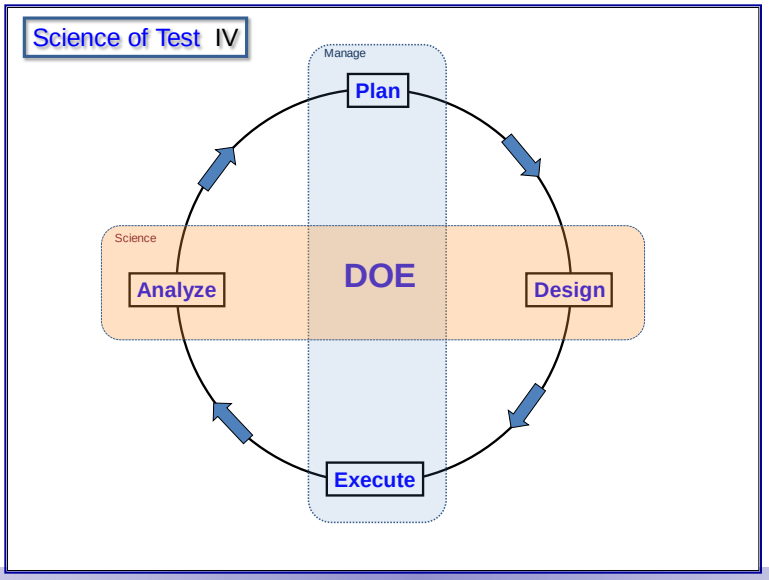


Systems Engineering: Connecting Test To Requirements with Scientific Test and Analysis Techniques (STAT)

Including Design of Experiments (DOE)

presented to:

NDIA National Systems Engineering Conference

24 Oct 2012

Greg Hutto 96Test Wing (DT) – Jim Simpson, 53d Wing (ACC OT)

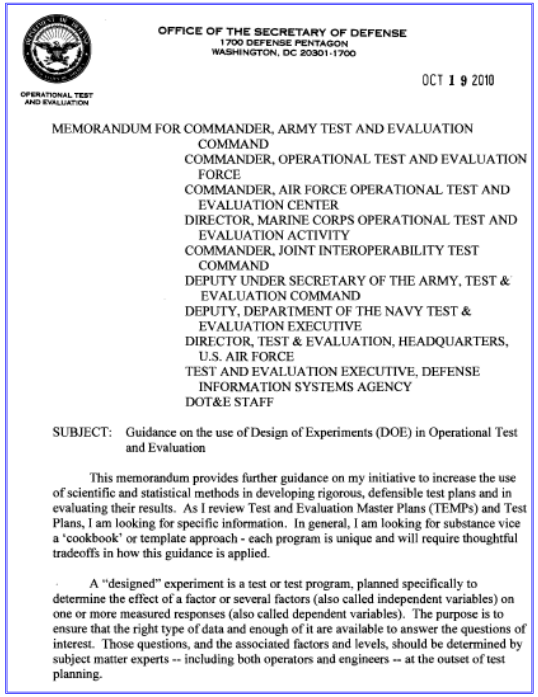
Gregory.hutto@eglin.af.mil

Disclaimers

- **This paper reflects the experiences and opinions of the authors and is not the official position of the 96 Test Wing, AFTC, the 53d Wing, or AFSOC**
- **We surely will improve our methods and words as we continue down this road …**

Why? Dr Gilmore's DOT&E Checklist Raised the Bar

- **19 Oct 2010 DOT&E Guidance Memorandum**
- **Program TEMPs should address the following questions:**



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

OPERATIONAL TEST AND EVALUATION

OCT 19 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

Checklist for TEMP Excellence:

- **What are your *Goal(s)*?**
- **How many *Trials*?**
- **Under what *conditions*?**
- **How to measure *success* (MOPs and MOSs)**
- **Statistical metrics – *risk* of wrong decisions**
- **Link to DOT&E Memo**

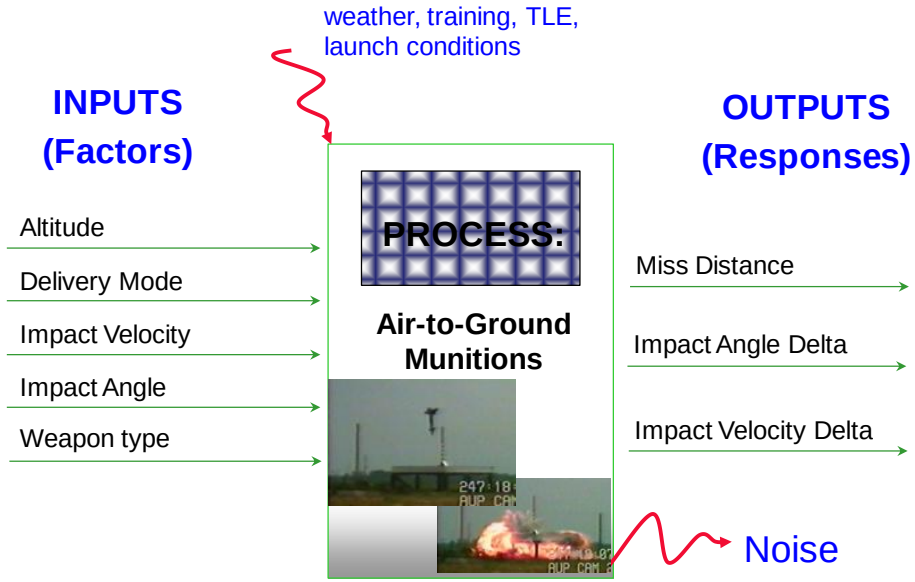
Why? As do our draft DT&E words in AFI and AFMCI 99-103

- **As agreed by Edwards, Arnold, Eglin, AFMC/A3 Feb 2012**
- **STAT Summit, Wright Patterson AFB**

Whenever feasible and consistent with available resources scientifically-based test and analysis techniques (STAT) and methodologies must be used for designing and executing tests, and for analyzing … data. The top level process and test objectives must be described in the first issuance of the test and evaluation master plan (TEMP), the systems engineering plan (SEP), and in more detail <later>… The integrated test team (ITT) should consult a STAT practitioner whenever test designs are considered.

Whenever appropriate selected approach must address …in detailed test plan:
• Define the test objective(s) of the test (or series of tests, when appropriate).
• Identify the information required from the test to meet the test objective(s).
• Identify the important variables that must be measured to obtain the data required for analysis. Identify how those variables will be measured and controlled. Identify the analysis technique(s) to be used.
• Identify the test points required and justify their placement in the test space to maximize the information obtained from the test.
• If using a traditional hypothesis test for data analysis, calculate statistical measures of merit (power and confidence) for the relevant response variables for the selected number of test events. If using another statistical analysis technique, indicate what <metric> will be used. If a statistical analysis technique is not being used, discuss the analysis technique that is being used and provide rationale.
… In all cases, the PM shall be responsible for the adequacy of the planned series of tests and report on the expected decision risk remaining after test completion.

What are Statistically Designed Experiments?

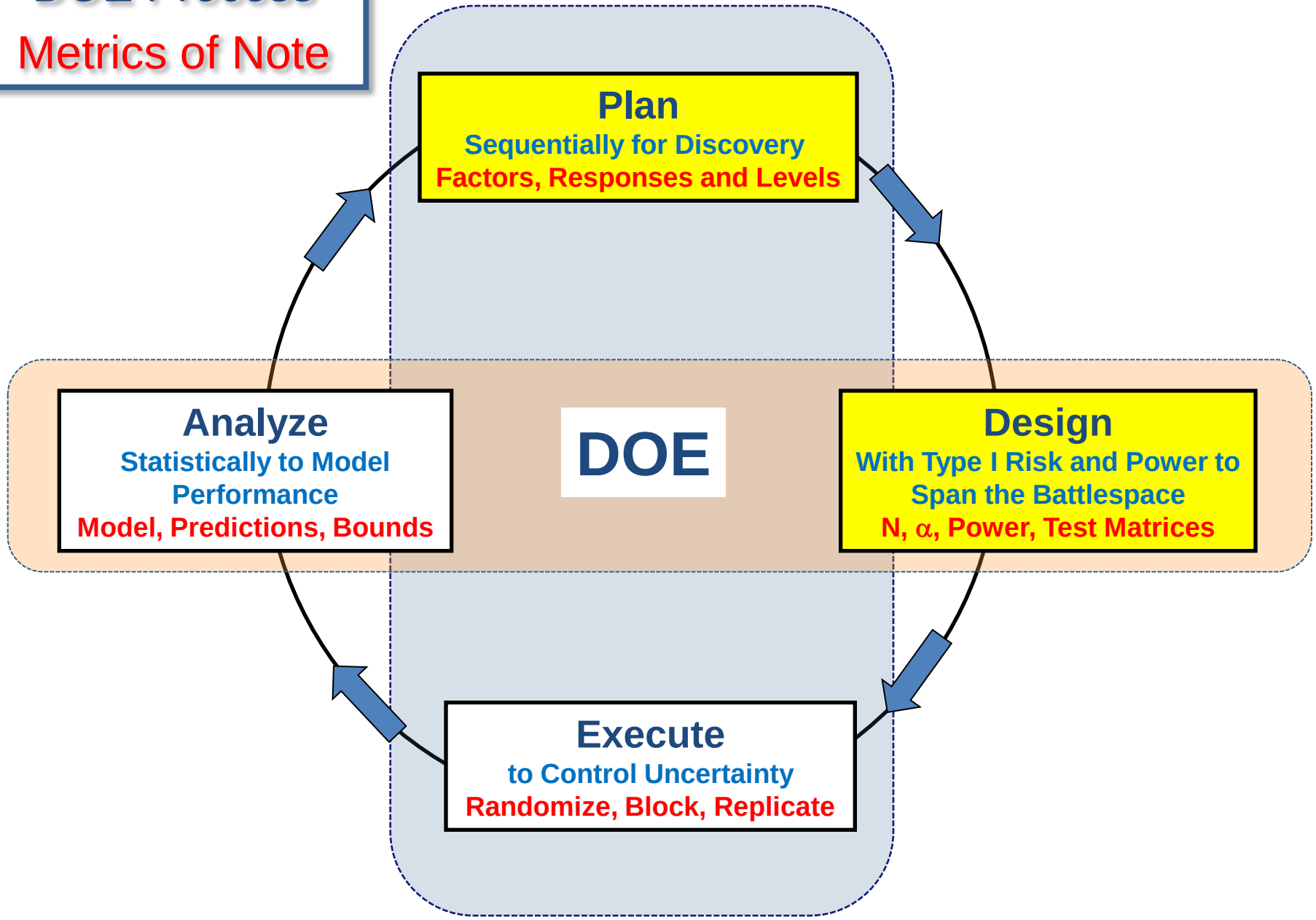


- Purposeful, systematic changes in the ***inputs*** in order to observe corresponding changes in the ***outputs***
- Results in a mathematical model that predicts system responses for specified factor settings

$$\text{Responses} = f\left(\text{Factors}\right) + \varepsilon$$

DOE Process

Metrics of Note



Dragon Spear (AC-130W) Integrated Tests

18 FLTS AFSOC 11 Dec 11 - Present

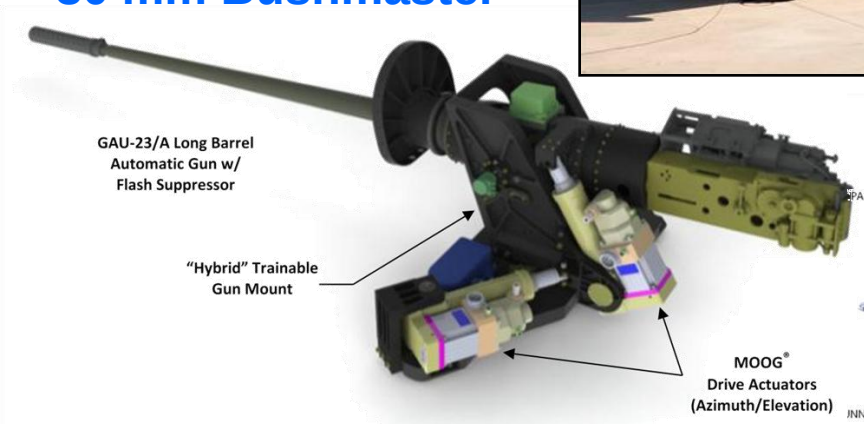
EO/IR/Laser Sensor Balls (2)

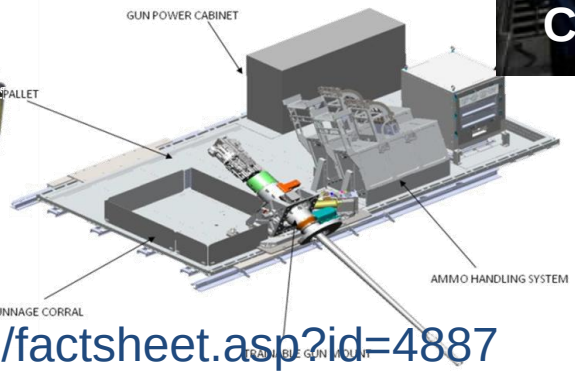
BRU-61 w/ SDB

Viper

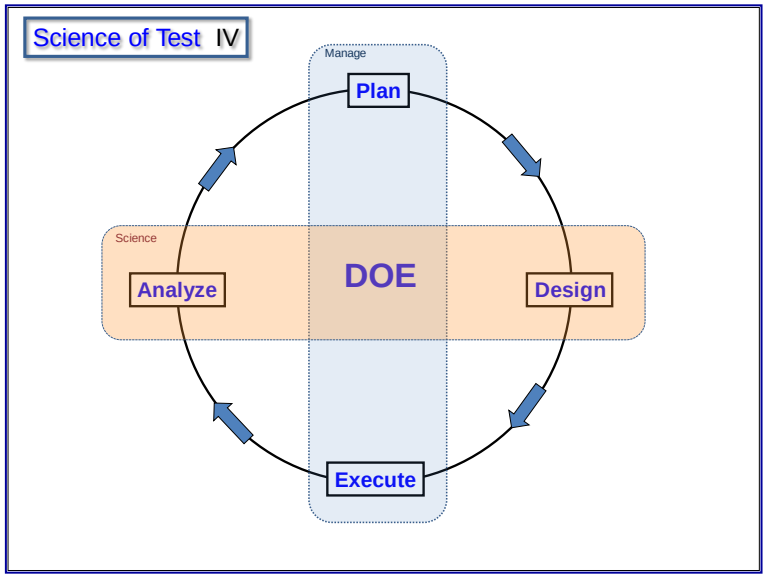
Combat Sys Operators

30 mm Bushmaster



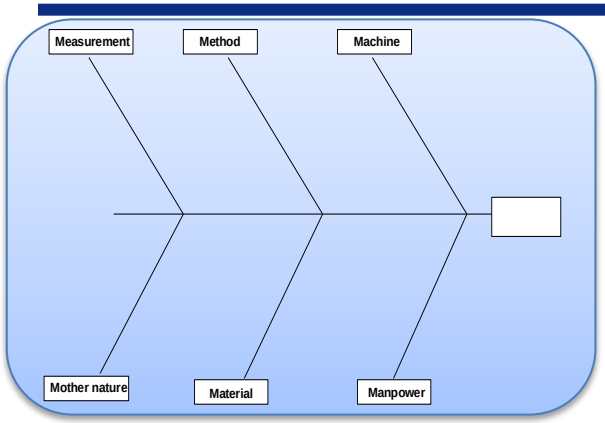


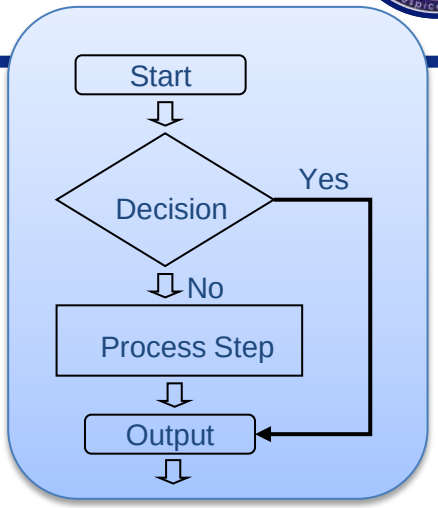
http://www.af.mil/information/factsheets/factsheet.asp?id=4887


