

Design of Experiments Applied to Flight Testing

Leslie L. Bordelon

U. S. Air Force Senior Executive Service Retired
38470 Shadow Valley Lane
Yucaipa, California 92399-7010
USA

lbordelon@adelphia.net

ABSTRACT

Every testing organization should formulate testing philosophy and testing approaches to assure scientific validity and maximum efficiency in accomplishing ground and flight tests. Test design is not an art...it is a science. Decisions are too important to be left to professional opinion alone...they should be based on mathematical fact. DOE is a systematic approach to investigation of a system or process. A series of structured tests are designed in which planned changes are made to the input variables of a process or system. The effects of these changes on a pre-defined output are then assessed. DOE is important as a formal way of maximizing information gained while minimizing resources required. It has more to offer than 'one change at a time' experimental methods, because it allows a judgment on the significance to the output of input variables acting alone, as well input variables acting in combination with one another. 'One change at a time' testing always carries the risk that the experimenter may find one input variable to have a significant effect on the response (output) while failing to discover that changing another variable may alter the effect of the first (i.e. some kind of dependency or interaction). This is because the temptation is to stop the test when this first significant effect has been found. In order to reveal an interaction or dependency, 'one change at a time' testing relies on the experimenter carrying the tests in the appropriate direction. However, DOE plans for all possible dependencies in the first place, and then prescribes exactly what data are needed to assess them i.e. whether input variables change the response on their own, when combined, or not at all. DOE experience in support of flight testing is reviewed with benefits and advantages shown.

1.0 INTRODUCTION

The mission of a test center is to conduct and support the research, development, test and evaluation of aerospace systems from concept to combat. Design of Experiments (DOE) is not new to United States Air Force (USAF) test and evaluation (T&E). The 53d Wing (53 WG) of Air Combat Command (ACC) at Eglin Air Force Base (AFB), Florida, has used experimental design on over 25 operations in the past 14 years. A sample of these operations includes IR sensor predictions, ballistics six degree of freedom initial conditions, threat and receiver testing, camouflage target flight tests, SCUD hunting tactics, medium range air-to-air missile hardware in the loop test facility validation, 30 millimeter ammunition over-age lot acceptance testing, 30 mm gun loading tests, missile simulation validation, helmet mounted sights and night vision goggles testing, fighter and bomber operational flight program flight tests, weapons accuracy testing, and electronic countermeasures development ground mounts [1,2]. These were operational test and evaluation (OT&E) test programs. During OT&E, the “ultimate customer” or the “warfighter,” conducts field tests, under realistic conditions, on any item (or key component) of weapons, equipment, or munitions for the purpose of determining the effectiveness and suitability of the weapons, equipment, or munitions for use in combat by typical military users; and the evaluation of the results of such tests [3]. Experimental design is aptly suited to operational testing because of the large number of factors (i.e.,

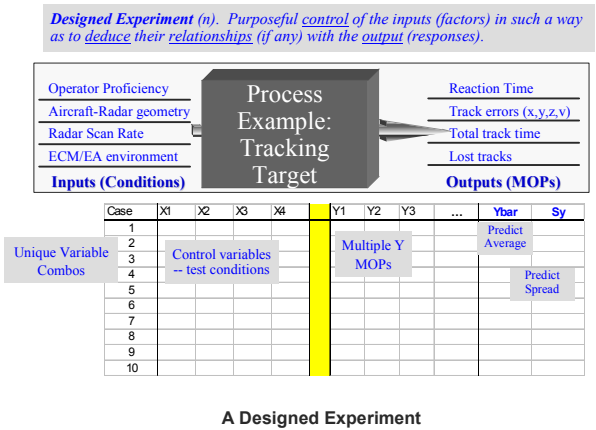
Bordelon, L.L. (2006) Design of Experiments Applied to Flight Testing. In *Mission Systems Engineering* (pp. 3-1 – 3-12). Educational Notes RTO-EN-SCI-176, Paper 3. Neuilly-sur-Seine, France: RTO. Available from: <http://www.rto.nato.int>.

