INSTITUTE FOR DEFENSE ANALYSES

Test Science Website Videos - Test Planning and Design

Rebecca Medlin, Project Leader

Kelly Avery
Caitlan Fealing
Rebecca Medlin
Rachel Haga
Matthew Avery

April 2021

Approved for Public Release.

Distribution Unlimited.

IDA Document NS D-16393

Log: 2020-000425

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task C9082 "Cross-Divisional Statistics and Data Science Working Group" and Task 229990 "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Stephen DeVito, Daniel Hellman, Rachel Haga, Rebecca Medlin, Jason Sheldon, Keyla Pagan-Rivera, and Courtney Au-Yeung from the Operational Evaluation Division.

For more information:
Rebecca Medlin, Project Leader
rmedlin@ida.org •C9082

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-16393

Test Science Website Videos - Test Planning and Design

Rebecca Medlin, Project Leader

Kelly Avery
Caitlan Fealing
Rebecca Medlin
Rachel Haga
Matthew Avery

Executive Summary

Understanding the test design process is fundamental to conducting efficient and effective tests of systems. Planning is a critical element of this process. Test planning involves identifying the test objectives, determining the appropriate outcome measure(s), and identifying what factors may influence those outcomes. The first set of website videos introduces the overarching test design process and provides best practices and examples for each of the planning steps.

Test design is the process of actually creating and evaluating a set of experimental test runs. It entails (1) carefully considering the test goals to ensure objectives can be met by the data collected, (2) building the corresponding type of design, and (3) evaluating the design across multiple criteria. The test design should (1) have a large enough sample size to be useful, (2) be executable in the real world, and (3) support the desired analysis. The second set of videos describes and provides recommendations for each element of the test design process, and also introduces considerations for how to manage and store data collected from tests.

A. Test Planning

The test design process, which is essentially an example of the scientific method, represents the gold standard for experimentation. It comprises seven steps:

1. Identify the test objectives
2. Identify the mission-oriented outcome metrics, or response variables
3. Identify the factors or variables that may affect those outcomes
4. Develop the test design
5. Conduct the test and collect the data
6. Analyze that data using statistical methods
7. Draw conclusions and inform your evaluation

This analytical framework for collecting and analyzing data is applicable to nearly every domain and industry. Successfully performing this process requires iteration and collaboration among subject matter experts.

Clearly defined objectives are essential when designing any operational or live fire test, because they will drive the associated test design and analysis. The two most common test objectives used in test and evaluation are (1) screening for influential factors and (2) characterizing a system's performance across a set of conditions. These objectives are best applied sequentially: first a screening experiment narrows down the set of factors that are likely to matter, and then a characterization design details the performance of the system under each of the relevant conditions.

Outcome metrics, also frequently called response variables, are what testers use to assess their test objectives. A good response variable should (1) provide a meaningful measure of performance, (2) support evaluation of requirements, and (3) lend itself well to experimental design, meaning it is measurable, valid, and informative. Oftentimes more than one response variable is necessary to capture the full picture of system performance.

Because defense systems rarely have uniform performance in every situation, it is important for testers to identify factors that may drive changes in performance. These typically relate to the environmental conditions, the threat level, the mission, the user type, or the system configuration. Good factors are significant, controllable, and informative. After brainstorming all potential factors, testers can use a factor management scheme to prioritize which factors should be controlled and systematically varied in a test design, and which will simply be logged, or fixed at a specific level. Modern Design of Experiments (DOE) techniques can investigate a large number of factors very efficiently.

B. Test Design

The best approach for collecting data from several factors is to conduct a statistically designed experiment. DOE is an experimentation strategy in which testers vary factors in a coordinated, strategic way, instead varying them one at a time, or arbitrarily choosing test cases. A test strategy that employs DOE will always provide the most powerful allocation of test resources for a given number of tests.

The specific choice of design should be tied to the test objectives and the anticipated analysis. For example, designs meant for factor screening have fundamentally different properties than designs meant to characterize or optimize. Sequential test design strategies, where the results of early tests inform the construction of future tests, are highly recommended and can provide additional resource savings.

Common test designs include factorial designs, response surface methodology, and optimal designs. The decision on which test design method is most appropriate depends on the question, metrics, types of factors (numeric or categorical), and available test resources. Many software packages are available to build and evaluate each of these designs.

The question of "How much testing is enough?" is persistent across Department of Defense test and evaluation endeavors. DOE can help decision makers "right-size" a test, providing a framework for building and evaluating test plans and assessing risk. Leveraging statistical techniques in both test design and test analysis helps decision makers reach correct conclusions about system performance.

Test designs should be realistic and executable. This might mean that factors cannot be perfectly randomized or that certain factor combinations will be disallowed. To avoid reporting artificially inflated metrics, design evaluations should always account for test structure and randomization.

Having a plan for collecting and managing test data is critical to a successful test program. A successful plan will include guidance on (1) how to collect data in the appropriate format, (2) where and how to store acquired data, (3) how to document, organize, and potentially automate data cleaning, reduction, and analysis, and (4) how to handle data access and distribution.

TestScience
Data . Driven . Defense

IDA

Test Planning

Lesson 1: Introduction to Test Planning

Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

What is the Purpose of Test and Evaluation?

To provide **warfighters** and **decision makers** with information about military systems that is

RELEVANT

CREDIBLE

OBJECTIVE

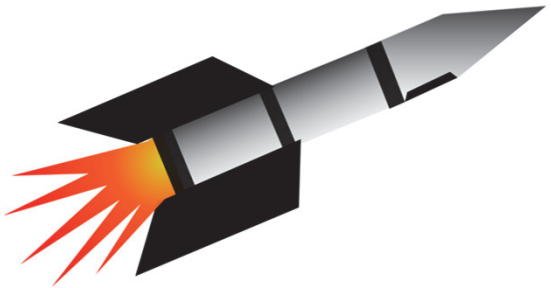
What is the Purpose of Test and Evaluation?

To provide **warfighters** and **decision makers** with information about military systems that is

RELEVANT

CREDIBLE

OBJECTIVE



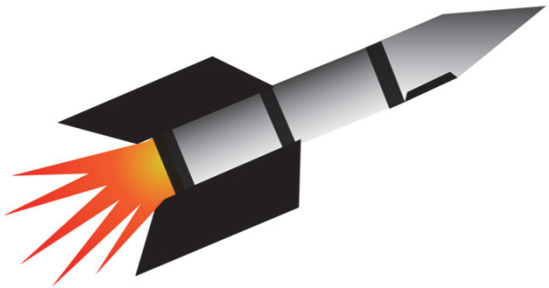
What is the Purpose of Test and Evaluation?

To provide **warfighters** and **decision makers** with information about military systems that is

RELEVANT

CREDIBLE

OBJECTIVE



TestScience
Data . Driven . Defense

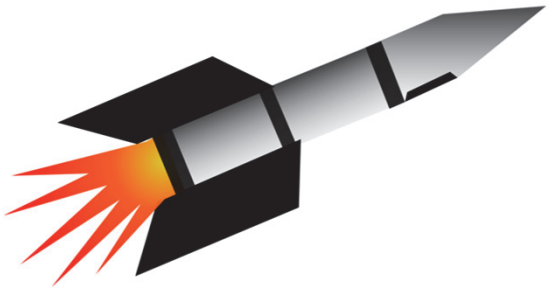
What is the Purpose of Test and Evaluation?

To provide **warfighters** and **decision makers** with information about military systems that is

RELEVANT

CREDIBLE

OBJECTIVE



TestScience
Data . Driven . Defense

What is the Purpose of Test and Evaluation?

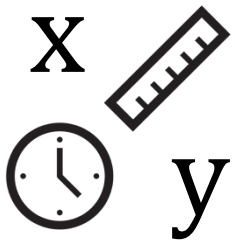
The **Design of Experiments** process
provides a framework
and an analytical basis
to help us do T&E
systematically and **efficiently**.

Testing Framework:
The Gold Standard of Experimentation

- The DOE process, at its core, is the scientific method.

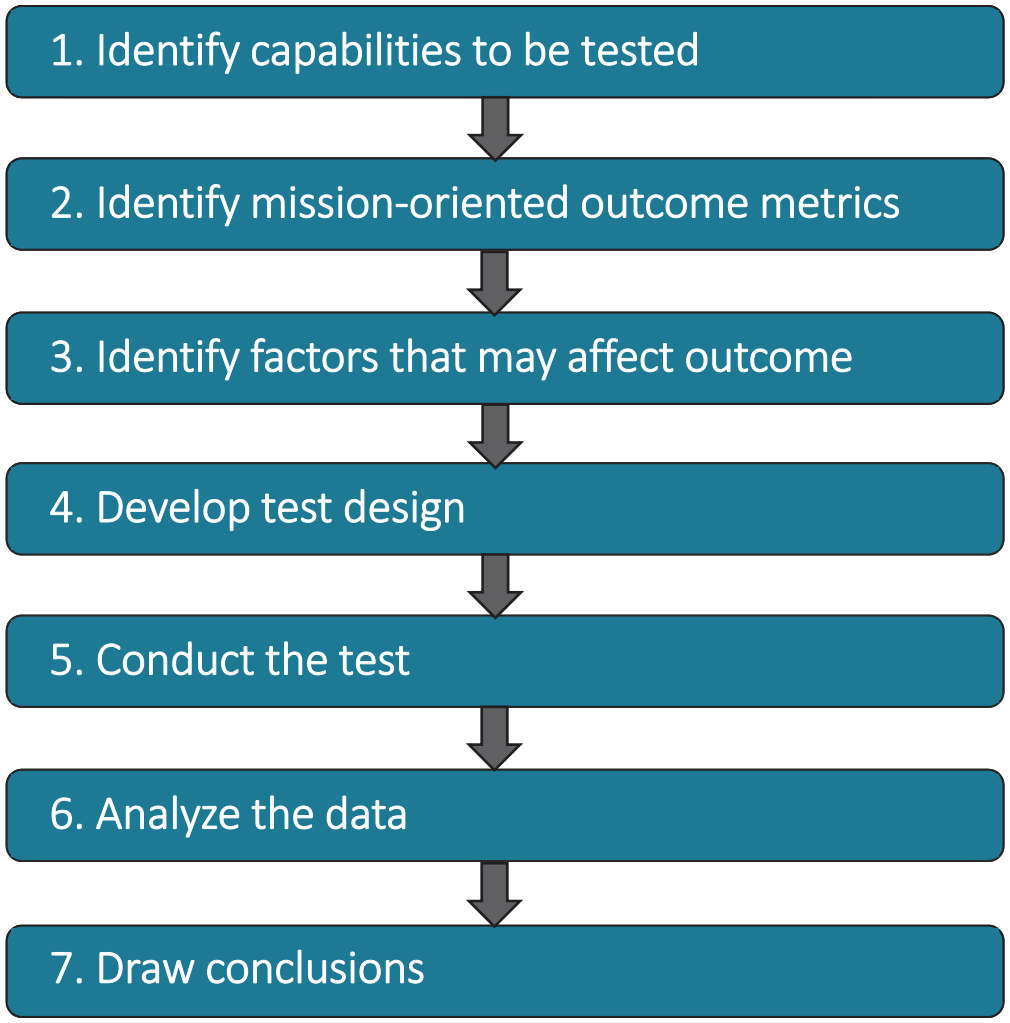
- It is the gold standard for performing experiments and determining cause and effect.

- What we mean by “experiment” goes well beyond the ‘chemistry lab’ scenario it often evokes.



- An experiment includes any test where variables are purposefully manipulated or inputs are purposefully set.
- So, the types of operational and live fire tests we conduct would certainly count as experiments.

