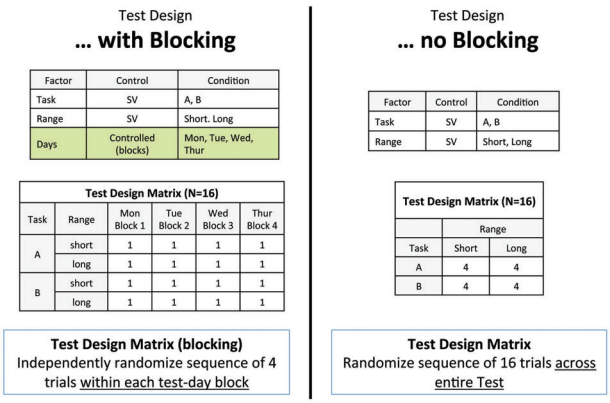

Figure 3: Test Design With and Without Blocking

Uncontrolled but measured (recordable conditions): Some factors expected to impact system performance are difficult to control during testing, but can be measured. An example is inter-visibility between an aircraft and its target; difficult to control but measurable. Measuring atmospheric visibility with range instrumentation allows analysis techniques called analysis of covariance (ANCOVA) to statistically remove impacts of measurable uncontrolled factors; thereby reducing performance variation noise.

There is always a tug in OT between allowing tactical conditions to flow and record them as they occur to promote operational realism (discussed later); or strategically controlling these tactical conditions to enhance test rigor and robustness. When the goal of the test is to characterize system performance, recording free-flowing tactical conditions is rarely useful. If these tactical conditions will have a strong impact on system performance then use of ANCOVA to statistically remove them from consideration is not a good option. Disadvantages to test rigor and robustness when critical conditions are allowed to tactically vary are discussed under parameters #6 and #12.

Uncontrolled and <u>not</u> measured: These are conditions impacting system performance that can neither be controlled nor measured during testing. These noise factors might include different levels of operator proficiency and fluctuations in weather. Tests with many uncontrolled, unmeasured factors increase test noise making it more difficult to detect true performance differences. *Blocking* test trials is a primary technique to reduce impacts of this noise.

Blocking test trials is a strategic design technique to reduce impacts of artificial noise induced incidental to test-execution, thus increasing test efficiency. Blocking

does this by attributing some of the noise to blocking factors thereby reducing our estimate of overall test noise, residual error.

Blocking test trials is best illustrated with an example, presented in Figure 3. In the blocking example, test days serve as a blocking factor to remove impacts of day-to-day uncontrolled, changing conditions (weather, visibility, player learning, etc.) that might cause day-to-day fluctuations in system performance, thus potentially masking performance differences due to the systematically-varied trial conditions. In this example, blocking would remove day-to-day noise from residual test error but would not remove trial-to-trial noise within each individual day of testing. Appropriate application of blocking test trials can greatly reduce estimates of test noise.

4. Like trial conditions are alike.
Systematically varied trial conditions require repeatable conditions during execution. All planned test trials designated to be a high level of a factor (e.g. high network loading, high jamming, or high altitude) should be executed with similar high levels of network loading, jamming, and altitude. Similarly, execution of the low levels of these factors needs to be, in fact, lower and the low level consistently applied to subsequent low-level trials. Inconsistent execution of like-trial conditions increases performance variability making it difficult to detect performance differences between factor levels as illustrated in the following examples:

- *Factors with categorical levels:* All designated night trials should be conducted during periods of darkness and not allowed to start early or end late incorporating day-light conditions. If execution of day versus night trials is not distinct and consistent, it may be difficult to detect true differences in performance due to day-versus-night conditions.
- *Factors with continuous levels:* Tests employing threat jammers to disrupt Blue Force communications should include an independent measure of jamming signals received at intended areas in order to verify actual levels of jamming impacting Blue Forces. While jammer operators can record jammer output levels; without an independent receiver located down range, testers do not know if intended high or low jamming actually arrived at target areas; or whether received jamming levels fluctuated during jamming intervals. Inability to deliver consistent jamming intensities (high or low) for multiple like-trial conditions decreases test rigor.

Similar difficulties can occur when stimulations (via modeling and simulation (M&S)) are employed to create trial conditions such as network loading or higher levels of threat activity. Part of verification, validation, and accreditation (VV&A) for M&S use in OT ensures consistent levels of stimulation are recreated for similar subsequent trials. Stimulations should be monitored real-time during trial execution and recalibrated between trials, if necessary, to ensure like trials are indeed alike.

5. Tested system and operators are stable.
A stable system yields similar outputs given similar inputs. Inconsistent output can result from unstable systems (software or hardware) or inconsistent operators. Variation due to either can mask trial-to-trial performance differences. Rigorous test design strategies avoid "fixing" systems after each trial when the test goal is to assess impacts of trial conditions on system performance. If system fixes must be incorporated during test execution, it is best to group system fixes at specific breaks between test trials and add another factor (pre- and post-fix blocking factor) to the test design.

Stable operators provide consistent operation of systems through successive trials. Variations in operator performance may be due to inexperience with new systems. Well-trained, consistently performing operators reduce performance variation, thereby increasing chances of detecting true shifts in system performance.

Test Indices of Ability to Detect Performance Signals

An early indication of how well test design and architecture will reduce test noise to make detection of systematic performance changes more efficient is to execute test trials with identical conditions during pilot testing (called *replication* trials). Any variation in system performance between replicated trials indicates noise caused by instability in data collection (P-2), insufficient factors and factor levels to characterize system and test architecture (P-3), inability to hold like test conditions alike (P-4), or instability in system and operators (P-5). System performance variation between replicated trials is an early index of lower test rigor.

It should be noted that increasing sample size (P-1) will allow tests with larger noise to still display detectable signals. Efficient test designs however, find performance signals by decreasing noise, rather than increasing test size.

