Scientific Test and Analysis Techniques: Statistical Measures of Merit

Laura J. Freeman
Research Staff Member
lfreeman@ida.org

IDA

Cleared for open publication.

The Science of Test & Evaluation

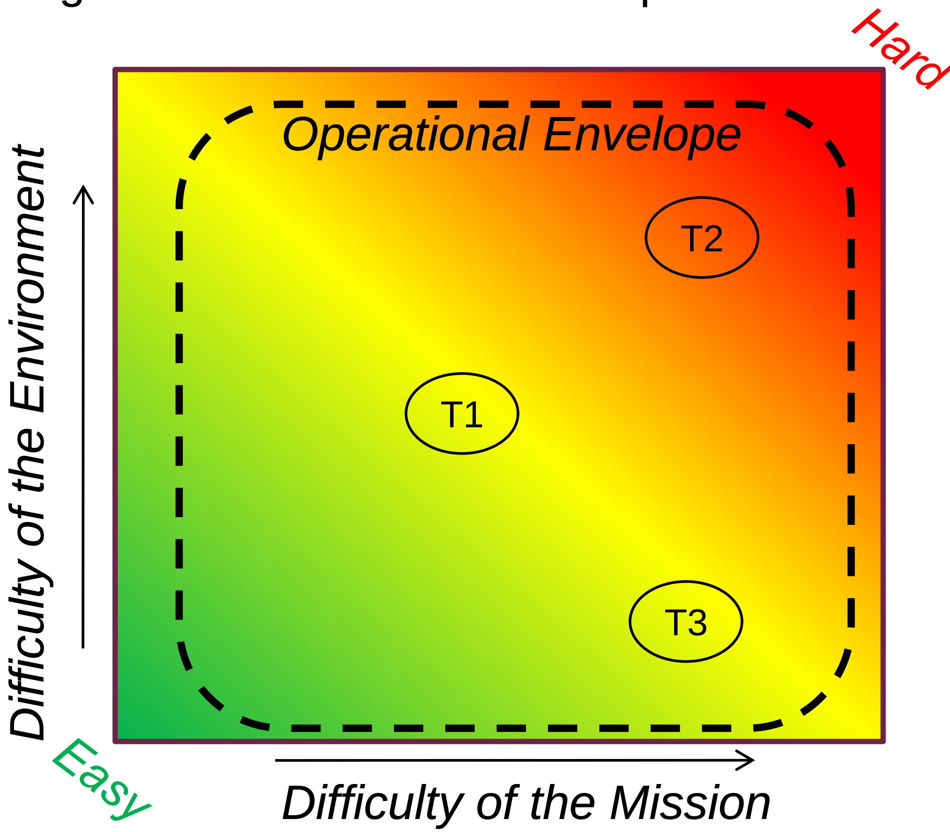
- **Statistics is the science of data analysis**
- **Design of Experiments (DOE) – a structured and purposeful approach to test planning**
  - Ensures adequate coverage of the operational envelope
  - Determines how much testing is enough
  - Provides an analytical basis for assessing test adequacy
  - Results:
    - More information from constrained resources
    - An analytical trade-space for test planning

Cleared for open publication.

Motivation for DOE

- **The purpose of testing is to provide relevant, credible evidence with some degree of inferential weight to decision makers about the operational benefits of buying a system**
  - DOE provides a framework for the argument and methods to help us do that systematically

- **Statistical thinking/DOE provide:**
  - a scientific, structured, objective test methodology answering the key questions of test:
    - How many points?
    - Which points?
    - In what order?
    - How to analyze?



DOE changes "I think" to "I know"

DOE is an Industry Best Practice

Design of Experiments has a long history of application across many fields.

- **Agricultural**
  - Early 20th century
  - Blocked, split-plot and strip-plot designs
- **Medical**
  - Control versus treatment experiments
- **Chemical and Process Industry**
  - Mixture experiments
  - Response surface methodology
- **Manufacturing and Quality Control**
  - Response surface methodology
  - DOE is a key element of Lean Six-Sigma
- **Psychology and Social Science Research**
  - Controls for order effects (e.g., learning, fatigue, etc.)
- **Software Testing**
  - Combinatorial designs test for problems

- **Pratt and Whitney Example**
  - Design for Variation process DOE
  - Turbine Engine Development
- **Key Steps**
  - Define requirements (probabilistic)
  - Analyze
    - Design experiment in key factors (heat transfer coefficients, load, geometric features, etc.)
    - Run experiment through finite element model
  - Solve for optimal design solution
    - Parametric statistical models
  - Verify/Validate
  - Sustain
- **Results**
  - Risk Quantification
  - Cost savings
  - Improved reliability

DOT&E Guidance

Dr. Gilmore's October 19, 2010 Memo to OTAs

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.

for when I approve TEMPs and

t evaluation of end-to-end
tic environment.

es for effectiveness and
Parameters but most likely there

ess and suitability.
y, develop a test plan that
tors across the applicable levels
mation in order to concentrate

ss both developmental and
interest.

dence) on the relevant response
stical measures are important to
can be evaluated by decision-
e off test resources for desired

lentify the metrics, factors, and
nd suitability and that should be
reflected in detailed test plans. DOT&E is working with other members of the test and evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. M. Gilmore

J. Michael Gilmore
Director

cc:
DDT&E

2

- **The goal of the experiment**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- **Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, taking into account known information in order to concentrate on the factors of most interest.
- **A method for strategically varying factors** across both developmental and operational testing with respect to responses of interest.
- **Statistical measures of merit** (power and confidence) on the relevant response variables for which it makes sense. These statistical measures are important to understand "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

IDA

Statistical Measures of Merit

- **The appropriate statistical tools depend on the goal of the test.**
  - What conclusion does the test need to support?
  - What statistical analysis will be used?

Statistical analysis methodology and model are essential!