Testing Framework:
The Gold Standard of Experimentation



TestScience
Data . Driven . Defense

Testing Framework:
The Gold Standard of Experimentation

1. Identify capabilities to be tested

Testing Framework:
The Gold Standard of Experimentation

2. Identify the mission-oriented outcome metrics

Testing Framework:
The Gold Standard of Experimentation

3. Identify factors that may affect outcome

Testing Framework:
The Gold Standard of Experimentation

4. Develop the test design

Testing Framework:
The Gold Standard of Experimentation

5. Conduct the test

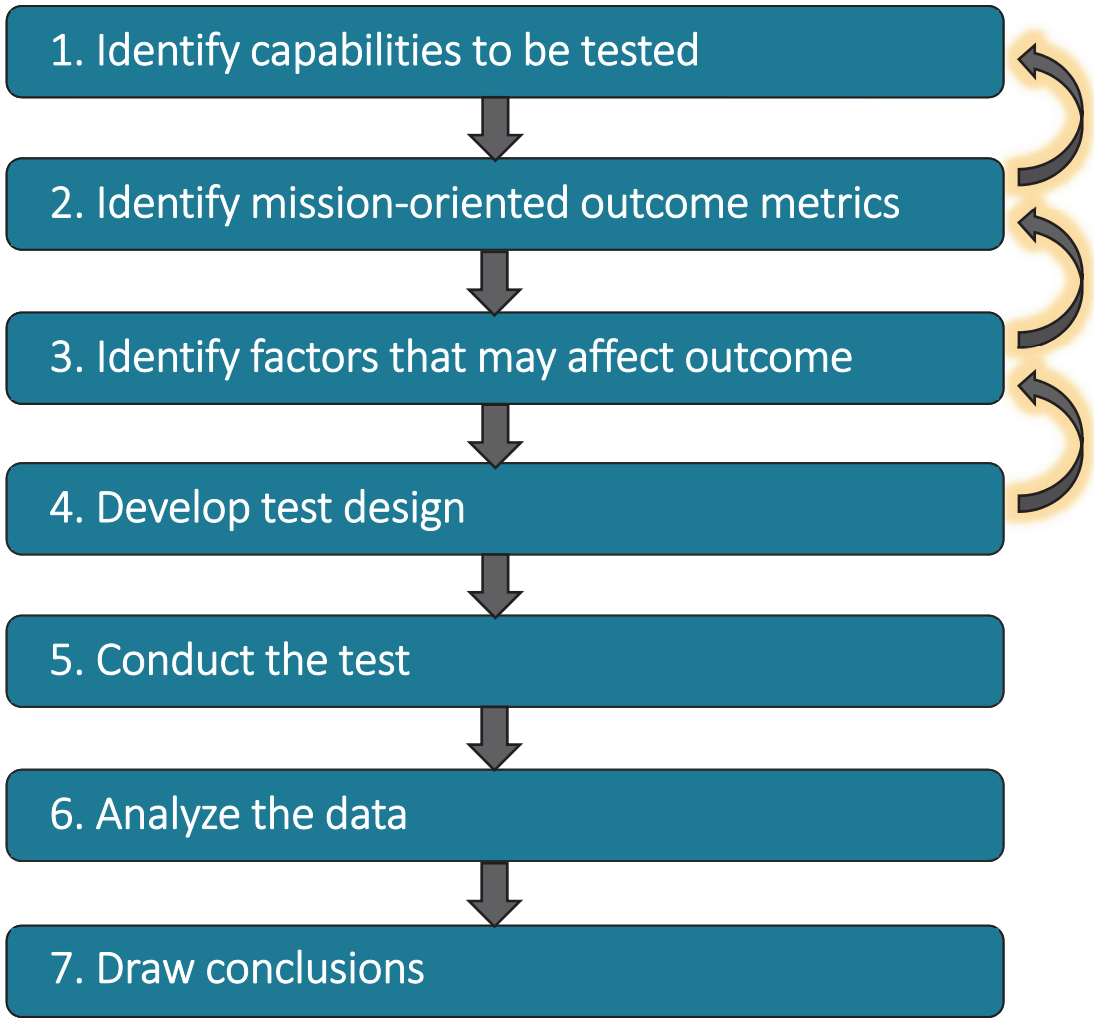
Testing Framework:
The Gold Standard of Experimentation

6. Analyze the data

Testing Framework:
The Gold Standard of Experimentation

7. Draw conclusions

Testing Framework:
The Gold Standard of Experimentation



TestScience
Data . Driven . Defense

Applying the Testing Framework

The test design process is applicable to any type of test:

- System Performance
- Human Factors
- Reliability
- Cybersecurity
- Live Fire
- Etc.

Applying the Testing Framework

Keys to success:

- Frequent **communication**
- **Collaboration** among technical, operational, and statistical experts
- Openness to **iteration**

The Three 'Test Planning' Steps are Crucial for Successful T&E

1. Identify capabilities to be tested

- Objectives of the test
- Questions you can ask about the system

2. Identify mission-oriented outcome metrics

- How you should measure system performance
- Response variables

3. Identify factors that may affect outcome

- Conditions or configurations under which the system may operate
- Operational envelope

The Three 'Test Planning' Steps are Crucial for Successful T&E

1. Identify capabilities to be tested

- Objectives of the test
- Questions you can ask about the system

2. Identify mission-oriented outcome metrics

- How you should measure system performance
- Response variables

3. Identify factors that may affect outcome

- Conditions or configurations under which the system may operate
- Operational envelope

The next three lessons cover each of these planning steps in more detail…

TestScience
Data . Driven . Defense

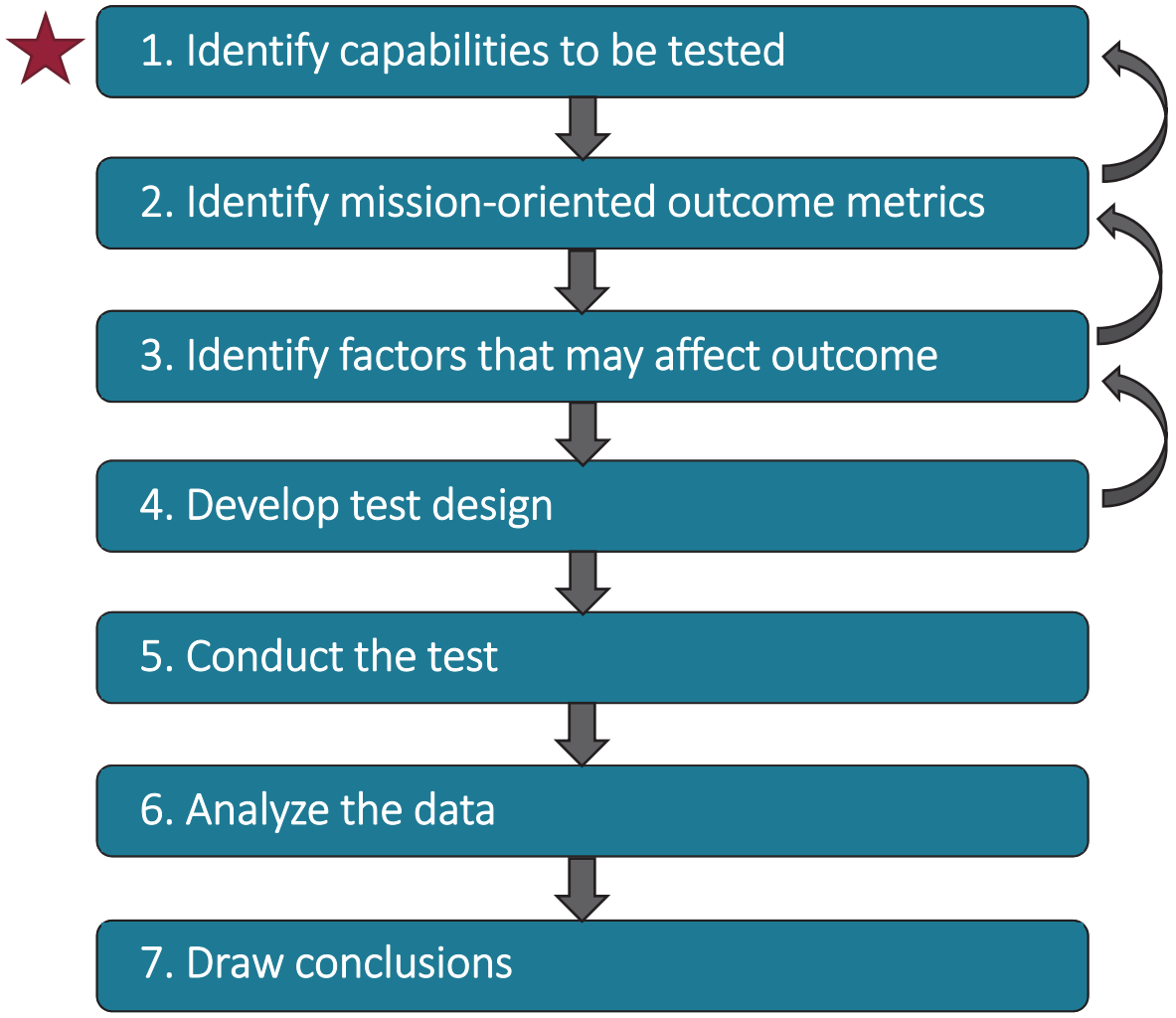
IDA

Test Planning

Lesson 2: Identify Test Objectives

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

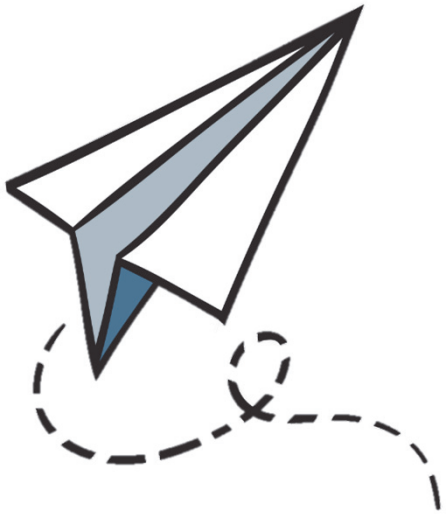
The Test Design Process



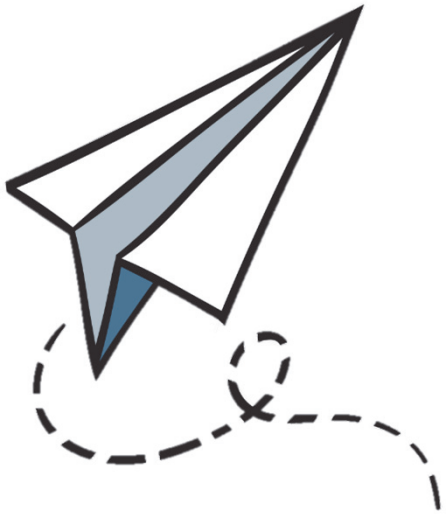
TestScience
Data . Driven . Defense

Suppose I ask you to make a paper aerial delivery vehicle

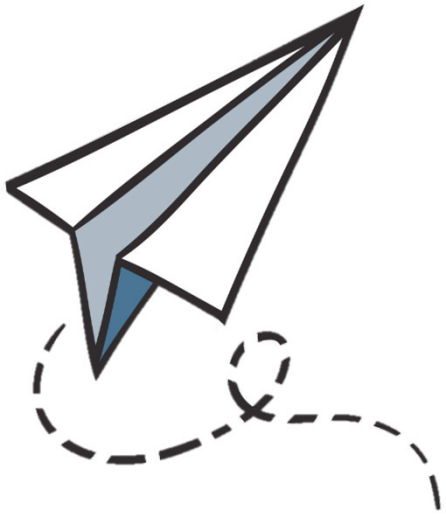
Suppose I ask you to make a paper aerial delivery vehicle



Suppose I ask you to make a paper aerial delivery vehicle



Suppose I ask you to make a paper aerial delivery vehicle



TestScience
Data . Driven . Defense

Suppose I ask you to make a paper aerial delivery vehicle

It is essential to have a clear goal when designing an experiment

1. Identify capabilities to be tested
   - Objectives of the test
   - Questions you can ask about the system

It is essential to have a clear goal when designing an experiment

1. Identify capabilities to be tested
   - Objectives of the test
   - Questions you can ask about the system

Test objectives will drive all future steps, including the test design and statistical analysis.

Common Test Objectives

- Screen for influential factors driving performance.
- Characterize performance across an operational envelope.
- Compare two systems (or more) across a variety of operating conditions.
- Identify problems that degrade system performance.
- Optimize system performance with respect to a set of conditions.

Common Test Objectives

- **Screen** for influential factors driving performance.
- **Characterize** performance across an operational envelope.
- **Compare** two systems (or more) across a variety of operating conditions.
- **Identify problems** that degrade system performance.
- **Optimize** system performance with respect to a set of conditions.

Common Test Objectives

- **Screen** for influential factors driving performance.
- **Characterize** performance across an operational envelope.
- **Compare** two systems (or more) across a variety of operating conditions.
- **Identify problems** that degrade system performance.
- **Optimize** system performance with respect to a set of conditions.

> DoDI 5000.02: The test program should produce "data to **characterize** combat mission capability across an appropriately selected set of factors and conditions."

Screen for Influential Factors

- Screening is useful early on in testing when we may know little about how the system responds to various factors.
- General approach:
  - Identify all potential factors that are thought to affect the response variable.
  - Choose an initial experimental design that uses minimal test resources.
  - Through analysis, identify the factors that have the largest impact on the response.
  - Update next test design to characterize the response (performance) as a function of only the important factors.

Screen for Influential Factors

- **Screening** is useful early on in testing when we may know little about how the system responds to various factors.
- General approach:
  - Identify all potential factors that are thought to affect the response variable.
  - Choose an initial experimental design that uses minimal test resources.
  - Through analysis, identify the factors that have the largest impact on the response.
  - Update next test design to characterize the response (performance) as a function of only the important factors.

(Screening is essential to integrated and sequential testing.)

Characterize performance across conditions

- Characterization tests describe performance at all relevant conditions and configurations in which the system may operate (i.e., across the operational envelope).
- They are also useful for determining if a system meets requirements across a variety of operational conditions.
- Characterization designs facilitate building precise mathematical models of performance based on test data.

Characterize performance across conditions

- Characterization tests describe performance at all relevant conditions and configurations in which the system may operate (i.e., across the operational envelope).
- They are also useful for determining if a system meets requirements across a variety of operational conditions.
- Characterization designs facilitate building precise mathematical models of performance based on test data.

> Characterization is the most important and common goal for operational testing.