variables) and levels (i.e., values) necessary to validate an aerospace system’s effectiveness and suitability to combat. Design of Experiments can also be used during DT&E where engineering-type tests are used to verify the status of technical progress, verify that design risks are minimized, substantiate or verify achievement of contract technical performance, and certify readiness for initial operational testing. These engineering-type tests generally require instrumentation and measurements and are accomplished by engineers and technicians in a controlled environment to facilitate failure analysis [3]. Although experimental design in most cases is aptly suited during DT&E, particularly when factors and levels are large, what is generally unknown is whether experimental design would play an equally large role during “gradual build-up tests” in which factors and levels are tightly controlled and limited such as during the envelope expansion of an aircraft’s performance and flying and handling qualities. During these tests, test teams have traditionally used one factor at a time (OFAT) as the preferred test strategy. Senior technical leaders at the U. S. Air Force Flight Test Center are assessing the use of experimental design as a possible test strategy for use in developmental test and evaluation (DT&E) flight test programs. As an example, this envelope expansion may be equally applicable when conducting a DT&E on an aircraft’s avionics systems or on a system of systems. This paper describes design of experiments and its application to overall flight testing, both developmental and operational.

2.0 WHY CONSIDER THE USE OF DESIGN OF EXPERIMENTS

Every testing organization should establish policy and procedures and should formulate testing philosophy and testing approaches to assure scientific validity and maximum efficiency in accomplishing ground and flight tests. Despite the talent of the typical test engineer, clearly they are generally experts in their specific engineering disciplines, more overall experience in test design needs to be present. Test design is not an art...it is a science. Decisions cannot be left to professional opinion alone. These decisions should be based on mathematical fact. DOE provides the means to bring this fact into the testing process.

Design of Experiments is a systematic approach to the investigation of a system or process. A series of structured tests are designed in which planned changes are made to the input variables of a process or system. The effects of these changes on a pre-defined output are then assessed. DOE is important as a formal way of maximizing information gained while minimizing resources required. It has more to offer than 'one change at a time' experimental methods, because it allows a judgment on the significance to the output of input variables acting alone, as well input variables acting in combination with one another.



'One change at a time' testing always carries the risk that the experimenter may find one input variable to have a significant effect on the response (output) while failing to discover that changing another variable may alter the effect of the first (i.e. some kind of dependency or interaction). This is because the temptation is to stop the test when this first significant effect has been found. In order to reveal an interaction or dependency, 'one change at a time' testing relies on the experimenter carrying the tests in the appropriate direction. However, DOE plans for all possible dependencies in the first place, and then prescribes exactly what data are needed to assess them, i.e. whether input variables change the response on their own, when combined, or not at all. In terms of resource the exact length and size of the experiment are set by the design (i.e. before testing begins).

DOE can be used to find answers in situations such as "what is the main contributing factor to a problem?", "how well does the system/process perform in the presence of noise?", "what is the best configuration of factor values to minimize variation in a response?" etc. In general, these questions are given labels as particular types of study. In the examples given above, these are known as problem solving, parameter design and robustness studies. In each case, DOE can be used to find the answer, because the only thing that marks them different is which factors would be used in the experiment.

The order of tasks to using this tool starts with identifying the input variables and the response (output) that is to be measured. For each input variable, a number of levels are defined that represent the range for which the effect of that variable is desired to be known. An experimental plan is produced which tells the experimenter where to set each test parameter for each run of the test. The response is then measured for each run. The method of analysis is to look for differences between response (output) readings for different groups of the input changes. These differences are then attributed to the input variables acting alone or in combination with another input variable.

DOE is team oriented and needs a variety backgrounds (e.g. design, manufacturing, statistics etc.) to be involved when identifying factors and levels and developing the matrix as this is the most skilled part. Moreover, as this tool is used to answer specific questions, the team should have a clear understanding of the difference between control and noise factors.

In order to draw the maximum amount of information a full matrix is needed which contains all possible combinations of factors and levels. If this requires too many experimental runs to be practical, fractions of the matrix can be taken dependent on which effects are of particular interest. The fewer the runs in the experiment the less information is available.

DOE refers to the process of planning the experiment so that appropriate data that can be analyzed by statistical methods will be collected, resulting in valid and objective conclusions. The 53d TMG has "decomposed" this process in what Mr. Greg Hutto describes as "DOE Process Steps—12 Steps – 4 Blocks." [1, 9]:

- I. Project Description and Decomposition
 1. Statement of the Problem—why this, why now?
 2. Objective of the Experiment—screen, model, characterize, optimize, compare....
 3. Response Variables—the process outputs (continuous are greatly to be desired).
 4. Potential Causal Variables—brainstorm and look for underlying physical variables.
- II. Plan the Test Matrix
 5. Establish Experimental Constraints—duration, analysis cycle time, reporting milestones, etc.
 6. Rank Order Factors—experimental, fixed, noise (both controlled and uncontrolled).
 7. Select Statistical Design—factorial, 2^k , fractional factorial, nested, split plot....
 8. Write test plan to include sample matrices, sample data and sample output.
- III. Produce the Observations
 9. Randomize run order and block as needed.
 10. Execute and control factors. Respond to unexpected. Record anomalies.
- IV. Ponder the Results
 11. Acquire, reduce, look, explore, analyze....
 12. Draw conclusions, redesign, assess results and plan further tests.