Post-test data analysis can provide a more comprehensive indicator analogous to signal-to-noise detection. All DOE analysis programs provide a numerical

index of total performance variability that can be attributed to test factors (signal). This R-Squared (R^2) index[iv] varies from 0 to 1.0 with high values (0.7 - 0.9) indicating that most system performance variability during testing was due to test factors, which is analogous to a high signal-to-noise detection. Lower R^2 values indicate most system variability could not explain by test factors, indicating we did not learn why our tested system performed as it did. Low R^2 values would prompt testers to review one or more of the first five rigor parameters.

Ability to Correctly Interpret Cause of System Performance Peaks and Valleys

> **Confounding factors must be avoided.**
> I have observed cases where the test plan varies factors in such as a way as to confound the effects, even when it was eminently feasible to avoid doing so.… Test strategies that confound factors are uninformative, inefficient, eliminate the ability to predict performance in other conditions, and preclude the ability to determine which factor affects performance the most. [[DOT&E, June 2013, op.cit. page 3]

Given that test designs have sufficient rigor to detect peaks and valleys in system performance, the next logical question is "What caused these variations?" Determination of causality is critical to explaining where and when performance peaks and valleys will occur on battlefields.

Inability to identify causes of system variation is the problem of *confounding*. Confounded results yield two or more potential explanations for an individual performance peak or valley. Confounded results are not useful; system developers will not know what to fix and Warfighters will not know what to expect on battlefields.

Test designs posit factors and conditions as potential causes of system performance variation. To determine whether systematically varied trial conditions (day versus night or target X versus target Y) actually caused variations in system performance, test designs require the capability to estimate un-confounded factors as illustrated in Figure 4. The left side represents *un-confounded, high-resolution, independent* factors, making it easy to determine, after test data is analyzed, that Factor A, not Factor B, was primary cause of variations in performance. In contrast, the right side represents *confounded, low resolution, aliased* factors, making it difficult to conclude which factor was responsible for system variations.

When factors are confounded, application of test results is reduced: "We can say how much system performance will vary; but are not sure when and where

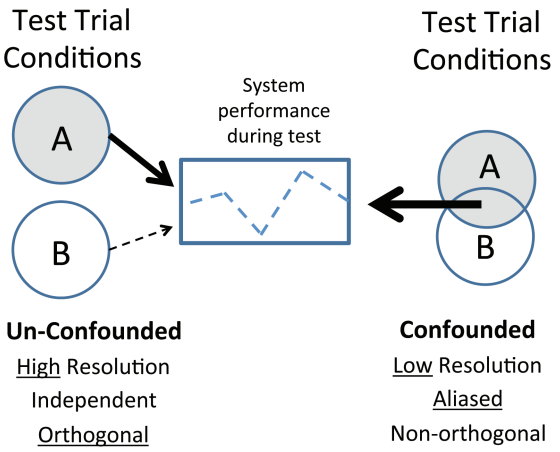

Figure 4: Confounded and Un-confounded Causes of System performance variations

you will see this performance difference on battlefields." Test parameters 6 through 10 minimize factor confounding. Careful test planning coupled with execution discipline can help avoid confounding factors.

6. Factors are systematically varied.
Failure to *systematically* vary factors and conditions for successive test trials can reduce test robustness when critical conditions fail to occur (discussed later under parameter #12); and also lead to factor confounding due to factor levels overlapping as illustrated in Table 5. The uncontrolled, tactically-varied test design does not have a test trial where *Attack* is paired with *High* threat; nor where *Defend* is paired with *Low* threat. Consequently, any performance variation between *Attack* and *Defend* could be due to performance differences resulting from *Low* and *High* threat. Mission and Threat factors are 100% confounded yielding two competing explanation for a single performance differences.

When both factors are systematically varied, there is no confounding of conditions. Each mission level is paired with a different threat level. Now any differences between *Attack* and *Defend* can be solely attributed to differences in missions, because each mission had both threat levels.

Table 5: Test Factors Uncontrolled or Systematically Varied

Uncontrolled		Test Trial	**Systematically Varied**	
Mission	Threat		Mission	Threat
Attack	Low	1	Attack	Low
Defend	High	2	Defend	High
Attack	Low	3	Attack	High
Defend	High	4	Defend	Low

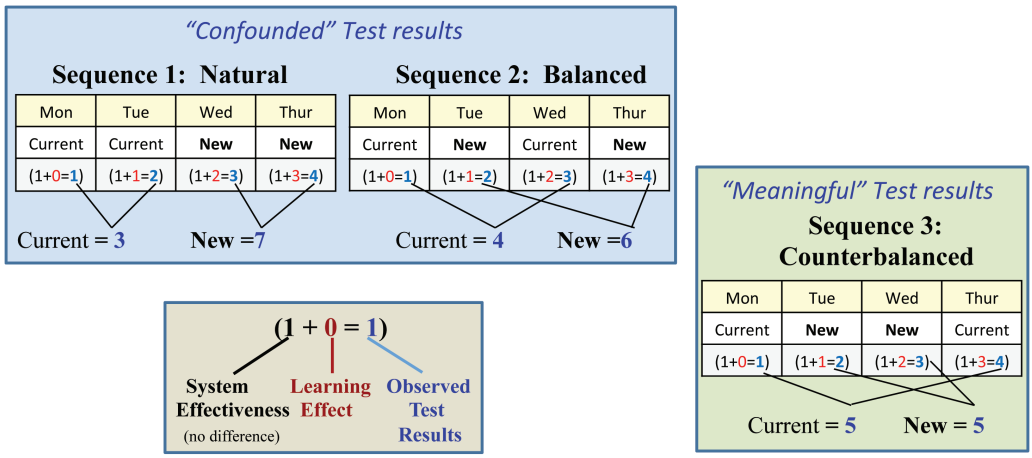

Figure 5: Results of Three Test-Trial Sequences

If Threat had been *held constant* at high in the uncontrolled design, comparisons between Attack and Defend would be un-confounded because both mission levels would experience the same high threat level. Holding some conditions constant can eliminate potential confounding when systematically varying those conditions is not an option. Unfortunately, this reduces test *robustness*, discussed later.