TestScience
Data . Driven . Defense

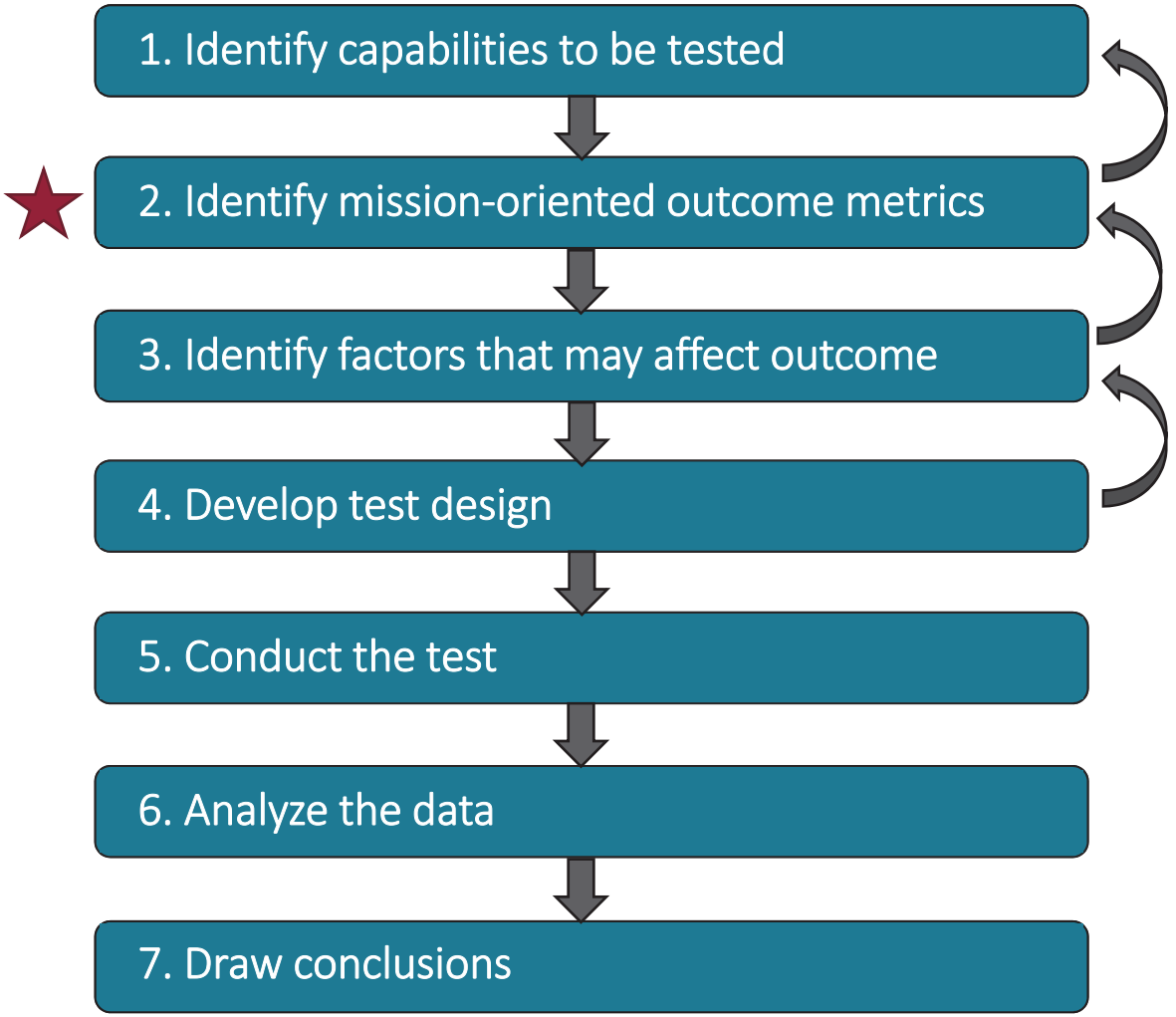
IDA

Test Planning

Lesson 3: Select Outcome Measures

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

The Test Design Process



TestScience
Data . Driven . Defense

Response variables measure the outcome of your test

2. Identify mission-oriented outcome metrics
   - How you should measure system performance
   - Response variables

Response variables measure the outcome of your test

2. Identify mission-oriented outcome metrics
   - How you should measure system performance
   - Response variables

TestScience
Data . Driven . Defense

Good response variables…

- Provide a meaningful measure of system performance or mission capability
  - These measures should encapsulate reasons for procuring the system

Good response variables…

- Provide a meaningful measure of system performance or mission capability
  - These measures should encapsulate reasons for procuring the system

Good response variables…

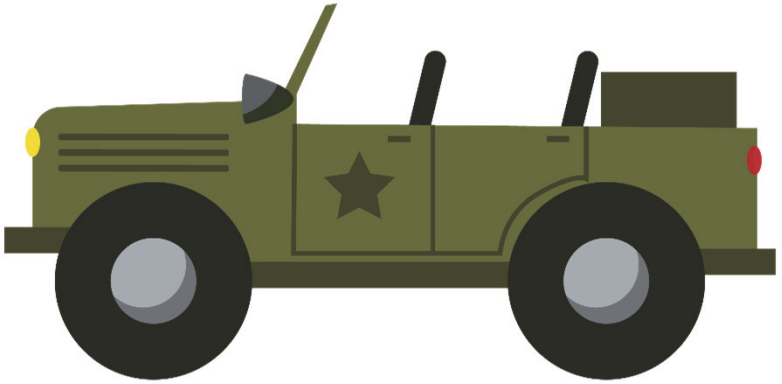
- Provide a meaningful measure of system performance or mission capability
- Provide adequate data to evaluate requirements
  - Selection of response variables is influenced by, but not limited to, the requirements.

Good response variables…

- Provide a meaningful measure of system performance or mission capability
- Provide adequate data to evaluate requirements
  - Selection of response variables is influenced by, but not limited to, the requirements.

Good response variables…

- Provide a meaningful measure of system performance or mission capability
- Provide adequate data to evaluate requirements
  - Selection of response variables is influenced by, but not limited to, the requirements.



Good response variables...

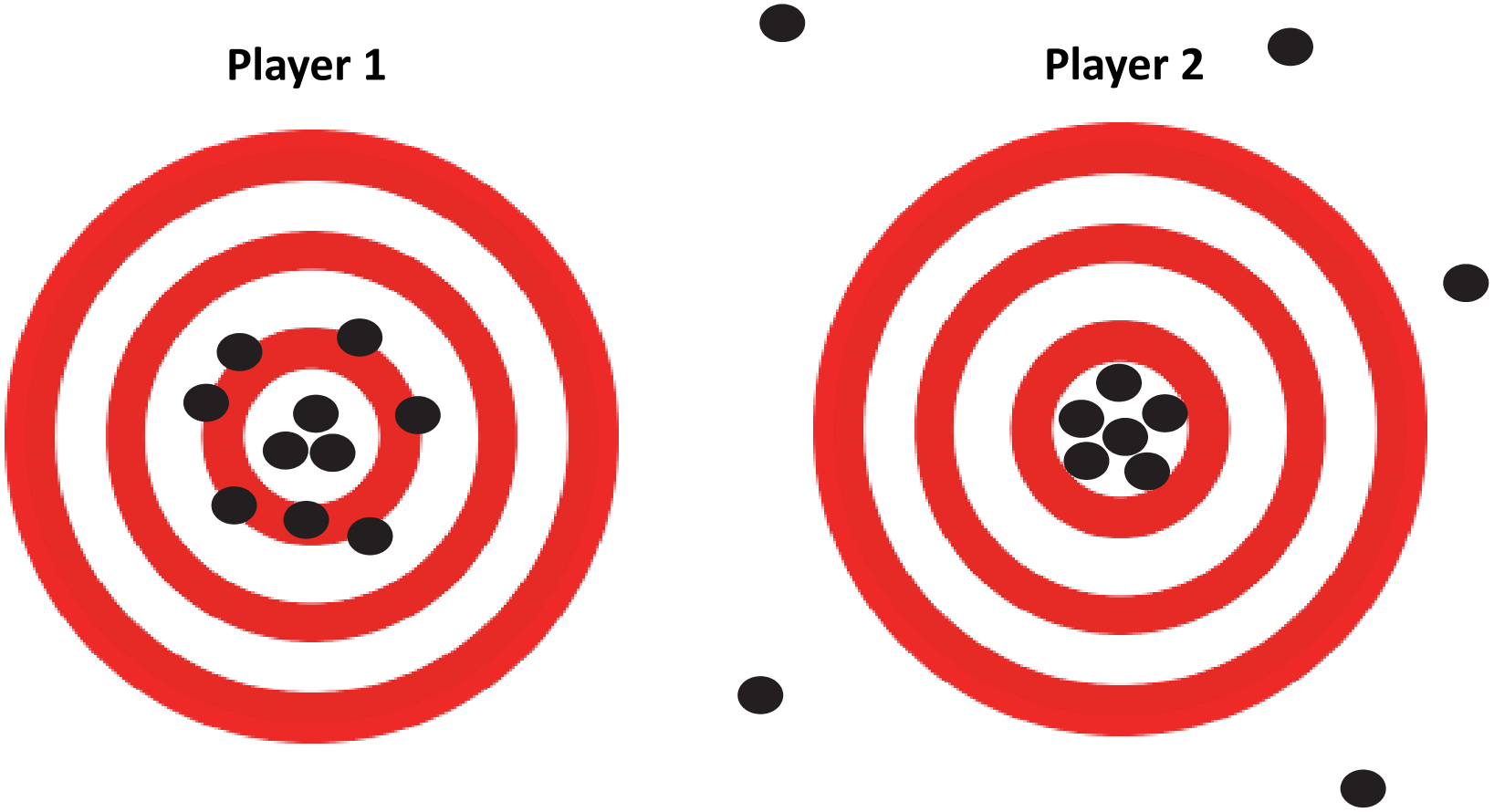
- Provide a meaningful measure of system performance or mission capability
- Provide adequate data to evaluate requirements
- Lend themselves well to defensible experimental design, which means they are:
  - Measurable. They can be measured at a reasonable cost and without affecting the test outcome.
  - Valid. They directly address the test objective.
  - Informative. Continuous responses always provide more information per test point than pass/fail metrics.

Good response variables…

- Provide a meaningful measure of system performance or mission capability
- Provide adequate data to evaluate requirements
- Lend themselves well to defensible experimental design, which means they are:
  - Measurable. They can be measured at a reasonable cost and without affecting the test outcome.
  - Valid. They directly address the test objective.
  - Informative. Continuous responses always provide more information per test point than pass/fail metrics.

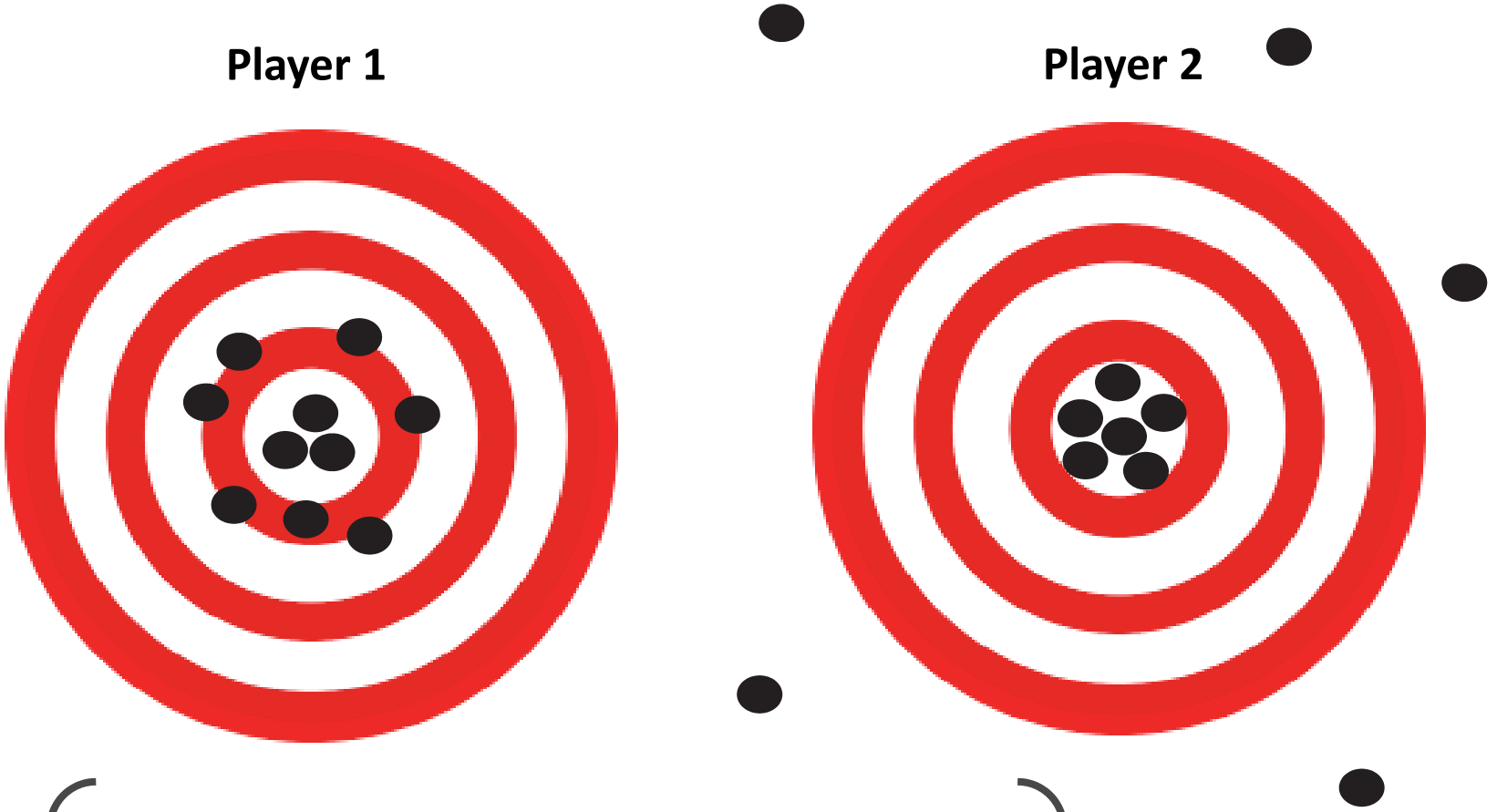
Multiple responses are common and almost always necessary!

