



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

JUN 26 2013

OPERATIONAL TEST
AND EVALUATION

MEMORANDUM FOR COMMANDER, OPERATIONAL TEST AND EVALUATION
FORCE (COMOPTEVFOR)

SUBJECT: Flawed Application of Design of Experiments (DOE) to Operational Test
and Evaluation (OT&E)

In October 2010 I communicated my expectations regarding the use of DOE for developing rigorous, adequate, and defensible test programs and for evaluating their results. Over the past several years, all of the operational test agencies have implemented DOE practices to varying degrees and have offered training to their staff on the statistical principles of DOE. However, I am concerned that OPTEVFOR is not complying with the intent of the use of DOE as a method for test planning, execution, and evaluation. I find that most test designs focus exclusively on verifying threshold requirements, rely too heavily on hypothesis tests for test sizing, and all too often do not embrace the statistical tenets of DOE. Furthermore, OPTEVFOR has not updated its data analysis practices to capitalize on the benefits of using DOE.

One of the most important goals of operational testing is to characterize a system's (or system of systems') end-to-end mission effectiveness over the operational envelope. Such characterization of performance informs the Fleet and the system operators of its capabilities and limitations in the various conditions that will be encountered during combat operations. 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 experimental design (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.

Test designs and integrated evaluation frameworks (IEFs) developed by your staff will improve by following the direction provided in the remainder of this memorandum.

1. A clear test goal must be created for each phase of test.

As I state in previous guidance, as well as in the recently promulgated Test and Evaluation Master Plan (TEMP) Guide, a successful test plan must identify the goal of the test. Goals should be clearly identified in the TEMP as well as the test plan, and should be specific. Future test plans must state clearly that data are being collected to measure a particular response variable (possibly more than one), in order to characterize the system's performance by examining the effects of multiple factors. Test plans must also clearly delineate what statistical model (e.g., main effects and interactions) is motivating the strategic factor variation of the test.



An example of a clearly stated test goal is the following: “The purpose of the operational test is to characterize miss distance under a range of operationally realistic conditions. The characterization of mission performance through the primary response variable miss distance will be determined using a main-effects and second-order interaction model; the following factors will be strategically controlled during the test:...”

2. 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 response variables are limited to the technical requirements of the system under test when, in fact, the mission capability requires a system of systems to succeed. Ideal OT&E metrics should provide a measure of mission accomplishment (not technical performance for a single subsystem), lend themselves to good test design (i.e. be continuous in nature), and in general comprehensively cover the reasons for procuring the system.

Another pitfall to avoid is relying on binary metrics as the primary response variable on which to base the test design. I expect that even if the requirements document defines a probability-based metric, that great effort is expended to find a related continuous measure on which to base the test design. We cannot afford tests based solely on the evaluation of probability-based metrics. I note that several OPTEVFOR test designs have successfully found continuous metrics on which to base the test design (such as LCS, P-8A, Virginia class submarine, and LHA-6); such practice should continue and extend to all test designs where feasible.

3. Factors varied in the test must not be limited to those defined in the requirements documents.

Fleet 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 system will actually be employed. For example, consider a threshold that is only based on one threat type; it is not acceptable to base the operational test on that one threat, especially if performance is expected to vary against other threats. The threat type must be considered as a factor in the design and all operationally relevant threats should be accounted for in the levels of that factor. Using DOE techniques will allow for a fiscally-sound, robust evaluation of the requirement *alongside* of a relevant and feasible characterization across the operational envelope. Replicating runs/events under one condition (such as that identified in the requirements document) is much less informative than a strategic spread of runs/events across all relevant operational conditions.

For example, the P-8A Increment 2 events should include all relevant operationally realistic environments and threats despite the fact that the requirements document specifies some environmental conditions as an objective, or does not mention specific threats. We know that the Fleet will employ P-8A sensors against threats and in environments other than those defined by the threshold requirement; hence, the test program must characterize that performance. Similarly, for the Aegis Baseline 9, the experimental design for modeling and simulation (M&S)

runs to evaluate probability of raid annihilation (P_{ra}) must not be constrained solely to the engagement and environmental cases described in the requirements document. While the design is robust for purposes of evaluating compliance with requirements, it will not support the characterization of P_{ra} performance across the varying levels of all the operationally relevant factors. For the draft Evolved Sea Sparrow Missile (ESSM) test design, performance against single target raids is the proposed focus of the test because of the limited definition of the requirements; yet, we know the Navy uses ESSM to engage multi-target raids of two or more missiles.

4. Important factors must be strategically controlled in test design and execution.

Although OPTEVFOR's mission-based 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. As an additional benefit, controlling significant factors helps us to better explain the variance in the data, leading to more meaningful test results. For factors that truly cannot be controlled, OPTEVFOR should evaluate the effect on the confidence and power for other factors (since choosing not to control a factor often decreases the power to observe other factor effects of the test), and should ensure the recordable conditions are included in the analysis.

As an example, in the draft LHA-6 IEF, although the test design varies threat presentation between single axis and multiple axes, the actual arrival angle of the threat is listed as a recordable condition. This ignores the fact that some threat presentations, depending on the arrival angle, may produce different results due to the asymmetry in the ship's gun configuration. Furthermore, the targets are being remotely controlled during the test, so target aspect should be strategically controlled (not simply recorded) to obtain the conditions we desire.

