air academy
ASSOCIATES

Knowing the International Language of T&E

Mark J. Kiemele, Ph.D.

ITEA Annual Symposium
Lihue, HI
November 13, 2019

Mark J. Kiemele, Ph.D.
Air Academy Associates
12295 Oracle Blvd, Ste 340
Colorado Springs, CO 80921
Phone: 719-531-0777
email: mkiemele@airacad.com

Some Terms in the T&E Language

- orthogonality
- replication
- randomization
- blocking
- screening
- modeling
- verification
- validation
- robustness
- interaction
- run
- level
- effect
- test design matrix
- factor
- test debt

What is Design of Experiments (DOE*)

Design of Experiments (DOE) is the

Planning,

Designing,

Conducting,

and

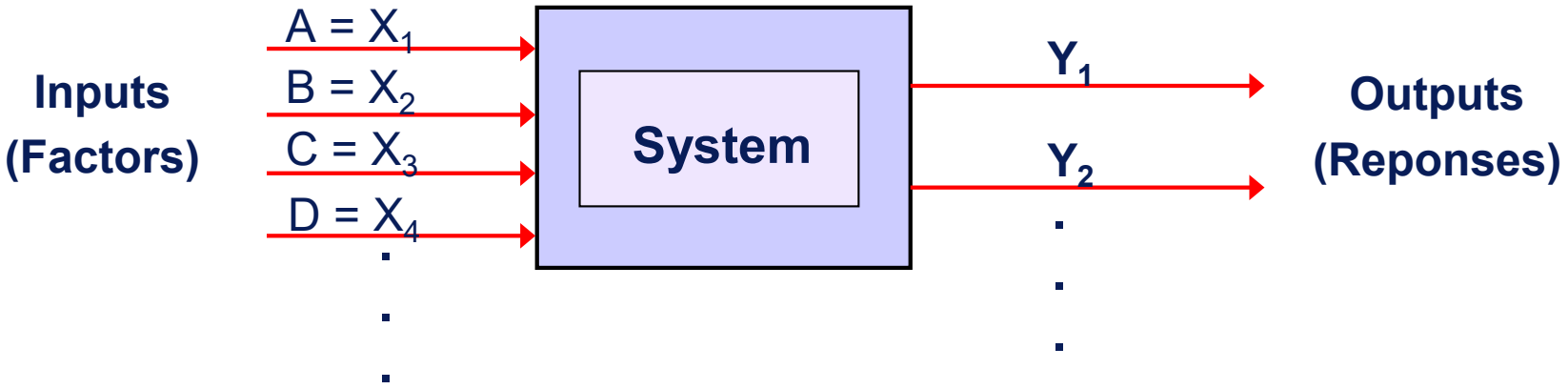
Analysis

of a **series of tests** in which purposeful changes are made to the input variables of a system or process and the corresponding effects on the output variables are measured.

* DOE may be implied by other names such as multi-variable testing, combinatorial testing, and statistical test optimization.