Who is the better dart player?



TestScience
Data . Driven . Defense

Who is the better dart player?



We have a much better picture of the players' abilities by looking at the miss distances from bullseye compared to just a bullseye hit/miss!

TestScience
Data . Driven . Defense

We can often convert probability requirements into continuous measures

	Chemical Agent Detector	Submarine Mine Detection	Missile/Bomb	Radar Detection
Requirement	Probability of detection greater than 85% after one minute of exposure	Probability of detection greater than 80% outside 200 meters	Probability of hit at least 90%	Probability of detection greater than 90% at 300 kilometers.
Original Test Measurement	Detect prior to 1 minute? (Yes/No)	Detect/Non-Detect	Hit/Miss	Detect/Non-Detect
Modified Test Measurement	Detection Time	Detection Range	Miss Distance	Detection Range

TestScience
Data . Driven . Defense

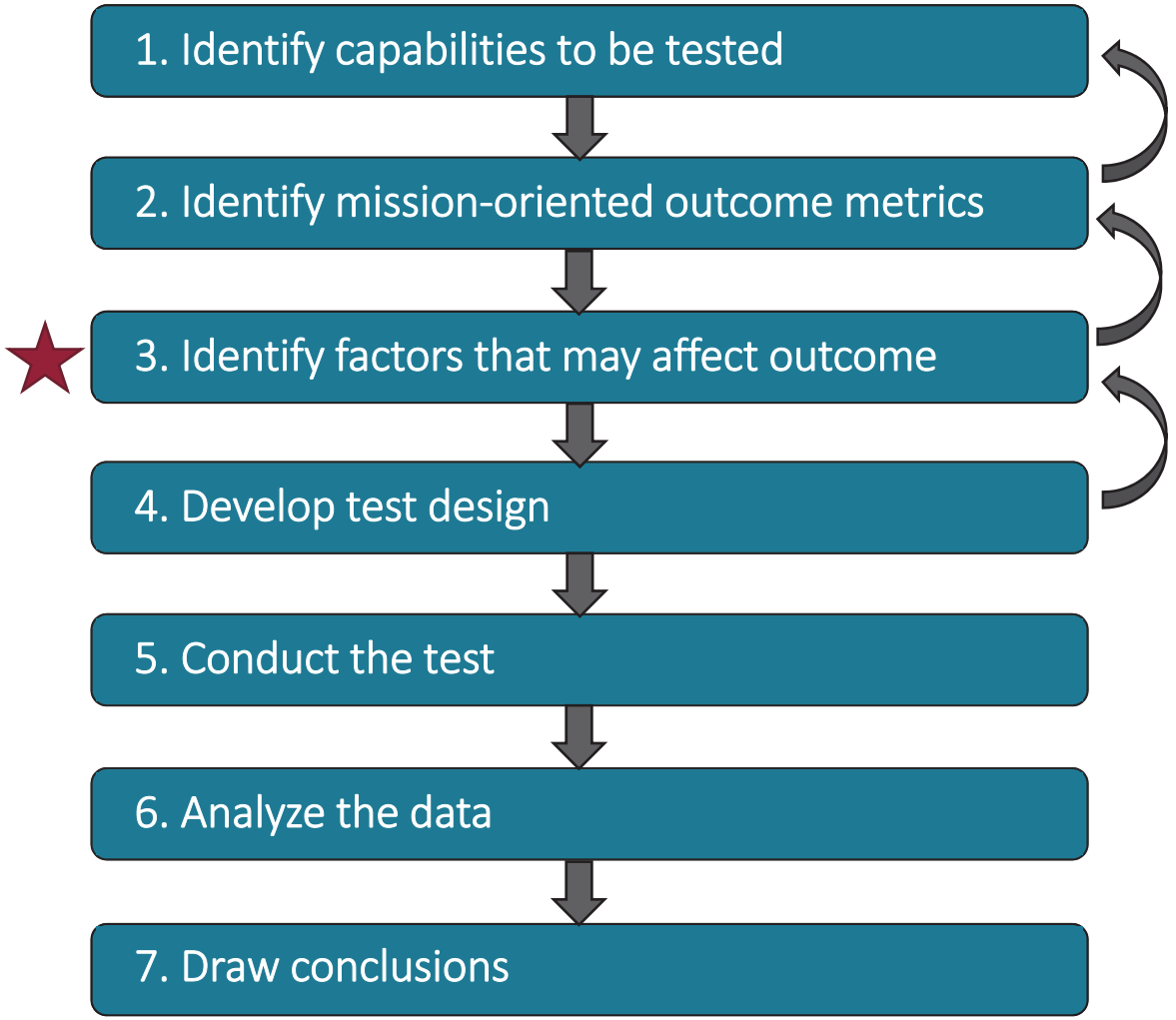
IDA

Test Planning

Lesson 4: Select Factors Affecting Outcomes

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

The Test Design Process



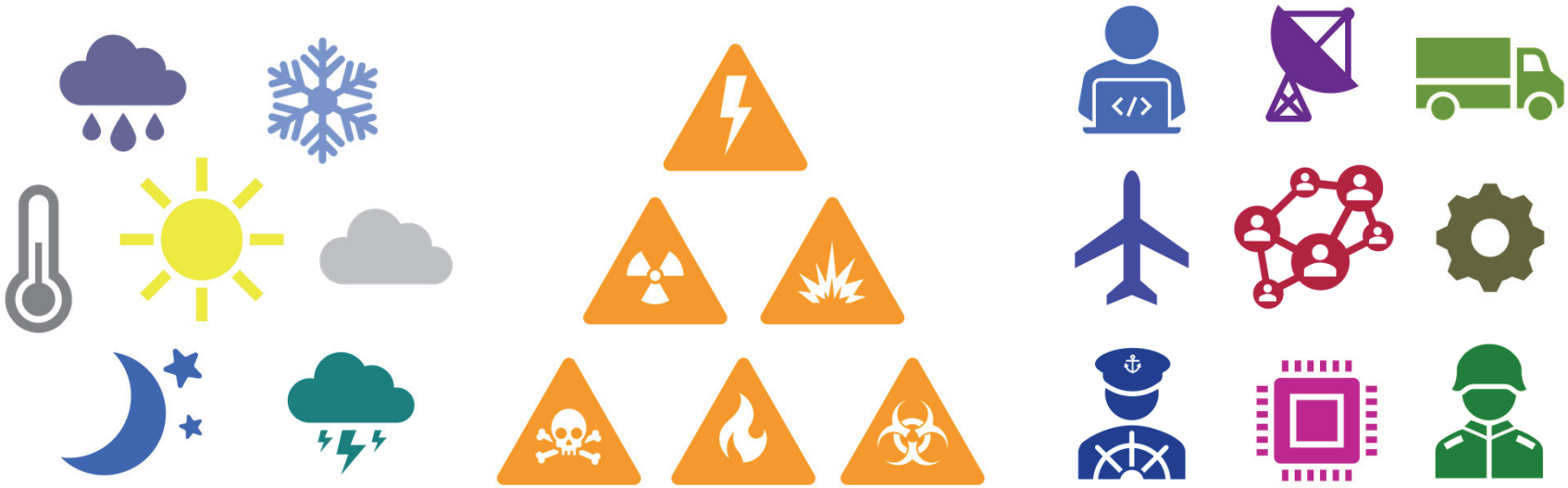
TestScience
Data . Driven . Defense

What might influence the performance of the process or system under test?

3. Identify factors that may affect outcome
    - Conditions or configurations under which the system may operate
    - Operational envelope

What might influence the performance of the process or system under test?

3. Identify factors that may affect outcome
   - Conditions or configurations under which the system may operate
   - Operational envelope

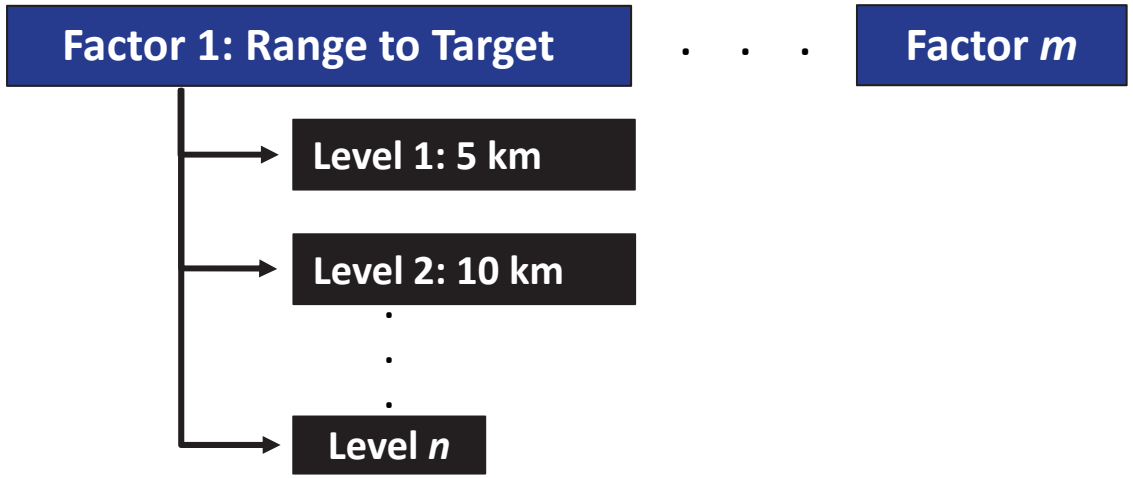


TestScience
Data . Driven . Defense

Terminology

Factors are independent variables that are expected to affect the outcome of a test.

Levels are the specific values that the factors assume. Factor levels are often referred to as the test conditions.



Good test design factors are…

Important: Factors are expected to have an effect on the test outcome.

Good test design factors are…

Important: Factors are expected to have an effect on the test outcome.

Controllable: Factors can be controlled (i.e., set to a specific level) at a reasonable cost.

Good test design factors are…

Important: Factors are expected to have an effect on the test outcome.

Controllable: Factors can be controlled (i.e., set to a specific level) at a reasonable cost.

Informative: Continuous factors are preferred to categorical factors.

Good test design factors are…

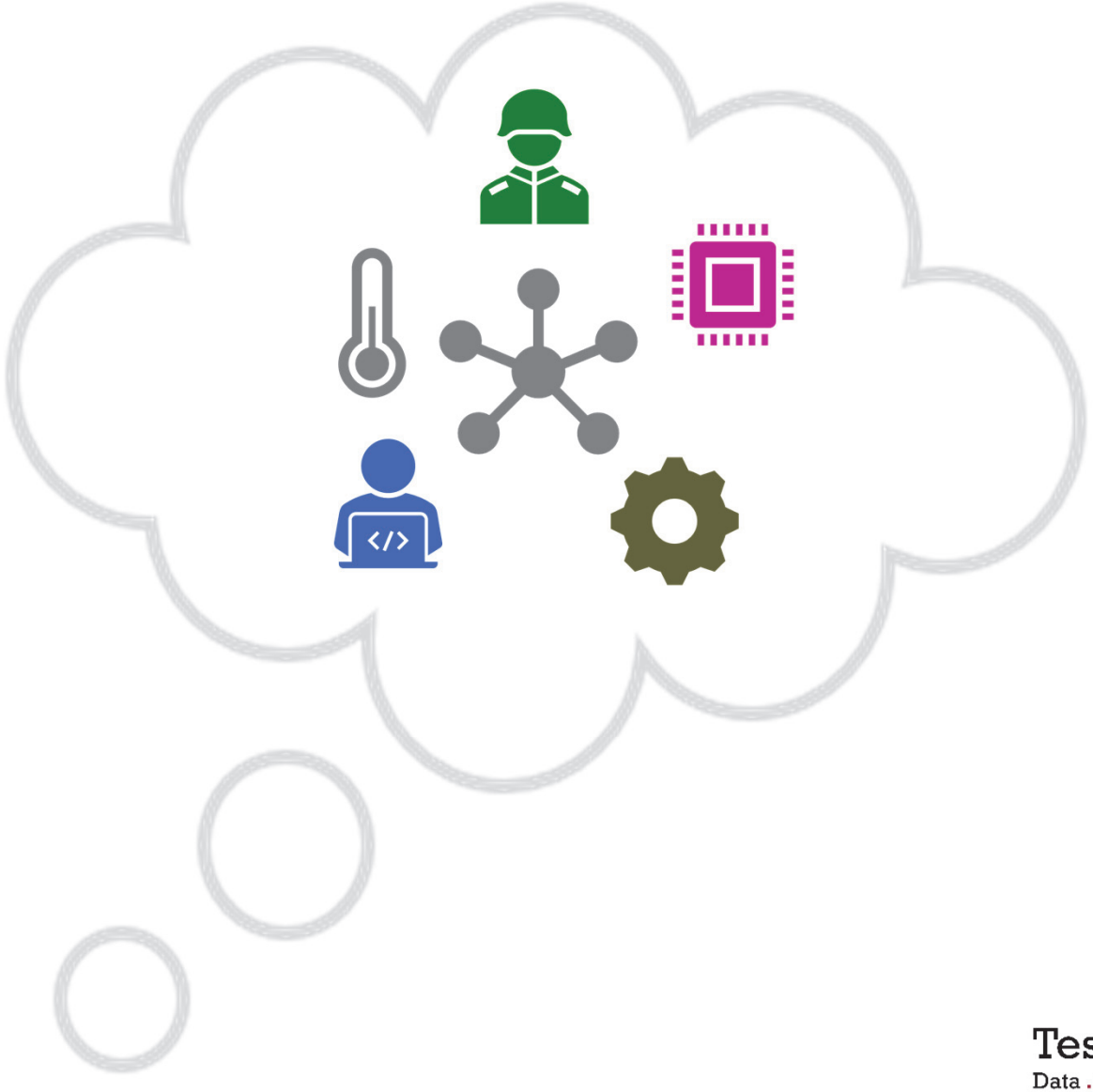
Important: Factors are expected to have an effect on the test outcome.