- *Mean*
- *Median*
- *Variance*
- *Models:*
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_1 x_2 + \beta^2 x_1$

Drives which tools are appropriate and how they should be used

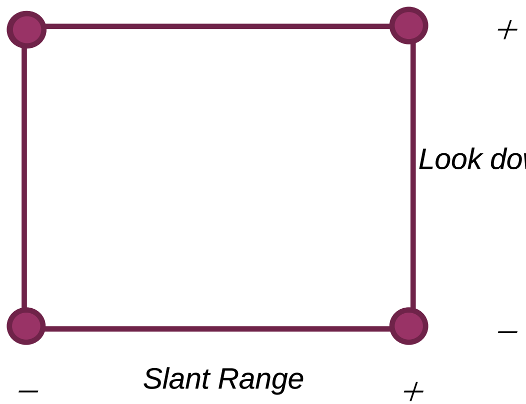
__Test Risks__
- *Confidence*
- *Power*

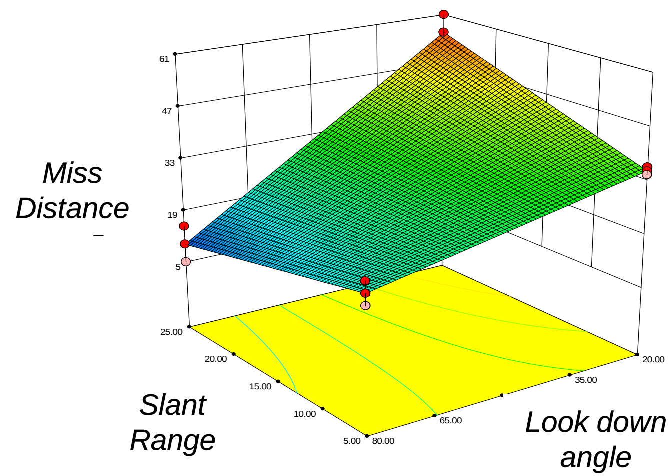
__Factor Relationships__
- *Correlation*
- *Variance Inflation Factors*

__Prediction Capabilities__
- *Scaled Prediction Variance*
- *Fraction of Design Space*

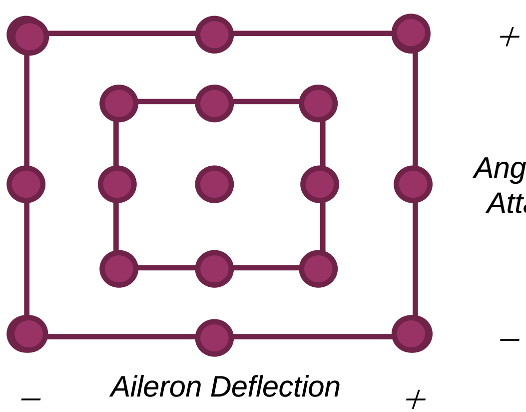
Test Design Supports the Model (The Analysis we expect to perform)

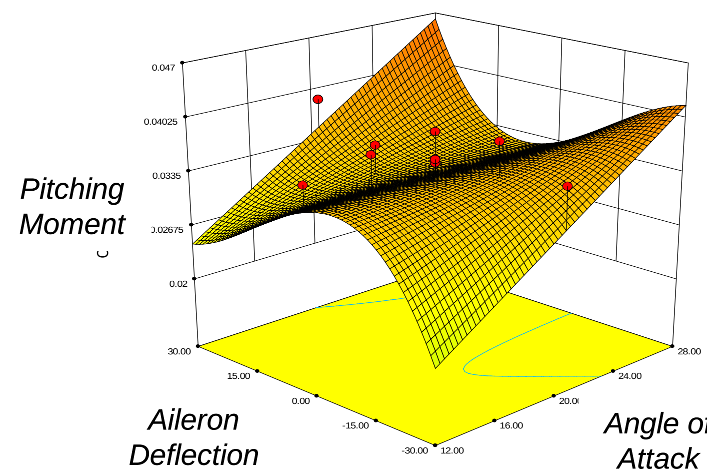
a)





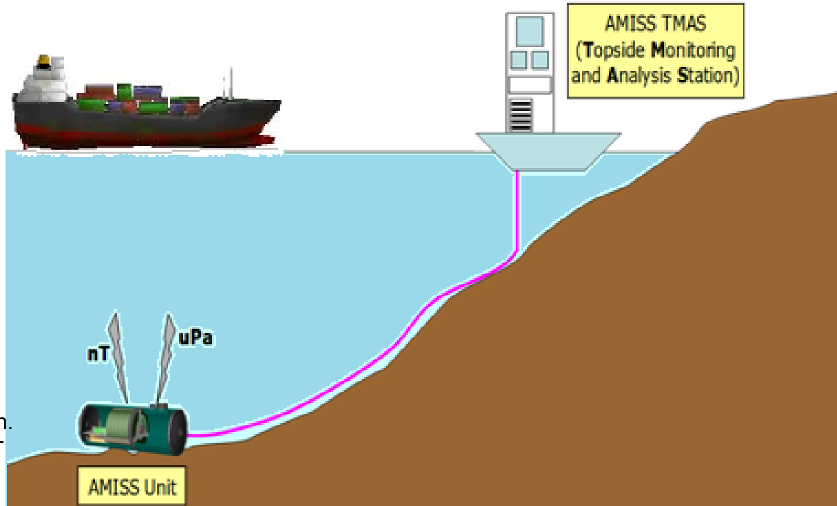
b)





Motivating Example: Test Plan for Mine Susceptibility

- **Goal:**
  – Develop an adequate test to assess the susceptibility of a cargo ship against a variety of mine types using the Advanced Mine Simulation System (AMISS).
- **Responses:**
  – Magnetic signature, acoustic signature, pressure
  – Slant range at simulated detonation
- **Factors:**
  – Speed, range, degaussing system status
- **Other considerations:**
  – Water depth
  – Ship direction