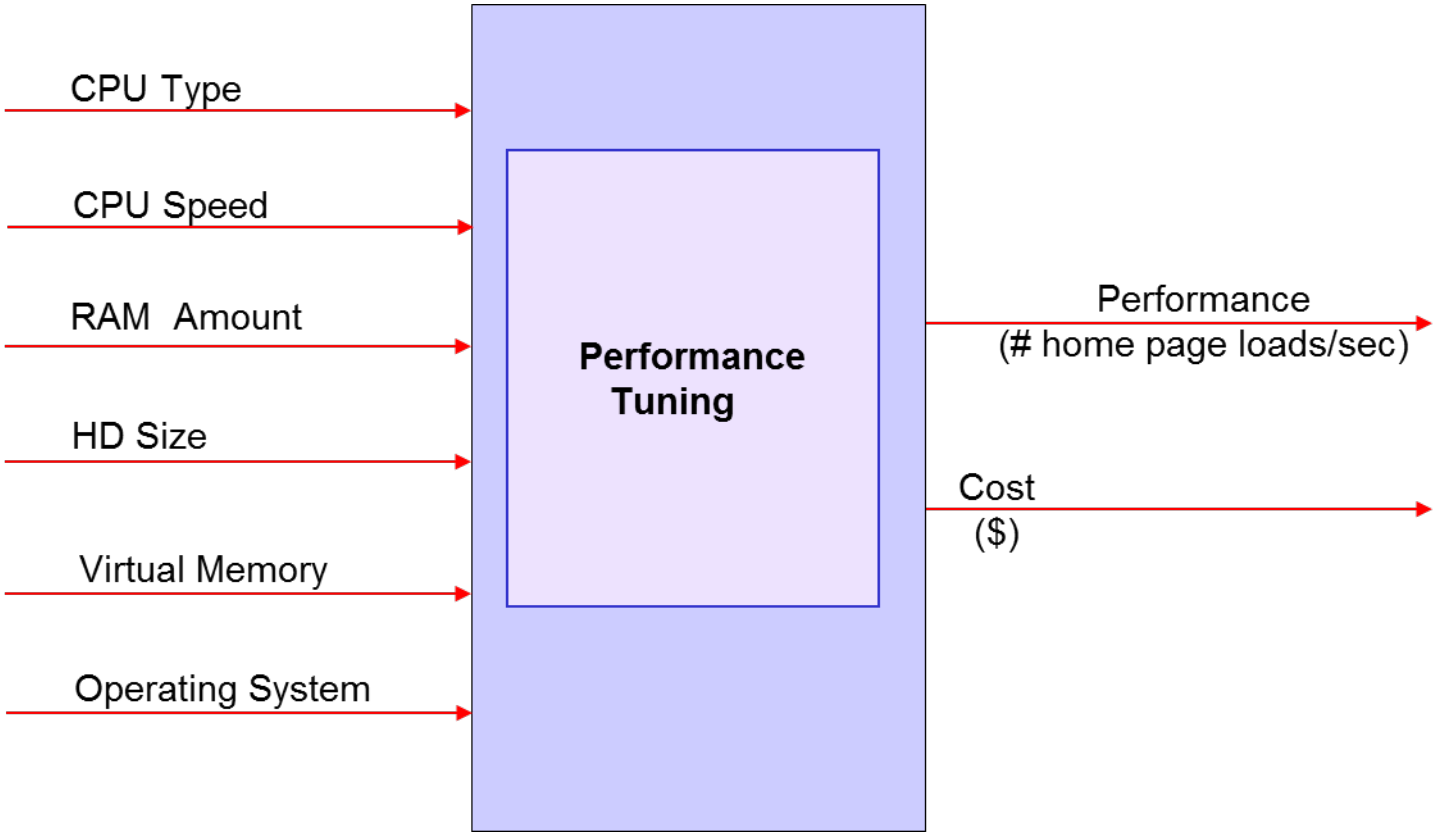
Who Would Benefit by Using DOE ?

Anyone who wants to understand the detailed relationships between the inputs to a system and the resulting outputs. That is, anyone who plans, designs, conducts, or analyzes tests.



DOE is applicable to both physical processes and computer simulation models.

Example of a Web-Based Test Scenario



Subject Matter Experts Must Be Involved

Factors/Inputs (X's)	Levels (Choices)	Response/Outputs (Y's)
CPU Type	Itanium, Xeon	# home page loads/sec
CPU Speed	1 GHz, 2.5 GHz	Cost
RAM Amount	256 MB, 1.5 GB	
HD Size	50 GB, 500 GB	
VM	J2EE, .NET	
OS	Windows, Linux	

Which factors are important? Which are not?

Which combination of factor choices will create operational problems?

How do you know for sure? Show me the data.

Terminology of DOE

Y: Output, response variable, dependent variable

X: Input, factor, test parameter, independent variable (a measurable entity that is purposely changed during a test)

Level: A unique value or choice of a factor (X)

Run: An experimental combination of the levels of the X's; a test case

Test Design Matrix: the collection of all test cases, also known as a covering array in software testing

Replication: Doing or repeating a test case

Effect: The difference or impact on Y when changing X

Interaction: When the effect of one factor depends on the level of another factor; also known as a combination effect

Test Design Matrix with Response Variables

Test Design Matrix

Run	CPU Type	CPU Speed	RAM Amount	HD Size	VM	OS
1	Itanium	1 GHz	256 MB	50 GB	J2EE	Windows
2	Itanium	1 GHz	256 MB	50 GB	J2EE	Linux
3	Itanium	1 GHz	1.5 GB	500 GB	.NET	Windows
4	Itanium	2.5 GHz	256 MB	500 GB	.NET	Windows
5	Itanium	2.5 GHz	1.5 GB	50 GB	.NET	Linux
6	Itanium	2.5 GHz	1.5 GB	500 GB	J2EE	Linux
7	Xeon	1 GHz	1.5 GB	500 GB	J2EE	Windows
8	Xeon	1 GHz	1.5 GB	50 GB	.NET	Linux
9	Xeon	1 GHz	256 MB	500 GB	.NET	Linux
10	Xeon	2.5 GHz	1.5 GB	50 GB	J2EE	Windows
11	Xeon	2.5 GHz	256 MB	500 GB	J2EE	Linux
12	Xeon	2.5 GHz	256 MB	50 GB	.NET	Windows