Controllable: Factors can be controlled (i.e., set to a specific level) at a reasonable cost.

Informative: Continuous factors are preferred to categorical factors.

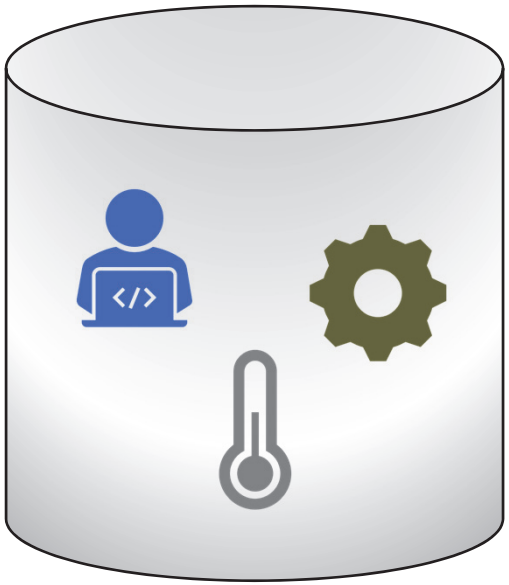
- Continuous factors allow for:
  - Interpolation
  - Higher power to detect significance of a factor
  - Strategic point placement

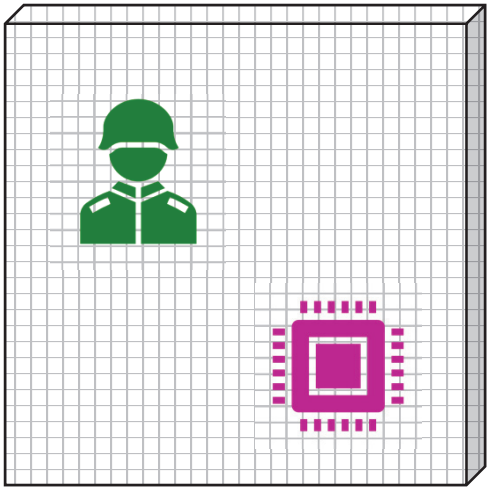
Testers should brainstorm ALL potential factors that could affect test outcomes



TestScience
Data . Driven . Defense

Testers should brainstorm ALL potential factors that could affect test outcomes



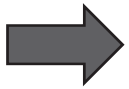


Factors can be prioritized and sorted using a Factor Management Scheme

TestScience
Data . Driven . Defense

Testers should brainstorm ALL potential factors that could affect test outcomes

- Factors of particular interest to the experimenter
- Factors that are measurable, controllable, and thought to be (very) influential



Strategically Vary

Factors can be prioritized and sorted using a Factor Management Scheme

TestScience
Data . Driven . Defense

Testers should brainstorm ALL potential factors that could affect test outcomes

- Factors of particular interest to the experimenter
- Factors that are measurable, controllable, and thought to be (very) influential

→ Strategically Vary

- Factors that are not of particular interest for the experiment
- Factors that seldom change levels
- Factors that are unlikely to have a large effect on the outcome
- Factors that cannot be controlled, but have too much of an influence on the response variable to simply ignore

→ Fix or Record

Factors can be prioritized and sorted using a Factor Management Scheme

TestScience
Data . Driven . Defense

Modern experimental designs can investigate a large number of factors efficiently

Testers should err on the side of strategically varying factors

Fixing factors limits conclusions about the system to the one condition tested

For recorded factors there is no guarantee that all levels of interest will be observed

- Potential tradeoff with operational realism

Adding factors does not cause the test size to grow exponentially!

TestScience
Data . Driven . Defense

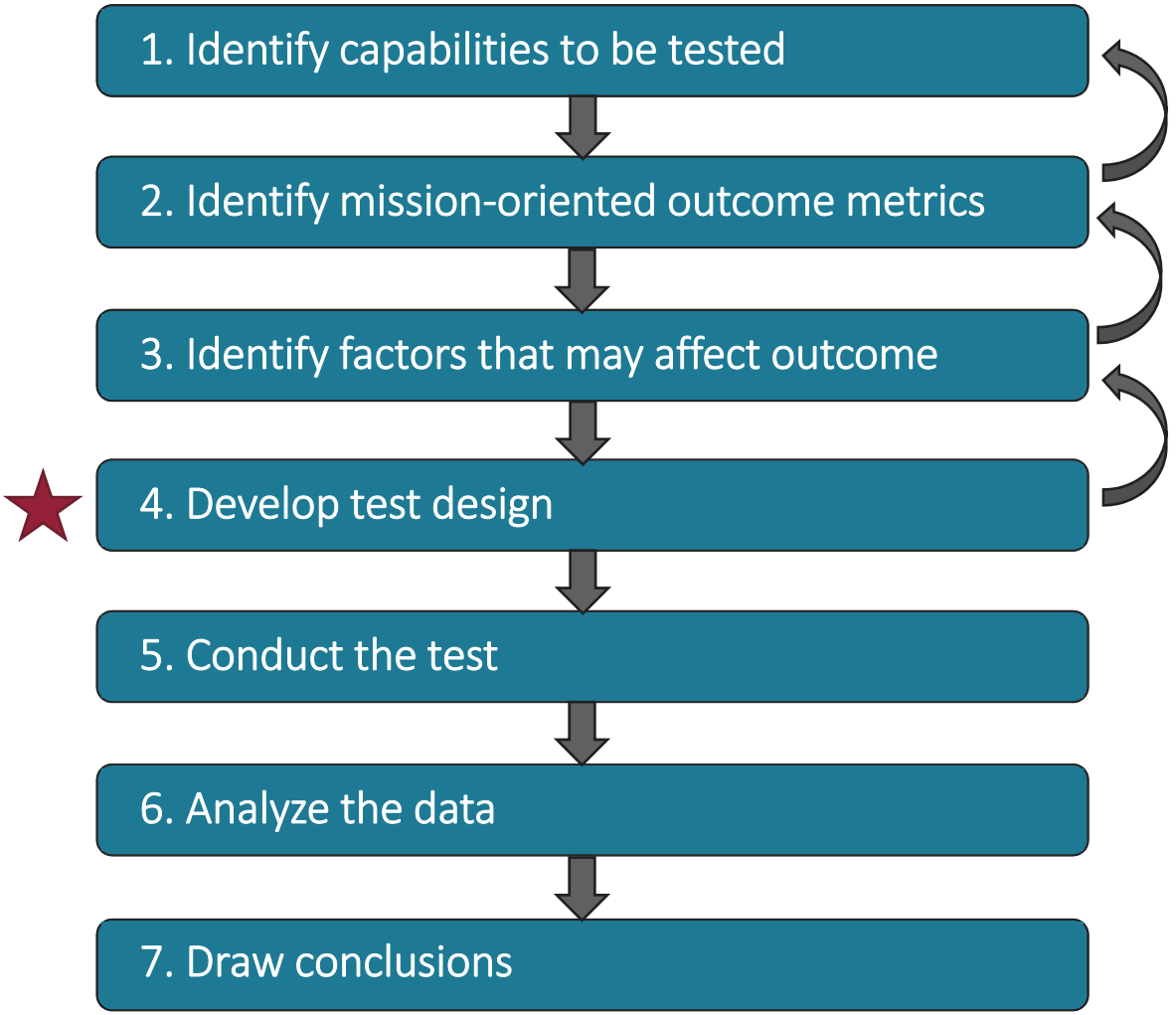
IDA

Test Design

Lesson 1: Introduction to Test Design

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

Testing Framework: The Gold Standard of Experimentation

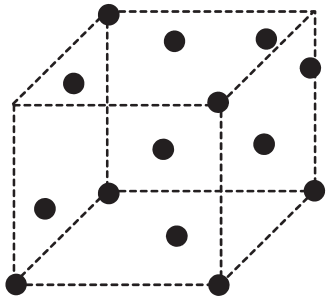


TestScience
Data . Driven . Defense

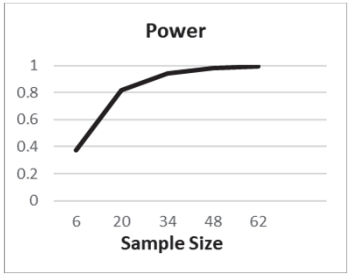
Common Challenges to Test

What Goal(s)?

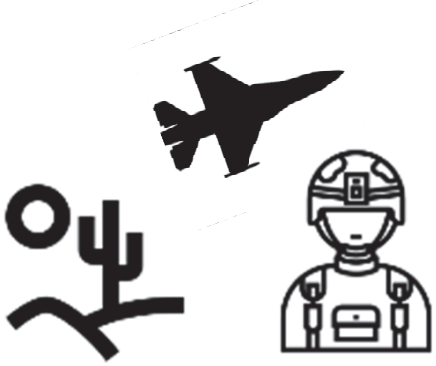
Which Points?



How Many Points?



How to Execute?



How to Capture?

What Conclusion(s)?

TestScience
Data . Driven . Defense

Test *Goals* Influence Test *Designs*

Screen

Characterize

Optimize

Compare

Identify Problems

Predict

Improve

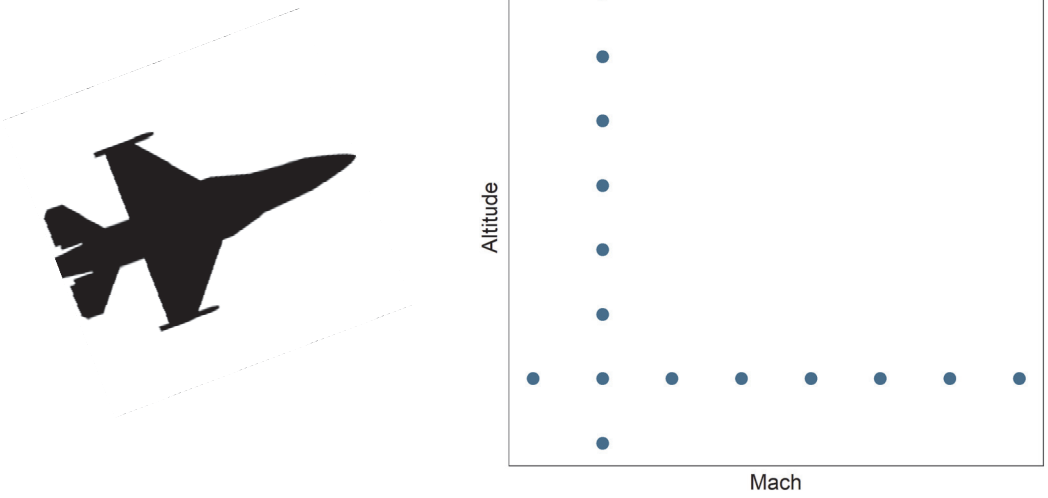
Demonstrate

All tests are designed, many poorly.

All tests are designed, many poorly.

All tests are designed, many poorly.

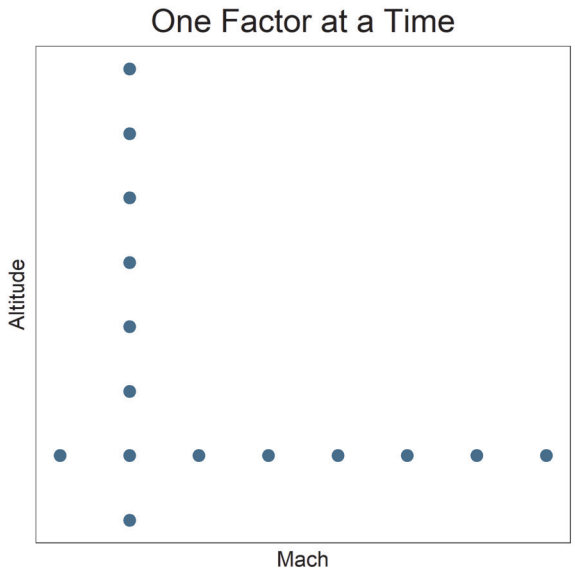
A one-factor-at-a-time test strategy is inefficient and fails to detect interactions

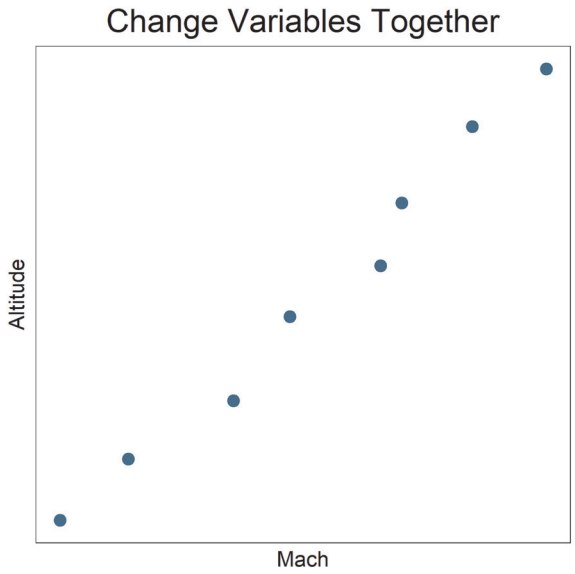


All tests are designed, many poorly.