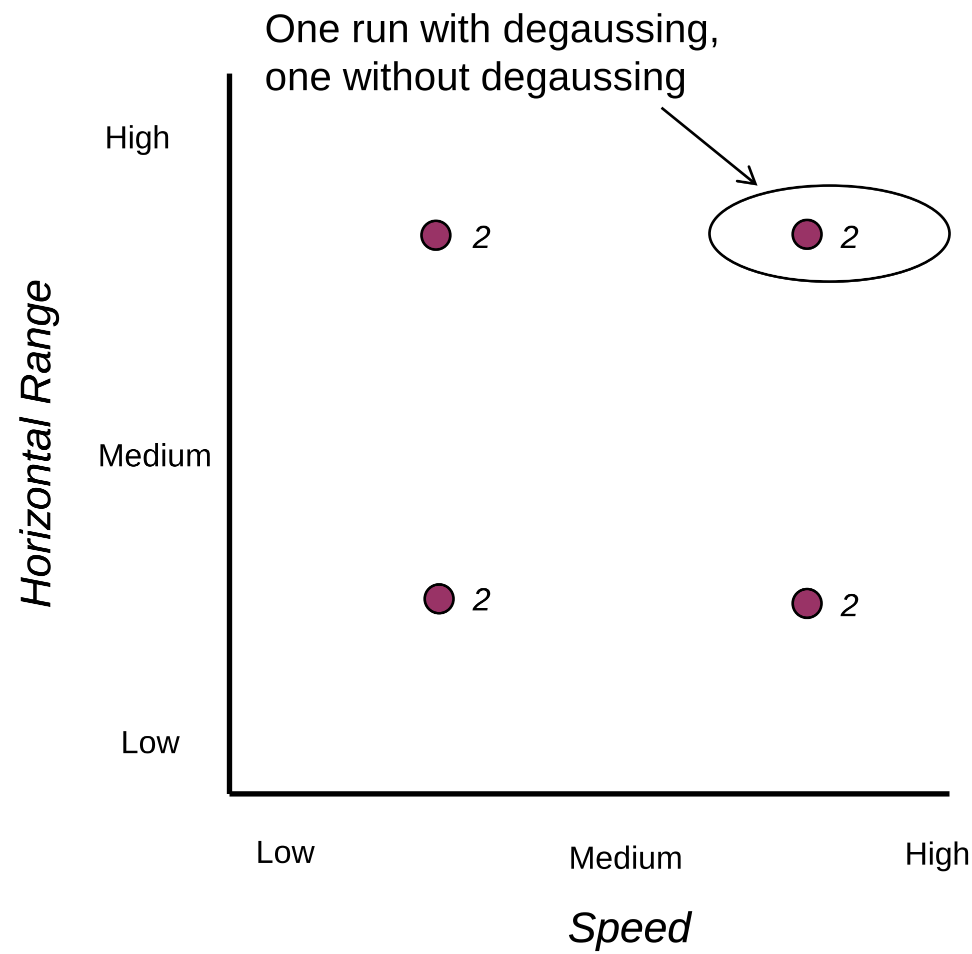


Cleared for open publication.

Test Designs

- **DOE provides a vast library of test design types**

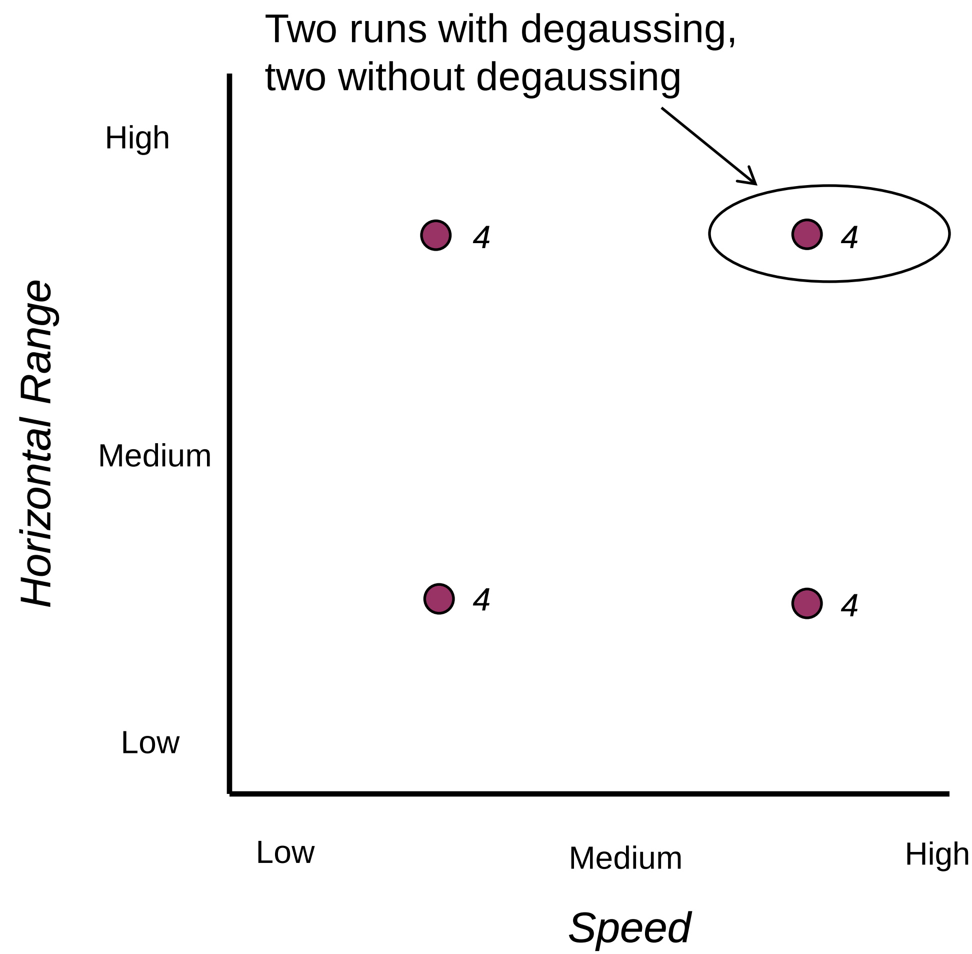
	Design Type	**Number of Runs**
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