Based on the above, DOE can provide a another tool in the tester's arsenal to ensure that their testing philosophy and testing approaches to are scientifically valid and they maximize efficiency in accomplishing ground and flight tests

3.0 WHAT'S THE BEST METHOD OF TEST

In 1906, W. T. Gosset, a chemist at the Guinness Brewing Company drew a culture sample to determine how much yeast should be used in the brewing process. If he guessed too little, the result would be incomplete fermentation. If he guessed too much, the result would be bitter beer. He wanted to get it right. In 1998, Mike Kelly, an engineer at a large contact lens company, drew a sample for a 15K lot of contact lens to help determine how many defective lens existed in the lot. If he guessed too little, the result would be mad customers. But if he guessed too many, the result would be the destruction of good product. He, too, just wanted to get it right. What is the best method of test?

The use of DOE dates back into the 1800's with hypothesis tests and least squares tests done by Gauss and Legendre. The first statistician to consider a methodology for the **design of experiments** was Sir Ronald A. Fisher. He described how to test the hypothesis that a certain lady could distinguish by flavor alone whether the milk or the tea was first placed in the cup. While this sounds like a frivolous application, it allowed him to illustrate the most important means of experimental design:

- Randomization - The process of making something random
- Replication - repeating the creation of a phenomenon, so that the variability associated with the phenomenon can be estimated
- Blocking - the arranging of experimental units in groups (blocks) which are similar to one another
- Orthogonality - Means perpendicular, at right angles or statistically normal.
- Use of factorial experiments instead of the one-factor-at-a-time method

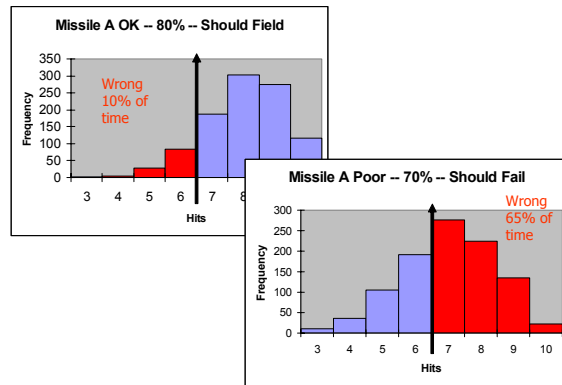
Analysis of the design of experiments was built on the foundation of the analysis of variance, a collection of models in which the observed variance is partitioned into components due to different factors which are estimated and/or tested.

3.1 The Challenges in Testing

The central challenge of test is to characterize what is to be tested. Testing a new or improved system is like reaching into a bowl (reality) and trying to draw a sample of operational performance. There is no advanced knowledge of what's in the bowl or even what bowl has been provided: the one where the system works or the one where the system does not. Sometimes, there is little in the way of characterization available and at other times there is too much characterization. Both situations can make decision making difficult, costly, and inefficient.

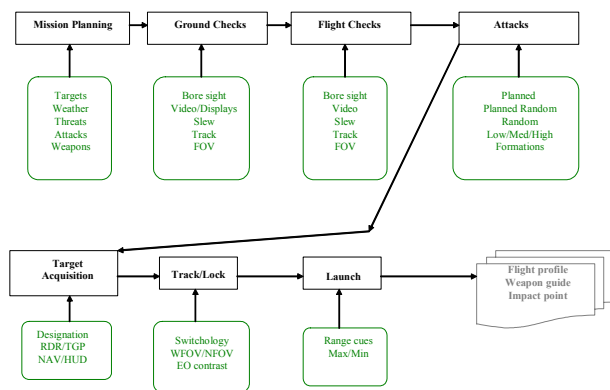
An example would be to consider an improved air-to-air missile system that requires testing—Missile A Improved. Suppose the original Missile A had an historical hit rate of 80%. The issue that prevails is whether the improved missile will be at least equal to or better than the original in hitting the target successfully. How many shots do we need to make to determine the performance of the improved Missile A?