A **one-factor-at-a-time** test strategy is inefficient and fails to detect interactions

Changing the variables together prevents the ability to detect cause end effect.

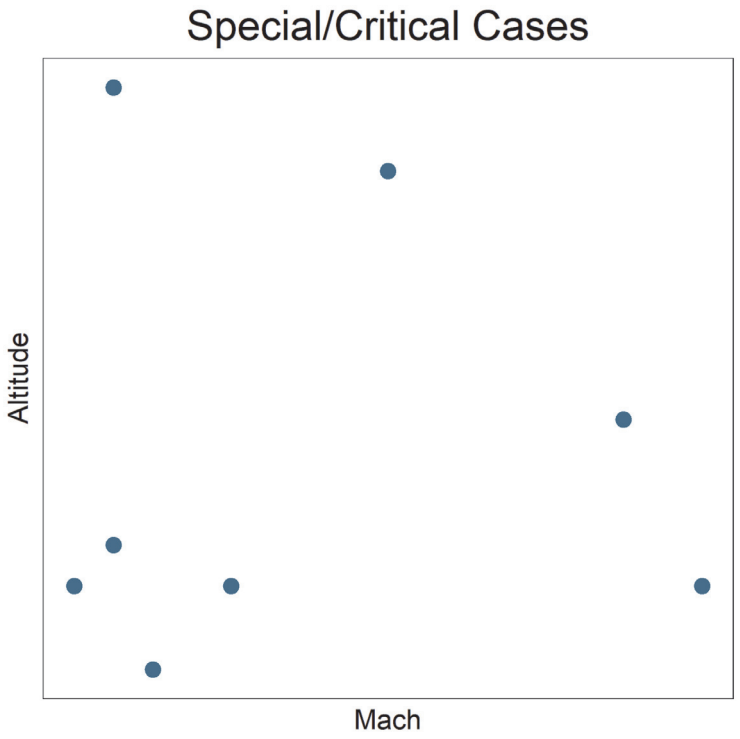




TestScience
Data . Driven . Defense

All tests are designed, many poorly.

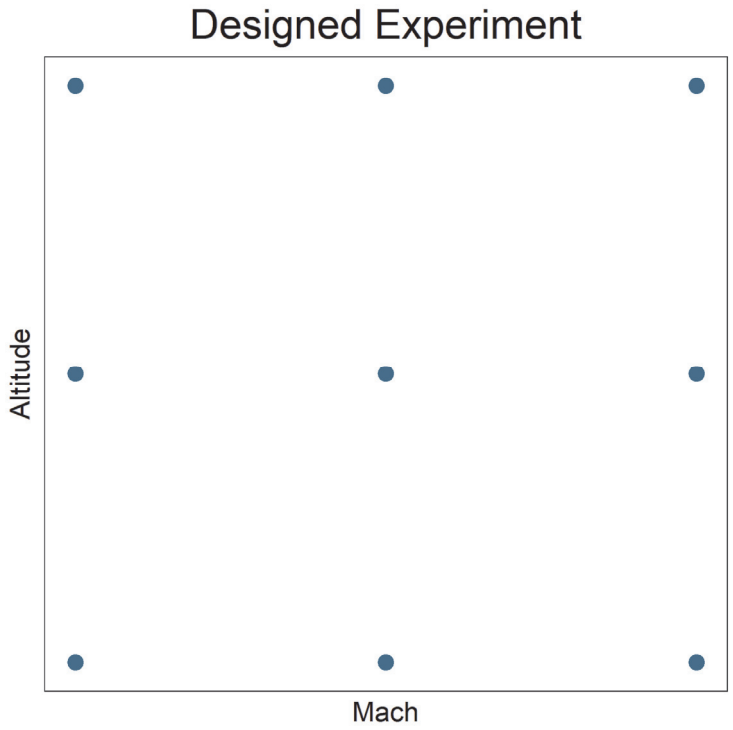
A test strategy of “We’ll just do what we did last time” or “We’ll test the special cases” might miss important performance shortfalls.



TestScience
Data . Driven . Defense

DOE changes “I think” to “I know”

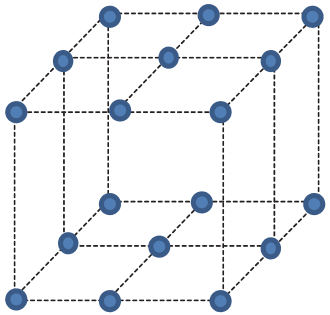
A test strategy that employs DOE will provide the most powerful allocation of test resources for a given number of tests.



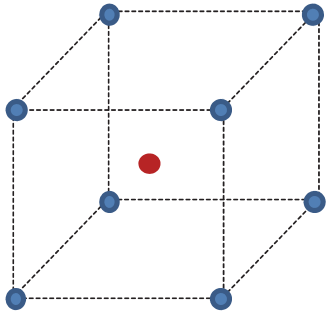
TestScience
Data . Driven . Defense

DOE Provides a Structured Approach to Picking Test Points

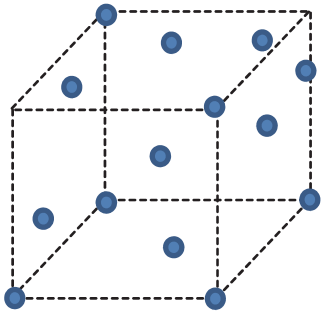
DOE Provides a Structured Approach to Picking Test Points



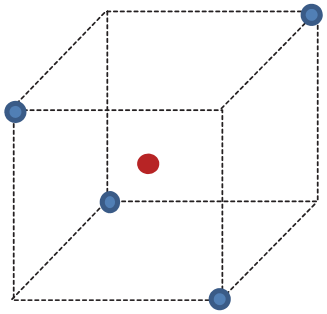
General Factorial
3x3x2 design



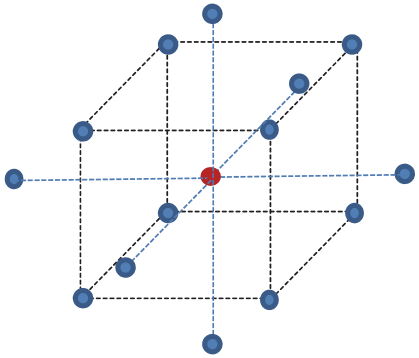
2-level Factorial
2^3 design



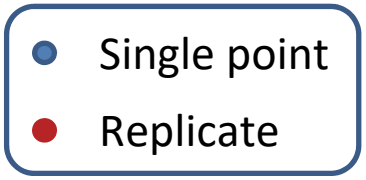
Optimal Design
D-optimal



Fractional Factorial
2^{3-1} design



Response Surface
Central Composite design



TestScience
Data . Driven . Defense

“How much testing is enough?”

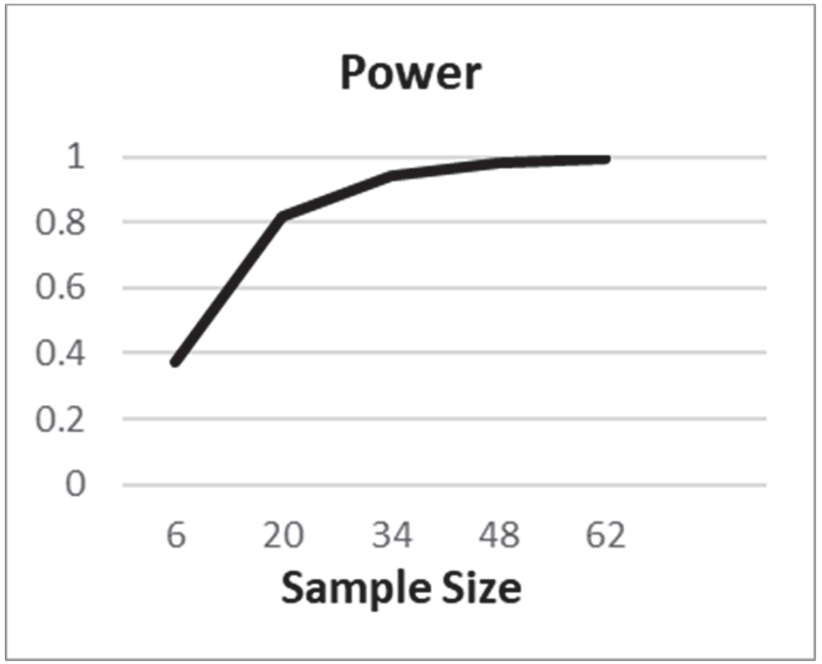
How many test points are enough to get it *right*?

3 ? That’s how much money/time we have.

8-10 ? That’s what we did last time.

30 ? From my high school stats class, something good happens at 30!

DOE methods provide tools to assess the risk of drawing incorrect conclusions



TestScience
Data . Driven . Defense

Design your test around how it will be executed

Executing an operationally realistic test means your factors might not be **randomized** or that certain factor combinations will be **disallowed**.

Design your test around how it will be executed

Executing an operationally realistic test means your factors might not be **randomized** or that certain factor combinations will be **disallowed**.

Response Variable

Miss Distance

Factor	Level
Illumination	Light, Dark
Range to Target	Short, Medium, Long
Fuze Type	A, B
Angle of Fire	Low, High

Design your test around how it will be executed

Executing an operationally realistic test means your factors might not be **randomized** or that certain factor combinations will be **disallowed**.

Response Variable

Miss Distance

CONOPS:

Fire → Move → Fire

Factor	Level
Illumination	Light, Dark
Range to Target	Short, Medium, Long
Fuze Type	A, B
Angle of Fire	Low, High

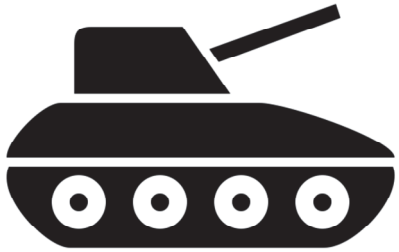
TestScience
Data . Driven . Defense

Design your test around how it will be executed

Executing an operationally realistic test means your factors might not be **randomized** or that certain factor combinations will be **disallowed**.

Response Variable

Miss Distance



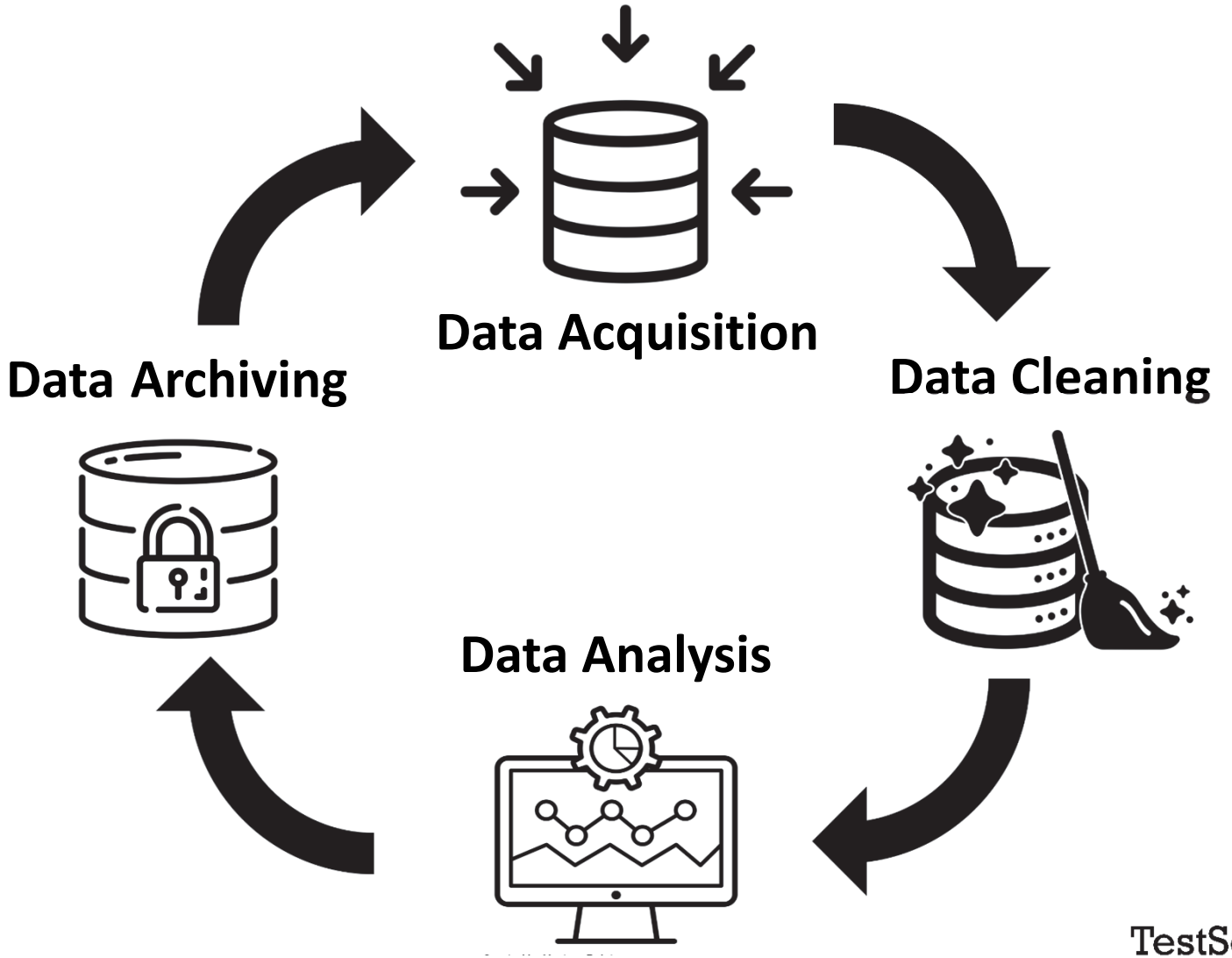
CONOPS:

Fire → Move → Fire

Factor	Level	Difficulty of Randomization
Illumination	Light, Dark	Hard
Range to Target	Short, Medium, Long	Hard
Fuze Type	A, B	Easy
Angle of Fire	Low, High	Easy

TestScience
Data . Driven . Defense

Data Management Plays a Key Role throughout the T&E Lifecycle



TestScience
Data . Driven . Defense

A Good Test Design Supports Credible and Clear Analyses

A successful analysis should follow
the structure of the test design.