Science of Test IV
Manage
Plan
Science
Analyze
DOE
Design
Execute

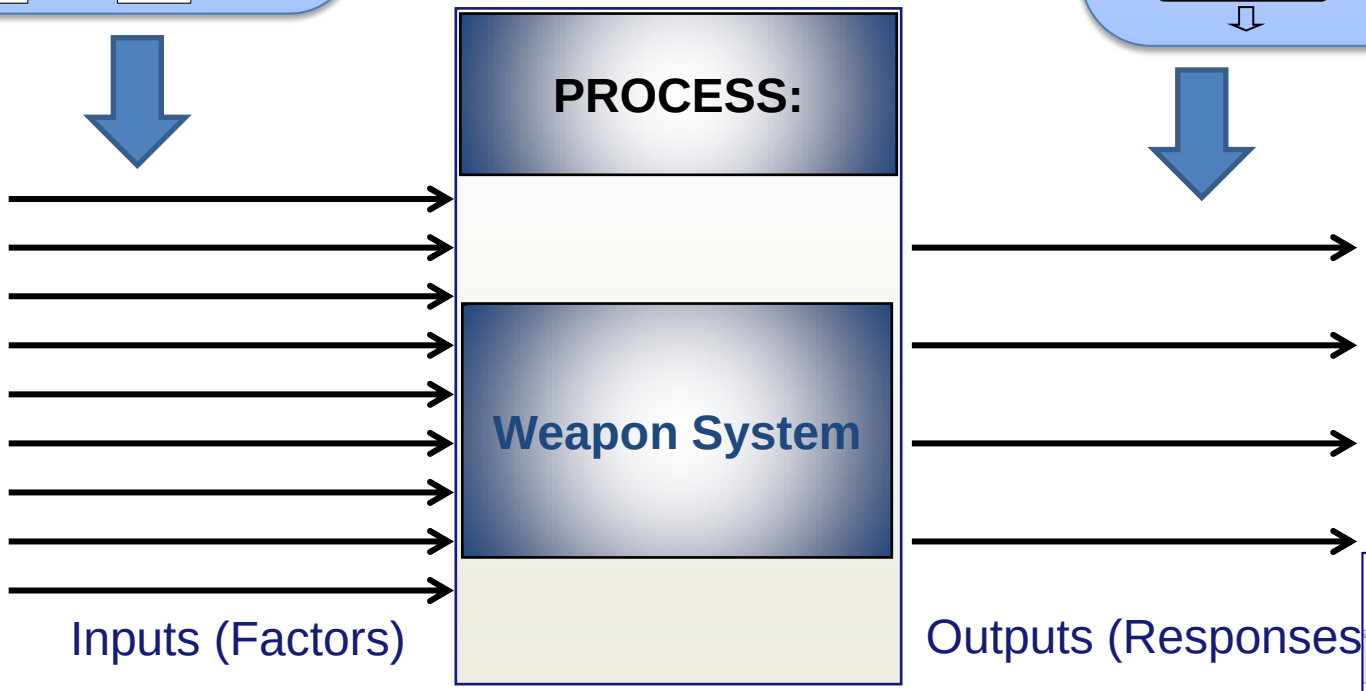
PLAN

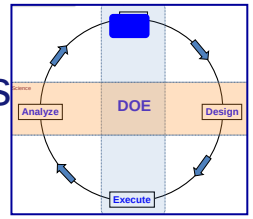
Planning Tools
Factors and Responses





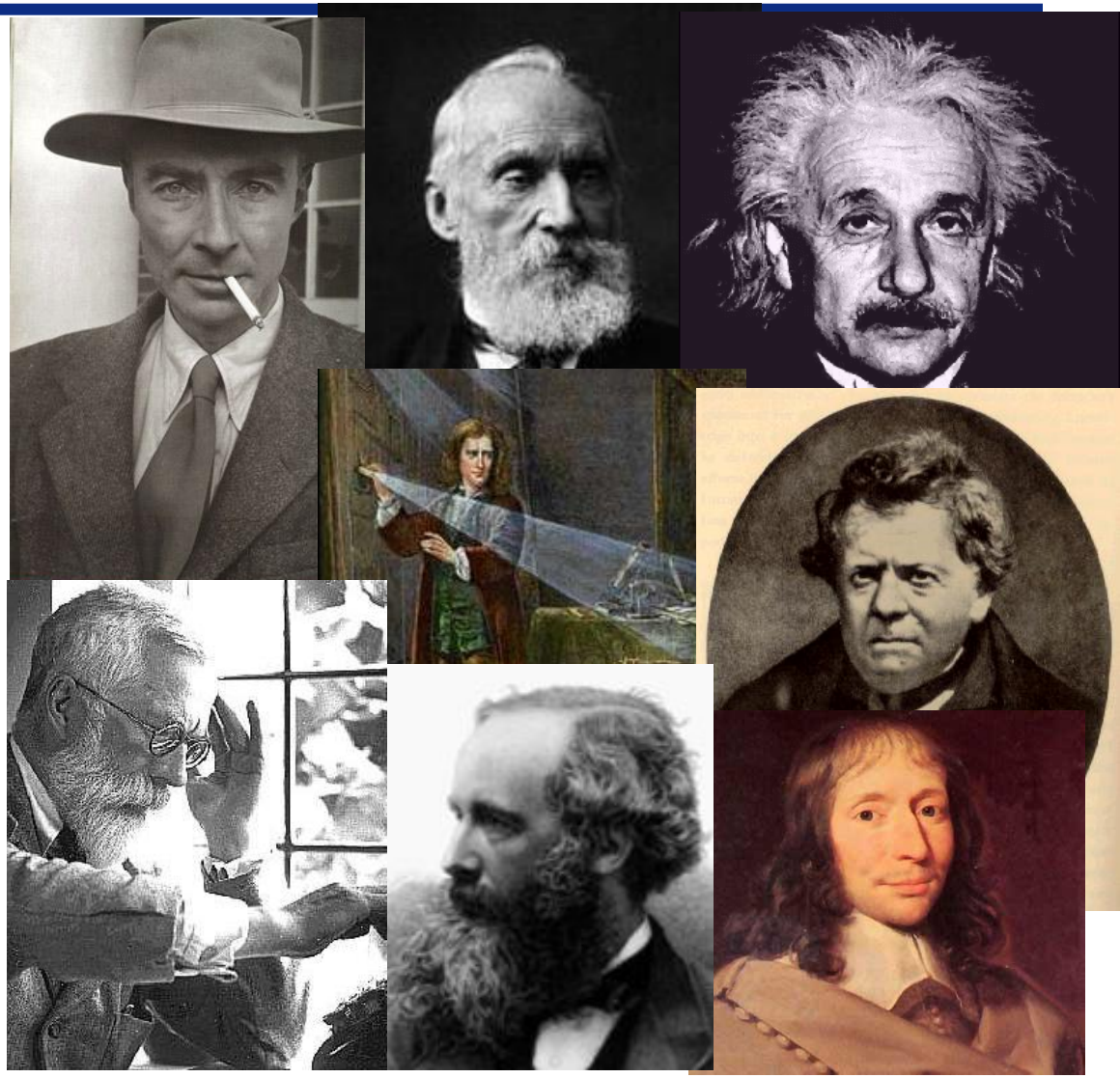
Input-Process-Output





In PLANNING, Science & Engineering has the lead role

- **Take-away: we already have good science in our DT&E!**
- **We understand sys-engineering, guidance, aero, mechanics, materials, physics, electromagnetics …**
- **To this great science, we introduces the *Science of Test***



Bonus: Match faces to names – Ohm, Oppenheimer, Einstein, Maxwell, Pascal, Fisher, Kelvin

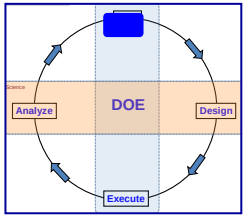
OT&E: Operations Skills are Vital to the Success of Integrated Test

- **Similarly, we already have good ops in our OT&E!**
- **We understand attack, defense, tactics, ISR, mass, unity of command, artillery, CAS, ASW, AAW, armored cav…**
- **DOE adds the *Science of Test***

Bonus: Match faces – Dolittle, Patton, Napoleon, Mitscher, Rickover, Francis Marion, Arnold, Puller

Plan: Define Test Objectives

- What is the point of these tests?
  - Screening: System has never been tested or never tested using DOE—need to identify important variables
  - Characterize: Need to understand system very well (predict)
  - Optimize: Ops testing – Tactics (what's the best thing to do)
  - Compare Performs as well or better than legacy system
  - Demonstrate Show that it works one time
- Objective may be as simple as:
- "Does this system (or tactic) work or not?"
- Definition of "Work"
  - Unfortunately the world is not that simple
  - How *well* does it work under a variety of battlespace conditions



Search the CPD and CONOPS

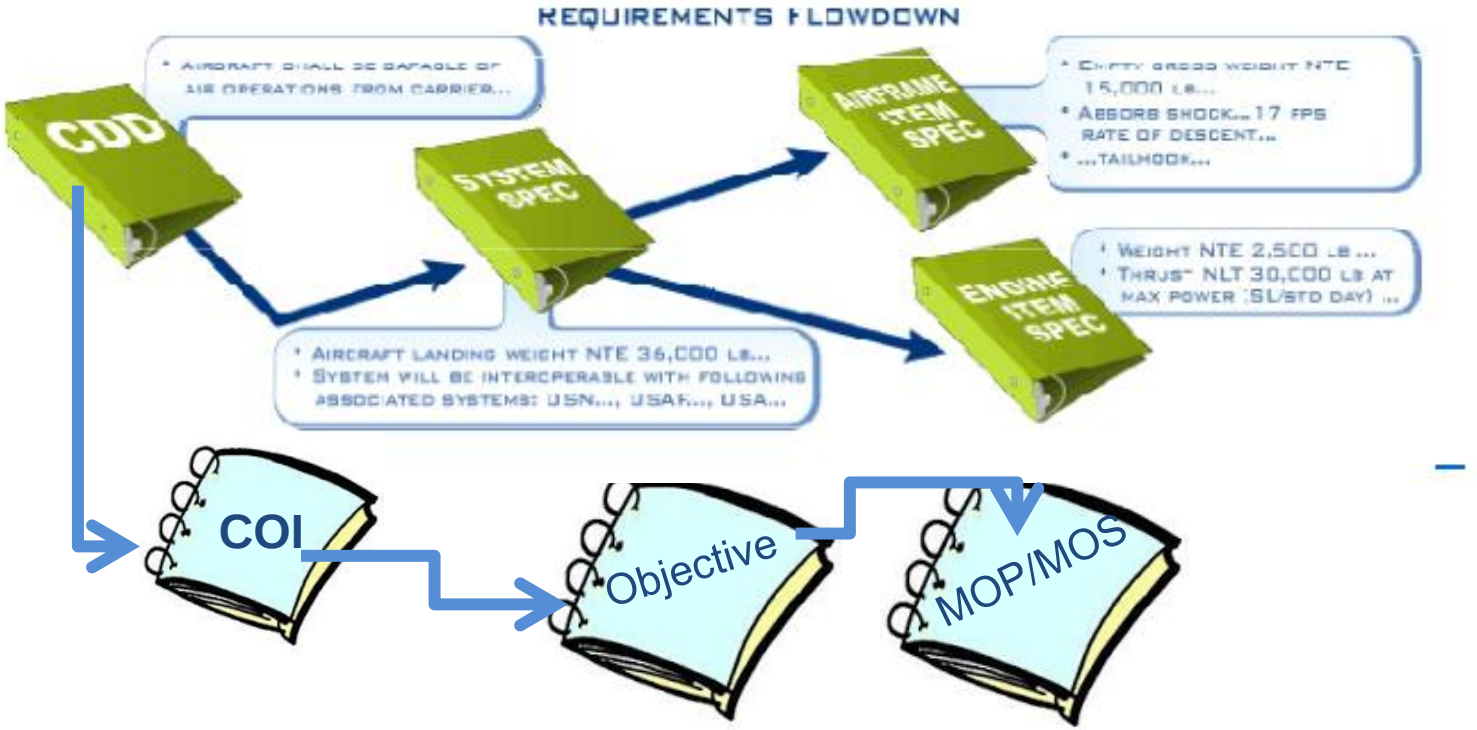
COI 1: Persistent strike ?		
Operational Capability	TEMP Measures	Reference
Lethality	-Time to wpn release -Accuracy -Appropriate Weapon -Target effects -Range to ID -Sensor track stability -Range of wpn release -TLE -Location SA (friendly/tgt)	CPD 6.1.2, 6.1.3, 6.1.4 AC Recap CONOP 4.1
Persistence	-Loiter time -Reattack time	AC Recap CONOP 4.1.4
Survivability	-Threat ID -Threat reaction -Threat avoidance/defeat	CPD 6.1.6 AC Recap CONOP 4.1.5
Interoperability (SII)	-Network compatibility	CPD 6.1.1
Connectivity		CPD 6.1.5 AC Recap CONOP 4.1.2
Limited Standoff	-Range -Noise signature -Visually signature	CPD Multiple

- Research can start the ball rolling before planning team

Today's Focus Capability: Combat Armed Escort & Strike

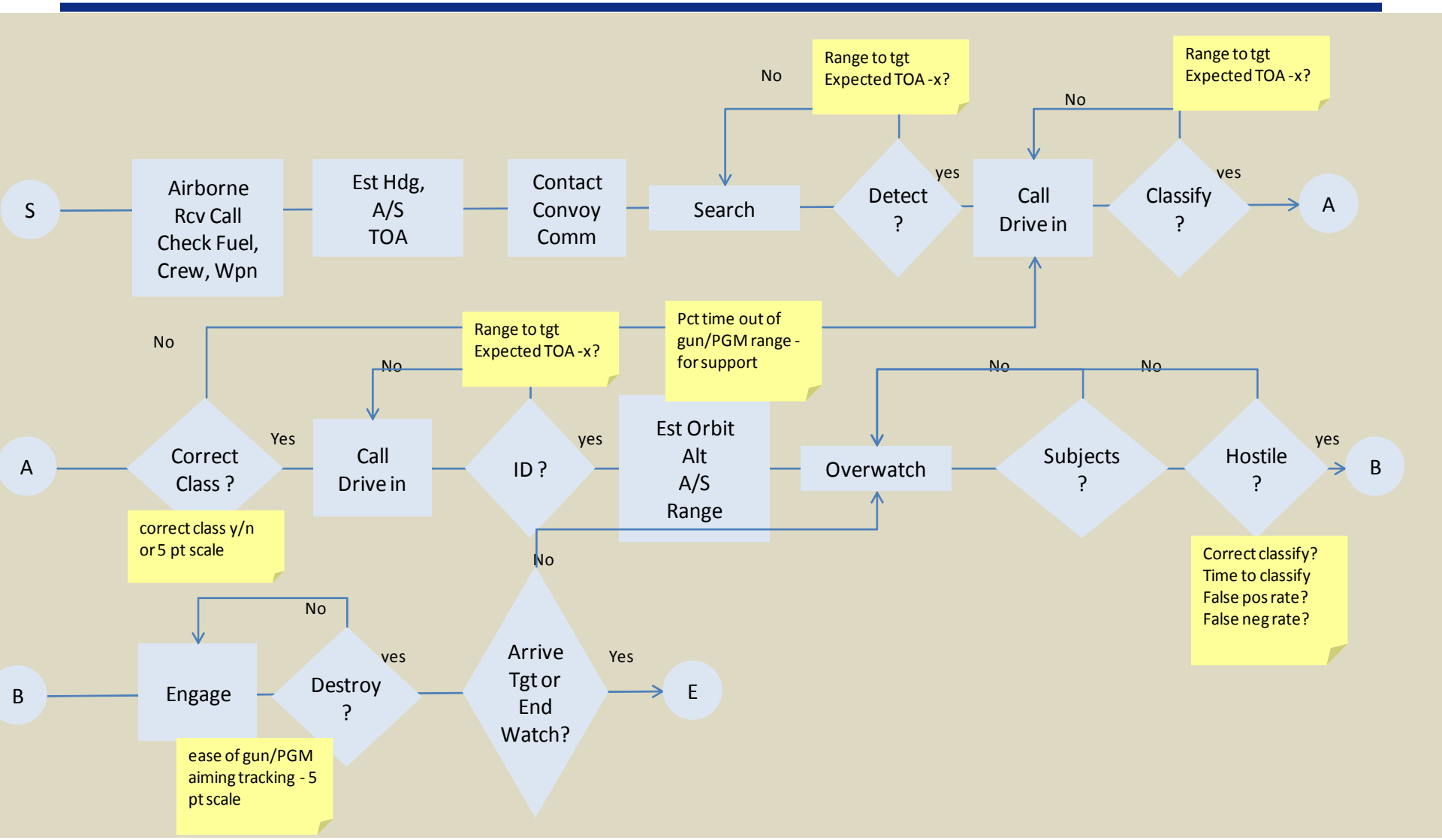
- What are the functionalities, success metrics, battlespace conditions, and required runs to prove?
- Easy answer: F2T2EA -- Find, Fix, Track, Target, Engage, Assess (targeting of hostile forces)