Test Designs

- **DOE provides a vast library of test design types**

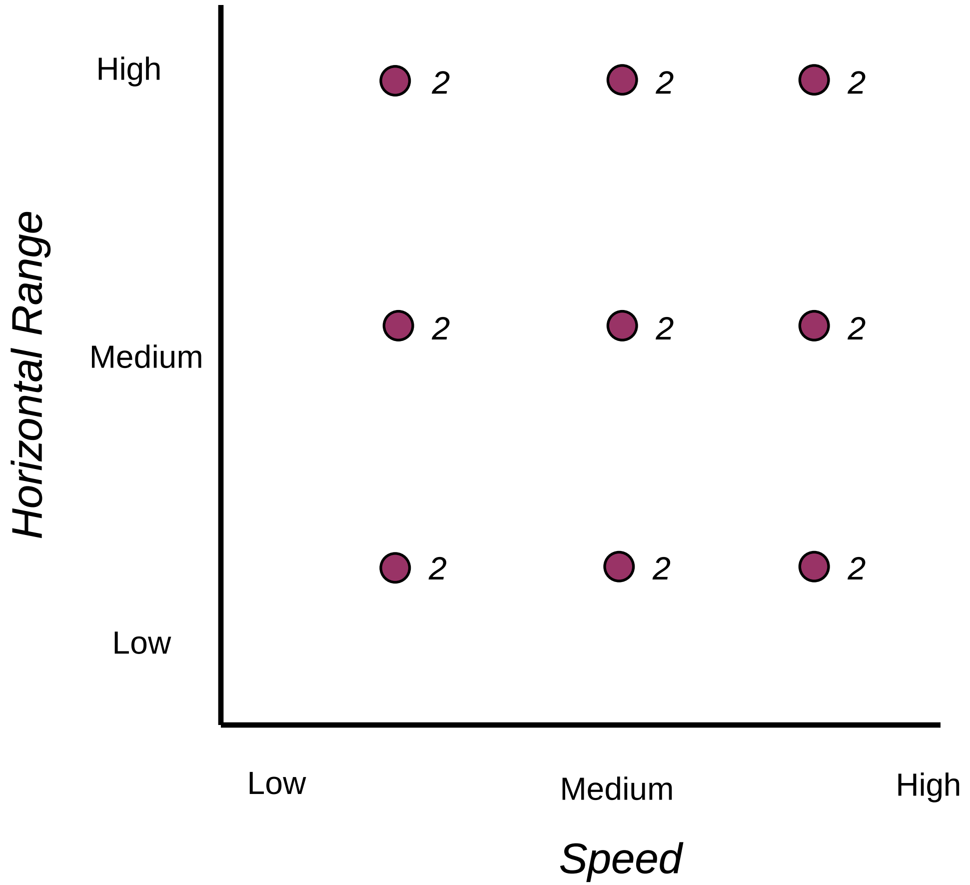
	Design Type	**Number of Runs**
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



Test Designs

- **DOE provides a vast library of test design types**

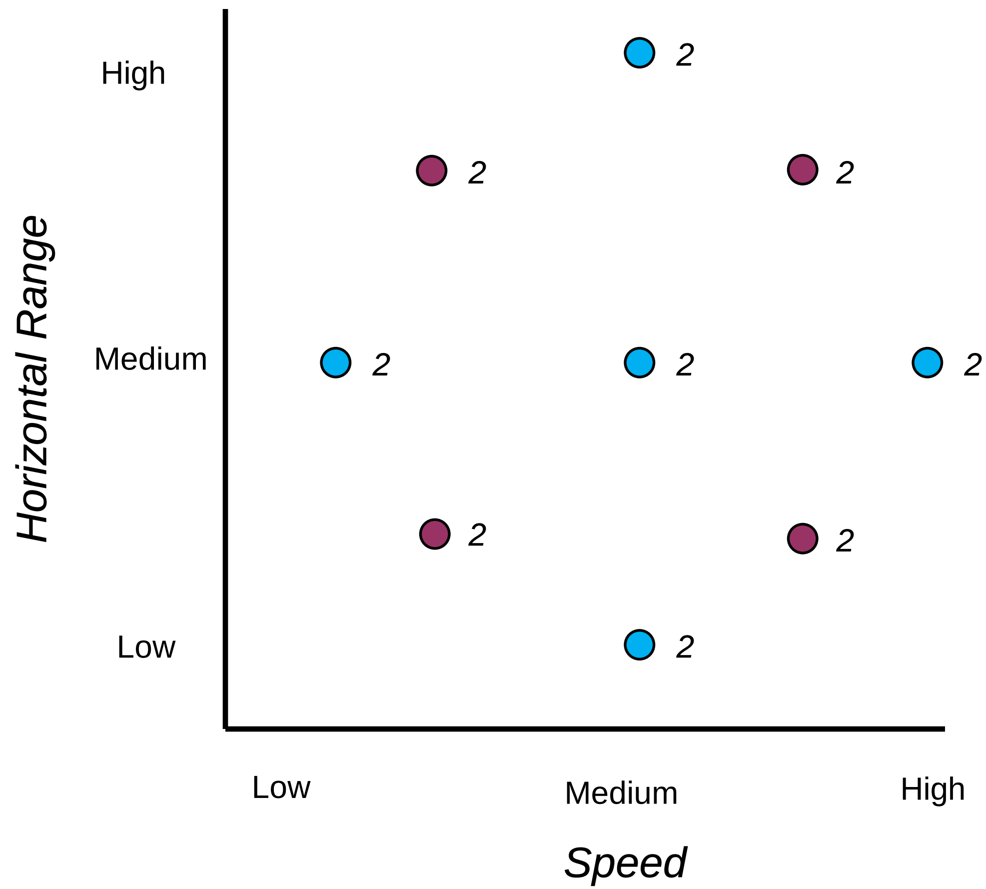
	Design Type	**Number of Runs**
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