Response Variables

Page Loads/second	Cost

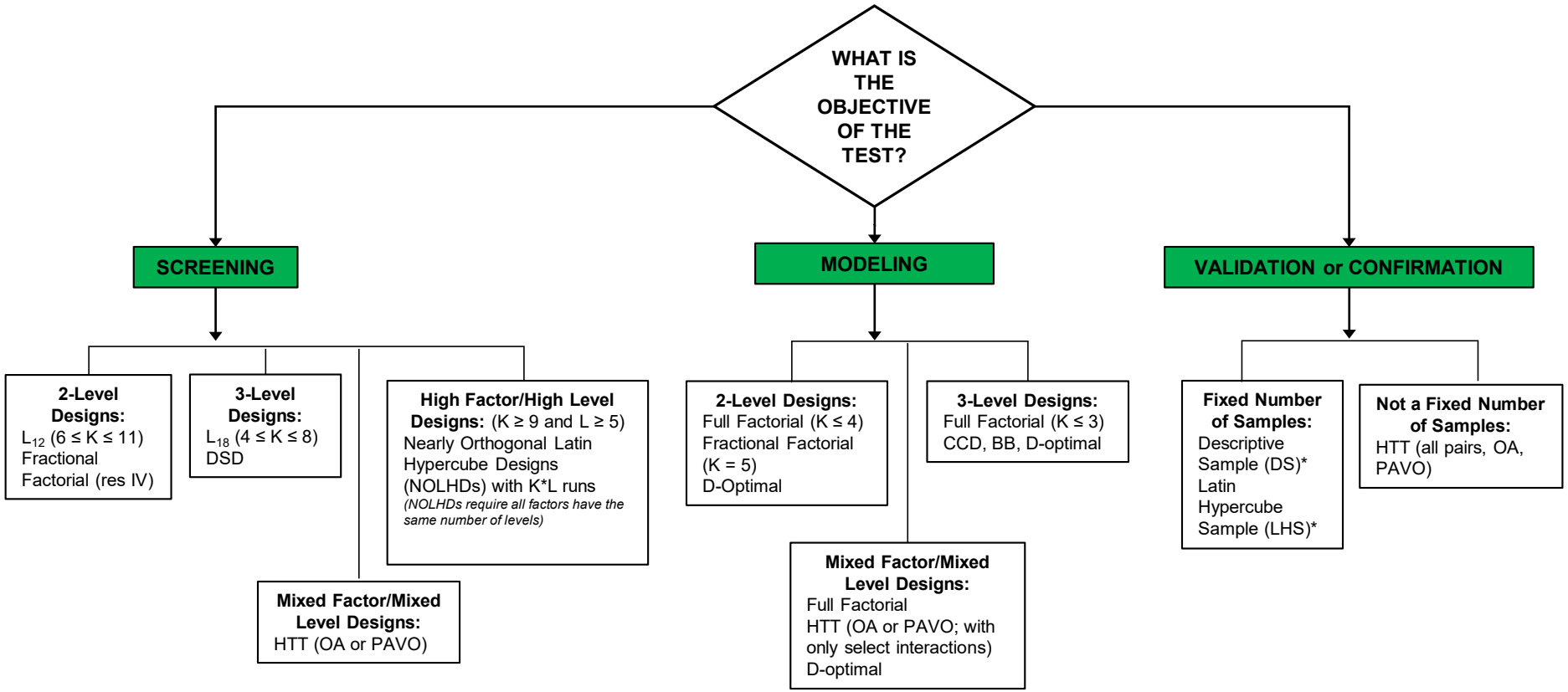
Major Test Objectives for which a DOE should be used

- **Screening Designs**
  - For testing many factors in order to **separate** the vital few factors from the trivial many.

- **Modeling Designs**
  - For building **functions** that can be used to predict outcomes, assess risk, and optimize performance. These **include** the ability to evaluate **interaction and higher order effects**.

- **Performance Verification and Validation Testing**
  - For **confirming** that a system performs in accordance with its specifications/requirements.