Some factors may be difficult to systematically change trial to trial. For example, a test team may need an entire administrative day to move test execution from desert to mountainous terrain. Sophisticated split-plot test designs are available when it is more economical to group test trials into discrete "hard to change" conditions.

7. Distribution of trials in test design maximizes factor resolution.

Even when all test conditions are systematically varied, factor confounding can still occur when there are empty test cells in the test design matrix, or there is an unbalanced distribution of trials (Table 6). This confounding is termed factor *aliasing*. *Design resolution* represents the extent of non-aliasing, independence among test factors.

Not all factor aliasing is equally harmful. We often require tests with large numbers of factors and levels to have some empty cells to economize test costs, as illustrated later in Table 10. These empty cells can be allocated in specific, balanced patterns to minimize factor aliasing for main effects and lower order factor interactions. DOE software programs provide an index of factor aliasing, called variance inflation factor[v] (VIF); and allocate placement of test points to minimize factor aliasing (test rigor) while maximizing coverage of factor main effects and important lower order interactions.

While factor aliasing is a source of factor confounding, this confounding could be fixed post-test by executing additional test trials to selectively fill-in missing or unbalanced test cells. Other sources of confounding (parameters 6, 8, 9, and 10) are not fixable post-test, meaning we are stuck with a non-rigorous test.

8. Trial sequence minimizes factor confounding.

Even when factors are systematically varied and their distribution supports sufficient factor resolution, confounding still occurs if the sequence of trial execution is inappropriate. Figure 5 illustrates impacts of trial sequencing on factor confounding.

Imagine a new test officer is to design a quick four-day test to assess whether a New System performs better than the Current System. Our test officer might employ Sequence #1 in Figure 5. In this "natural" design, the

Table 6: Test Design with Empty or Unbalanced Trials

Empty Test Cells					
		Threat X		Threat Y	
		Day	Night	Day	Night
Desert	Task A	1			1
	Task B		1	1	
Urban	Task A		1	1	
	Task B	1			1
Unbalanced Test Cells					
		Threat X		Threat Y	
		Day	Night	Day	Night
Desert	Task A	1	3	1	1
	Task B	2	1	3	1
Urban	Task A	1	2	1	1
	Task B	1	4	1	2

test unit operates Current System in first two trials and New System in last two trials.

To illustrate problems with Sequence #1, we quantify what can happen in each separate trial to produce test results. The three numbers below each test trial quantify *system effect, order effect* (learning effect), and *observed results.* In this example, system effect is held constant by making the first number equal to "1" for all four trials. Designating system effect constant indicates that Current and New systems are equivalent in capability. Consequently, any observed differences between systems during testing must have resulted from some other factor.

The other factor in this example is a "learning effect." Each test day our test unit gets better; represented by increasing the second number for each successive trial. The third number represents test results obtained by adding system and learning effects. When test analysts use this third number in Sequence #1, they conclude New System greatly outperformed Current System, 7 versus 3. This incorrect conclusion was due to a learning-effect factor, not system factor.

The *balanced* design in Sequence #2 helps some, but does not eliminate confounding. Only the *counterbalanced* design Sequence #3 provides interpretable results. In all three designs, test officers faithfully executed all trials and collected all data; but inadequate trial sequences in designs 1 and 2 rendered these two tests useless for comparing systems.

While it is easy to establish a best sequence in this small notional example, it is not so easy in tests with multiple factors and factor levels. In complex test designs where the number of test trials is large, using DOE software to *randomize* trial sequence for test execution works well.

9. Player assignment to baseline system or alternate vendors minimizes system factor confounding.

Parameters 6-8 discussed thus far are considerations for reducing confounding in every test design. Parameter #9 is only applicable to tests that involve *system-versus-system* comparisons:

(a) comparison of new systems with their legacy or baseline (BL) systems, (b) comparison of alternative system configurations, or (c) comparison of alternative vendor prototypes. In all three system-versus-system tests, assignment of test players (operators or test units) to alternate systems is critical to interpreting causes of system differences.

System-versus-system testing does not occur often. Most tests involve a single player unit operating a single new system (with multiple prototypes to outfit test units) across all test conditions as illustrated in Table 7. Assignment of a player unit to operate this system is not an issue for test rigor but is an issue for test robustness and test realism. For example, we can still have a *rigorous* test (detect effects and interpret causes) even if system contractors were to operate tested systems. However, we would not have *robust or realistic* testing because we could not apply these rigorous conclusions to intended users in an operational environment.

Table 7: Test Design matrix where Single Unit conducts all trials employing New System

	Low Threat	High Threat
Attack	4	4
Defense	4	4

Table 8. Non-rigorous Test Design comparing two alternate Vendors

	Attack Mission		Defense Mission	
	Low Threat	High Threat	Low Threat	High Threat
Vendor X, Unit A	2	2	2	2
Vendor Y, Unit B	2	2	2	2

In contrast, player assignment is a major challenge in system-versus system testing. A notional test to compare alternative vendors is displayed in Table 8. In this example, Vendor X and Y systems are totally confounded with test unit. If there is a performance difference between the two vendor rows, we would not know if differences were due to real differences between vendors or due to real differences between player units. There are two traditional attempts to justify rigor of this design, neither works well.

(1) *Random assignment was used to decide which of two units were matched to a particular vendor.* This is not a viable option as the problem of separating unit effects from vendor effects still exists, regardless which unit operates which system.

(2) *The units are "declared equivalent" for military skills, combat experience, prior test scores, etc.* This is also not a viable option as traits that determine mission success are not easily measurable.

Table 9 illustrates four rigorous designs for system-versus-system tests comparing alternative vendors. Each design minimizes unit-vendor confounding and each has advantages and disadvantages.