5. Confounding factors must be avoided.

I have observed cases where the test plan varies factors in such a way as to confound the effects, even when it was eminently feasible to avoid doing so. For example, during the Advanced Precision Kill Weapon System (APKWS) testing all of the low altitude runs were executed at short slant ranges and high altitude runs at long slant ranges. This made it impossible in the analysis to determine whether altitude or slant range was the primary cause of the observed difference in performance. 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. I understand that physical limitations, non-repeatable ocean conditions, and test infrastructure might force a confounded test design in some cases; however, these cases are to be strenuously avoided.

6. Single hypothesis tests must not be used to size the test program.

Although OPTEVFOR rightly focuses on the use of power and confidence to size a test, the intent of my guidance has not been achieved. If the goal of the test is to characterize

performance across the primary factors, then the number of runs/events necessary to accomplish this goal cannot be determined from a hypothesis test which compares a single average across all conditions to a single threshold requirement (or average historical performance). The hypothesis test is lacking in several respects:

- Most importantly, a single hypothesis test fails to identify differences in performance across the operational envelope. This is essential for meeting one of the primary objectives of OT---characterizing a system's end-to-end mission effectiveness across the operational envelope.
- The average performance across all conditions is typically a meaningless analytical result; not only should this be avoided in the analysis of the data, it should be avoided in test planning. The goal of the test is not to "roll up" all the results across all conditions and compare to a single number derived from a requirements document (which was likely defined under a different set of conditions). An extreme example of this would be if the performance in one set of conditions is zero and in another is perfect, the average of 50 percent across the environments would be meaningless.
- Similarly, the effect size selected for a single hypothesis test is not typically meaningful. This results in misleading characterizations of power and produces a false sense of adequacy for the test, especially when several factors are expected to result in performance differences.
- It is rare that the requirements document threshold, which is used as the null hypothesis, is a true threshold, in the sense that we must ensure our tests reject systems that cannot be proven to be above threshold.

In some cases a single hypothesis test might be warranted: tests where it is impossible to vary factors, or in cases where the threshold requirement must be met and we must be confident in correctly rejecting the null hypothesis (e.g., body armor). In those cases, a series of operating characteristic curves for various test sizes is warranted to determine test risk, as opposed to the use of a single poorly-defined effect size, fixed confidence level, and single power calculation. In all other cases, an estimate of the precision of the test must be provided for assessing the adequacy of characterizing the system's performance across conditions. This will typically be in the form of power estimates to detect factor effects; however, other methods might be acceptable provided they capture the accuracy of the test's ability to characterize performance. I also expect the size of the anticipated factor effects to be justified.

A "roll-up" power calculation should not be used except in rare cases. Unfortunately, nearly every test design from OPTEVFOR has used this approach. Although a few recent test designs have successfully avoided this practice, such as Integrated Defensive Electronic Countermeasures and Joint and Allied Threat Awareness System, they are the exception and not the norm.

7. Understand that adding a condition/level/factor does not necessarily increase the size of the test.

Although this effect is true for 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. A strategic spread of resources across all conditions, particularly if factors are continuous in nature, ensures that the precision of the test remains acceptable without undue replication in every test condition. Therefore, important factors should not be eliminated from the test design (or made recordable) 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. In cases where the level of a factor is not physically attainable due to resource limitations (e.g., a particular threat), your office should work with mine to determine how we can overcome that limitation. However, it is unacceptable to eliminate these cases completely from evaluation.

The OPTEVFOR Operational Test Director manual provides guidance that suggests it is always appropriate to reduce factor space, claiming that “adding conditions unnecessarily ... can exponentially grow the size of the test,” and “In general, more levels result in more test requirements.” This guidance is inappropriate and indicates an unfortunate misunderstanding of our responsibilities to conduct adequate operational testing, as well as a misunderstanding of the appropriate use of DOE.

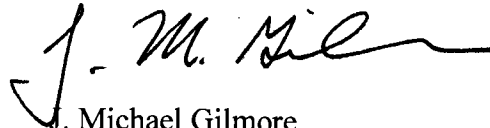
8. Understanding that removing a condition/level/factor does not significantly decrease the size of the test.

A test program based on DOE requires all of the test events across the various conditions to be accomplished to achieve the precision we seek in determining operational performance. In many test designs (but not all), the factors are continuous in nature; testing in one condition (level) helps to determine the common variance of the response variable across all conditions. Eliminating runs in one set of conditions, therefore, will reduce the precision of the measurement in other conditions when employing a statistical model (a key tenet of DOE analysis). In cases where it becomes apparent that a factor or level should be eliminated, the test design (including the originally determined sample size) must remain adequate to characterize performance across the remaining factors.

For example, if in the P-8A test design we agree to eliminate one of the many test environments for a test period, the overall sample size should not decrease; this will ensure that the desired precision for other factor effects is maintained.

By correcting these misconceptions in the implementation of my guidance on DOE, OPTEVFOR will improve the rigor of TEMPs and test plans, facilitating my office’s ability to approve those plans. I suggest that OPTEVFOR expand the statistical training available to its workforce, hire targeted government personnel with DOE expertise, and revise the Framework development process so that it is less focused on requirements verification and more on characterizing performance across the operational envelope.

I am dedicated to working with OPTEVFOR to resolve these concerns. Please provide me an update within 60 days regarding your specific proposals for improving the application of DOE at OPTEVFOR. My point of contact for these issues is Mr. John Allen, 703.372.3803, john.allen@osd.mil.

A handwritten signature in black ink, appearing to read "J. M. Gilmore". The signature is fluid and cursive, with a long horizontal stroke at the end.

J. Michael Gilmore
Director

cc:
Commanding General, ATEC
Director, MCOTEA
Commander, AFOTEC
Commander, JITC