Help: Requirements Management & Initial Test Design



- Some of this work should already have been done and documented; elsewise … assemble the team!
- Scientists, operators, contractor, and testers can brainstorm what needs to be done (as in this case)

Team-Derived Process Flow
for Ground Party Armed Escort (Ops/DT/OT)



Link the Combat Tasks to the Success Metrics (CTP & MOPs)

	Combat Missions				Sample Measures of Success (Responses)	
Combat Tasks	**Armed Escort**	**Close Air Spt**	**Intel, Surv, Recce**	**Time Sens Tgt**	**Crit Tech Param (CTP)**	**Meas of Perf (MOP)**
Detect	√	√	√	√	Tgt/Bkgd ^{o}K	Range to detect
Classify	√	√	√	√	Pixels on Target	time to classify
Identify	√	√	√	√	40x zoom, angle resolve	correct ID
(Fix) Locate	√	√	√	√	TLE	TLE
Track	√	√	√	√	RMS track error	Breaklock - time on tgt
Target	√	√		√	mil aim error	autotrack stability
Engage	√	√		√	miss distance	lethality (M&S) or scale
Assess	√	√	√	√	time to transmit video	assess accuracy (scale)

- Mapping tasks to missions reduces redundant testing
- It also clearly maps test to JCIDS capability needs, ensuring test is relevant to the acquisition
- Finally, it naturally sets us on a path to true Integrated Test (IT)

Steps in the Process: Responses

- **Brainstorm as many responses as possible**
  - **List Possible Responses**

Rounds	$Aircraft_{Prior}$	PGMs	ID	Fixed	Moving	$Aircraft_{Post}$
# Damaged	# Grounding Issues	# Damaged	Range to Detect	X Miss Dist. Laser	# Rounds on Target	# Post Issues
# New	MTBCF	# New	Time to Detec	Y Miss Dist. Laser	Hit or Miss	# Unused Rounds
# Used	MTA	# Used	Range to Classify	Radial Miss Dist. Laser	BDA (% destroyed)	# Unused PGMs
Time Stored	Tail #	Time Stored	Time to Classify	X Miss Dist. Live		Lbs of Fuel Left
	# Years Used		Range to Target ID	Y Miss Dist. Live		
			Time to Target ID	Radial Miss Dist. Live		
			Range to Acquire Aimpoint	Hit or Miss		
			Time to Acquire Aimpoint	BDA (% destroyed)		

Steps in the Process: Conditions

- **Brainstorm Factors:**

Prioritized List

Type	Potential Influences
X	1. fixed and moving targets (buildings, vehicles, personnel,
X	2. Light: day night transition
X	3. terrain – urban vs plains vs forested vs. mountainous
X	4. sensor type – IR or visual
X	5. firing mode – direct or offset?
C	6. gun ammo type PGU-13A/B, PGU-46/B, PGU-15B
C	7. gun firing rate: single, burst, continuous
C	8. Aircraft altitude fixed or vary?
Cov	9. Aircraft tail – one or two
C	10. Missions (omnibus factor) airland cargo, armed overwatch/CAS, air drop
Cov	11. NVG and cockpit light levels
N	12. Comms: TOC, ground forces (Call For Fire) SA Data Link (SADL) High Perf Waveform (HPW) Adaptive Network Waveform 2(ANW2)
Cov	13. Video transfer via datalink
N	14. Ground party – present, none
N	15. Load – Pax, cargo (interfere?)

Legend:
X = Variable in DOE Matrix
C = Held Constant
Cov = Covariate (uncontrolled variation which is measured)
N = Noise Variable (uncontrolled variation which is not measured)

Battlespace for Dragon Spear SDB Case: Factors and Responses

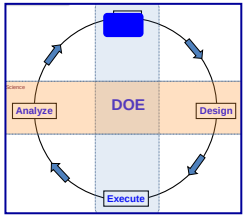
- **Systems Engineering Question: Does Dragon Spear perform at required capability level across the planned battlespace?**

Factors

Test Condition	Number	Lvls-Type	Num Levels
Target Type:	1	Vehicle, People, Building	3
Num Weapons	2	1, 2, 4	3
Target Angle on Nose	3	0, 45, 90	3
Release Altitude	4	10K, 15K, 20K	3
Release Velocity	5	180, 220, 240, 250	4
Release Heading	6	000, 045, 090, 135,180	5
Target Downrange	7	0, 2, 4, 8	4
Target Crossrange	8	0, 1, 2, 3	4
Impact Azimuth (°)	9	000, +45, +90	3
Fuze Point	10	Impact, HOB	2
Fuze Delay	11	0, 5,10, 15, 25	5
Impact Angle (°)	12	15, 60, 85	3
Total Combinations			2,332,800

Responses

Type	Measure of Performance
Objective	TOF Accuracy (%)
	Target Standoff (altitude)
	Launch range
	Mean radial arrival distance
	Probability of damage
	Reliability
Subjective	Interoperability
	Human factors
	Tech data
	Support equipment
	Tactics

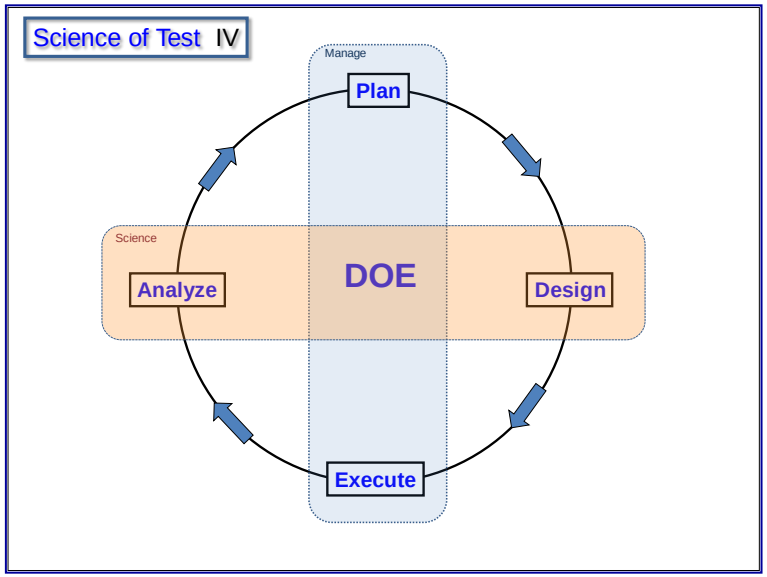


Plan Checklist

Excellence Checklist for:

Metric I. Plan a Series of Experiments to Accelerate Discovery

- Problem statement with a synopsis of historical information research
- Clear, concise, comprehensive (and ideally, numeric) objectives for each stage of test
- List of output performance measures to include the variable type, anticipated ranges, estimated precision, and potential for measurement error
- Evidence of in-depth and comprehensive brainstorming by all parties: fishbone diagrams, process flow diagrams, measures, factors and levels, and test point collection strategy
- Thorough list of relevant system factors or conditions to include those which will be experimental, held constant, and allowed to vary (noise)
- Discussion of likely re-design strategies following initial exploration (e.g. screen, explore, validate)
- Management reserve (confirmation) ~10% to 30% depending on assessed risk and technology maturity


Science of Test IV
Manage
Plan
Science
Analyze
DOE
Design
Execute

DESIGN

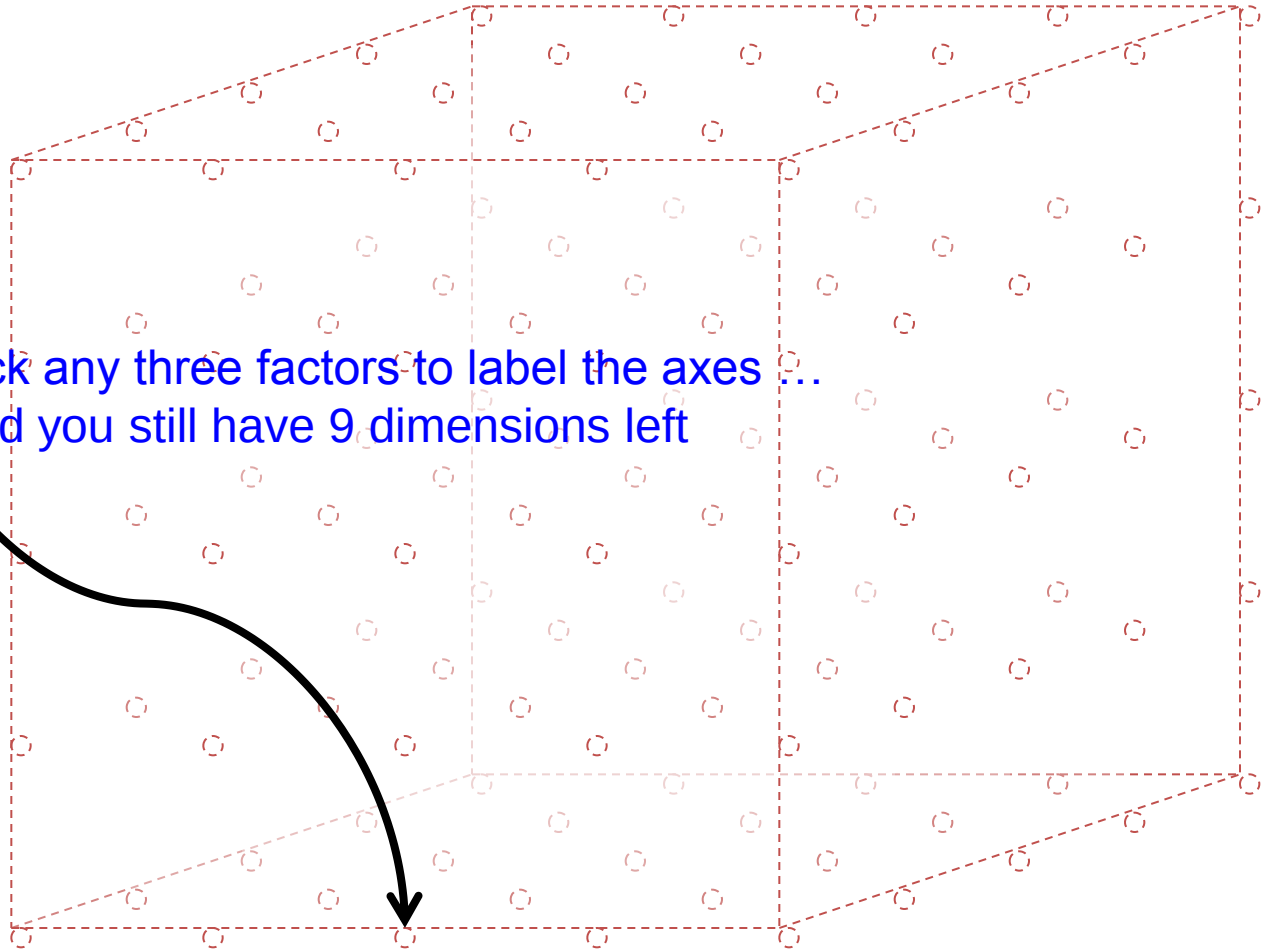
Visualizing an infinite 12-D Battlespace

Test Condition
Target Type:
Num Weapons
Target Angle on Nose
Release Altitude
Release Velocity
Release Heading
Target Downrange
Target Crossrange
Impact Azimuth (°)
Fuze Point
Fuze Delay
Impact Angle (°)

If each factor constrained to just two levels, you still have ...

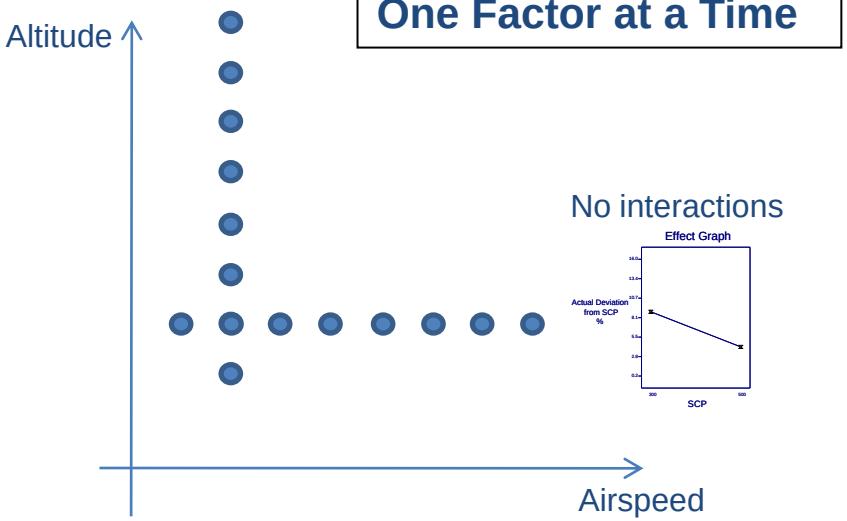
$2^{12} = 4096$

… lattice points!

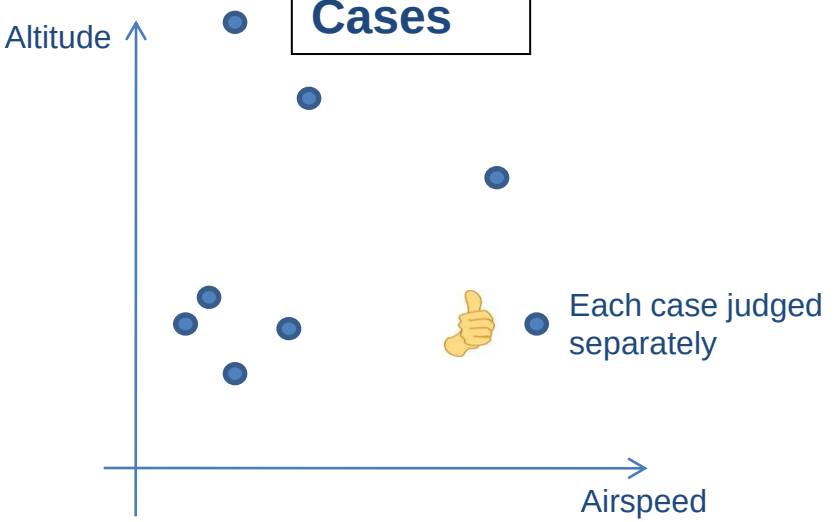


Design: Which Points?
Traditional Test Strategies

One Factor at a Time

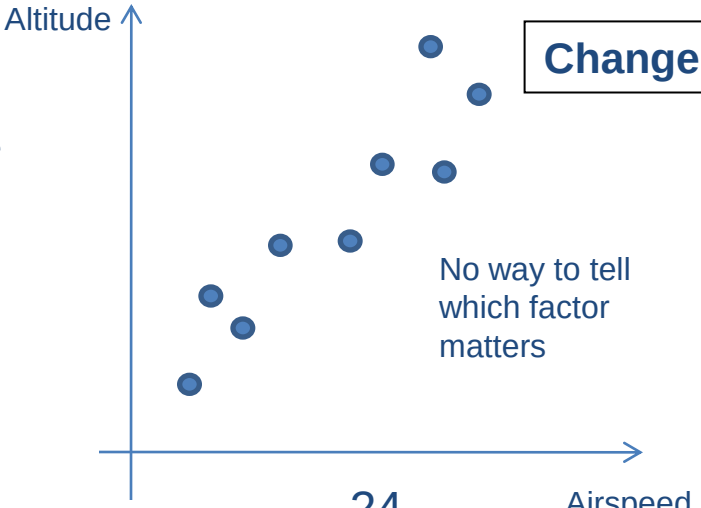


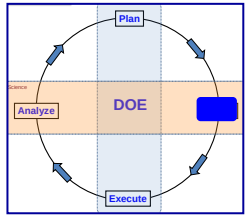
Cases



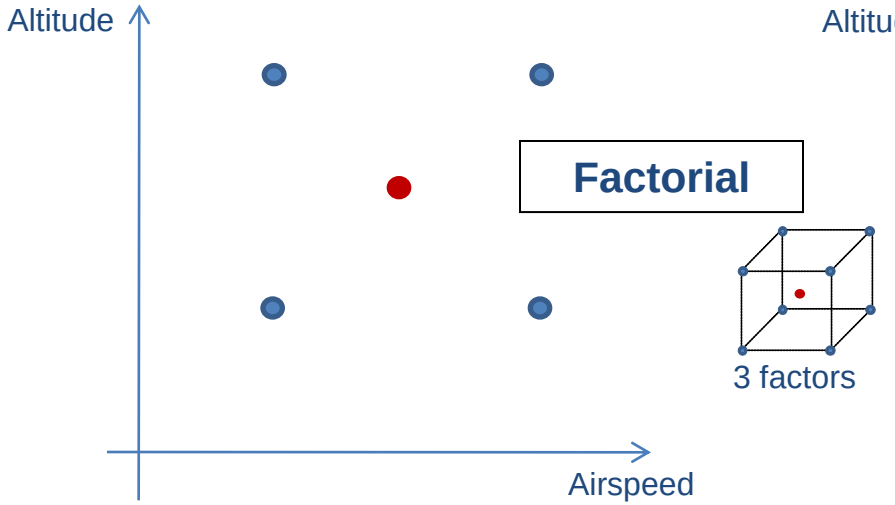
And the <u>always-popular</u> DWWDLT – do what we did last time