Building a statistical model using the factors
specified in our test design allows us to
quantify differences across relevant conditions
and draw conclusions with confidence.

All experiments are designed with an analysis methodology in mind:
to reap the benefits, we need to follow through to the analysis!

TestScience
Data . Driven . Defense

IDA

Test Design

Lesson 2: Determine Design Goals

Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Before creating a test design, ensure you have the answers to these questions:

Are the system, mission, and operating environment clearly defined?

Are the goals of the test clearly defined?

What are the question(s) to be answered in testing?

Does the overall test strategy support characterization of combat mission capability?

Before creating a test design, ensure you have the answers to these questions:

Are the system, mission, and operating environment clearly defined?

Are the goals of the test clearly defined?

What are the question(s) to be answered in testing?

Does the overall test strategy support characterization of combat mission capability?

Before creating a test design, ensure you have the answers to these questions:

Are the system, mission, and operating environment clearly defined?

Are the goals of the test clearly defined?

What are the question(s) to be answered in testing?

Does the overall test strategy support characterization of combat mission capability?

Before creating a test design, ensure you have the answers to these questions:

Are the system, mission, and operating environment clearly defined?

Are the goals of the test clearly defined?

What are the question(s) to be answered in testing?

Does the overall test strategy support characterization of combat mission capability?

What are some common design goals?

Screen	for influential factors driving performance
Characterize	performance across an operational envelope
Optimize	system performance with respect to a set of conditions
Compare	two systems (or more) across a variety of operating conditions
Identify Problems	that degrade system performance

TestScience
Data . Driven . Defense

What are some common design goals?

Screen	for influential factors driving performance
Characterize	performance across an operational envelope
Optimize	system performance with respect to a set of conditions
Compare	two systems (or more) across a variety of operating conditions
Identify Problems	that degrade system performance

TestScience
Data . Driven . Defense

What are reasons to screen for factors?

What are reasons to screen for factors?

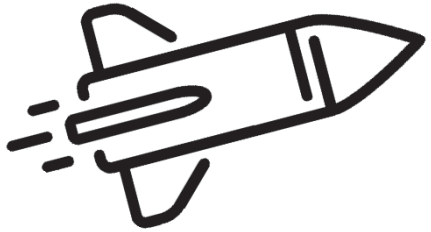
Screening is useful early in testing when we know little about how the system responds to various factors.

What are reasons to screen for factors?

Screening is useful early in testing when we know little about how the system responds to various factors.

Screening is essential to integrated & sequential testing to reduce the factor space prior to operational testing.

Consider a novel hypersonic missile



What factors will impact its performance?

Long list of factors that may or may not be relevant

Large design space is too expensive to optimize or characterize

We need to identify relevant factors with minimal test resources

TestScience
Data . Driven . Defense

How do you pick a screening design?

Design Space Features

Many factors

Typically 2 levels per factor

How do you pick a screening design?

Design Space Features

Many factors

Typically 2 levels per factor

Screening Design Features

Sparse designs

Cover large spaces efficiently

Some higher-order terms confounded

How do you pick a screening design?

Design Space Features

Many factors

Typically 2 levels per factor

Screening Design Features

Sparse designs

Cover large spaces efficiently

Some higher-order terms confounded

Common Screening Designs

Fractional factorial

Optimal

TestScience
Data . Driven . Defense

What are reasons to characterize performance?

What are reasons to characterize performance?

Characterization tests describe system performance across the operational envelope

What are reasons to characterize performance?

Characterization tests describe system performance across the operational envelope

Characterization tests determine if a system meets requirements across operational conditions

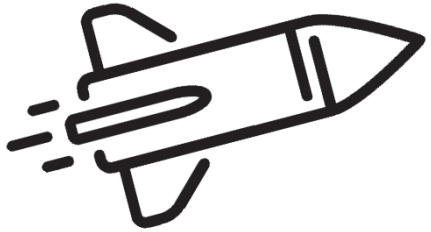
What are reasons to characterize performance?

Characterization tests describe system performance across the operational envelope

Characterization tests determine if a system meets requirements across operational conditions

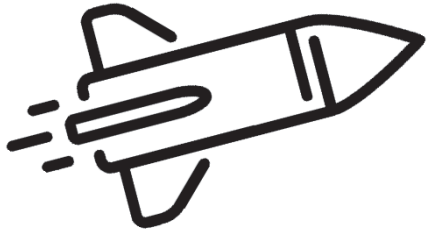
DoDI 5000.02: The test program should produce "**data to characterize combat mission capability across an appropriately selected set of factors and conditions.**"

Consider our screened hypersonic missile



How will our factors impact its performance?

Consider our screened hypersonic missile

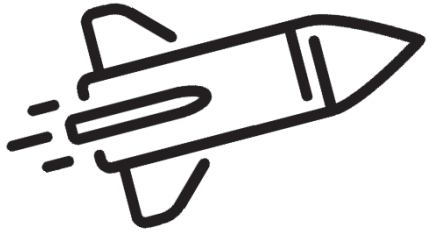


How will our factors impact its performance?

We identified factors relevant to performance through screening

TestScience
Data . Driven . Defense

Consider our screened hypersonic missile



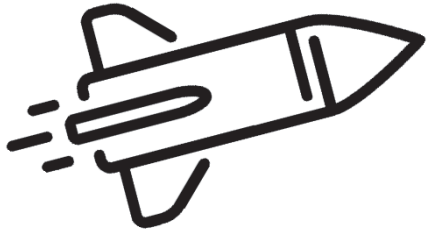
How will our factors impact its performance?

We identified factors relevant to performance through screening

We need to test across a variety of operational conditions …

TestScience
Data . Driven . Defense

Consider our screened hypersonic missile



How will our factors impact its performance?

We identified factors relevant to performance through screening

We need to test across a variety of operational conditions …

… to adequately understand performance using minimal resources

TestScience
Data . Driven . Defense

How do you pick a characterizing design?

Design Space Features

Fewer factors; more levels

Anticipated variation in performance

How do you pick a characterizing design?

Design Space Features

Fewer factors; more levels

Anticipated variation in performance

Characterizing Design Features

Determine factor impact

Allows for adequate precision

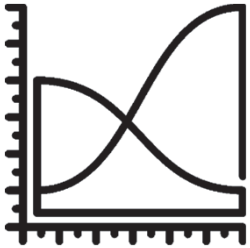
No/little confounding

How do you pick a characterizing design?

Design Space Features	Characterizing Design Features	Common Characterizing Designs
Fewer factors; more levels	Determine factor impact	Factorial
Anticipated variation in performance	Allows for adequate precision	Optimal
	No/little confounding	

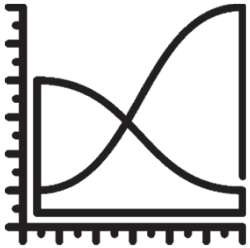
TestScience
Data . Driven . Defense

What are reasons to optimize performance?



Optimization test designs seek the combination of controllable factors that optimizes test outcomes

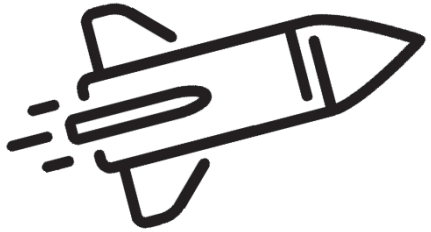
What are reasons to optimize performance?



Optimization test designs seek the combination of controllable factors that optimizes test outcomes

Optimization tests are useful in design, manufacturing, and the development of tactics, techniques, & procedures

(Re-)consider our screened hypersonic missile



What factors will optimize its performance?

We identified factors relevant to performance through screening

We need identify the optimal flight profile settings …

… for minimizing missile miss distance

TestScience
Data . Driven . Defense

How do you pick an optimization design?

Design Space Features

Few factors;
many levels

Precision needed

TestScience
Data . Driven . Defense

How do you pick an optimization design?

Design Space Features

Few factors; many levels

Precision needed

Optimizing Design Features

Dense designs

Often iterative

Searching for a local or global optimum (min/max)

How do you pick an optimization design?

Design Space Features	Optimizing Design Features	Common Optimizing Designs
Few factors; many levels	Dense designs	Response Surface Methods (RSM)
Precision needed	Often iterative	Optimal
	Searching for a local or global optimum (min/max)	

TestScience
Data . Driven . Defense

An effective test depends upon good, clearly articulated design goals

We have learned about 3 common design goals and how to approach them:

- **Screening**
- **Characterizing**
- **Optimizing**

An effective test depends upon good, clearly articulated design goals

- We have learned about 3 common design goals and how to approach them:
  - **Screening**
  - **Characterizing**
  - **Optimizing**
- We have learning that screening is a great tool in sequential and integrated testing to better constrain the design space for characterization and optimization tests.

An effective test depends upon good, clearly articulated design goals

- We have learned about 3 common design goals and how to approach them:
  - **Screening**
  - **Characterizing**
  - **Optimizing**
- We have learning that screening is a great tool in sequential and integrated testing to better constrain the design space for characterization and optimization tests.
- In the next lesson, *Creating the Design*, we will learn more about common experimental designs.

TestScience
Data . Driven . Defense

TestScience
Data . Driven . Defense

IDA

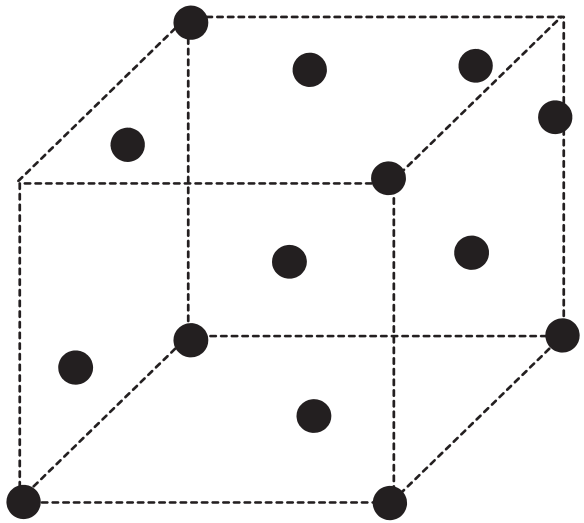
Test Design

Lesson 3: Create the Test Design

Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Design of experiments (DOE)

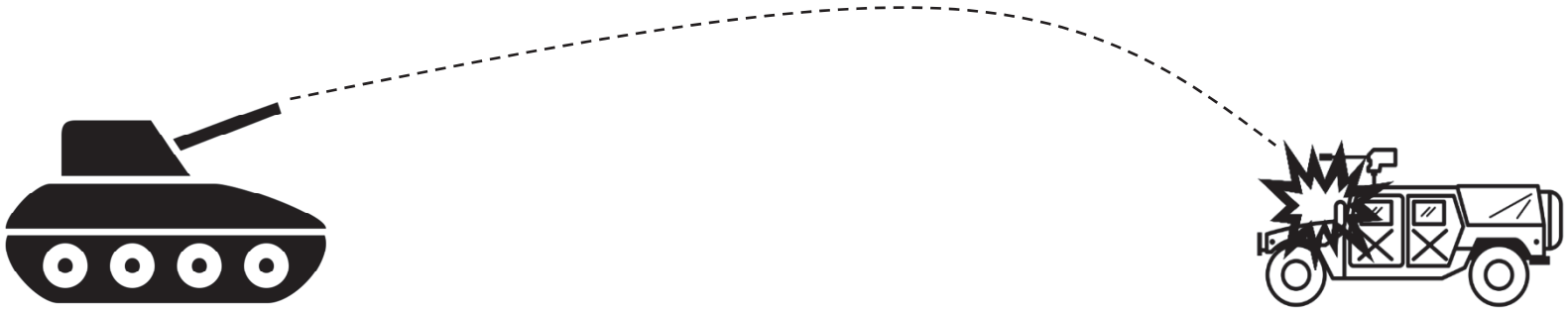
Which Points?