Design 1: Multiple Units crossover to Alternate Vendors: Each unit operates each vendor's systems. Unit-vendor

Table 9. Rigorous Test Designs comparing two alternative Vendors

		Attack Mission		Defense Mission	
		Low Threat	High Threat	Low Threat	High Threat
Design 1: Multiple Units Crossover to Alternate Vendor (total trials = 16)					
Week 1	Vendor X, Unit A	1	1	1	1
	Vendor Y, Unit B	1	1	1	1
Week 2	Vendor X, Unit B	1	1	1	1
	Vendor Y, Unit A	1	1	1	1
Design 2: Single Unit Crossover to Alternate Vendor (total trials = 16)					
Weeks 1- 4	Vendor X, Unit A	2	2	2	2
	Vendor Y, Unit A	2	2	2	2
Design 3: Multiple Units Crossover to BL (total trials = 16)					
Weeks 1-2	Vendor X, Unit A	1	1	1	1
	BL System, Unit A	1	1	1	1
	Vendor Y, Unit B	1	1	1	1
	BL System, Unit B	1	1	1	1
Design 4: Multiple Operators random assignment to Alternate Vendors with multiple independent prototypes (10 test scores for each vendor in each condition)					
Weeks 1-2	Vendor Y, operators 1-10	10	10	10	10
	Vendor X, operators 10-20	10	10	10	10

confounding is neutralized because unit differences impact each vendor equally. This design works best when there are only two or three competing vendors and little additional training is required to move from one vendor system to another. The trial sequence would not be completely randomized. To ensure comparability between vendors, first randomize the sequencing of 16 trials for Vendor X and then mirror this sequence for Vendor Y.

Design 2: Single Unit crossover to Alternate Vendor: A single player Unit operates each vendor's system eliminating player-system confounding. This design works best when the transition from vendor to vendor after each trial is easy. Unfortunately, total test time is increased, since a single unit cannot execute both vendor trials simultaneously. As above, trial sequence would be randomized for one vendor, and the second vendor mirrors this sequence. The overall test sequence would inter-mingle vendor trials to ensure each vendor had equal opportunities to "go first." This mitigates potential player learning enhancing the vendor performance that always went second.

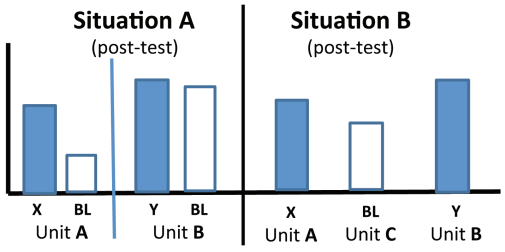

Figure 6: Example Results of Testing Vendors X and Y with Baseline (BL)

Design 3: Multiple Units crossover to BL: Each unit operates a single vendor system and the baseline (BL) legacy system. Unit-vendor confounding is neutralized as long as comparisons between vendor and BL are within a unit to determine which vendor produced a larger delta from its BL, as in Situation A in Figure 6. A rigorous conclusion can be made that Vendor X provides better improvement over our BL system even though Vendor Y had more successes.

In Situation B, the addition of a baseline operated by an independent unit does not provide test rigor. Each vendor and BL is confounded with performance of its test unit. One cannot conclude how much of each system-unit result (X, Y, or BL) is due to inherent differences between units or due to real differences between systems.

Crossover designs illustrated in Situation A are useful when vendor training is complex. Each test unit only trains on a single vendor system. There are disadvantages. Direct comparisons between vendors are confounded and only comparison between a vendor and its BL are valid. Additionally, there are 50% fewer vendor trials for the same total trials in Designs 1 and 2. Finally, test results may indicate a vendor with the fewest overall successes is the better performer (as in Situation A) – a difficult result to explain.

Design 4: Multiple Player Random Assignment to Vendors. While random assignment of participants to alternative treatments is a "gold standard" in rigorous medical research, it is seldom applicable to acquisition testing. Random player assignment is only meaningful in a subset of system-versus-system designs where each vendor has multiple prototypes that are operated independently or for a single prototype that can be operated by alternate operators sequentially and independently in relatively short vignettes. Generally this occurs when testing new systems for individuals, such as modifications to individual weapons, body armor, hearing protectors, parachutes; and each individual soldier-system receives an independent score.

For random assignment to work in our example, we would need an initial pool of 20 or more potential operators/players to randomly assign at least 10 to each vendor's prototype to ensure that one or two outliers in a group of 10 would not greatly impact the group average for an individual vendor. Advantage of this design is that each operator only needs to be trained on a single vendor's system. The disadvantage is that each vendor needs sufficient prototypes to accommodate all 10 operators and each operator is expected to operate individually and independently, not as a cohesive unit. The sequence of test conditions for each vendor should mirror each other.

10. Data collection is not biased.
Data collection inconsistencies and lack of discrimination were discussed earlier in parameter #2. Here we are concerned with convergent validity (see Table 4) to counter measurement *bias.* A *biased* measure is one that is "*consistently* off mark" (consistently too high or too low).

Over or under measurement of system performance is of greatest concern in system-versus-system testing where alternative systems or vendors are assessed with different data collection capabilities: different instrumented collection, different manual data collectors, or different SMEs. Data collection bias makes it difficult to determine if system differences are due to system or data collection differences. For example, when new collection instrumentation is developed for "new" tested systems and legacy systems are instrumented with "old" collection devices, potential measurement bias exists. Bias can also occur when test instrumentation is tailored for alternative vendors or when data collectors and SMEs are different for alternative systems and vendors.

To minimize collection bias, data collectors and SMEs should be systematically rotated among alternative vendors and between new and legacy systems. Certification of test instrumentation should ensure that collection devices for alternative systems provide equivalent results for equivalent system performance.

Robust

Test Scope Represents Full Battlefield Domain for System

Rigorous tests, as just discussed, have strong *internal validity* (Shadish, et. al. 2007). Test rigor permits conclusions about cause and effect to characterize system performance. The remaining 11 parameters pertain to tests' *external validity.* Test *robustness*, discussed here, indicates the extent that rigorous cause and effect conclusions will apply across the entire operational envelope. Test *realism* indicates how well cause-and-effect conclusions from an individual test environment represent actual battlefield causes and effects.