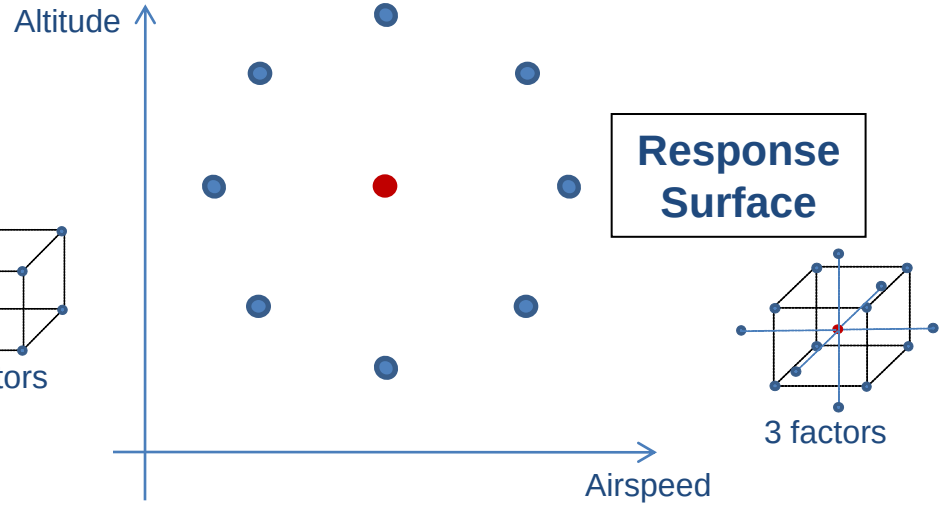
Change variables together





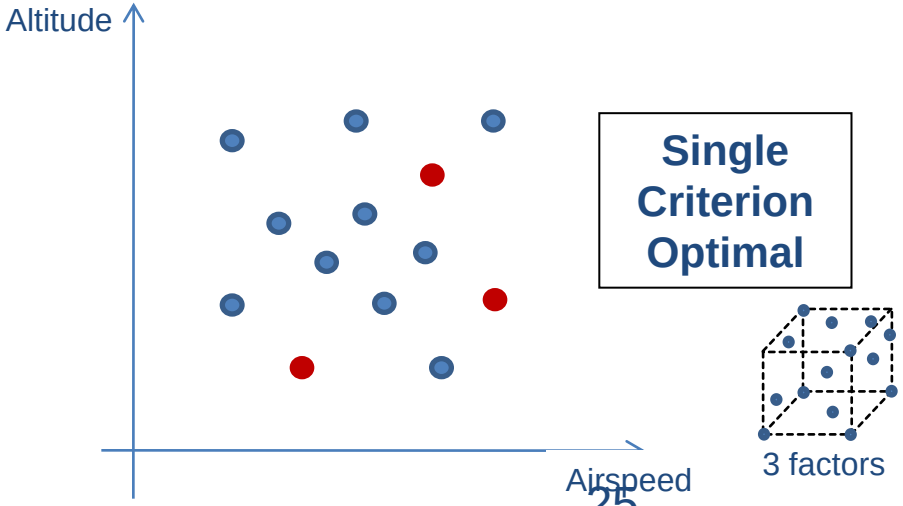
Test Space – Some experimental design Choices



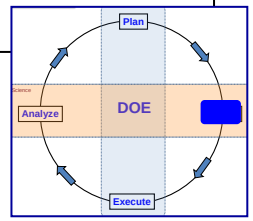


single point

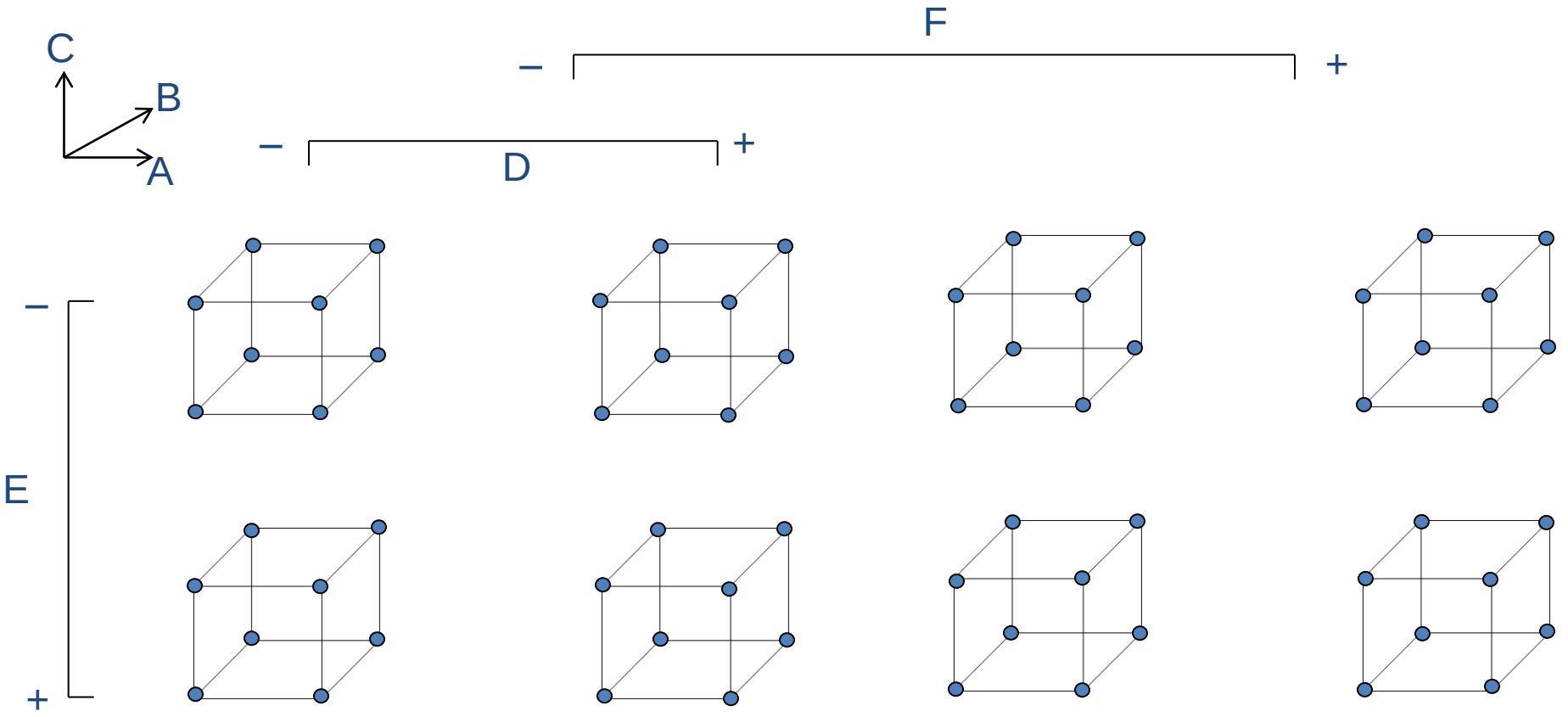
replicates

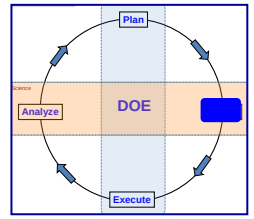


As the number of factors increases, fractional factorials are efficient solutions

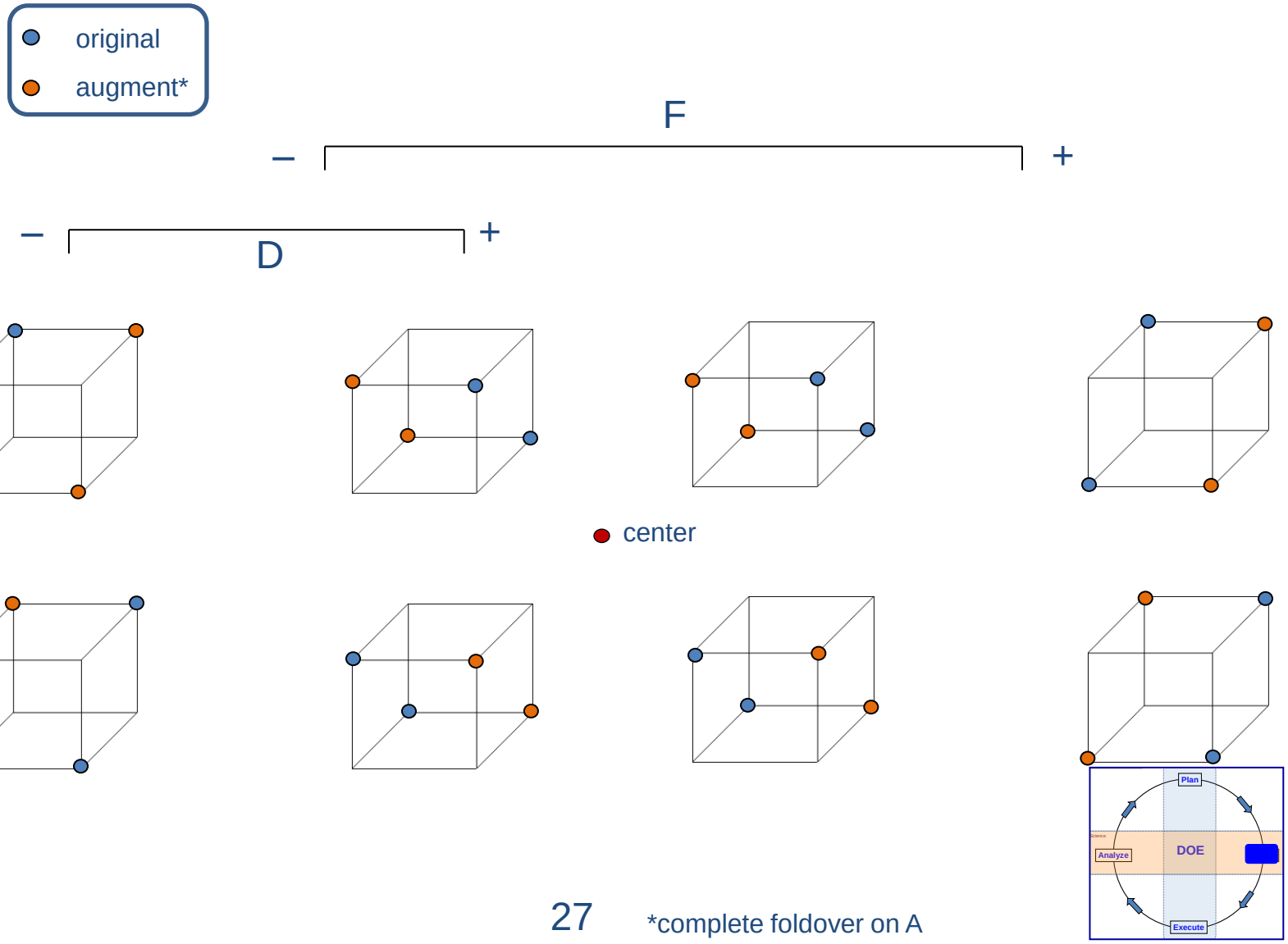


More Variables



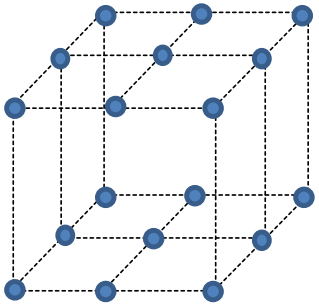


Fractions & Sequential Design

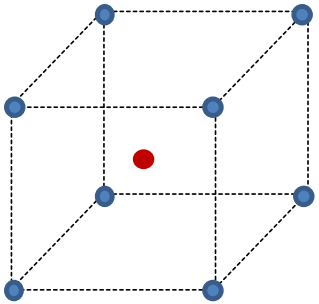


*complete foldover on A

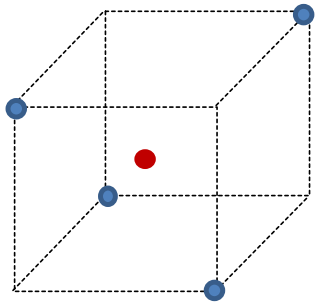
Classic Experimental Designs



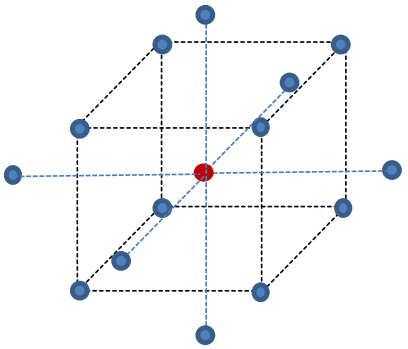
General Factorial
3x3x2 design



2-level Factorial
2^3 design

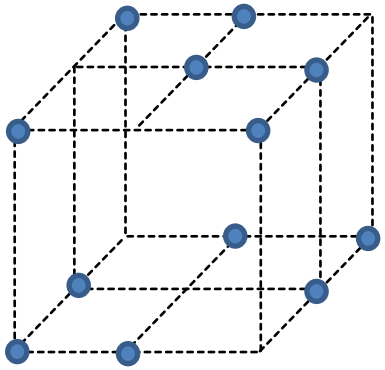


Fractional Factorial
2^{3-1} design

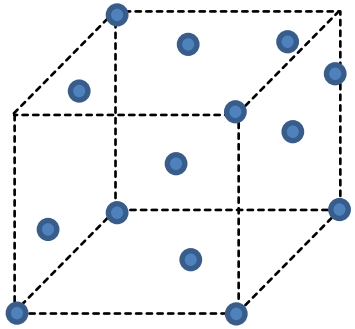


Response Surface
Central Composite design

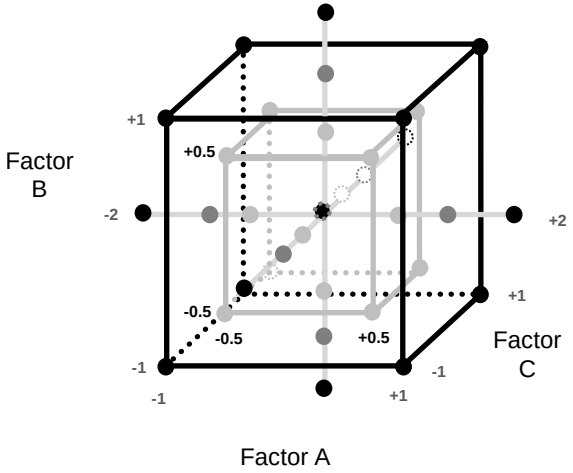
Other Designs



Mixed-level Fractional
3x3x2 fraction



Optimal Design
IV-optimal



Higher Order Design
Nested CCD

Metrics for Evaluating Designs

Characterize

Objectives	Criteria
Characterize	Sample Size
Compare	Power
Screen	ME
	2FI
Assume	Replicates - Pure Error
Same factors	Orthognonality
General Model = ME + 2FI	Terms aliased
Type I Error = 0.05	Word length count
	VIF
Response type: Numeric	Categoric balance
	Interaction balance (GBM)
	Partial aliasing
	Model misspecification (lack of fit)
	3FI
	Curvature
	Quadratic
	Range of Inputs
	Robustness to outliers/missing data
	Points total for rep/LOF
	Functionality
	Levels per factor - intended
	Levels per factor - design
	Ease of augmentation
	foldover strategy
	additional levels ease