Test Designs

- **DOE provides a vast library of test design types**

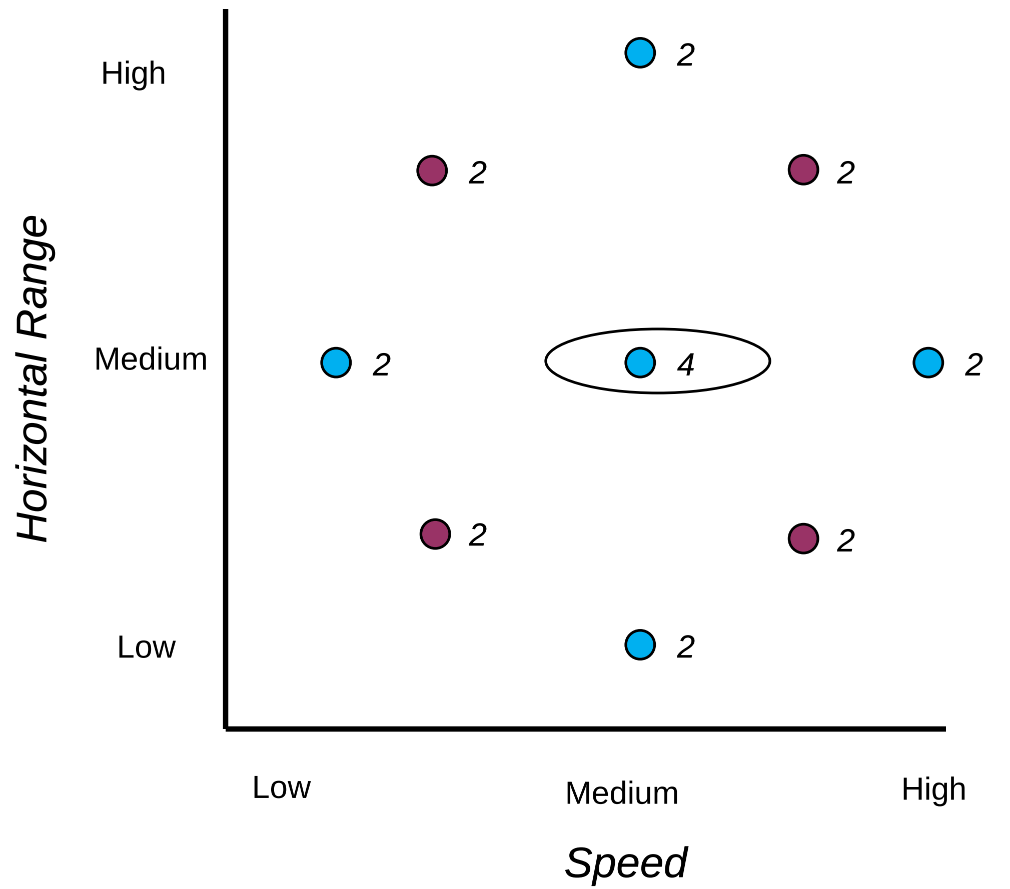
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



Test Designs

- **DOE provides a vast library of test design types**

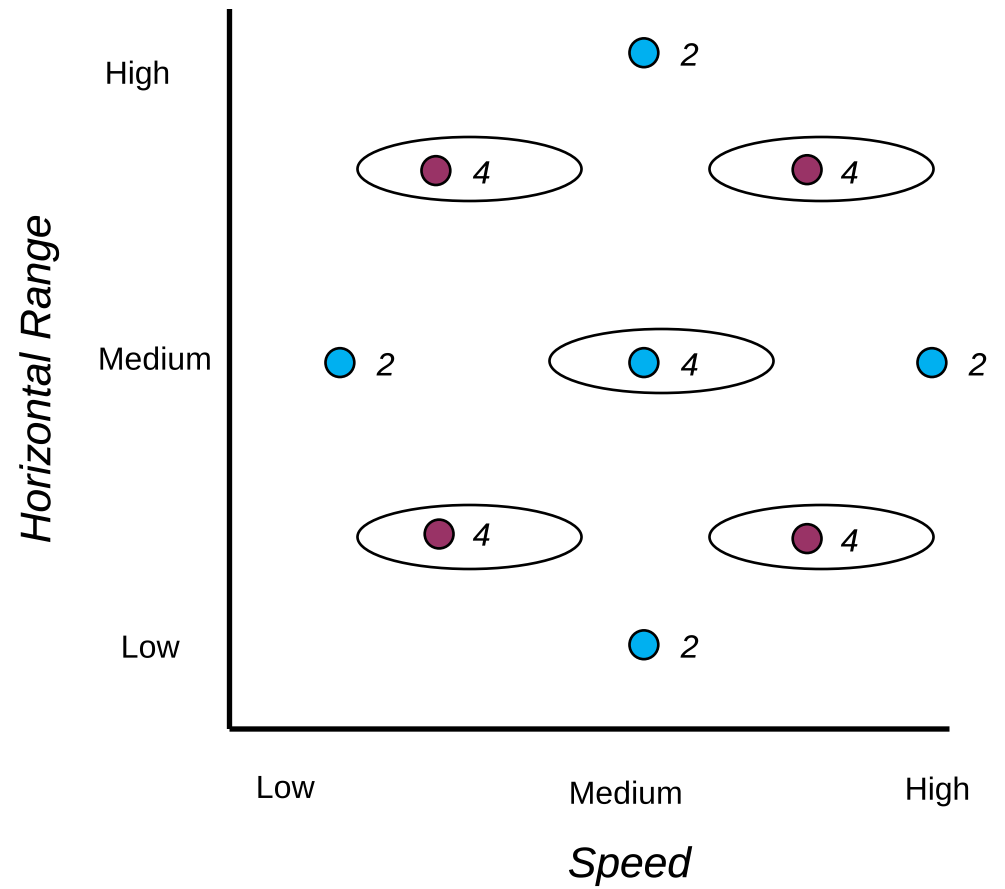
	Design Type	**Number of Runs**
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