Starting with a blank sheet of paper, the test engineer must define the appropriate number. But what is the number of shots necessary to verify the improved Missile A. Maybe the number is 3, because that is what the available time or money will support. Maybe the number is 8 because the engineer just likes 8. Maybe the number is 10 because the engineer is challenged by fractions. Or maybe the number is 30 because in life something good happens at 30! There is no statistical backing for any of these numbers, but all remain possibilities. For no particular reason, the engineer chooses 10. What are the results?



Decision to Field Or Fail

The test engineer now faces the second challenge in seeking to balance the chance of errors. Basically, this will be a trade of one error for the other to determine the success or failure within acceptable level of testing risk. Increasing the sample size will decrease the risk but may result in cost and schedule increases. If the engineer just considers the number of hits versus misses, the decision to field the system is based on equalling previous performance of 80% hits. Is this the best criteria for the decision? If the engineer accepts a degradation of 10% in targeting accuracy as meeting the requirement, then the number of shots required is very large. However, if the miss distance becomes the main focus, then the number of shots can be reduced considerably, by a factor of approximately 10.



Battlespace Breadth: Missile A Attack Process

These two challenges, the effect of the sample size on errors or depth of test and measuring best response or measuring what is rich and relevant, must be faced by the test engineer. It clearly matters how many tests are done and what is measured. But there is a third challenge. The breadth of testing must also be determined. To define breadth, the engineer must search the employment environment or the battlespace for the system. This becomes more difficult

because the number of variables increases. But the focus must remain on getting it right: confidence in stating results and power to find small differences.

Design of Experiments can provide a method to reduce the complexity and define a more focused scope of the tests required of the system to help determine whether ready to field. DOE can provide purposeful control of the inputs (factors) in such a way as to deduce their relationships (if any) with the output (responses). In the case of Missile A, the inputs are the missile itself, the slew sensor (test points and radar), the target type, and the launch platform. The outputs are the trajectory, hits/misses, target damage, and miss distance.

Battlespace breadth must also be considered in determining whether the improved Missile A is indeed equal to or better the original. The test team (operations, analysts, program manager, and engineer) must break the process down. They must determine the events, steps, and outcomes of the test. The need to understand the operations choices and conditions of each step. Furthermore, they must know the size of the operations test space and how to measure success. The result will be a definition of the battlespace conditions.

Design of Experiments Applied to Flight Testing

In addressing the third challenge, that of understanding the breadth of testing required, the test engineer must now do a systematic search of the relevant battlespace. Using the method of factorial designs, the engineer can learn more from the same number of assets. Factorials can also be used to reduce assets while maintaining confidence and power. Additionally, the engineer can combine both methods.

The actual tests were a great success. The extensive captive carry tests included 22 sorties with approximately 100 simulated shots. The old and new seekers were included on each wing of the carrier aircraft to equalize weather conditions. Three different launch platforms were used with operations at both wet and dry ranges. The results showed an improvement of approximately two times the original acquisition and targeting range. There were 9 shots comparable to the current Missile A performance.

Mission: MAV-1						
Wet Range	Range Time: 0700-0900 (Dawn)	Target: Point (Tank)	Weather: As Scheduled			
Launch	Airspeed: 400 KIAS					
F-16 - #1	Left Wing: Missile A-1 on LR-2	Right Wing: Missile A-2 on LR-2				
Run#	Target Type	Altitude	Sun Angle	Missile Type	Cueing	Comments*
1	Tank/Truck	18,000'	Sun at 6	V2	Visual	
2	Tank/Truck	18,000'	Sun at 6	V3	"	
3	Tank/Truck	1500'	Sun at 3/9	V2	NAV/GPS	
4	Tank/Truck	1500'	Sun at 3/9	V3	"	
5	Tank/Truck	1500'	Sun at 6	V2	Radar	
6	Tank/Truck	1500'	Sun at 6	V3	"	
7	Tank/Truck	18,000'	Sun at 3/9	V2	Infra-Red	
8	Tank/Truck	18,000'	Sun at 3/9	V3	"	

F-16 - #2 Typical Missile A Run Card						
Run#	Target Type	Altitude	Sun Angle	Missile Type	Cueing	Comments*
1	Tank/Truck	18,000'	Sun at 6	V2	Visual	
2	Tank/Truck	18,000'	Sun at 6	V1	"	
3	Tank/Truck	1500'	Sun at 3/9	V2	NAV/GPS	
4	Tank/Truck	1500'	Sun at 3/9	V1	"	
5	Tank/Truck	1500'	Sun at 6	V2	Radar	
6	Tank/Truck	1500'	Sun at 6	V1	"	
7	Tank/Truck	18,000'	Sun at 3/9	V2	Infra-Red	
8	Tank/Truck	18,000'	Sun at 3/9	V1	"	

* After simulated pickle, simulate missile flyout by overflaying target and recording seeker video.