In this proposed framework test robustness pertains to breadth of battlefield envelopes represented in individual test environments. Parameter #3 focused on including all conditions resident in a test environment. Test robustness is wider as it includes all conditions found in the system's operational envelopes, *even those that do not naturally occur in test environments.* Adequate coverage of systems' operational employment environments means all relevant factors and factor levels are identified *and assessed.* Robustness of an individual test is a comparison of its design space to the tested system's future *operational space.* After establishing a robust set of factors and conditions, sophisticated DOE space-filing designs are available as well as fraction of design space (FDS) metrics to ensure adequate trial coverage *within* a robust test design space.

11. Comprehensive set of factors

> **Factors varied in test must not be limited to those defined in the requirements documents.** …users will employ the system in conditions that are different from those identified for system development and specification compliance. As operational testers we are responsible for evaluating a system across the conditions under which the system will actually be employed.…Replicating runs/events under one condition (such as that identified in requirements document is less informative than a strategic spread of runs/events across all relevant conditions. [DOT&E June 2013, op.cit. page 2]

OT should reflect a system's full potential operational environment. There are two keys to representing

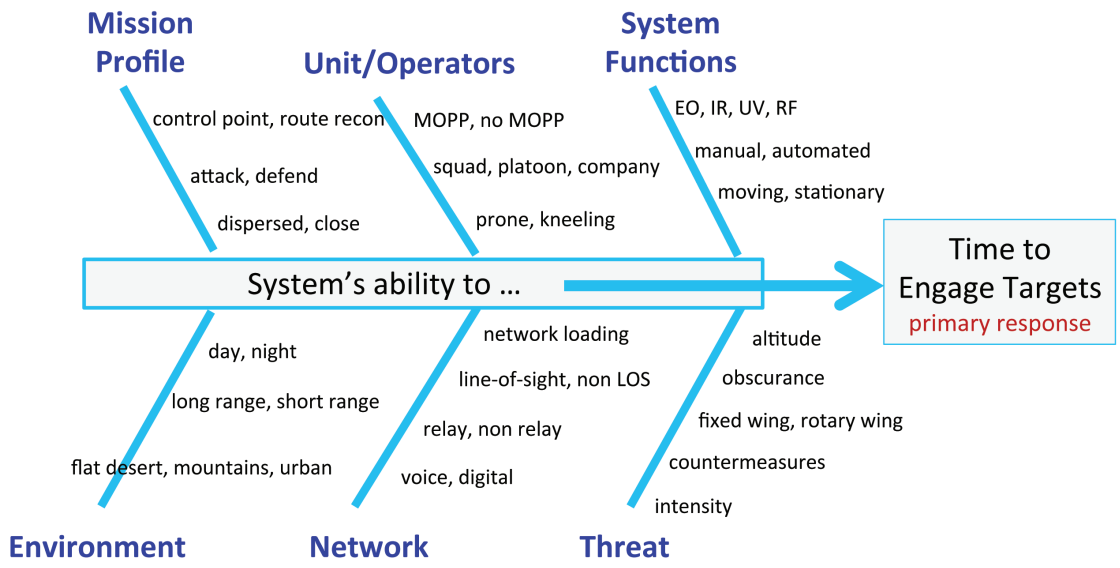

Figure 7. Cause and Effect Fishbone Diagram

all intended battlefield conditions in test events. First, gather experts from across a Service to augment and refine factors and factor levels defined in system requirement documents and test support packages (TSPs). Secondly, these experts should assist in determining the subset of factors that might cause variations in system performance on its *primary responses*. Results of this effort can be summarized in a fishbone cause and effect diagram (Figure 7). Common mistakes in developing a robust set of factors are to list only those used in prior tests of similar systems, only list those in player unit training scenarios, or only list those available at test sites.

Results from tests with limited robustness only apply to a portion of the intended operational envelope. However, a single robust test encompassing all conditions may not be affordable. Strategies for reducing the number of factors and factor levels in a single test are discussed under parameter #13.

12. Categorical and continuous factors have full range of levels.

Ensuring testing across the full range of factor levels for categorical and continuous factors requires an initial identification of the full range of conditions for each factor and then ensuring the full range occurs during test execution.

Categorical factors have discrete levels — different missions, different system configurations, different types of targets, or different types of terrain. For example, if a new system will engage six different target types, but testing includes only four, our test is not robust. Some tests might require two levels of light (day and night) and some tests four levels (day, dusk, night, and dawn); if four different levels could impact system performance differently.

Continuous factors have naturally continuous conditions — e.g., target speed 0-40 mph, range to target 3-10 km, azimuth and elevation, network loading, and jammer power levels. Robust tests include low and high endpoints that represent true high and low operational conditions. Having established robust endpoints, testers should not convert naturally continuous factors to categorical factors. For example, converting target speed that ranges from 0-40 mph to two categorical levels of "stationary and moving" should be avoided. This

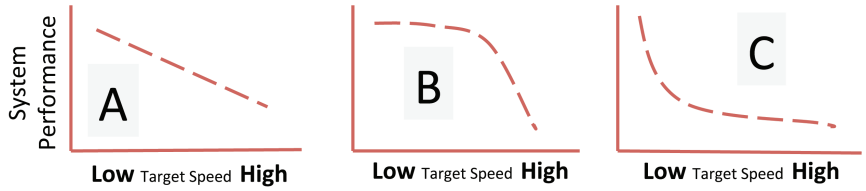

Figure 8. Linear and Non-Linear Continuous Factors

categorical conversion reduces consistency of the high end-point for target speed. For example, some trials designated as *moving* may have target speeds as low as 10 or 15 mph or as high as 30 or 40 mph. This violates test rigor parameter #4, calling subsequent like conditions to be alike.

Additionally, continuous factors should often include mid points. Testing only at low and high end points assumes that any causal relationship between factor levels and system performance is linear as in Figure 8, graph A. If potential non-linear relationships between factors and system response are present, such as in graphs B and C of Figure 8, testing without mid points would not be sufficiently robust to assess non-linear effects.