Optimize

Objectives	Criteria
Estimate	Sample Size
Predict	Prediction Variance
Optimize	50% FDS
Map	90% FDS
Assume	95% FDS
Same factors	G-eff (min max prediction)
General Model = ME + 2FI+PQ	I-eff (avg pred variance)
Type I Error = 0.05	Replicates - Pure Error
	Orthognonality
Response type: Numeric	Condition number
	VIF
	Prediction Uniformity
	Rotatability
	Uniform precision
	Model misspecification (lack of fit)
	3FI
	Pure Cubic
	Range of Inputs
	Sensitivity to outliers/missing data
	Influence / Leverage
	Points total for rep/LOF
	Space Fill Properties
	entropy
	minimize Euclidean distance among pts
	Design Functionality
	Levels per factor
	Number of evenly spaced levels

Recall Battlespace for Dragon Spear: Factors and Responses

- **Systems Engineering Question: Does SDB perform at required capability level across the planned battlespace?**

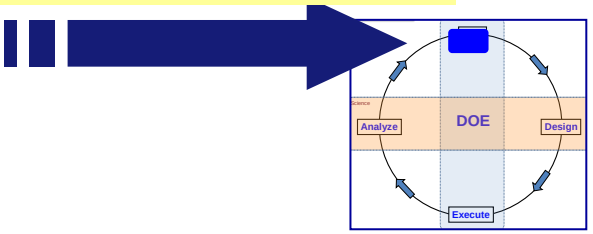
Factors

Test Condition	Number	Lvls-Type	Num Levels
Target Type:	1	Vehicle, People, Building	3
Num Weapons	2	1, 2, 4	3
Target Angle on Nose	3	0, 45, 90	3
Release Altitude	4	10K, 15K, 20K	3
Release Velocity	5	180, 220, 240, 250	4
Release Heading	6	000, 045, 090, 135,180	5
Target Downrange	7	0, 2, 4, 8	4
Target Crossrange	8	0, 1, 2, 3	4
Impact Azimuth (°)	9	000, +45, +90	3
Fuze Point	10	Impact, HOB	2
Fuze Delay	11	0, 5,10, 15, 25	5
Impact Angle (°)	12	15, 60, 85	3
Total Combinations			2,332,800

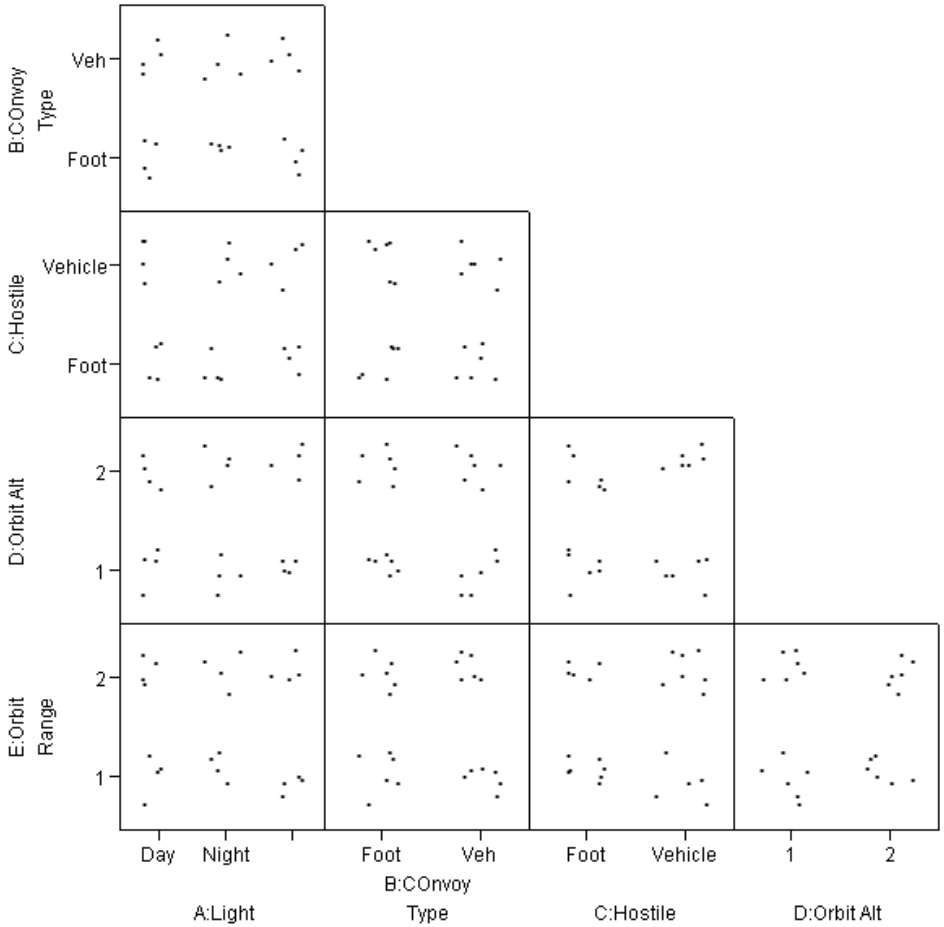
Responses

Type	Measure of Performance
Objective	TOF Accuracy (%)
	Target Standoff (altitude)
	Launch range
	Mean radial arrival distance
	Probability of damage
	Reliability
Subjective	Interoperability
	Human factors
	Tech data
	Support equipment
	Tactics

12 dimensions –A large test space… how to search it?



Overview of a Target Engagement Design 5 Vars-24 runs



Statistical Risks (Power) for Terms A,B,C,D,E

Power at 5 % alpha level to detect signal/noise ratios of				
Term	StdErr**	VIF	1 Std. Dev.	2 Std. Dev.
A[1]	0.29		35.5%	91.5%
A[2]	0.29			
B	0.20	1	63.7%	99.6%
C	0.21	1.028571	62.5%	99.5%
D	0.20	1	63.7%	99.6%
E	0.21	1.028571	62.5%	99.5%

- This design based on mixed level D-optimal algorithm

Design Checklist

Excellence Checklist for:

Metric IIa. Design to Span the Battlespace

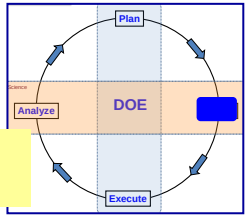
- Thorough list of candidate factors. Table should include:
  - Name of factor
  - Unit of measurement (real (physical) values much preferred over labels)
  - Range of physical levels; Levels chosen for experimental design
  - Estimated priority of factor in describing system performance
  - Expense/difficulty of controlling level: easy, hard, very hard to change
  - Proposed strategy of experimental control for each factor: constant, matrix variable, or noise. If noise, specify covariate, randomized, or random effect
- Name of chosen design strategy
  - Type of statistical design selected (e.g. factorial, fractional factorial, or response surface)
  - Size of design – number of factors and levels planned
  - *N* – number of total trials
  - Anticipated statistical model and discussion of how design fits the model

Statistical Power: Factorials

Which Points & How Many?

			Statistical Power to Detect Issues	
Factors	**Tests**	**Confidence**	**1 std dev**	**2 std dev**
2	**5**	0.95	0.07	0.13
	8	0.95	0.19	0.57
	16	0.95	0.45	**0.95**
3	**8**	0.95	0.09	0.17
	12	0.95	0.19	0.57
	16	0.95	0.42	**0.93**
4	**9**	0.95	0.09	0.17
	12	0.95	0.19	0.57
	16	0.95	0.39	**0.91**
	32	0.95	0.77	**0.99**
8	**16**	0.95	0.12	0.24
	24	0.95	0.41	**0.92**
	32	0.95	0.75	**0.99**

Statistical Power only slightly affected by number of test variables

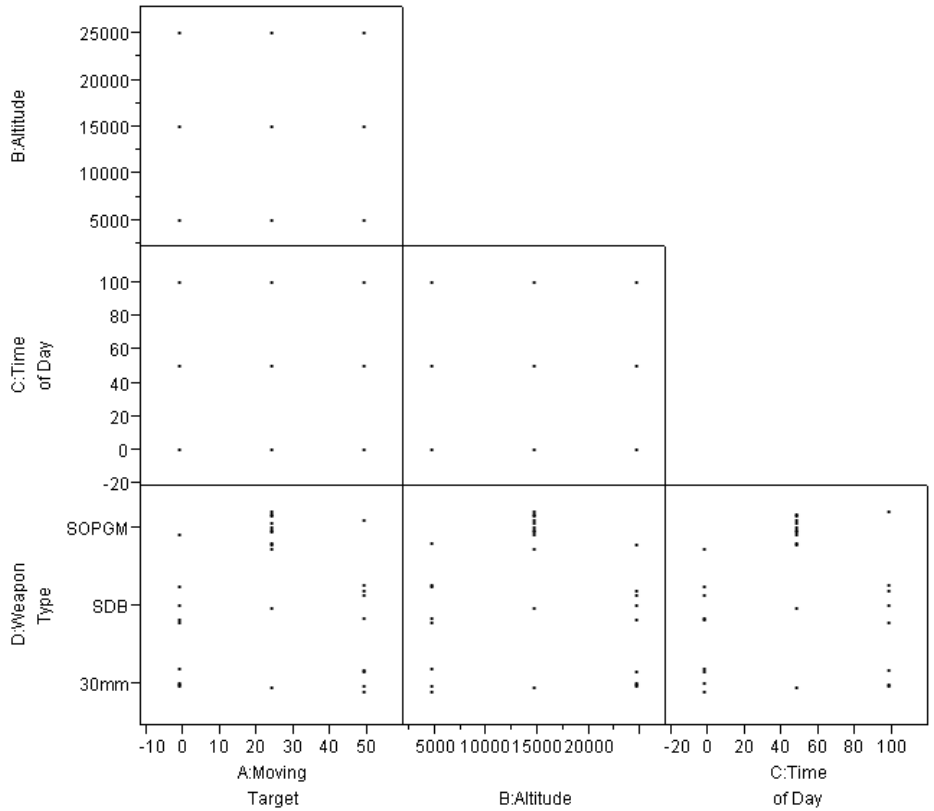


Design Strategy plus Power
The Experts again in the driver's seat

- Determining the size of the test requires that a number of parameters be estimated or declared, including the decision risks α and β
- The number of factors, assets for estimating error, size of the change in MOP/MOS of operational interest, and the magnitude of system variability also play a role

Parameter	Description	How Obtained	Relevance in Planning
***k*: factors**	Number of factors in the experiment	Determined in process decomposition	Key finding from process decomposition
df_{error}: model error	Amount of data reserved for estimating system noise	Replication and extra model terms	Estimate of complexity of input-output relation
α: alpha	Probability of incorrectly declaring a factor matters	Set by test team	Fix and leave alone
δ: delta	Size of response change expert wants to detect	Experts and management determine	Some ability to vary
σ: sigma	System noise – run-to-run variability or repeatability	Historical data; pilot tests; expert judgment	System driven but can be improved by planning
1-β: power	Probability of correctly declaring a factor matters	Lower bound set by test team	Primary goal is to set *N* to achieve high power
***N*: test size**	"how many"	Usually computed based on all above parameters	Direct, should modify to satisfy power

A Second Design – 4 Vars, 3 levels 30 runs (live fire)



Statistical Risks (Power) for Terms A,B,C,D

Power at 5 % alpha level to detect signal/noise ratios of				
Term	StdErr**	VIF	1 Std. Dev.	2 Std. Dev.
A	0.29	1.5	51%	98%
B	0.29	1.5	51%	98%
C	0.29	1.5	51%	98%
D[1]	0.27		39%	94%
D[2]	0.25			
AB	0.25	1	47%	96%
AC	0.25	1	47%	96%
AD[1]	0.35		19%	62%
AD[2]	0.50			
BC	0.25	1	47%	96%
BD[1]	0.35		19%	62%
BD[2]	0.50			
CD[1]	0.35		19%	62%
CD[2]	0.50			

- This 3-level design based on a Face-Centered CCD

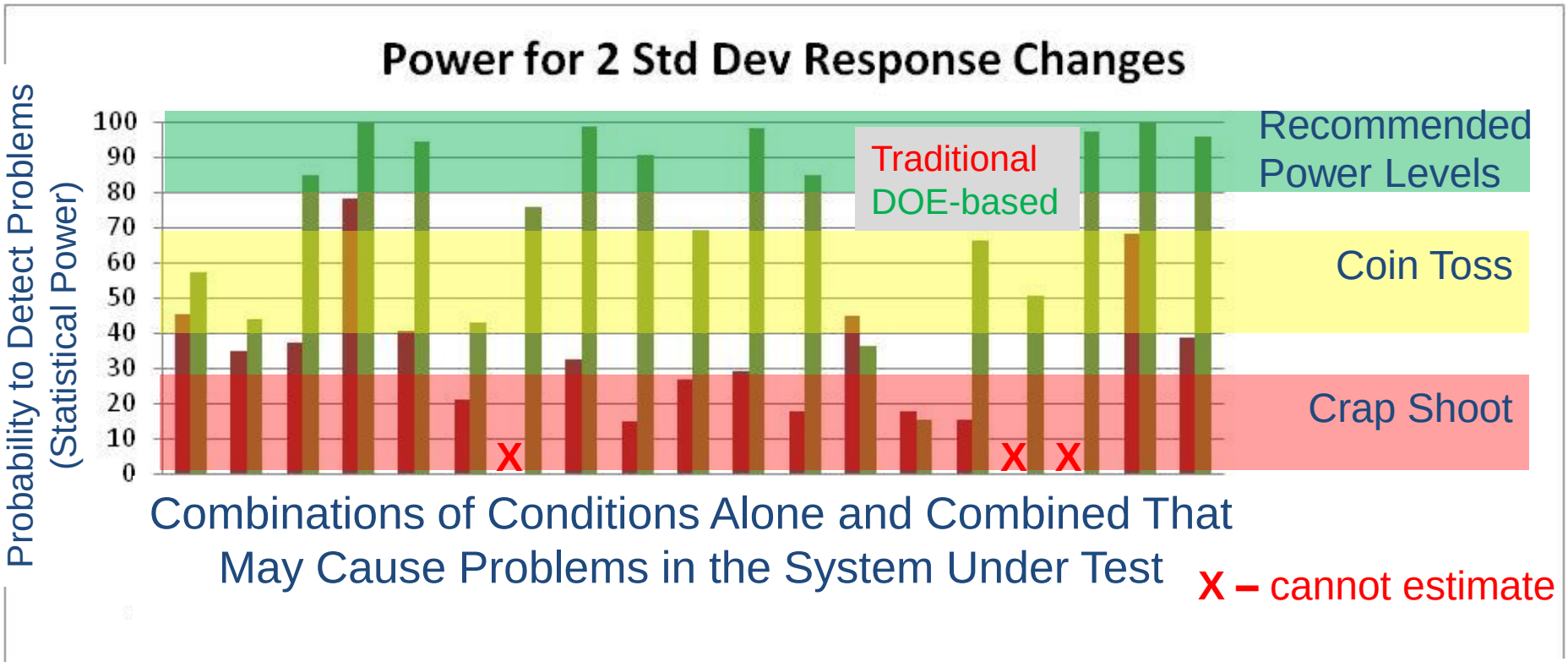
Design Checklist - Power

Excellence Checklist for:

Metric IIb. Design to Control the Risk of Wrong Conclusions

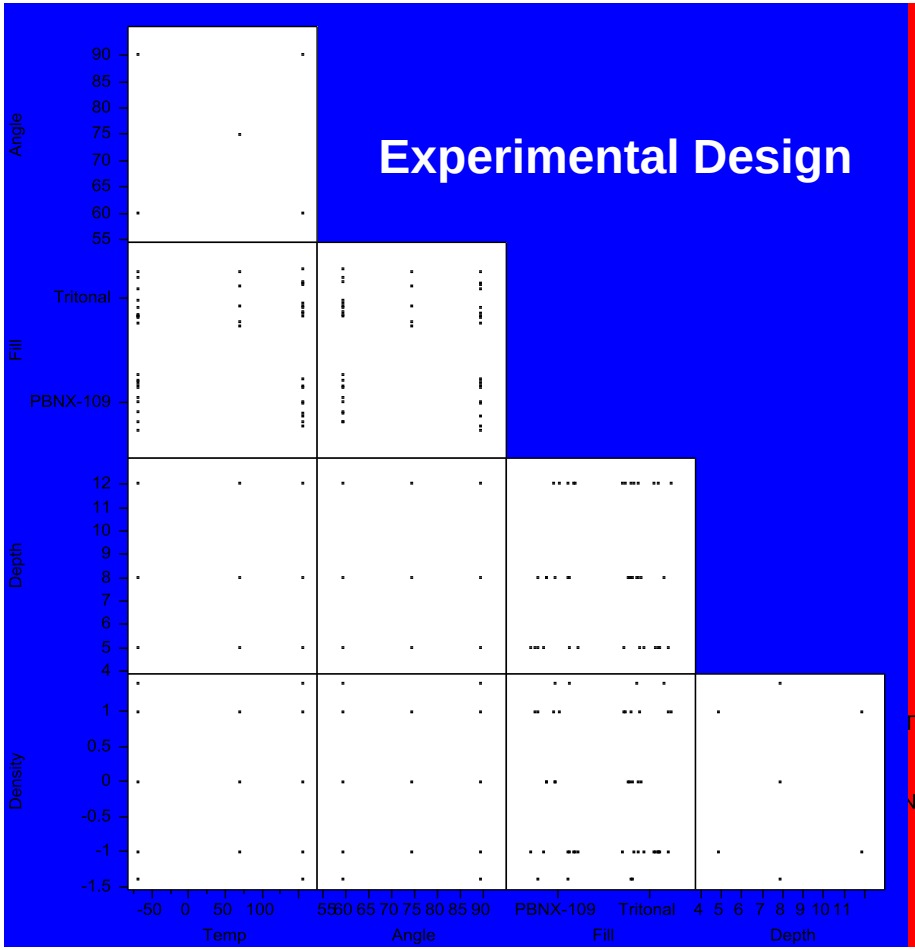
- Design reports power for one or more key performance measures (e.g. accuracy, task time, etc)
  - Best practice – power reported for a range of total events (*N*)
- Reported power also lists other power analysis parameters
  - Name of experimental design strategy
  - Size of design – number of factors (*k*) and levels planned
  - *N* – total number of trials
  - Chosen level of α error for power values
  - Expected process noise and basis for estimate (σ)
  - Shift in process performance to be detected and basis for importance (δ)
- Power reported with alpha risk across all tests in test plan