Various Options for Design Selection



Notes:

1. "Mixed" factors means a combination of quantitative and qualitative (categorical)
2. "Mixed" levels means that not all factors have the same number of levels (settings)
3. "K" = Number of Factors and "L" = Number of Levels
4. HTT = High Throughput Testing
5. DSD = Definitive Screening Design
6. "OA" stands for Orthogonal Array; "PAVO" = Pairwise Value Ordering
7. Software such as HD Tools™, rdExpert™ Lite, Pro-Test™ and Quantum XL™ generate some or all of these designs

* DS and LHS are sampling techniques to generate representative samples according to a specified distribution and a specified sample size

* Representative samples do <u>not</u> give orthogonal designs. They are often used for getting test coverage, validating performance/ determining capability, or creating noise combinations for test

Testing 47 Factors in a BMD Scenario
(using an Air Defense Simulation Model)

- Threat radar cross section
- Satellite cueing system probability of detection
- Satellite cueing system network delay
- Satellite cueing system accuracy
- Satellite cueing system time to form track
- GB upper tier time to acquire track
- GB upper tier time to discriminate
- GB upper tier time to commit
- GB upper tier time to kill assessment
- GB upper tier probability of correct discrimination
- GB upper tier probability of kill (Pk) assessment
- GB upper tier launch reliability
- GB upper tier reaction time
- GB upper tier Pk
- GB upper tier burnout velocity (Vbo)
- GB lower tier time to acquire track
- GB lower tier time to discriminate
- GB lower tier time to commit
- GB lower tier probability of correct discrimination
- GB lower tier 1 launch reliability
- GB lower tier 1 reaction time
- GB lower tier 1 Pk
- GB lower tier 1 Vbo
- GB lower tier 2 launch reliability
- GB lower tier 2 reaction time
- GB lower tier 2 Pk
- GB lower tier 2 Vbo
- SB lower tier time to acquire track
- SB lower tier time to discriminate
- SB lower tier time to commit
- SB lower tier time to kill assessment
- SB lower tier probability of correct discrimination
- SB lower tier Pk assessment
- SB lower tier launch reliability
- SB lower tier reaction time
- SB lower tier Pk
- SB lower tier Vbo
- Network delay
- Lower tier minimum intercept altitude
- Upper tier minimum intercept altitude
- ABL reaction time
- ABL beam spread
- ABL atmospheric attenuation
- ABL downtime
- GB upper tier downtime
- GB lower tier downtime
- SB lower tier downtime

Famous Quote

**"All experiments (tests) are designed;
some are poorly designed,
some are well designed."**

George Box (1919-2013), Professor of Statistics, DOE Guru

Approaches to Testing Multiple Factors

- **Traditional Approaches**
  - One Factor at a Time (OFAT)
  - Oracle (Best Guess)
  - All possible combinations (full factorial)

- **Modern Approach**
  - Statistically designed experiments (DOE) … factorial designs plus other selected DOE designs, depending on the situation

Statistically Designed Experiments (DOE):
The Modern Approach

- FULL FACTORIALS (for modeling a small numbers of factors)
- FRACTIONAL FACTORIALS (for screening or modeling)
- PLACKETT - BURMAN
- LATIN SQUARES
- HADAMARD MATRICES