Once a robust set of categorical and continuous conditions are included in test designs, we need to ensure all specified factor levels occur during test execution by strategically controlling factors and sometimes augmenting tests with simulations.

Important factors must be strategically controlled in test design and execution.
Although … test design process identifies factors, the proposed run plan often relegates them to recordable conditions. This is a risky test strategy because all levels of the factor of interest might not be observed in the operational test. Significant factors must be controlled whenever possible; this will ensure coverage of the full operational envelope. [DOT&E, June 2013, op.cit. page 3]

Relegating factors to "recordable conditions" (DOT&E text box) allows test units to tactically vary their operations and record these tactical variations, *if and when* they occur. Recordable conditions are sufficient to promote test rigor as discussed under parameter #3 since their impact is statistically removed from system noise variation when they occur. However if recordable conditions do not occur, envelope coverage is reduced; reducing test robustness.

Augmenting tests via virtual methods is a means to achieve robustness in the face of limited assets. Battle command systems may require both live and simulated operational traffic to ensure full levels of system stress when it is cost prohibitive to including all "live" subordinate and adjacent units and systems in a test. Similarly, tactical networks may require stimulation to test operationally high levels of network loading when there are insufficient transmitters for full communications architectures.

13. Factors held constant are minimal.
What if the list of robust factors and factor levels is quite long? There are at least three strategies for reducing the number of factors for individual tests.

(1) Holding some test conditions constant is a common strategy to reduce test scope and save costs. For example, test only desert terrain instead of desert and forest, or test only benign conditions instead of benign and jamming. Holding conditions constant reduces test cost and maintains test rigor even though it reduces test robustness.

Holding conditions constant makes sense when conditions are expensive to achieve or are simply unique conditions. For example, extreme environmental conditions may not be easily reproduced in field tests and are more economically assessed in chamber testing. In these situations, it is usually more cost effective to address the full spectrum of factors across separate assessment events (OT, DT, DT/OT, demonstrations, and M&S). While individual test events may not be fully robust, a campaign of assessment events may still provide a robust milestone assessment.

(2) Factor screening with sequential testing is a strategy for reducing the number of factors in subsequent tests. This strategy loads early tests with all, or most, factors and assesses (screens) these factors statistically. Any factors found to be statistically insignificant is eliminated from consideration in subsequent testing. Thus "factor-loading" early tests may serve to reduce scope requirements for subsequent testing.

Constructing sequential test programs where early tests screen-out unimportant factors to focus and reduce the size of subsequent tests have worked well in industry but their application to acquisition testing may be limited. In acquisition testing, systems-under-test continually evolve in functionality while the level of mission participation also increases from early single prototypes to fully equipped units for final testing. These evolving characteristics make it less likely that factors eliminated in early prototype testing would not apply to subsequent full-up unit testing. Additionally, results of early screening tests may not be available until just weeks prior to start of subsequent testing, making it practically difficult to apply information about non-relevant factors to the next, near-term sequential tests.

(3) Removing factors to reduce number of test trials has minimal impact on reducing test trials for traditional weapon system testing (see DOT&E text box). Adding or removing factors or factor levels does have some impact on sample size requirements. However, impacts are relatively gradual as factors and levels increase as illustrated in the following example.

Understand that adding a condition/level/factor does not necessarily increase the size of the test.
Although this effect is true for *[full]* factorial designs, many other design types exist to overcome these difficulties and ensure testing adequately covers the operational envelope without significantly increasing the necessary test resources.… Therefore, important factors should not be eliminated from the test design … in an attempt to avoid increasing the test size. Instead, the test design should include all operationally relevant factors and capitalize on the statistical tenets of DOE which enable the test size to remain manageable while ensuring the operational envelope is covered adequately. [DOT&E June 2013, op.cit. page 5]

Table 10: Test Design with Six Factors

			D1				D2			
			E1		E2		E1		E2	
			F1	F2	F1	F2	F1	F2	F1	F2
A1	B1	C1	1			1		1	1	
		C2		1	1		1			1
	B2	C1		1			1			1
		C2	1			1		1	1	
A2	B1	C1	1			1			1	
		C2		1	1			1		
	B2	C1		1	1		1			1
		C2	1			1			1	

Table 2 indicated a sample size of 20 is sufficient to achieve 80% confidence and 80% power for two factors with four combinations. The notional test design in Table 10 has six factors with 64 combinations of conditions. Here, an increase of only 8 trials, a sample size of 28 instead of 20, is sufficient for 80% confidence and power to evaluate all six main effects and two-factor interactions. Moreover, factor aliasing (parameter #7) is minimal even with unbalanced and missing cells.

This 6-factor test with only 28 trials is robust. DOE statistical modeling allows estimates of system performance for all factors and factor levels even when most individual combinations are not tested! Test efficiency through DOE implementation.

Realistic

Test Environment Represents Actual Battlefield Conditions

Test results are only useful to the extent they say something about actual combat operations. *Generalizability* is the scientific term for applying results outside of test context. Test realism enhances generalizability of test results to actual operational battlefield conditions. A realistic test requires realism in four areas: representative systems under test, test players, performance measures, and test environment.

System Representative

14. System is production representative.
Experienced test officers know it is an exception when truly "production representative" systems show up for test. The system test and evaluation master plan (TEMP) provides details on production representative capabilities and software versions. It is important to document system software and hardware capabilities in the version available for test. Services have procedures for controlling system configuration during test execution. Testing immature prototypes yields less realistic results. Initial Operational Test and Evaluation (IOTE) events require fully *realistic*, representative systems.

15. Appropriate number and distribution of systems in unit
TEMPs and System Basis of Issue Plans (BOIP) provide information on distribution of new systems within and between operational units. The level of player unit (team, section, platoon, company, and battalion/squadron) appropriate for a realistic "fighting unit" depends on the unit level that will independently employ the new system.