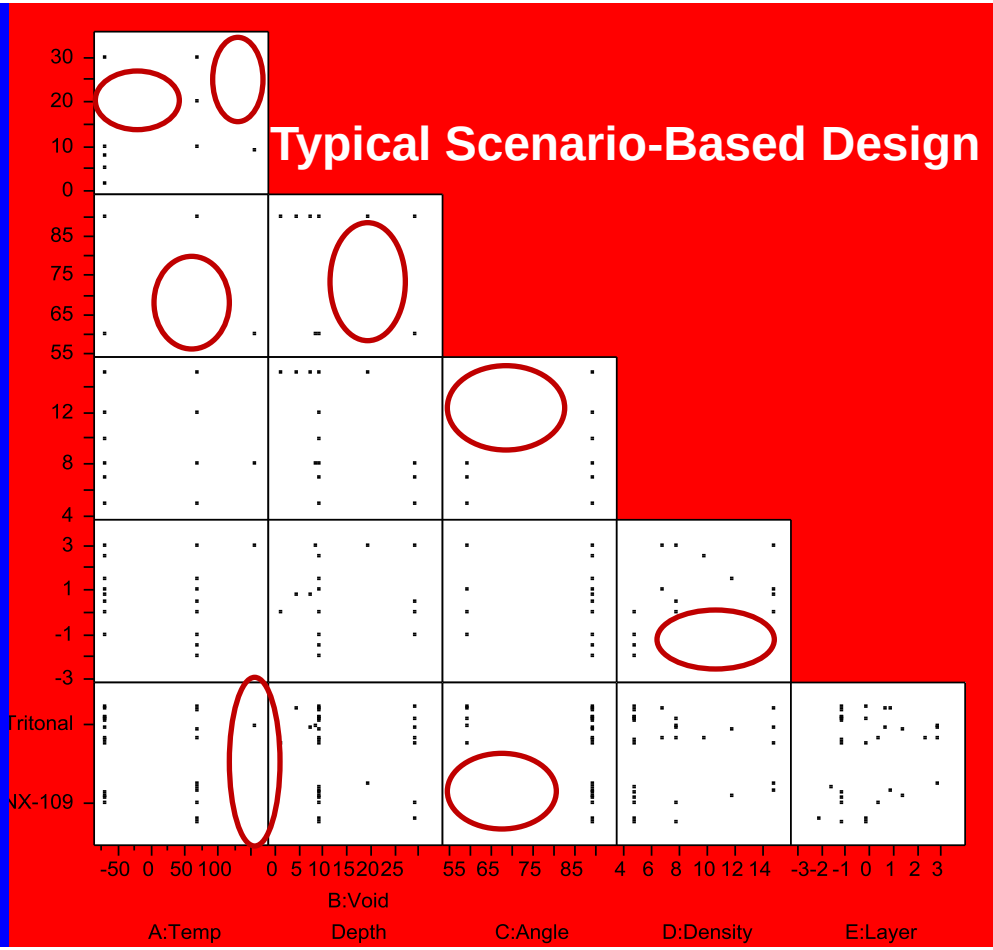
We are not OK Now: Traditional Test Design Likely to Miss Important Problems



- **Setting: Sled test for new system – 10 shots for BLU-109 warhead**
- **Traditional design used real world–like "scenarios" or "vignettes"**
- **As RTO, 46 TW signed TEMP, but cautioned Program Office client**
- **46 TW recommends further work to improve test design**
- **DOE Designs cost-effective: "Pay me now … or pay me later."**
- **Note: despite some claims – number of runs held constant both cases**

Statistical Power depends not only on *how many* points but *where* they are placed in the space

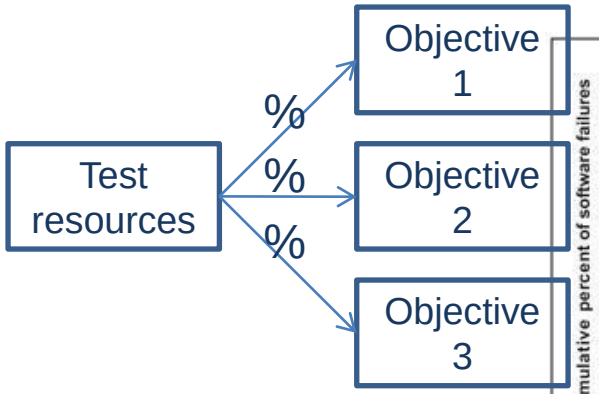




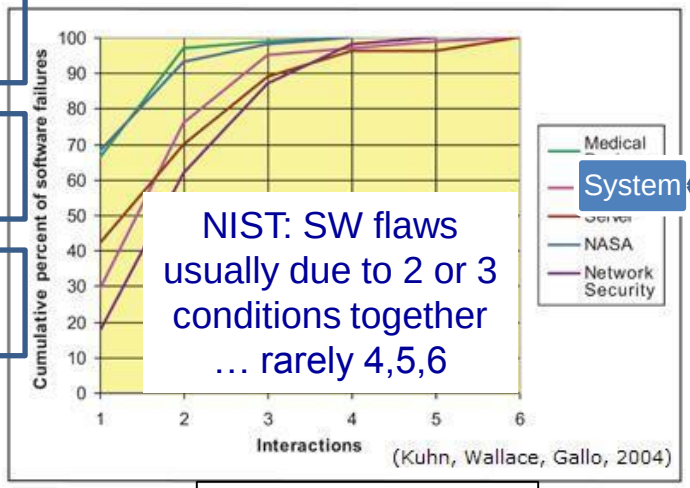
Note even placement of points to cover all targets in the battlespace (repeated points are “jittered” to show number of points)

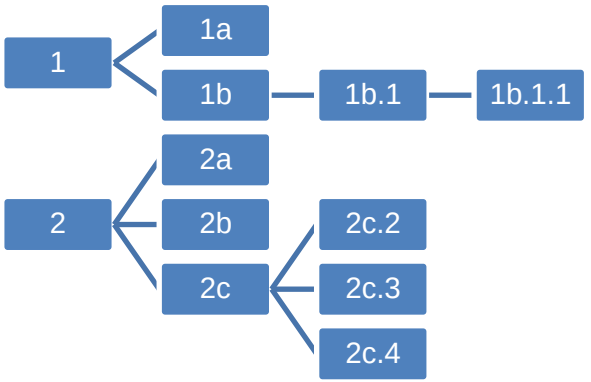
Note gaps and concentrations from choosing “typical targets”

DOE for SDB Integration: Not Stochastic - Deterministic



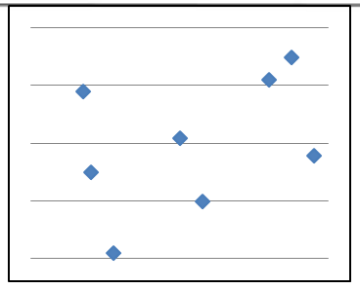
Decision Analysis





Factor Covering Arrays

Space Filling

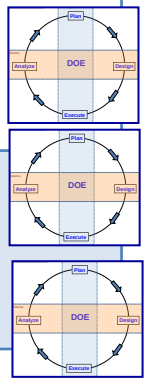


- **In software functionality (vice performance) combinatoric and space-filling designs fit the problem; analysis "by inspection"**
  - **How to spread out test resources effectively/efficiently**
  - **How to test configurations effectively/efficiently**
  - **How to fill a space effectively/efficiently**

See: http://csrc.nist.gov/groups/SNS/acts/index.html

Multiple Experiments for a Test
Dragon Spear AC/MC-130*

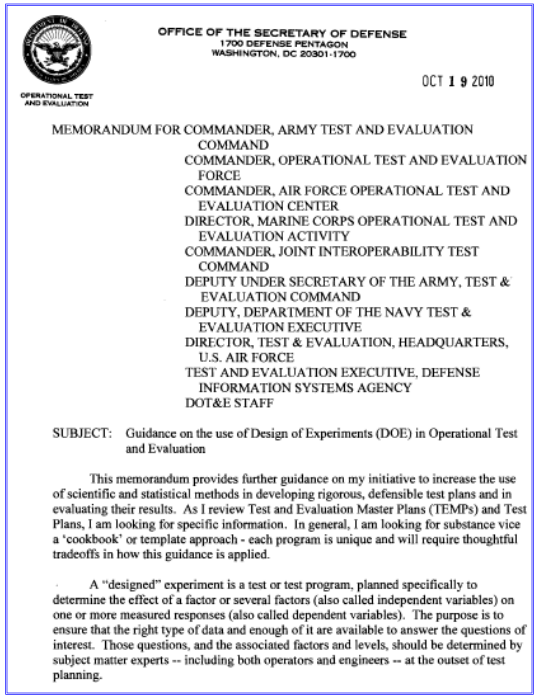
Mission or Capability	Purpose	Operator Interest	Initial Design	Factors / Points
CAS/AI	Tgt Det/Trk/Engage (Dry)	Primary	Fractional Factorial	8 / 40+20
	Aircraft Survival	Primary	D-Optimal	6 / 35
	Live Shot (all weaps)	Primary	RSM FCD	4-5 / 28-30
	Move/Man'ver Tgt (Dry)	Excursion	RSM FCD	5 / 30-35
	Sensor TLE (EO-IR-SAR)	Primary	RSM FCD	6 / 95 + TLE90
Suitability	Aircraft Generation	Primary	Full Facotrial w/ rep pts	4 /20
	Weapons Loading	Primary	Mixed Level D Optimal	5 / 32-35
A/C Integration	SDB Stores S-W Mode Exploration (Bug Hunt)	Excursion	Factor Covering Array (Str 4)	8 / 78
Demonstrations	Some information	Minor	GPS Jx, Austere Generation, Max range SDB	0 / 1-2



*Designs Representative – not Sqdn-Approved as of 9 Feb 12

We can now answer Dr Gilmore's Checklist for TEMP Excellence

- **19 Oct 2010 DOT&E Guidance Memorandum**
- **Program TEMPs should address the following questions:**



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

OPERATIONAL TEST AND EVALUATION

OCT 19 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

Checklist for TEMP Excellence:

- ☑ What are your *Goal(s)*?
- ☑ How to measure *success*? (MOPs and MOSs)
- ☑ How many *Trials*?
- ☑ Under what *conditions*?
- ☑ Point placement *strategy*?
- ☑ Statistical metrics – *risk* of wrong decisions
- ❑ Execution decisions?
- ❑ Method of *analysis*?

Summary: SE – Connecting to Battlespace

- All test programs are not equal & not all are well-designed
  - Inform each stage of testing from previous tests
  - Span the battlespace - careful placement of enough points
  - Best success is planning a campaign of experimentation
- STAT/DOE is not another 3-letter, 4-letter word
  - DOT&E/DDT&E asking us to Raise the Bar of excellence in test
  - Objective measures of excellence – not a matter of opinion
- In summary the science of test provides:
  - the most powerful allocation of test resources for:
    - A well-chosen number of test events
    - If budget-constrained, for a given number of tests
  - a scientific, structured, objective way to plan tests
  - an efficient approach to integrated testing

STAT and other DOE Resources

Texts:

Design and Analysis Of Experiments Douglas C. Montgomery (2008), 7th edition, hardcover with 704 pp.; ISBN 0-471-15746-5; $81.95

Response Surface Methodology Process and Product Optimization Using Designed Experiments Raymond H. Myers and Douglas C. Montgomery (1995), hardcover with 700 pp.; $59.95

Statistics for Experimenters George E. P. Box, William G. Hunter, and J. Stuart Hunter (2004), 2nd Ed. hardcover with 638 pp.; ISBN 0-471-09315-7; $69.95

Links:

http://www.nap.edu/catalog/6037.html?send An NRC-directed study paid for by DOT&E and the service OT Agencies that recommends DOE for military acquisition testing. Online-Adobe format.

http://www.minitab.com/resources/articles/StatisticsArticles.aspx Contains Minitab's (a stats package) take on some statistical topics of interest.

www.statease.com Is an accessible site with good references. Stat Ease writes the software Design Ease and its big brother Design Expert. Best DOE-dedicated software in the business with sound caution.

http://www.itl.nist.gov/div898/handbook/index.htm Maybe the best online source – an engineering statistics handbook at the National Institute of Science and Technology (NIST.)

http://www.stat.uiowa.edu/~rlenth/Power/index.html My vote for the best power-analysis site on the Web. Authored by a well-know reseacher and practitioner, Dr Russ Lenth – Java code from U of Iowa.

www.jmp.com Excellent general stats package that has the best implementation of I-Optimal and Space-Filling designs. Very good at DOE, but interface has its' little ways. Caution and expertise advised.

https://extranet.dote.osd.mil/about/workinggroups/index.html DOT&E Extranet with OSD working group minutes and products

https://afkm.wpafb.af.mil/community/views/home.aspx?Filter=OO-TE-MC-79 Current USAF DOE CoP

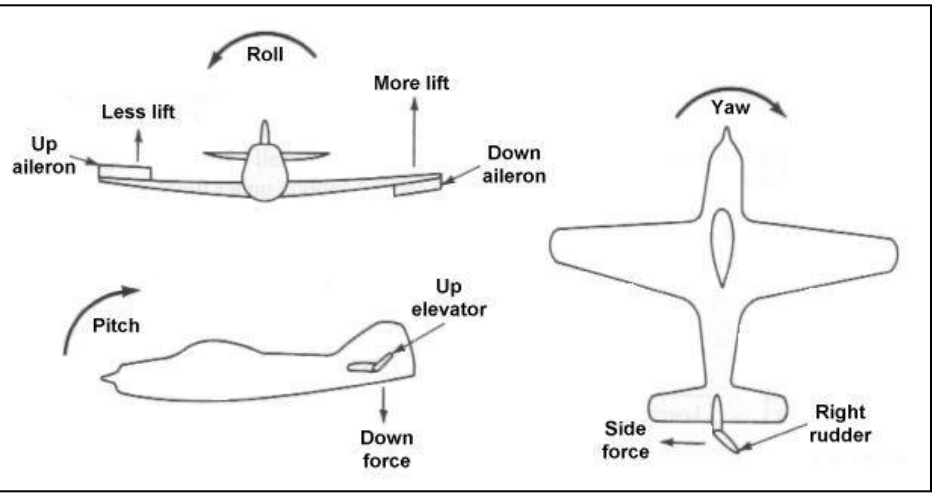
https://eglin.eis.af.mil/53tmg/DOE/default.aspx New USAF CoP location – need SharePoint account

Here are some sound references and tools to practice STAT & DOE

STAT: The *Science* of Test

Questions?