Actual Missile A Test – A Great Success

3.2 Which Method Is Best

The question of “What’s best” always arises because many methods have been tried with successes and failures attributed to each method. Characterizing, as an example, a Radar Target Location Error test can provide additional insight into the benefits of DOE as a tool to the flight testers.

The process to be tested is radar ground mapping. The inputs are altitude, target radar cross-sectional area, angle off-nose, aircraft tail number, target range, calibration date of the radar, time of last doppler update, target elevation, and operator skill level. Output is target location accuracy. The test team’s methods include intuition, one factor at a time (OFAT), scenario (best guess or case), do what we did last time (DWWDLT), or rent-a-test info sheet (TIS). A question immediately arises, “How big is the test space?” Attempting to test each possible combination when considering the nine factors above at two levels each (angle off nose at 15 and 30 degrees), the result would be 2^9 or 512 tests. There has got to be a better way to examine the test space. Which method is the best?

One way to examine the test space is to rely on intuition or subject matter expertise. Intuition can greatly benefit a test, however, it requires deep subject matter knowledge that is not always available. Typically, using this approach, the engineer only “discovers” what is already believed to be true. Many times this method lacks objective proof of conclusions reached from the test results. Intuition can assist with test design, but should never be the total strategy, as it lacks proof. Intuition always plays a part in any test, for example, it is how we come up with variables and conditions. Intuition is also critical to successful analysis effort.

Another methodology is referred to as “One Factor at a Time or OFAT.” Basically, OFAT begins with finding a nominal configuration for all factors, 9 in our example, and performing the initial testing based on these parameters. The

We're going to run 10 because we're 10-guys and that's what we do.

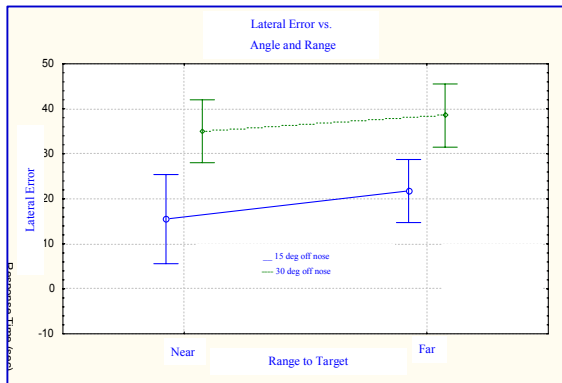
Case	A	B	C	D	E	F	G	H	J	Y1	... Y10	Avg
1	low	low	low	low	low	low	low	low	low			
2	hi	low	low	low	low	low	low	low	low			
3	best	hi	low	low	low	low	low	low	low			
4		best	hi	low	low	low	low	low	low			
5			best	hi	low	low	low	low	low			
6				best	hi	low	low	low	low			
7					best	hi	low	low	low			
8						best	hi	low	low			
9							best	hi	low			
10								best	hi			
11									best			

Total number of combinations = ?
At two levels each variable we have $2 \times 2 \times 2 \dots = 2^9 = 512$ combinations