Test Unit Representative

16. Full spectrum of units/operators in test
How well do test players represent operators and actual units that will eventually employ the new capability? A good practice to ensure representative test players is to select players directly from those operational units among the first to receive the new capability. A challenge to player representativeness occurs when new systems are fielded to different types of units and only one type of unit is assigned to test. In this situation, a campaign of multiple tests needs to show that the full range

of units will be tested in different events. If not, multiple units need to participate in a single test event.

17. Unit/operators not over trained (golden crew) nor undertrained

Realism requires appropriate levels of player training. If test units are under-trained or over-trained, true capabilities of new systems in the hands of typical users will be misrepresented. Under-training can result from compressed schedules as well as new equipment training (NET) that focuses on *operating* new systems rather than *employing* them to maximize their capability. Unfortunately, too often initial pilot tests and first days of record testing are used to improve system employment procedures. This is further rationale for randomizing test-trial sequences as discussed under test *rigor*.

Over-training arises when player units undergo extensive training not planned for typical units receiving fielded systems. The temptation is to over-train test units to ensure success. Over-trained test units are "golden crews." Our realism challenge is to produce well-trained, typical users as test players rather than over- or under-trained unique players.

Primary Measures Representative

18. Primary measures reflect system contribution to mission/task success.

> **Metrics must be mission-oriented, relevant, informative, and not rigidly adhere to the narrowest possible interpretation of definitions in requirements documents.**
> Too often the goals of OT&E are not captured by technical performance requirements. This is especially true when responses are limited to the technical performance requirements of the system under test when, in fact, the mission capability requires a system of systems to succeed. Ideal OT should provide a measure of mission accomplishment (not technical performance for a single system), lend themselves to good test design (i.e. to be continuous in nature), and in general comprehensively cover the reason for procuring the system. [DOT&E Memorandum, June 2013, op.cit. page 2]

Ensuring that primary measures *directly* reflect mission accomplishment is easiest when system performance responses and mission measures are complimentary. This occurs rather easily for combat systems that engage enemy forces such as sensors and shooters: targets detected, identified, engaged, and destroyed. Measure realism is more difficult when (a) tested systems are not expected to have large impacts on mission success; and (b) when tested systems impact complex military processes.

Some tested systems do not have large impacts on units' capability to succeed. It is unrealistic to expect a software change to unit radios to have large impacts on combat outcomes. In these cases, testers should look for meaningful "sub-mission" tasks that can be expected to be more directly impacted.

Other tested systems may have a major impact on mission success, but mission success is a complex outcome such as increased information superiority, increased situational awareness, or improved command and control. In these instances, system requirements tend to focus on technical, sub-aspects of the complex process supported. For example, increased message completion rate is often an approximate measure for improved command and control. When technical system measures are "too distant" to measure complex operational process outcomes, testers add expert raters to assess mission or process success. Ensuring realism of collected measures is discussed next.

19. Primary measures adequately represented in data collection

Table 4 summarized three calibration goals for test realism in data collection. Test agencies have established procedures to calibrate and certified test instrumentation to ensure it does not adversely impact the tested system (non-intrusive validity), it provides interpretable data (face validity), and its output agrees with expectations (concurrent validity). The following discussion focuses on player surveys, data collectors, and SMEs.

> Surveys are a key mechanism to obtain needed data to aid the operational evaluation. Properly designed surveys, which measure the thoughts and opinions of operators and maintainers, are, therefore, essential elements in the evaluation of a system's operational effectiveness and suitability. A substantial body of scientific research exists on survey design, analysis, and administration that we should leverage in OT&E. [DOT&E Memorandum, June 2014 op.cit. page 1]

Similar to parameter #2, (DOT&E June 2014, op.cit.) pre-calibrated survey rating scales are useful to measure workload reduction or increase in system usability. These pre-calibrated surveys have accepted administration procedures to ensure non-intrusiveness and academically verified face and concurrent validity

as accurate measures of system workload and usability. DOT&E also recommends best practices for developing custom designed surveys to adequately collect diagnostic information. These best practices include developing short, confidential surveys with clear, singular, independent, and neutral questions that are within operators' knowledge.

With adequate training, data collectors (DCs) can meet all three test realism goals listed in Table 4. DCs may need to be monitored for fatigue effects during long test periods. When DC observations are critical to measuring primary performance responses, realism goals are better assured by aggregating the observations of two independent DCs.

Discussions on *subject matter experts (SMEs)* ratings usually focus on the number of rating scale options, typically employing 3, 4, 5, or 7 options. This discussion is not sufficient for realistic ratings of complex outcomes. SME ratings of mission success may not always clearly connect to systems under evaluation. Techniques for better connecting SME ratings to system-related mission performance include the following:

- SMEs should minimize verbal interactions with test players during mission execution. (non-intrusive)
- Ensure SMEs are truly knowledgeable of the activities to be rated. Use SME rating scales with observable "anchor" points -- for example "1= not completed within time and multiple steps missed, 4= all steps completed within allotted time;" rather than "1= very poor, 4=very good." (face and concurrent validity)
- Have two or more SMEs observe the same event and provide their ratings independently. (concurrent validity)
- Correlate SME rating to additional data-driven quantitative metrics related to system and unit performance. Evaluators can use these technical (but still mission-focused) metrics together with SME ratings to form an overall determination of effectiveness. (concurrent validity)

Scenario/Site Representative

20. Blue operations not artificially augmented nor constrained

Realistic Blue operations depend on implementing realistic tactics, techniques, and procedures (TTP). Testing artifacts can make it difficult for test units to develop and acquire realistic TTPs. Modifying current Blue force tactics to incorporate new capabilities often follows, rather than precedes new capability development. Even when new techniques and procedures have been developed, adequate training is difficult due to untried new TTP. Implementation of trained TTP may be hampered by range terrain restrictions, collecting and harvesting data collection, or safety restraints.