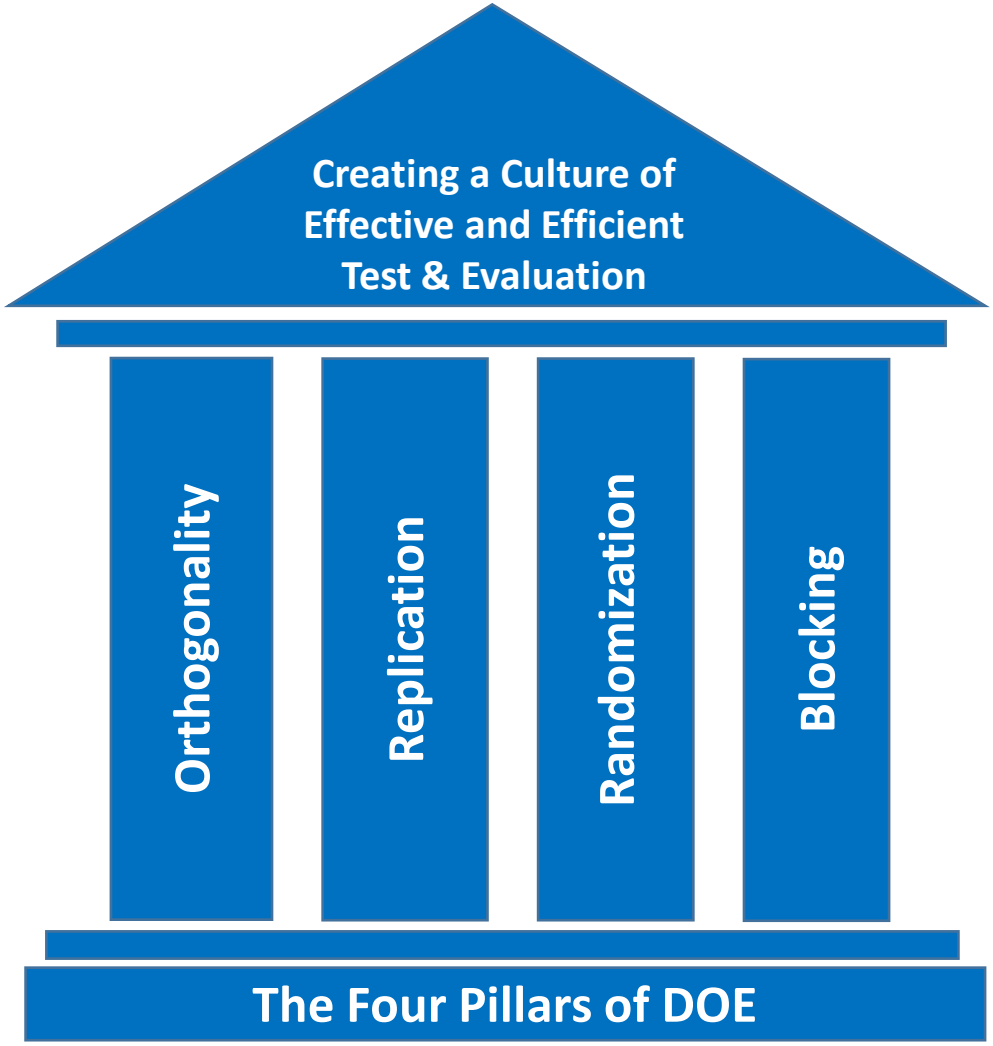
  } Taguchi Designs (for screening)

- BOX - BEHNKEN DESIGNS
- CENTRAL COMPOSITE DESIGNS

  } Response Surface Designs (for modeling nonlinear effects)

- HIGH THROUGHPUT TESTING (ALL PAIRS) (for high test coverage at low cost)
- NEARLY ORTHOGONAL LATIN HYPERCUBE DESIGNS (for screening or modeling computer simulators)

The Foundations of Design of Experiments (DOE)



Orthogonality

- This is the feature of a test design that allows for the ***independent*** evaluation of the effects of factors and their interactions – and nonlinear effects as well, depending on the type of design chosen.
- Why is independent evaluation so important? It gets us much closer to cause and effect relationships, and it makes the subsequent analysis of the data much easier.
- The difference between DOE and an observational study (historical data analysis) is the ability to do independent evaluation and arrive at causal relationships.
- Most leaders of organizations do not know this. They may know that DOE is important but they really don't know why. Orthogonality is a major reason why.

Replication

- Since DOE is about the study of variation, replicating or getting repeated measures for the same test condition allows us to study variation.

- One of the most common questions in all of statistics and process improvement is, “what should my sample size be?” That is, how much data do I need?

- The power of a statistical test is based on the sample size or the number of replications.

- Replication is a major attribute of any test design and thus a stalwart in DOE, because it increases the precision of the test.

Randomization

- The basic reason we randomize the test cases is to spread the noise (from factors we cannot control) as evenly as possible across the entire design space.
- In that way, noise factors do not become confounded or correlated with factors that are involved and controlled in the experiment.
- Randomization minimizes bias in the testing process.
- Depending on the test scenario, complete randomization may not be achievable because it may cost too much. Factors that are difficult to change or very expensive to change make randomization a real challenge.

Blocking

- Blocking is the arranging of test cases in groups (blocks) that are similar to one another. There are a variety of reasons why we might want to block.

- Oftentimes the runs in an experiment are completed under different conditions. This may lead to the consideration of variables that are not part of the designed experiment but still could be important and influence the results. These are called nuisance variables. Nuisance variables include things like operator, time of day, room temperature, and lot number when they themselves are not factors in the experiment.

- Blocking can be used to remove or estimate the effect of a nuisance variable.

- A factor in a DOE may be very difficult or expensive to change. Thus, we block on that variable while doing all of the runs or tests cases (randomly) at one level of the factor before changing the level and conducting all of the remaining runs (randomly) when that factor is at a second level.