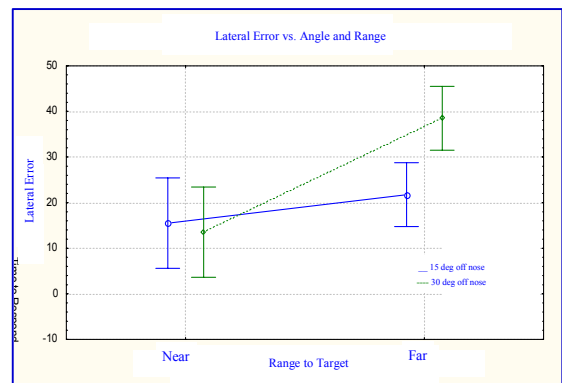
Using “One Factor At a Time” Design

first factor is then varied with the other factors held constant and the test is performed again. With these results, the best setting for the first factor is determined and changed for the remainder of the test. The next factor is varied and testing continues to determine its best value. The process continues until all factor settings have been determined. Intuition is used many times to define the number of repetitions. In this case, for example, 10 repetitions have been decided upon. Applying OFAT to the radar test would result in 11 combinations with 10 repetitions for each combination. The problem with this approach is that there are 501 other combinations untested with no mathematics model to predict their responses. Therefore, only samples a small percentage of the possible combinations are tested with no math model to help point to those combinations which might be most meaningful. Clearly, interactions also existing in all test situations, but OFAT cannot detect interactions. As a simple example from safety/test control, varying all the factors all of the time introduces safety concerns into any test and many times the changes preclude determining what went wrong if a test failed.

As an example, OFAT can be applied to baking cookies. There are two factors, time and temperature. The temperature can be set with the time varied or the time set with the temperature varied. The objective is to maximize “cookie goodness.” Graphing the results shows the contours of the various temperature versus time data. If contours are not aligned with the axes, then the optimum result is missed. The problem is not simply to find the best time and the best temperature—it is also to find if the variable interact as well. Response contours are often ridges, saddles, and other shapes at an angle to the control axes. Using orthogonal DOE and factorial design, the true optimum turns out to be 90+% versus the OFAT result of 80%.



Aircraft A Radar: No interaction



Aircraft A Radar: Interaction

In returning to the radar example, the weakness of OFAT in detecting interaction between variables can be shown. Where variables are independent of each other, as seen by parallel slopes of A with B—the response at level of B does not depend on level of A and the variable settings can be addressed independently. “15 degrees off nose always best” is the initial conclusion and can be modeled simply by $Y=b_0+b_1*Range+b_2*Angle$. With nonparallel slopes of A with B—the response at the level of B does depend on the level of A. The variable settings must be addressed together—“long range targets detected best at 15 degrees off nose.” A more complex model with interaction term is needed— $Y=b_0+b_1*Range+b_2*Angle+b_3*Range*Angle$.

The last methodology to consider is scenario or case design. Factors are defined based on scenarios. Examples are easy versus hard, low versus high, large versus small, etc. Taking the 11 combinations and assigning scenario based results raises new questions. Many times the determination of easy versus hard, or high versus low, etc., is based on a best guess, a magic number, or good feeling of someone on the team. In these cases, determination of what is right or best becomes very difficult to conclude. As an example, building scenarios using the radar target location error test results in a considerable number of

cases or scenarios to be evaluated in order to address all variable combinations. It is not efficient or effective to test all combinations, therefore only a fraction of the combinations are tested and the effects of combined variables remains unknown.

3.3 How To Build a Developmental or Operational Test

DOE helps the test team to interrogate the overall process and to improve the knowledge of how the process works. The goal is a systematic method to efficiently and unambiguously improve the outcomes. Compared to any other systematic method, current experience within the U. S. Air Force test community shows that DOE designs yield better process understanding, can be planned and analyzed faster, and are cheaper, using 20-80% of usual resources.

Designing the test using DOE is done in four basic steps as described earlier: 1) Develop the Project Description and Decomposition, 2) Plan the Test Matrix, 3) Produce the Observations, and 4) Ponder the Results.

Step 1 begins with defining the test objective. To do this, the problem must be clearly understood. The process flow needs to be defined including the inputs and the outputs. Casual variables that can affect the results must be defined as well. Test objectives directly influence test designs. The following considerations such be included: comparing the system to a fixed standard (ORD, specification, demo goal), comparing the system to older version (OFP or software/hardware upgrade, new mission data, etc.), characterizing system performance (test first, then evaluate later), optimizing the performance of the system under test (maximize or minimize), making the system more robust to environmental conditions (weather, threat actions, countermeasures, operator experience), minimize the variability of system performance, and troubleshooting faulty system performance (false alarms).

Step 2 is the plan or test matrix. This needs to include the identification of constraints—duration, analysis cycle time, events/sorties, etc. The factors need to be prioritized—control, fixed, noise, etc. A statistical design method needs to be decided upon—factorial, 2^k , etc. Basically a test design needs to be formulated and an algorithm built for the test event under consideration. An easy algorithm constructs simple factorial designs. Designs can be simple or complex, depending on the number of variables, the interaction between them, and how these affect the outputs.