Test Designs

- **DOE provides a vast library of test design types**

	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36

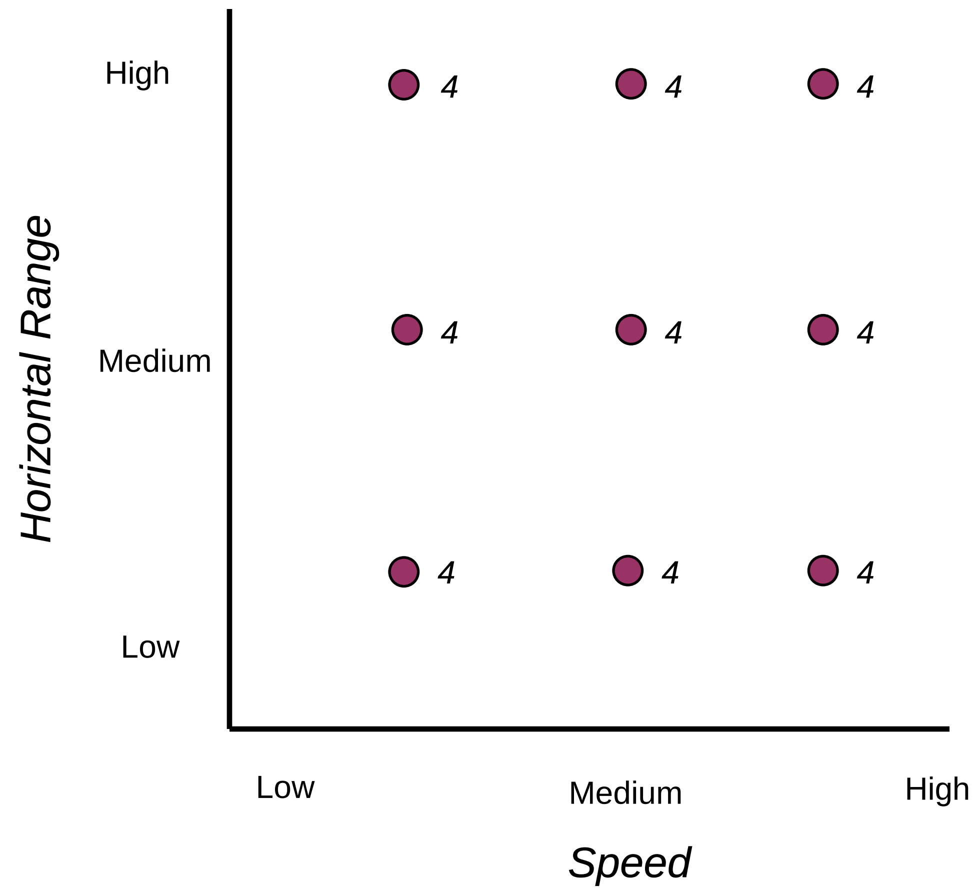


Cleared for open publication.

Test Designs

- **DOE provides a vast library of test design types**

	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



Cleared for open publication.

Power and Confidence

- **Power and confidence are only meaningful in the context of a hypothesis test!**

- **Statistical hypotheses:**

H_0: *Detonation slant range is the same with and without degaussing*

H_1: *Detonation slant range differs when degaussing is employed*

$$H_0: \mu_D = \mu_{ND}$$
$$H_1: \mu_D \neq \mu_{ND}$$

- **Power is the probability that we conclude that the degaussing system makes a difference when it truly does have an effect.**

- **Similarly, power can be calculated for any other factor or model term**

Test Decision	*Difference*	*No Difference*
Accept H_O	*False Negative (β Risk)*	*Confidence (1-α)*
Reject H_O	*Power (1-β)*	*False Positive (α Risk)*

Real World

We need to understand risk!

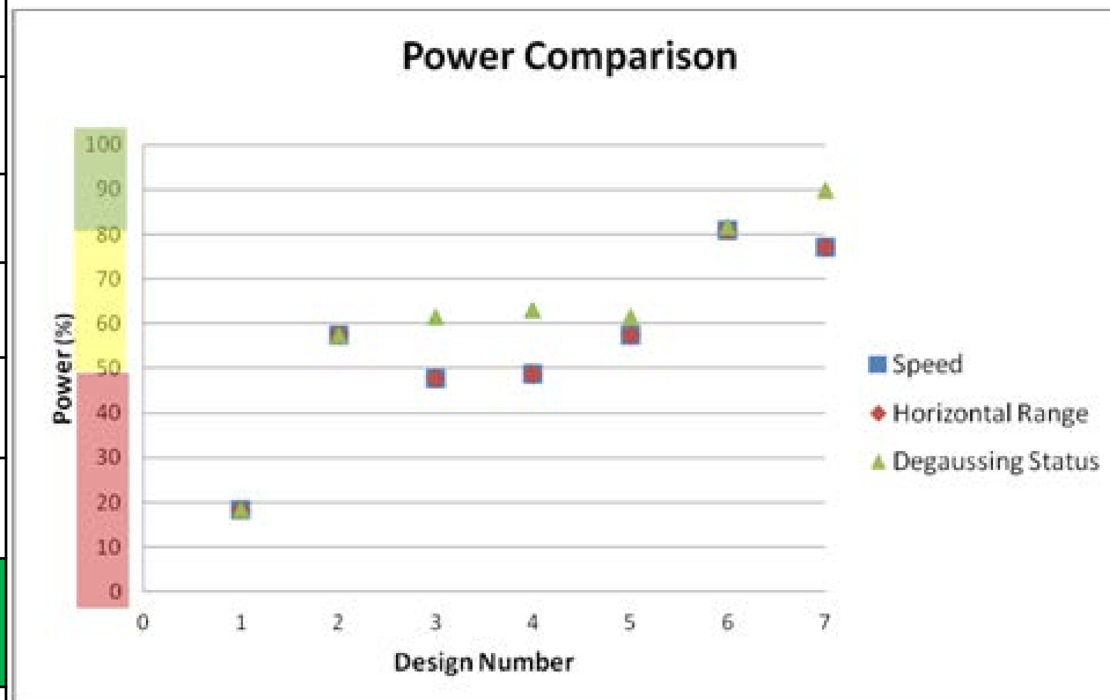
Cleared for open publication.