NTSB's Investigation of Boeing 737 MAX Crashes*

- NTSB chairman Robert Sumwalt said, "We saw in these two accidents that the crews did not react in the ways Boeing and the FAA assumed they would."

- "Those assumptions were used in the design of the airplane and we have found a gap between the assumptions used to certify the MAX and the real-world experiences of these crews, where pilots were faced with multiple alarms and alerts at the same time."

- The NTSB has recommended that system safety assessments should consider the effects of alerts on pilot responses and address gaps in design, procedures and training and that robust tools and methods for validating assumptions about pilot responses to airplane failures should be made as part of design certification.

*Source: Article in ***Aerospace Testing International***, 30 Sep 2019, authored by Ben Sampson

Prerequisites for Successful Testing

- **Remove excessive variation from the system.**
  - Look at each step in the process and see where we can reduce variation (Process Flow).
  - Document all factors that could possibly impact the results (Cause and Effect).
  - Use Standard Operating Procedures (SOPs) to remove as much noise as possible.

- **Perform a Measurement System Analysis (MSA)**
  - To determine how much variation is coming from the measurement system itself.
  - To ensure the measurement system is capable.
  - To improve the measurement system if it is not capable.

Electronic Warfare Case Study
Parameter Design Space (8 Factors)

- Platform Type (4 choices)
  - Missile, Aircraft, Ship, Land
- Frequency (5 choices)
  - Band1, Band2, Band3 Low, Band3 High, Band4
- Frequency Type (2 choices)
  - Constant, Agile
- PRI (5 choices)
  - CW, Very Low (<10μs), Low (10-100μs), Medium (100-1000μs), High (>1000μs)
- PRI Type (5 choices)
  - CW, Constant, Switcher, Jitter, Stagger
- PW (4 choices)
  - CW, Narrow (<1μs), Medium (1-10μs), Wide (>10μs)
- Scan (4 choices)
  - None (0s), Fast (<2s), Medium (2-10s), Slow (>10s)
- Scan Type (4 choices)
  - Steady, Circular, Conical, Sector

Total Number of Combinations = 64,000

Combinatorial Test Constraints*

CW-related constraints

- PRI [CW] with PRI Type [Constant, Switcher, Jitter, or Stagger]
- PRI Type [CW] with PRI [Very Low, Low, Medium, or High]
- PW [CW] with PRI Type [Constant, Switcher, Jitter, or Stagger]
- PW [CW] with PRI [Very Low, Low, Medium, or High]

Scan-related constraints

- Scan [None] with Scan Type [Circular, Conical or Sector]
- Scan Type [Steady] with Scan [Fast, Medium, or Slow]

* Since it is often the case for certain test factor level combinations to be infeasible, a constrained solution algorithm is utilized.

Electronic Warfare Case Study (cont.)

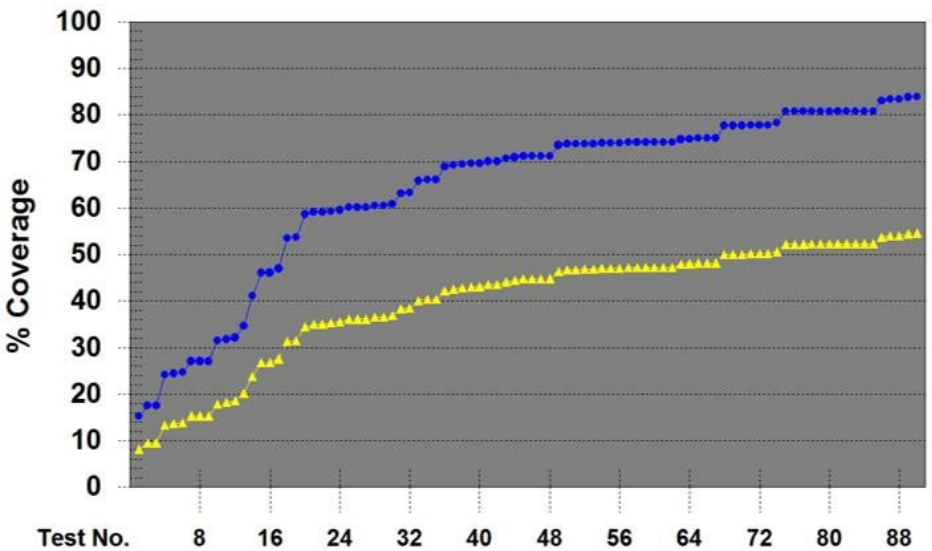
Coverage Analysis

Original Test Plan:

90 Test Cases

84% Critical Coverage (blue)

54% Overall Coverage (yellow)



Coverage Analysis Tool (CAT) results:

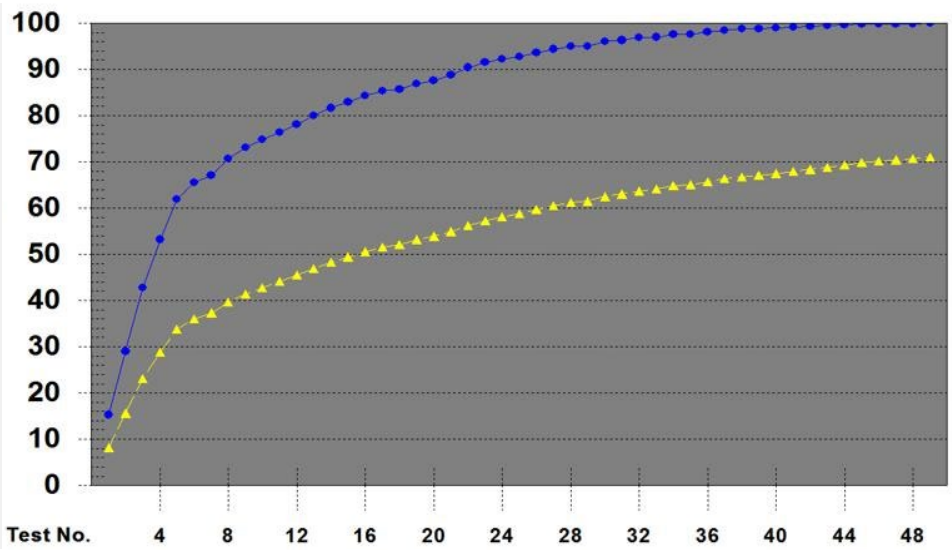
- Critical Coverage (Mains and 2-Ways): 84%
- Overall Coverage (Mains and 2-, 3-, 4-Ways): 54%
  - 2-Ways: 68%
  - 3-Ways: 35%
  - 4-Ways: 16%
  - Missing 2-Ways: 168

Optimized Test Plan Using DOE:

49 Test Cases (45% reduction)

100% Critical Coverage (blue)

71% Overall Coverage (yellow)



Coverage Analysis Tool (CAT) results:

- Critical Coverage: 100% (guaranteed by the algorithm used)
- Overall Coverage: 71%
  - 2-Ways: 100%
  - 3-Ways: 61%
  - 4-Ways: 23%
  - Missing 2-Ways: 0

Cyber Penetration Testing Example*

Putting these factor names and their levels, along with the test constraints, into Pro-Test software yields these 14 optimal test configurations:

Factors and Levels

	Factor 1	Factor 2	Factor 3
Factor Name:	**Port**	**OS**	**IS Type**
Level 1	21	Linux	network
Level 2	22	Windows	storage
Level 3	23		server
Level 4	25		
Level 5	80		

With Constraints

	If Factor ...	is at level ...	then Factor ...	can't be ...
Constraint 1	Port	25	IS Type	network
Constraint 2	Port	25	IS Type	storage
Constraint 3	IS Type	network	OS	Windows
Constraint 4	IS Type	storage	OS	Windows

Produces 14 Optimized Test Configurations

	Factor_A	Factor_B	Factor_C
Factor Name	Port	OS	IS Type
Case 1	25	Windows	server
Case 2	21	Linux	storage
Case 3	22	Linux	storage
Case 4	80	Windows	server
Case 5	21	Windows	server
Case 6	23	Windows	server
Case 7	80	Linux	network
Case 8	23	Linux	storage
Case 9	21	Linux	network
Case 10	22	Windows	server
Case 11	22	Linux	network
Case 12	23	Linux	network
Case 13	25	Linux	server
Case 14	80	Linux	storage