Step 3 is to produce the observations. This is simply running the tests. Randomizing the run order is important in that it protects the results from unknown background changes within an experimental period (according to Fischer), the experimental period being the time of the test event. The design should also be blocked to protect from unknown background changes between experimental periods.

Step 4 is pondering the results. The data acquired needs to be gathered, reduced and reviewed, fully analyzing to fully understand the test results. Results need to be fully assessed and conclusions need to be drawn to support decisions to end testing, continue testing or redesign the test and continue.

DOE can interrogate the process and improve the tester's knowledge of how the process works. The goal is a systematic method to efficiently and unambiguously improve our outcomes. Applying DOE to an example can show more clearly the benefits of its use by the test team. The example that follows shows the DOE Test Design Process for testing a modification to an existing radar system to precisely locate and identify targets:

3.4 DOE Test Design Process Example

I Project Description and Decomposition.

1. Statement of the Problem. A concise description of desired outcome and factors affecting it.

These modifications are a set of hardware and software updates to the signal processing algorithms of existing radars... The modifications allow the system to range and angle-locate threats much more accurately than previously, using one of three or four algorithms. The results are a displayed geo-location of the threat with an uncertainty region that changes as the system's confidence in the threat location changes. The current technology demonstration is intended to demonstrate the feasibility of this geo-location feature of the system.

2. Define the objective(s) of the test.

- screening to find important factors affecting response.
- comparing results to a known baseline or standard.
- fitting a function of the variables to the response (modeling).
- finding an optimum setting of variables.
- finding factors that affect dispersion vs. location of response.
- In this step, the precision and confidence desired in the test results is defined.

In the test the mean and variance of the location error with respect to several variables thought to affect the location accuracy need to be characterized.

3. Define the potential response variables, including expected noise levels in measurements.

A number of response variables were considered and ultimately rejected for the system. Some examples --

<i>Response Variable (MOE)</i>	<i>Comment</i>	<i>Range Expected</i>	<i>Precision and Accuracy of Measurements</i>
<i>Location accuracy</i>	<i>Perceived minus actual, radial distance in feet</i>	<i>1% to 10% of range to threat</i>	<i>Actual precise to 5 feet, perceived TBD. Aircraft position uncertainty unknown.</i>
<i>Response time</i>	<i>Time from threat radiate to display</i>	<i>2-15 seconds</i>	<i>Not representative of real time operating system</i>
<i>Threat ID</i>	<i>Correlation of de-interleaved stream and threat name.</i>	<i>0 or 1</i>	<i>Not addressed in this flight test – CPU load too high for breadboard system. Treat each threat as a beam. Address in ground mount</i>
<i>Age out, detection range, others</i>	<i>Not considered technical challenges</i>		

Design of Experiments Applied to Flight Testing

As a comment, the test plan listed about 6 or 7 responses with goals for each. During DOE flight test planning, it turned out there was only one real response of concern to the technology demonstration.

- 4. Define all potential test factors and levels for each factor.** Steps 2 and 3 are often combined through process flowcharting using Ishikawa (fishbone) diagrams. Draw the process flow – inputs and outputs and basic states and transitions

<i>Test Variable</i>	<i>Definition</i>	<i>Range of interest</i>	<i>Method of Control (accuracy)</i>
<i>Range to threat (signal strength)</i>	<i>Range from aircraft position to threat antenna</i>	<i>5 to 30 miles</i>	<i>Flight path +/- 5%. Continuously changing</i>
<i>Receive antenna</i>	<i>One of four antennas primary</i>	<i>For and aft port and starboard</i>	<i>Flight path quadrant to threat</i>
<i>Closure rate (signal strength)</i>	<i>Positive or negative</i>	<i>At 45 degree aspect</i>	<i>Flight path – no radial paths.</i>
<i>Doppler change</i>	<i>Rate of radial distance change to threat</i>	<i>TBD</i>	<i>Not fixed levels – changing or constant</i>
<i>Threat beam/mode</i>	<i>5 threats and several modes each</i>		<i>Some more difficult than others – short PRI and acq radars for example</i>

II Plan the Test matrix.

5. Establish the experimental constraints:

- duration
- cost
- setup/teardown
- control of factors
- background environment
- instrument calibration, precision and accuracy

The team had up to four flight tests with a current fighter aircraft, for about 1.5 hours of range time each. The threats available and their modes and activity were up to the team. So were the profiles and the length of each leg, within the constraints of range space. The system could be reset as often as wanted, at about 3-5 seconds to reboot. Only one threat should be viewed at a time due to processor limitations in the brass board version.