There will be restrictions on full Blue "free tactical play" in order to execute designed trial conditions. Systematically varied trial conditions may stipulate specific tactical maneuvers, movement rates, or route directions instead of allowing player units to freely adjust. There are always trade-offs between full Blue free-play and requirements for test rigor and robustness.

Completely independent free play is not in the interest of efficient testing. Experienced testers permit *sufficient* force-on-force free-play to promote *test realism* while adhering to assigned systematically-varied conditions for that trial *(test rigor)* and ensuring the full range of conditions occur in the trial (test rigor). This is "realistic free-play within a box." Requirements for *realistic* Blue and Threat tactical operations need to also be balanced with requirements for data collection, data harvesting, and test instrumentation calibration during extended trials.

Realistic Blue operations are dependent on minimizing adverse impacts of test operations during test execution. For example, directing players to an assembly area to calibrate instrumentation during continuous operations provides advance warning to units that a battle will soon occur. Additionally, test players undergoing similar scenarios over successive trials know what to expect.

Realistic Blue situational awareness may also require virtual augmentation. It was previously noted that virtual augmentation supports test robustness to achieve full ranges of conditions (parameter #12). Here, virtual and constructive simulations can provide a more realistic context for test unit operations by supplying adjacent units and systems not available as "live" components. Full, live joint context for system-of-system testing is costly. Virtual and constructive simulations, interlaced to live test players, can efficiently provide full realistic tactical context.

21. Independent, reactive, current threat

Realistic tests include current threats anticipated in the intended battlespace, not limited to threats describe in requirement documents written 7-10 years previously. Robust tests (parameter #12) ensure threat representation spans the space of current threats (different target types, electronic attack types, computer attacks, etc.) that can stress the system differently. Threat experts from national agencies can assist identifying current threat capabilities and scenarios. Full representation of

threat tactics and weapons is difficult. Most tests approximate threat realism by employing operational units to emulate threat tactics augmented with live actual threat systems available from the Service's Threat Systems Management Offices (TSMOs). Test realism may also require real-time virtual augmentation to achieve adequate levels of threat forces and activities.

Force-on-force engagements during test execution can take place within *restricted free-play* discussed above. Conducting field tests at Service combat training centers (CTCs) employing their dedicated, well-trained, and equipped threats can enhance test realism. However, rigor and robustness are more difficult to achieve conducting trials during CTCs training missions.

R3 Summary

This paper proposes 21 parameters to design and conduct rigorous, robust, and realistic tests; and suggests no single event can achieve 100% of all three conditions. There are tradeoffs between enhancing one that places limits on the other two. It is a zero-sum situation. Enhancing test rigor necessarily constrains robustness and realism and vice versa. Understanding these 21 parameters coupled with careful planning and test execution can assist in test events with appropriate levels of rigor, robustness, and realism to support acquisition decisions. ❑

RICHARD A. KASS, Ph.D., is a Senior Analyst with Geeks and Nerds®. He currently provides technical support on Design of Experiments (DOE) to US Army Operational Test Command (USAOTC). Previously in civil service, Dr. Kass conducted all levels of operational testing for US Army Test and Evaluation Command (USATEC) for 18 years and developed methodology for joint experimentation in US Joint Forces Command (USJFCOM) for eight years. Dr Kass is a former Marine officer and graduate of Southern Illinois University where he earned his Ph.D. in Psychology. Email: rick.kass@geeksandnerds.com

Disclaimer

Views expressed in this article are those of the author and do not reflect official policy or position of the United States Army, Department of Defense, or U.S. Government.

Notes

[i] Figure 1 and discussion points adapted from chart and presentation by Dr. Mark E. Kiemele, Air Academy Associates, during Webinar on "Operationalizing Design of Experiments (DOE)" hosted by ITEA 20 June 2013.

[ii] Performance "peaks and valleys" as used here refers to the observed multiple highs and lows in performance results exhibited by tested systems during OT. This descriptive terminology is not meant to restrict OT analysis to response surface methodology (RSM) that focuses on finding performance minimum and maximums for a given response surface. The challenge in OT, rather, is to ensure true system highs and lows are allowed to exhibit themselves during testing and that appropriate causes of these variations are available and unambiguous.

[iii] Service Operational Test Agencies (OTAs) do not characterize hypothesis testing similarly for the two test objectives that compare results to thresholds or compare new systems to baselines. Army OTA traditionally uses the null hypothesis to express that new systems are less or equal to the threshold or baseline system; and alternative hypothesis to express that new systems are better than threshold or baseline system. Other OTAs use null hypothesis to express that new systems are better than or equal to threshold or baseline system; and alternative hypothesis to express that new systems are less than threshold or baseline system. This second hypothesis convention changes the tabled meaning of confidence and power for these two test objectives.

[iv] R^2 = ratio of sum of squares model to sum of squares total (SS_{model} / SS_{total}); while a post-test equivalent S/N = ratio of mean-square factor effect to mean-square error (MS_{effect} / M_{Serror}).

[v] VIF=1.0 indicates linearly independent, column-wise, pairwise orthogonal factors; while higher VIFs indicate correlated factors known as multicollinearity.

References

DOT&E Memorandum October, 2010. SUBJECT: Guidance on Use of DOE in OTE.

------ June 2013. SUBJECT: Flawed Application of DOE to OT&E.

------ July 2013. SUBJECT: Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in OT&E.

------ June 2014. SUBJECT: Guidance on Use and Design of Surveys in OT&E.

Hill, R., Ahner, D. K., and Gutman, A. J. 2014. What the Department of Defense Leaders Should Know about Statistical Rigor and Design of Experiments for Test and Evaluation. The ITEA Journal 35 (3): 258-265.

Montgomery, D. C. 2011. The Principles of Testing. The ITEA Journal 32 (3): 231-234.

------ 2012. The Design and Analysis of Experiments (8th ed.). Hoboken, NY: John Wiley & Sons, Inc.

Shadish, W. R., Cook, D. T., and Campbell, D. T. 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Boston, MA: Houghton Mifflin Company.