Cleared for open publication.

Test Design Comparison: Statistical Power

- **Compared several statistical designs**
  - Selected a replicated central composite design with 28 runs
  - Power calculations are for effects of one standard deviation at the 90% confidence level

	Design Type	**Number of Runs**
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



Cleared for open publication.

The wrong way to think about power

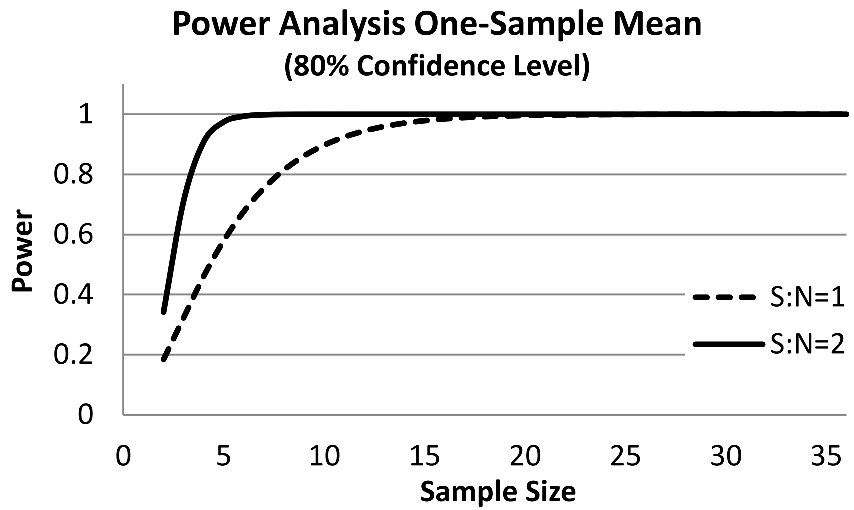
- **One sample hypotheses:**

*H_0: The system **doesn't** meet or exceed the threshold value*
H_1: The system exceeds the threshold requirement

Mathematically:

H_0: $\mu \leq 75$ *(notional requirement)*
H_1: $\mu > 75$

- **Power provides little insight to the adequacy of the test in this case**



Cleared for open publication.

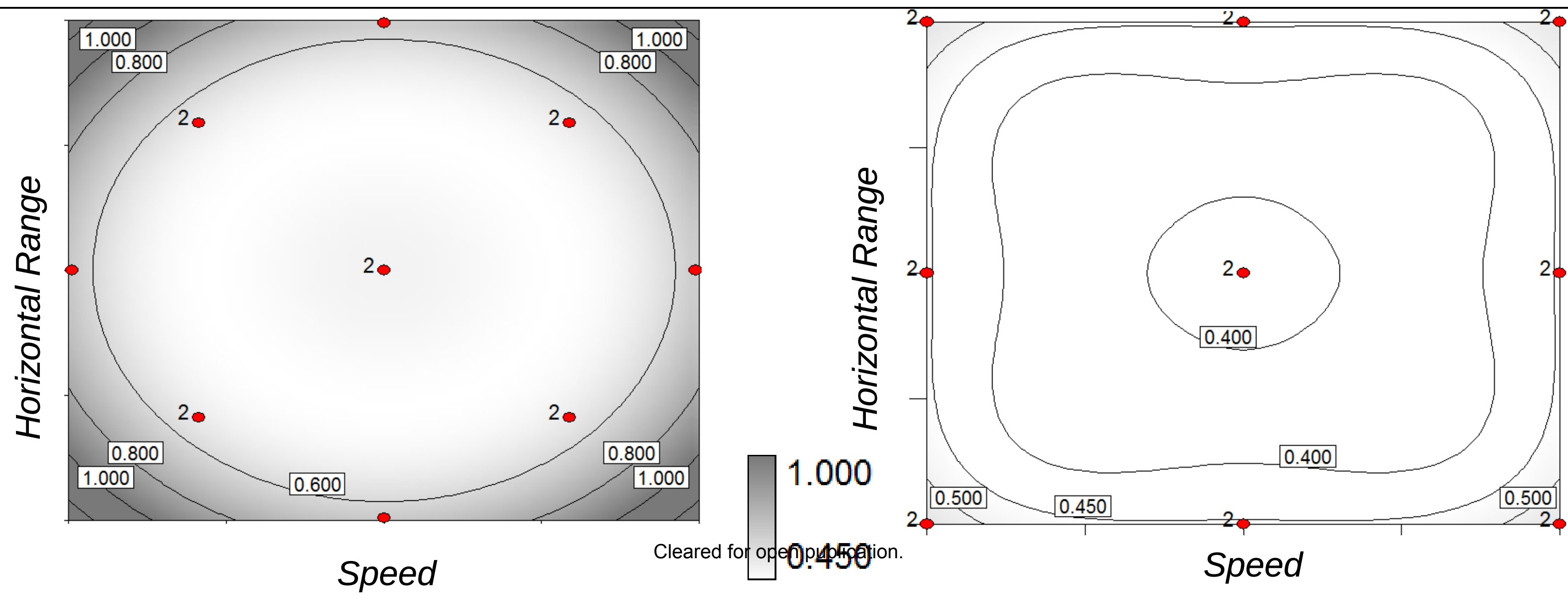
Factor Relationships, Prediction Capabilities