6. Rank-order factors as “priority”, “background” or “fixed”:

<i>Test Variable</i>	<i>Control method: Priority, Background, or Fixed</i>
<i>Range to threat (signal strength)</i>	
<i>Receive antenna</i>	
<i>Closure rate (signal strength)</i>	
<i>Doppler change</i>	
<i>Threat beam/mode</i>	

7. Define the statistical design layout and replications needed. Completely randomized, factorial, response surface, Plackett Burman are some of the possibilities. Consider the analysis method as well, including transforms, bounding, use of ranks, etc.

A full factorial with repetitions was chosen for the design: $2^4 \times 6 = 96$ cases. There was no need to run fractions as there were enough legs to cover most all the threats in each mission.

8. Write the test plan (instruments, data reduction, personnel, computers, schedule, etc.) Especially consider analysis throughput cycle time from execution to record conclusions.

To speed analysis throughput, the developer coded each leg with a unique code so that response data could be merged with the test conditions. The reduction should take about 4 hours for the developer and 30 minutes for the test team. The test team allocated 4 hours for initial analysis and decision. This left about a work day for analysis cycle time. To cancel a range mission, a 48 hours warning was needed. So the total analysis cycle time was about 3 days, or 2 if evenings were worked.

III Produce the Observations

9. Randomize the runs as far as possible, blocking whenever randomization can not or should not be used.

For the most part, the runs were randomized, with the only exception being that the boomerangs were flown in order. We did change threats and jinking on a boomerang by boomerang basis and cycled through long and short range boomerangs repeatedly

10. Execute the plan, recording the results and annotating test conditions and aberrations.

In actual testing, the team found that there were several problems with the test design. First, the team could not reset the system frequently since the location did not converge to a good value quickly as originally predicted. Secondly, the long ranges were too long for the short range threats to track, so the team had to forgo the longer range tracks for them.

IV Ponder the Results

11. Acquire, reduce, validate, and analyze the data. Be sure to LEAP but verify:

- Eye-ball the data – plots and tables
- Determine the effects and check for violations of assumptions.

Design of Experiments Applied to Flight Testing

- Remedy any violations and re-determine effects.
- Check for practical as well as statistical significance.

The system had many problems in the data set as run. Many of the cells were empty either because they were not, or could not be flown, or the system did not perform at those settings. In the cases where there were data values, only one estimate per leg was acquired since the location estimates were not independent. Finally, the record typically contained huge outliers as much as 103x greater than the usual values, exploding the team's variances.

12. Assess results, run confirmations and excursions, conduct more experimentation.

Had the data actually been reviewed between missions, the team would have certainly ended early, lengthened the legs of the runs to allow for time to converge, and moved in the longer range legs for the short range threats. The team might also have investigated the strange behavior of the one threat that was not done well.

5.0 SUMMARY

The central problem of test is inferring what the real world is like based on the test samples. Two errors can be made—false positive or false negative. The test designs need to minimize these risks. A designed experiment is about the “Design”—the pattern of test conditions run by the test team. Variables become confounded when their effects on the responses cannot be separated. DOE provides a means to avoid confounding two or more variables. DOE has had an eighty year track record in all areas of science, engineering, and test. DOE becomes a test strategy in the overall test big picture.

6.0 REFERENCES

- [1] Hutto, G. (2004). *Test For Leaders—What You Should Know*. 53d Test Management Group, 53d Wing, Air Combat Command.
- [2] Hutto, G. & Akerson, J. (2004). *Raising the Bar: Equipping 53d Wing Test Teams to Excel with DOE*. 53d Test Management Group, 53d Wing, Air Combat Command.
- [3] Montgomery, D. (2001). *Design and Analysis of Experiments*. Fifth Edition. John, Wiley & Sons, Inc.
- [4] Hutto, G. (2004). *Basic Statistics Review: Foundations for Designed Experiments*. 53d Test Management Group, 53d Wing, Air Combat Command.
- [5] Triola, M. (2005). *Elementary Statistics*. Ninth Edition. Pearson, Addison Wesley.
- [6] Hutto, G. (2004). *Applied DOE I & II: ANOVA and Factorial Experiments*. 53d Test Management Group, 53d Wing, Air Combat Command.
- [7] Kailiwai, G (2005). *The United States Air Force Flight Test Center Experiment in Experimental Design*, Air Force Flight Test Center