* Courtesy of Raytheon

Penetration Vulnerability Test Design Matrix*

Apply these known vulnerability tests

To each of these test cases

Vuln_Name				Weak Password	No Access Control	Insufficient Access Control	Blank Password	Anonymous FTP Access Allowed(Write Access)	Anonymous FTP Access Allowed(Read Access)	Weak Warning Banner	Missing Warning Banner	Missing or Outdated Patches	Mail Relaying Allowed	Mail Spoofing Allowed	Default Credentials Found	Unnecessary TCP Services are Open	Excessive File and Directory Permissions Granted to General Users	No Access Control on Remote Protocols	Cleartext Protocol
	Port	OS	IS Type																
Case 1	25	Windows	Server																
Case 2	21	Linux	Storage																
Case 3	22	Linux	Storage																
Case 4	80	Windows	Server																
Case 5	21	Windows	Server																
Case 6	23	Windows	Server																
Case 7	80	Linux	network																
Case 8	23	Linux	Storage																
Case 9	21	Linux	network																
Case 10	22	Windows	Server																
Case 11	22	Linux	network																
Case 12	23	Linux	network																
Case 13	25	Linux	Server																
Case 14	80	Linux	Storage																

* Courtesy of Raytheon

Why Use DOE?

- **More effective testing**
  - Performing the right test
  - For the right purpose
  - At the right time
- **More efficient testing**
  - Using resources (time, people, and test materials) wisely
  - Removing non-value-added activities
- **Minimizing the test debt**
  - Right-sizing the test for its intended purpose
  - Gaining the required knowledge at lowest cost

Examples of Test Debt

Test	Original Test Plan	Optimized Test Plan
Subsystem Testing	28 Tests	8 Tests (71% reduction)
Subsystem Simulation	100 Runs	40 Runs (60% reduction)
Near Field Range Testing	1036 Tests	632 Tests (39% reduction)
Software Subsystem Testing	90 Tests	63 Tests (30% reduction)
System Range Testing	826 Tests	473 Tests (42% reduction)
System Modeling & Simulation	26 Scenarios	22 Scenarios (15% reduction)
Electronic Warfare Testing*	90 Tests	49 Tests (45% reduction)
System Testing	332 Tests	238 Tests (28% reduction)

In each case, the reduction in number of test cases was achieved while maintaining or improving upon existing statistical test coverage.

* Recall this example showed an improvement of critical test coverage from 84% to 100%.

Design of Experiments (DOE)
(an elevator speech)

- An optimal data collection methodology
- “Interrogates” the process or product
- Used to identify important relationships between inputs (factors) and outputs (response variables)
- Identifies important interactions between input variables
- Can be used to characterize and optimize a process
- Can be used to asses risk
- Makes testing more effective and efficient and minimizes test debt
- Changes “I think” to “I know”
- Is the science of test and the key connector between test and evaluation

Excuses for NOT Using DOE?

- ~~It costs too much.~~
- ~~It takes too long.~~

These myths have been debunked by a century of data.

- We have not needed it before, and we are still doing fine. This is what Takata thought.

Takata Airbag Defect Findings:
Independent Testing Coalition (ITC) makes key findings

ITC says exposure to **heat** and **humidity,** and the use of **ammonium nitrate** are all required to produce what the commission and the National Highway Traffic Safety Administration (NHTSA) call an "energetic disassembly."

"You can't have the energetic disassembly without all three factors," David Kelly, leader of the ITC and former chief of the NHTSA told Automotive News Europe. "You have to have all three." In DOE, this is called a significant 3-way interaction effect.

- Ignorance is bliss!

The ITC has named three main technical factors that were all present in violent ruptures of Takata airbags linked to injuries and deaths.
PHOTO BY ADAM BARTLETT

Best Practices for "Operationalizing" DOE

(i.e., making it an integral part of the international language of T&E)

- Leadership actively promotes and expects DOE as a key test and evaluation competency.
- Trainers and coaches are experienced and motivational.
- A Keep-It-Simple-Statistically (KISS) approach with easy-to-comprehend materials and easy-to-use software.
- Gaining and propagating quick-hitting successes.
- Continuously generating transfer functions (predictive models) for the purpose of optimization, prediction, and risk assessment.

On Leading Change and Driving Business Results

"There is no way around it - we have to find ways to do more with less. The integrated program use of statistical techniques such as Design of Experiments, have proven themselves to be powerful enablers in our test optimization efforts to reduce cost and cycle time while providing our customers with confidence that our systems will perform."

Dr. Tom Kennedy
Chief Executive Officer
Raytheon Company

Thank You

Air Academy Associates, 12295 Oracle Blvd, Ste 340, Colorado Springs, CO 80921

air academy ASSOCIATES

For More Information Please Contact

12295 Oracle Blvd, Ste 340
Colorado Springs, CO 80921

Toll Free: (800) 748-1277 or (719) 531-0777
Email: aaa@airacad.com
Website: www.airacad.com