w/o DOE: Anonymous Aircraft TF/TA Performance DT&E – Not Executed



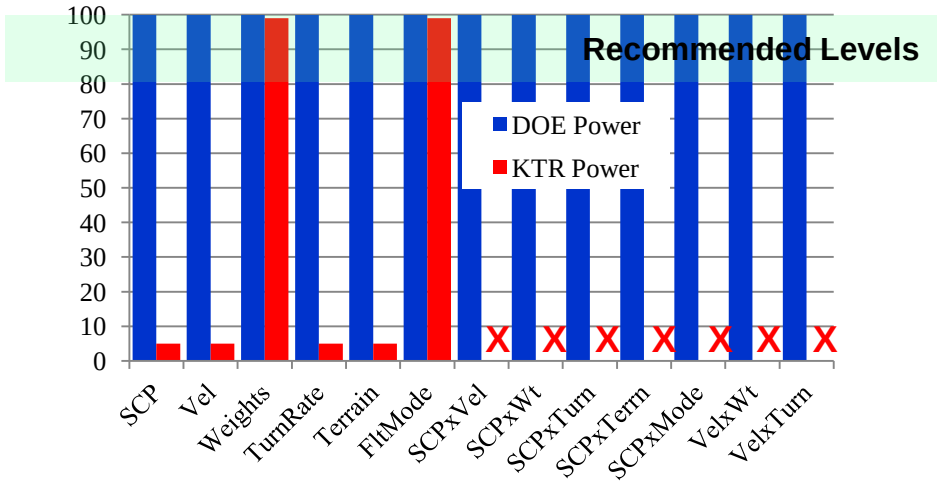
Test Objective:

- **Diverse stakeholders – KTR Team, 46 TW, other Services, two OTAs**
- **Low speed TF/TA performance to be evaluated**
- **Constrained sorties for aircraft test**
- **SPO/KTR solution – use the points used last time plus some expert choices**

DOE Approach:

- Chart at right shows two designs – KTR/DOE
- Team w/ SME worked several days to search same space with designed experiment
- Statistical Power* - 2 designs shown at right –
- KTR is 46 runs – DOE is 39 runs – ***20% savings*** with <u>much</u> better statistical power
- Effect of many test conditions can't be estimated w/ KTR runs (power=0.0)

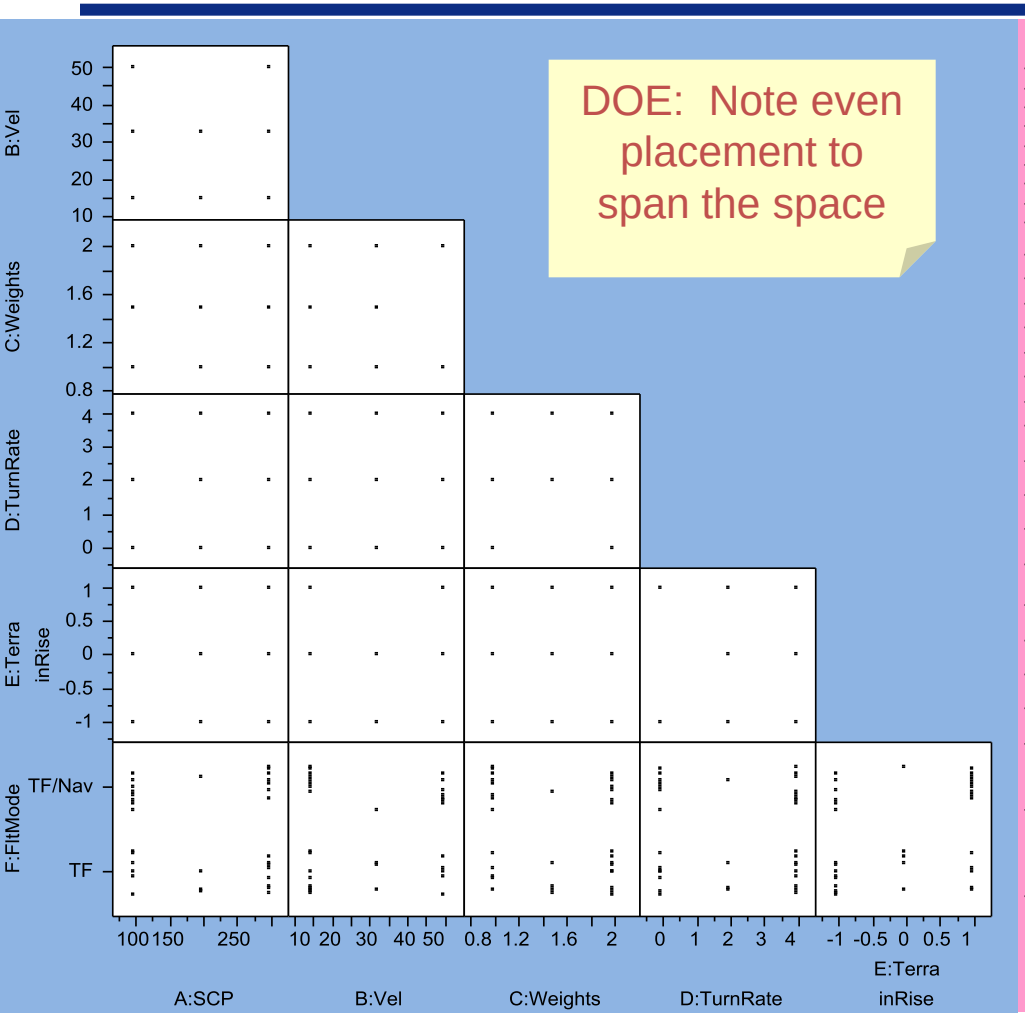
* Power = Prob(Det problem if problem exists)

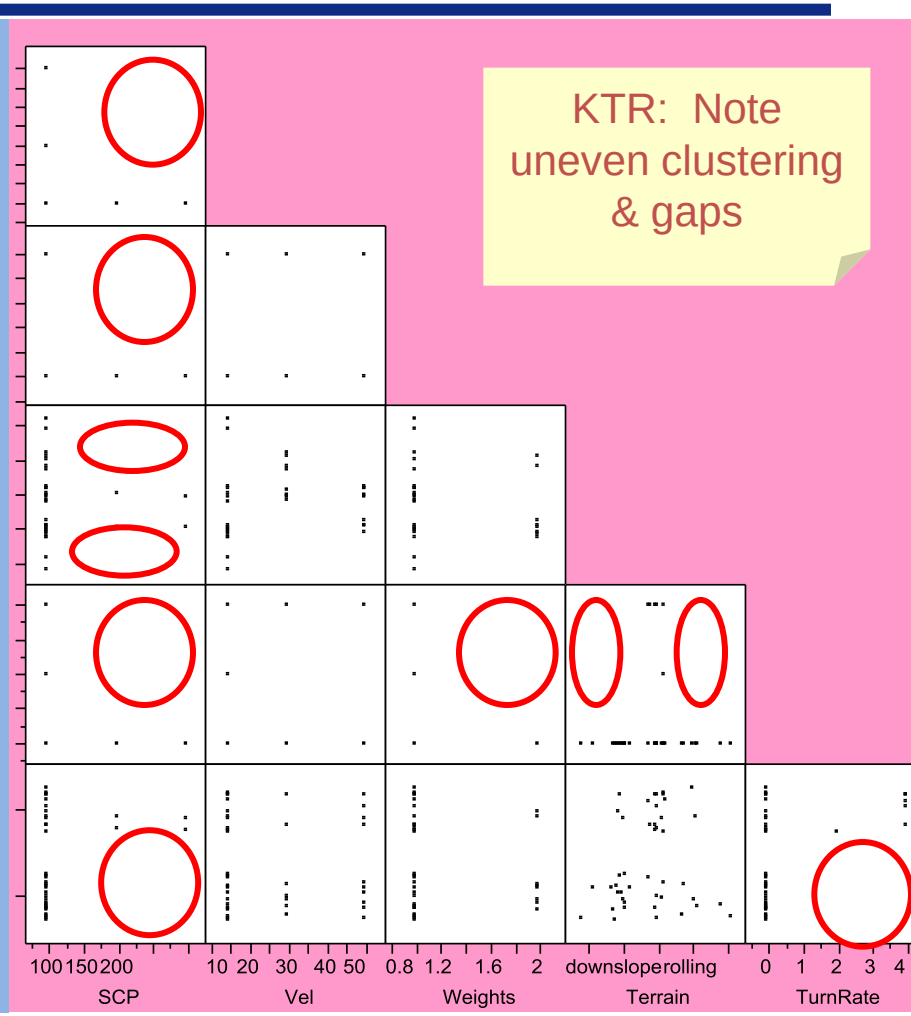


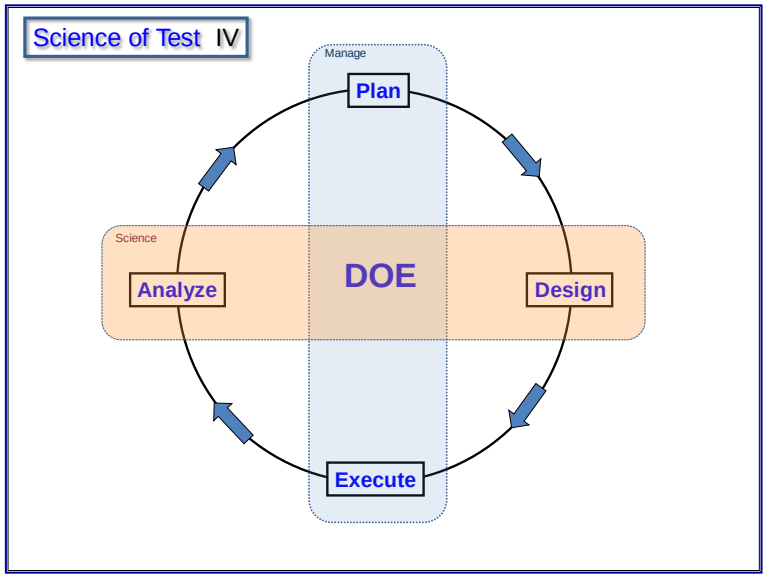
Results: **X –** cannot estimate

- Ran KTR design since run cards already made

Why so many "X's"? Varying just one condition at a time spoils statistical power






Science of Test IV
Manage
Plan
Science
Analyze
DOE
Design
Execute

EXECUTE

Execute: How to Sequence?

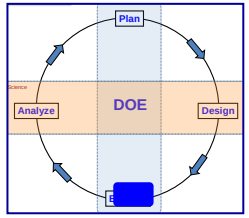
Standard Order

Case	A	B	C	Response
1	-1	-1	-1	
2	1	-1	-1	
3	-1	1	-1	
4	1	1	-1	
5	-1	-1	1	
6	1	-1	1	
7	-1	1	1	
8	1	1	1	

Randomized

Case	A	B	C	Response
2	1	-1	-1	
8	1	1	1	
5	-1	-1	1	
4	1	1	-1	
1	-1	-1	-1	
3	-1	1	-1	
7	-1	1	1	
6	1	-1	1	

- All possible combinations of three factors at 2 levels each
- If run in structured sequence, possible outside influence over time
- If run in random order, outside influences are averaged out – best strategy
- DOE can be effectively used if some factors are hard to change



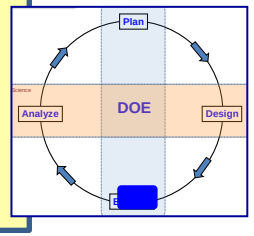
Execute

Design Blocked & Randomized …

	Mission	Gross Weight	SCP	Turn Rate	Airspeed	Ride	SCP Deviation	Pilot Ratings
Mission 1	1	47.5	500	0	230	Hard	15	2.8
	1	47.5	400	2	195	Medium	4	4.2
	1	47.5	500	4	230	Medium	16	2
	1	47.5	300	0	160	Medium	5.6	4.5
	1	47.5	300	4	160	Hard	5.2	4.2
	1	55	400	2	195	Hard	7.2	3.7
	1	55	500	0	160	Medium	2.3	4.8
	1	55	300	0	230	Hard	0.2	5.4
	1	55	300	4	230	Medium	1.9	5
	1	55	500	4	160	Hard	6.7	3.4
Mission 2	2	47.5	500	0	160	Hard	7.5	4.2
	2	47.5	300	0	230	Medium	4	4.8
	2	47.5	300	4	230	Hard	5.8	4.5
	2	47.5	500	4	160	Medium	12	3.2
	2	47.5	400	2	195	Hard	7.7	3.8
	2	55	300	0	160	Hard	0.5	4.8
	2	55	500	4	230	Hard	12	2.5
	2	55	300	4	160	Medium	1.2	4.6
	2	55	400	2	195	Medium	6.6	4.4
	2	55	500	0	230	Medium	8.3	3.2

20 Design Points

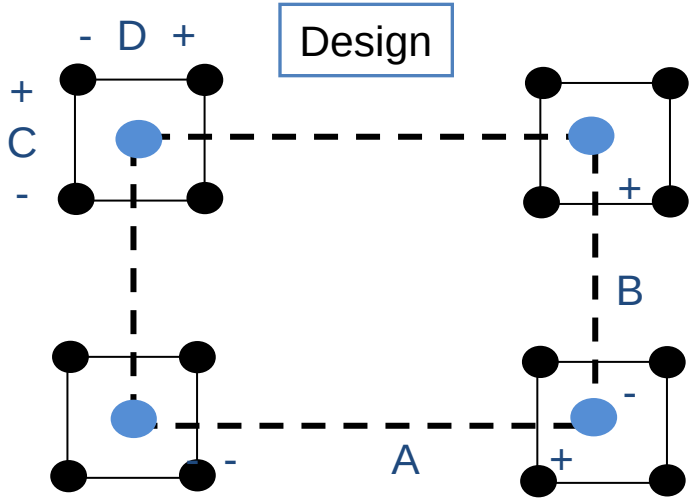
Responses (MOPs)



Some factors hard, expensive, unsafe to change: Split-Plot Designs

Assumptions
Hard to Change Factors
Numeric or Categorical

Design



Attributes
Replication
Orthogonal

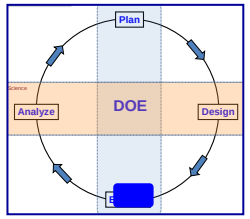
Assumptions
Two Independent Error Terms, both NID (0, σ^2)
Model is adequate
Y well behaved

Model

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \delta + \varepsilon$$

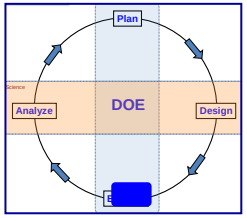
WP error

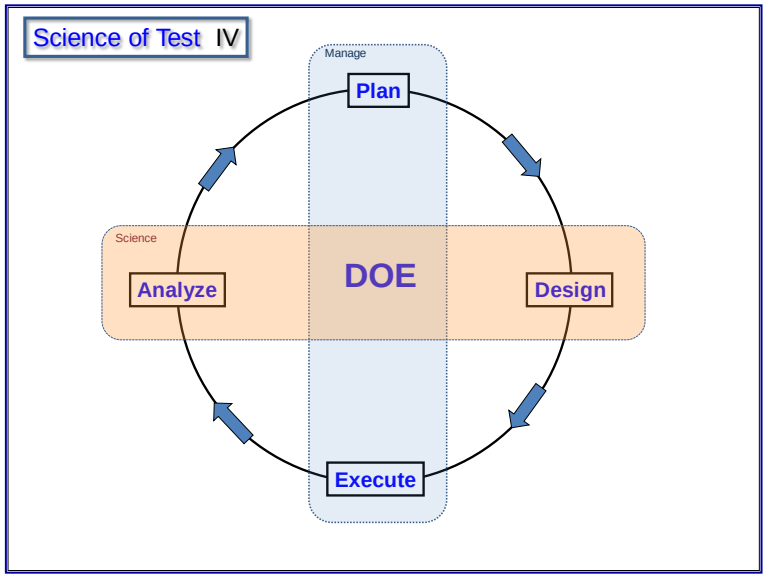
Attributes
All effects of interest
Limited WP error df
Independent β estimates



Execution Checklist

Excellence Checklist for: ***Metric III. Execute the Test***
❖ Name of the chosen execution strategy to account for background change. For example: ○ Completely randomized   ○ Factorial in blocks   ○ Split plot design with easy- and hard-to-change factors   ○ Analysis of Covariance   ○ With replication vs. repetition   ❖ Describe methods to control background variability   ❖ Describe approach to ensure independence of successive observations




Science of Test IV
Manage
Plan
Science
Analyze
DOE
Design
Execute

ANALYZE

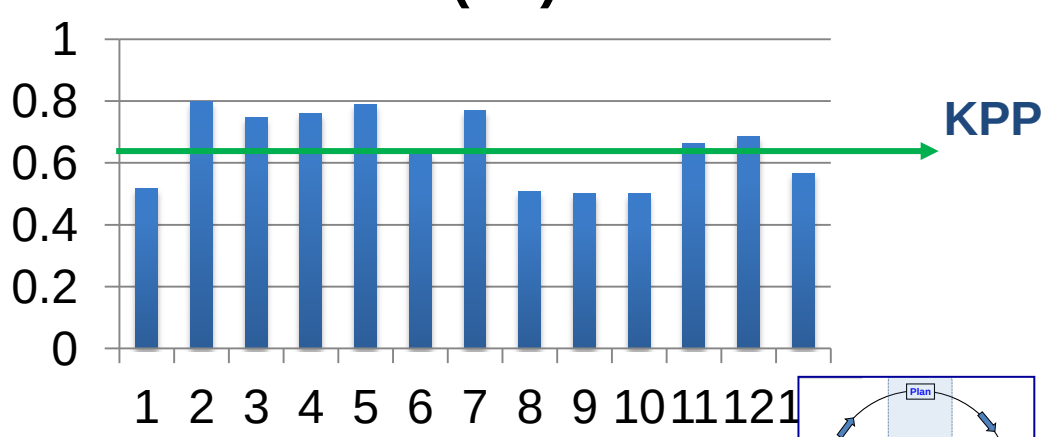
Analyze: What Conclusions?