- **All designs considered were orthogonal for main effects and two-way interactions**
  - Small correlations for quadratic terms in Central Composite Design

- **Predictive capabilities are very different for the two primary designs considered**

Standard Error of the Design

Central Composite Design | General Factorial Design

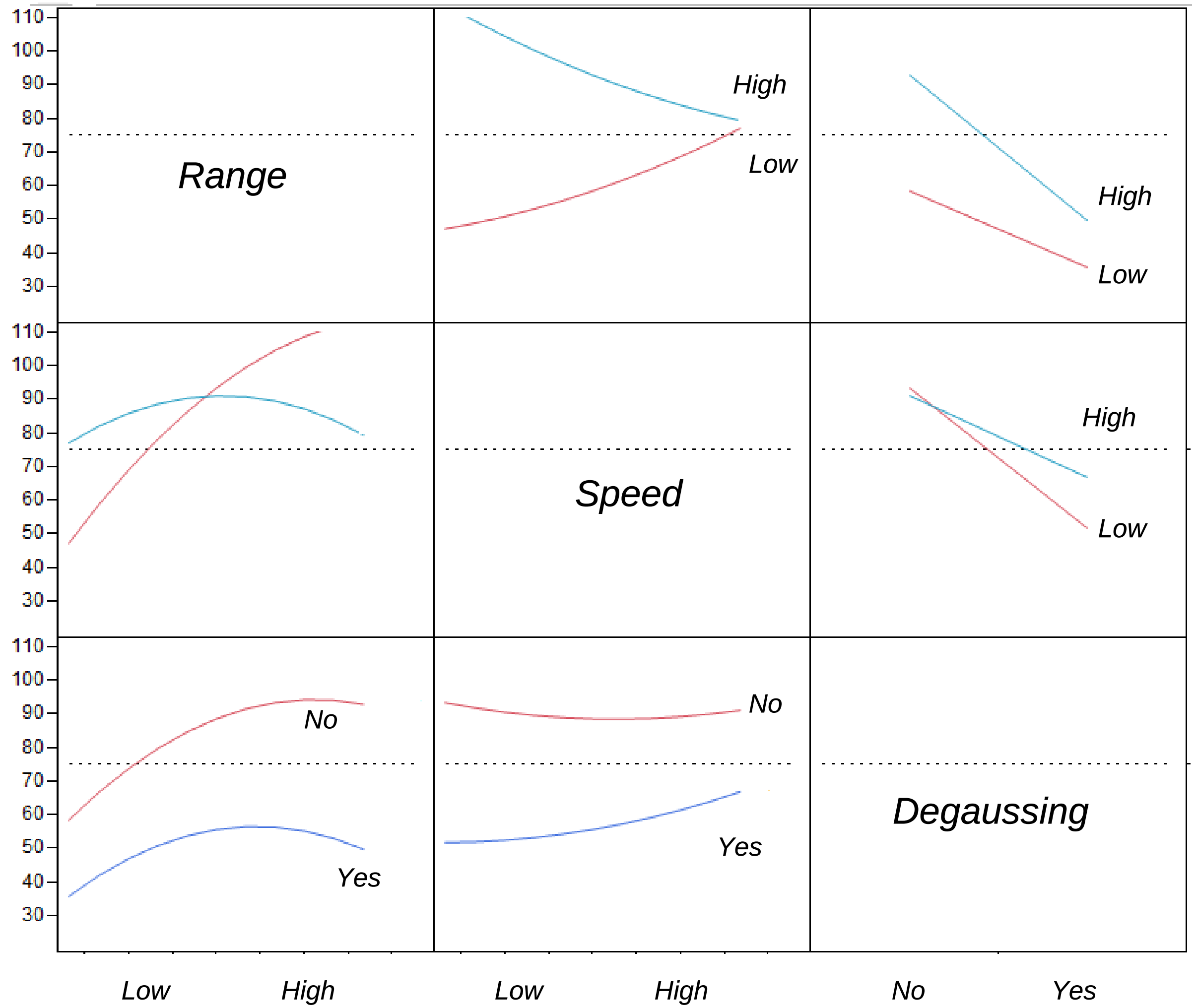


Cleared for open publication.

Notional Statistical Analysis

- **Statistical models support characterization of data across the operational envelope**
- **Power to detect factor effects also provides us with the ability to compare to the requirement across the operational envelope.**
  - Some regions are more powerful than others

Interaction Plots of Simulated Results



Conclusions

- **Design of Experiments (DOE) – a structured and purposeful approach to test planning**
  - Ensures adequate coverage of the operational envelope
  - Determines how much testing is enough
  - Quantifies test risks
  - Results:
    - » More information from constrained resources
    - » An analytical trade-space for test planning

- **Statistical measures of merit provide the tools needed to understand the quality of any test design to support statistical analysis**

- **Statistical analysis methods**
  - Do more with the data you have
  - Incorporate all relevant information in evaluations
    - » Supports integrated testing

Current Efforts to Institutionalize Statistical Rigor in T&E

- **DOT&E Test Science Roadmap – published June 2013**
- **DDT&E Scientific Test and Analysis Techniques (STAT) Implementation Plan**
- **Scientific Test and Analysis Techniques (STAT) Center of Excellence provides support to programs**
- **Research Consortium**
  - Navel Post Graduate School, Air Force Institute for Technology, Arizona State University, Virginia Tech
  - Research areas:
    - » Case studies applying experimental design in T&E.
    - » Experimental Design methods that account for T&E challenges.
    - » Improved reliability analysis.
- **Current Training and Education Opportunities**
  - Air Force sponsored short courses on DOE
  - Army sponsored short courses on reliability
  - AFIT T&E Certificate Program
- **Review of current policy & guidance**
  - DOD 5000
  - Defense Acquisition Guidebook

Cleared for open publication.