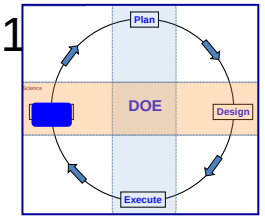
Traditional "Analysis"

- **Cases or Scenario settings and findings**

Sortie	Alt	Mach	MDS	Range	Tgt Aspect	OBA	Tgt Velocity	Target Type	Result
1	10K	0.7	F-16	4	0	0	0	truck	Hit
1	10K	0.9	F-16	7	180	0	0	bldg	Hit
2	20K	1.1	F-15	3	180	0	10	tank	Miss

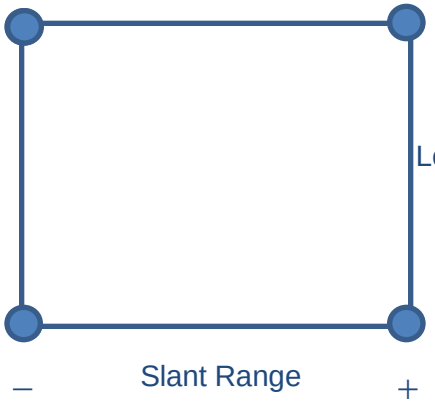
- **Summary performance, subject to**
  - **Change in scale**
  - **Subset the domain to show desired trend**

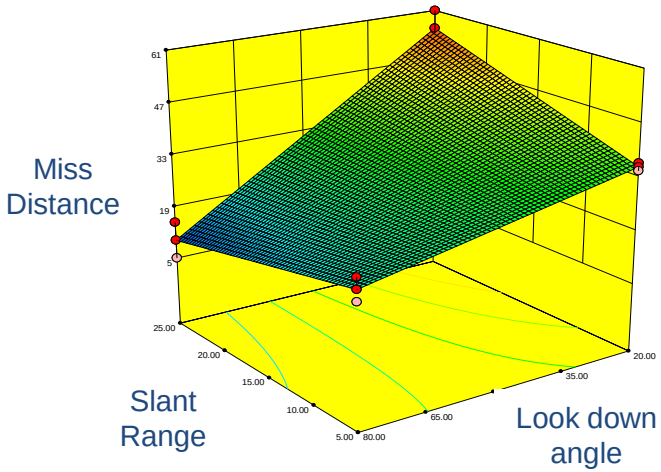




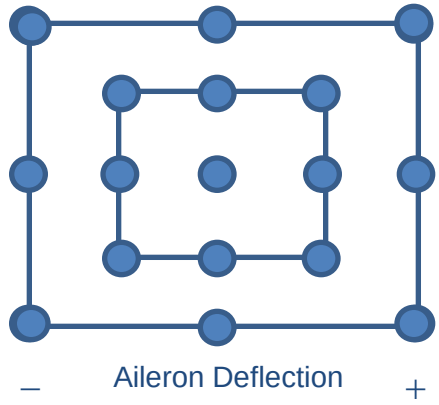
Designs should support the complexity of expected performance

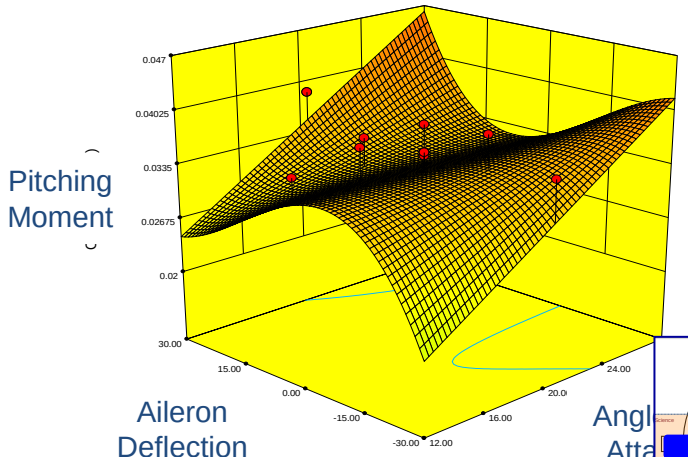
a)

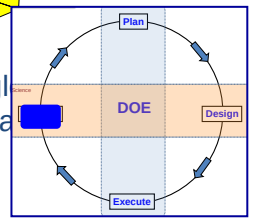




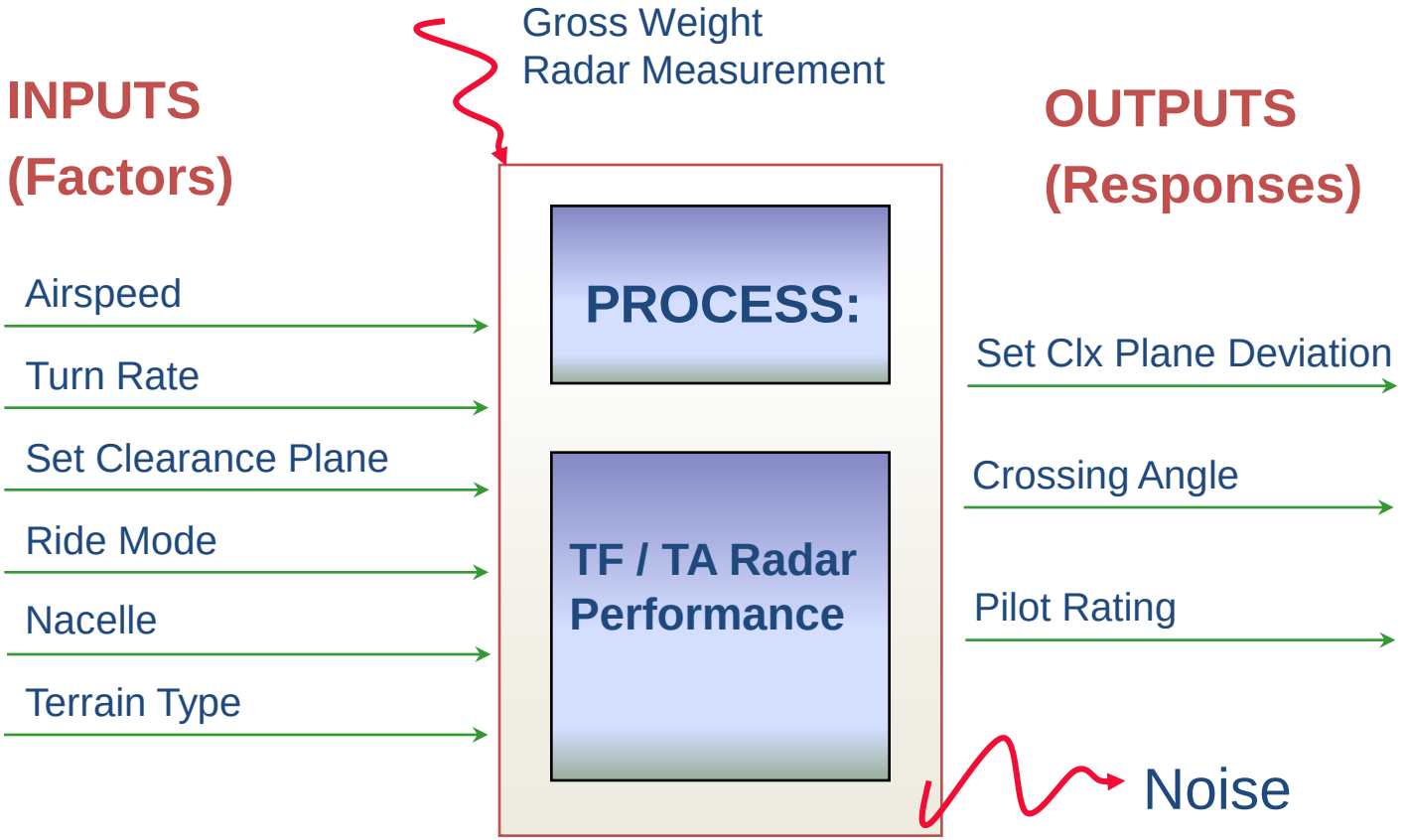
b)



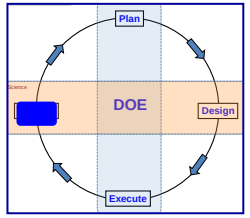




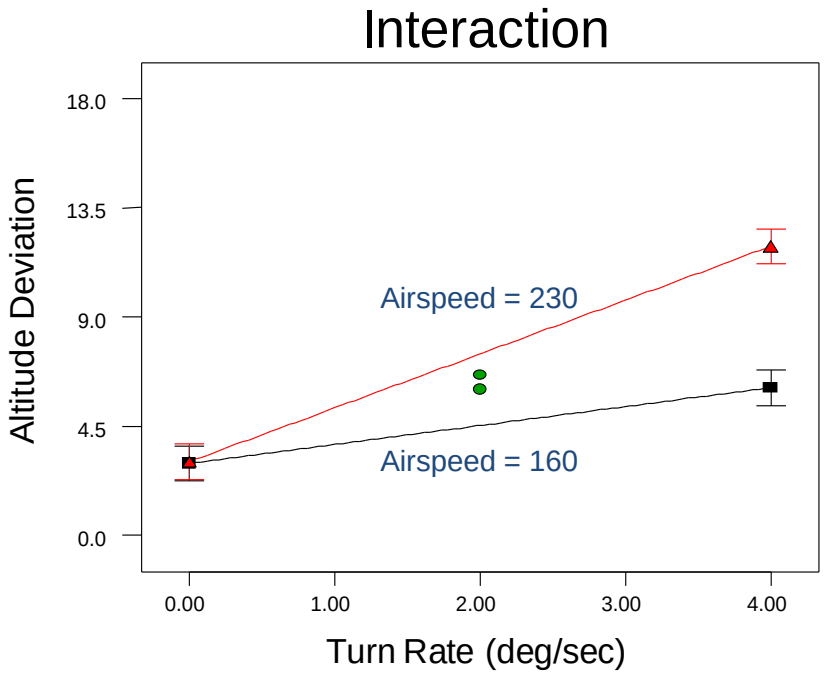
CV-22 TF Example
What Conclusions



$$\text{Responses} = f\left(\text{Factors}\right) + \varepsilon$$



Analysis: What the Data Reveals



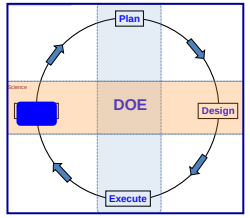
Note the uncertainty (error) bars and Linking cause to effect.

Response Model

Factor settings (-1=Low, +1=High)

Altitude Deviation = + 6.51
- 2.38 * Altitude
+ 3.46 * Turn Rate
+ 1.08 * Ride
+ 1.39 * Airspeed
+ 0.61 * Turn * Ride
+ 1.46 * Turn * Airspeed

Note the test conditions that had an effect, Magnitude and direction, alone and combined, As well as those that did not (Gross Weight).

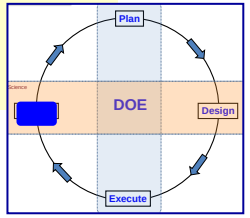


Performance predictions can validate model adequacy

Name	Setting	Low Level	High Level
SCP	**460.00**	300.00	500.00
Turn Rate	**2.80**	0.00	4.00
Ride	**Hard**	Medium	Hard
Airspeed	**180.00**	160.00	230.00

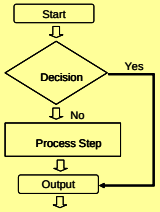
	Prediction	95% PI low	95% PI high
Deviation from SCP	**6.96**	4.93	8.98
Pilot Ratings	**3.62**	3.34	3.90

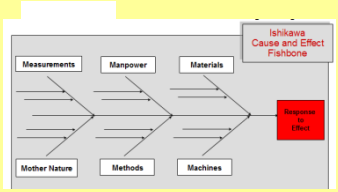
Our math model is capable of validation – set conditions in battlespace
Predict outcome and uncertainty
Run new points and compare.



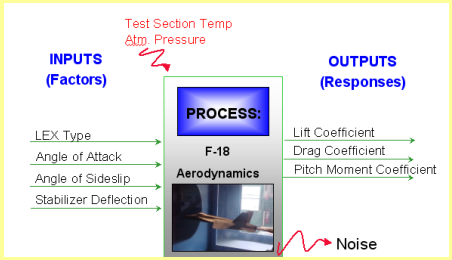
Design of Experiments Test Process is Well-Defined

Planning: Factors Desirable and Nuisance

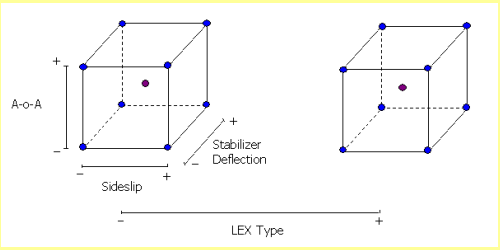




Desired Factors and Responses



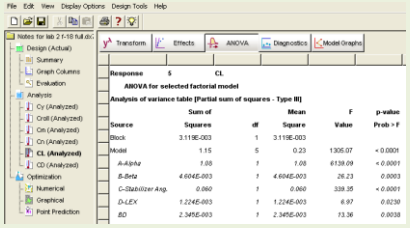
Design Points



Test Matrix

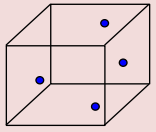
A-o-A	Sideslip	Stabilizer	LEX Type
2	0	5	-1
10	0	-5	1
10	8	5	-1
2	8	5	-1
2	8	-5	-1
2	0	-5	-1
10	8	-5	1
2	0	5	1
2	8	5	1
10	8	5	1
10	8	-5	-1
10	0	5	-1
10	0	-5	-1
2	8	-5	1
10	0	5	1
2	0	-5	1

Analysis and Model



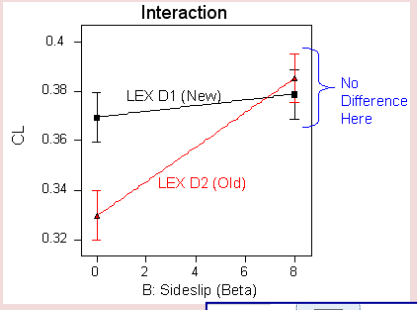
C_L = +0.38
+0.26 x A-o-A
+0.017 x Sideslip
+0.061 x Stabilizer Deflection
-.00875 x LEX Type
+0.012 x Sideslip x LEX Type

Validation

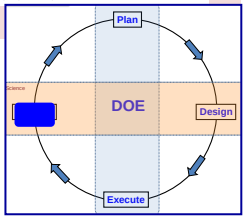


Actual	Predicted	Valid
0.315	(0.30 , .33)	✓

Discovery, Prediction



Just as in ISO 9000 and Software CMM – A solid, teachable *Process* does not leave excellence to chance or individual genius



Analysis Checklist

Excellence Checklist for:

Metric IV. Analyze the Experimental Design

- ❖ Ensure objectives of the test agree with the analysis objectives – screen, characterize, compare, predict, optimize, or map
- ❖ Describe the capability and intent to statistically analyze and model the measures
  - Explanation of modeling strategy
  - Intent to determine factor significance, quantify uncertainty, and provide intervals for estimation/prediction
- ❖ Compare the design strategy to the intended general model
  - State the general model intended for the design – linear, interaction, etc
  - Ensure adequate tests to enable fitting the general model, estimating error, and even fitting a model more complex than assumed (lack of fit)
  - Describe the confounding effects – e.g. resolution
- ❖ Detail the sequential model-building strategy and validation phase outlined
  - Describe strategy to augment initial design to resolve confounding – augmenting, foldover, predict-confirm, etc
  - Report estimate of number of augmenting runs required