Which Points? Testing a New Artillery Cannon

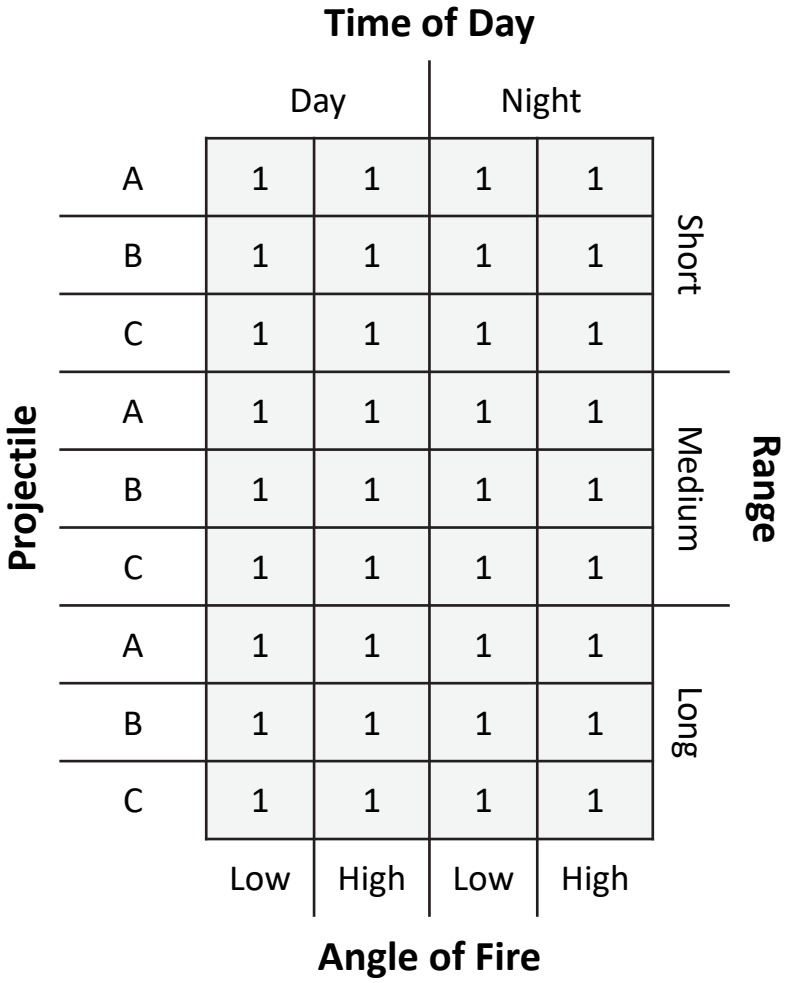
Response Variable

Accuracy of Fires - Miss Distance



Factor	Level
Time of Day	Day, Night
Range to Target	Short, Medium, Long
Projectile Type	A, B, C
Angle of Fire	Low, High

Factorial Designs Examine All Possible Combinations of Each Factor Level

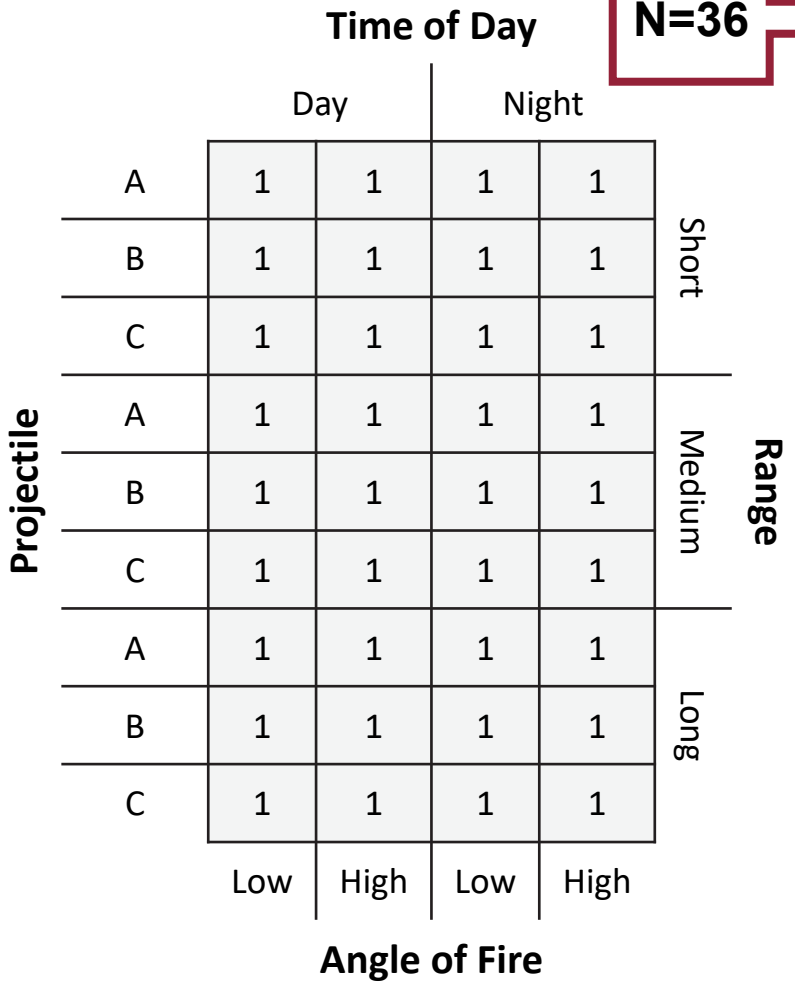


Full Factorial Designs:

- Support the test goals of characterize or compare.
- Examine every possible combination of each level.
- Allow for the estimation of all main effects and all possible interactions.
- Are highly efficient and informative, though potentially prohibitively costly.

TestScience
Data . Driven . Defense

Factorial Designs Examine All Possible Combinations of Each Factor Level



N=36

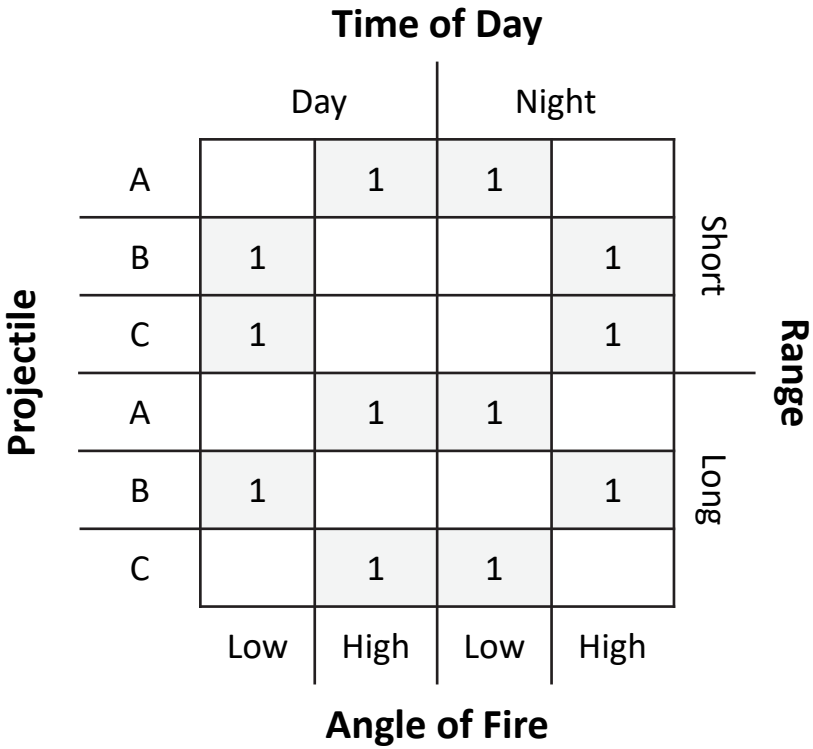
Tests all possible combinations. The first fire mission is conducted during the *Day* at *Short* Range for *Low* Angle using *Projectile A*

Full Factorial Designs:

- Support the test goals of characterize or compare.
- Examine every possible combination of each level.
- Allow for the estimation of all main effects and all possible interactions.
- Are highly efficient and informative, though potentially prohibitively costly.

TestScience
Data . Driven . Defense

Typically, Screening Or Characterization Experiments Involve a Fractional Factorial Design



Fractional Factorial Designs:

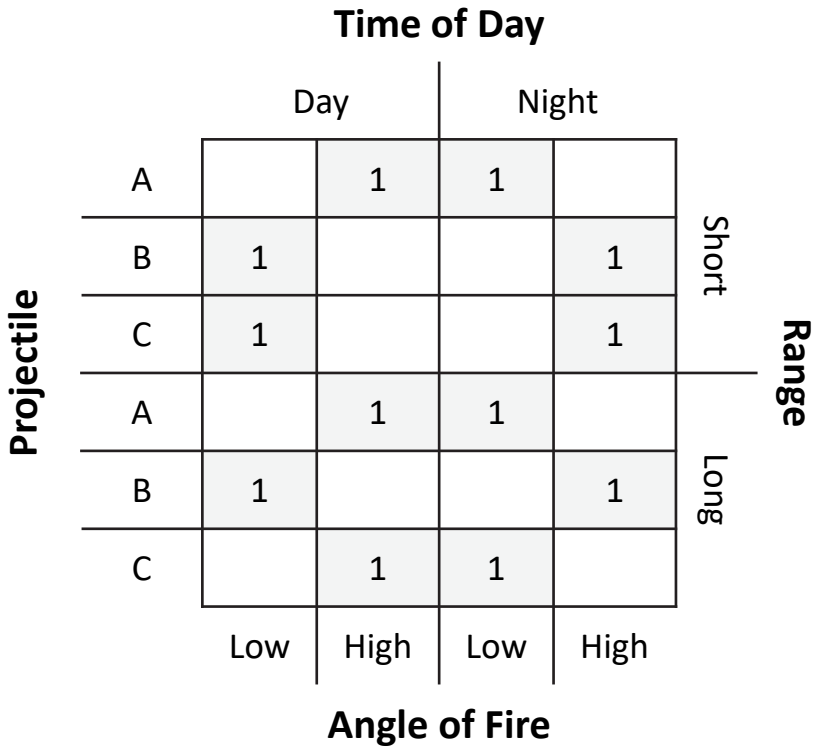
- Support the test goals of screen, characterize, or compare.
- Require a subset of the runs required for a full factorial.
- Achieve a large reduction in test points by trading off the ability to estimate high-order interaction effects.

TestScience
Data . Driven . Defense

Typically, Screening Or Characterization Experiments Involve a Fractional Factorial Design

N=12

Places test points at the right conditions to support a main effects model

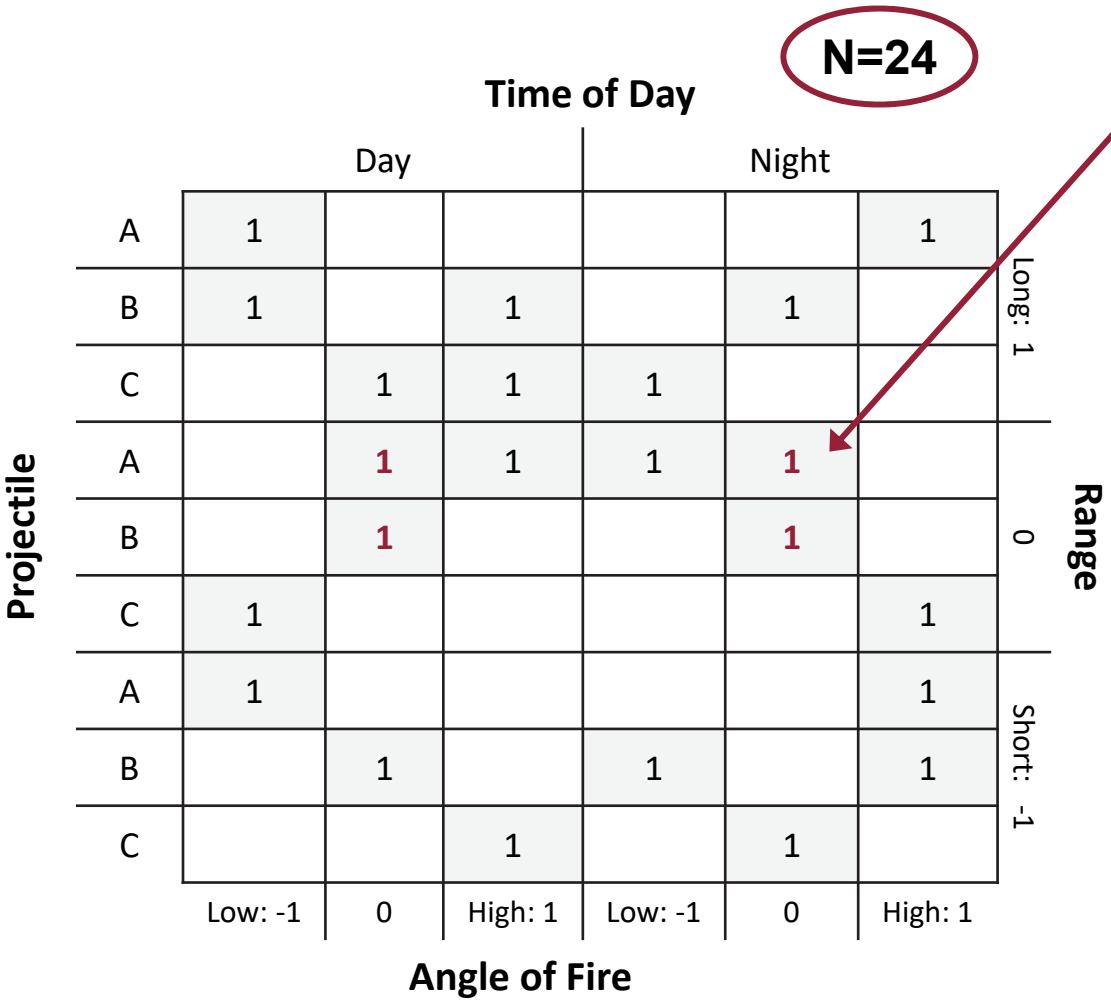


Fractional Factorial Designs:

- Support the test goals of screen, characterize, or compare.
- Require a subset of the runs required for a full factorial.
- Achieve a large reduction in test points by trading off the ability to estimate high-order interaction effects.

TestScience
Data . Driven . Defense

Response Surface Designs spread test points to collect data throughout the experimental region to support a detailed model of the response



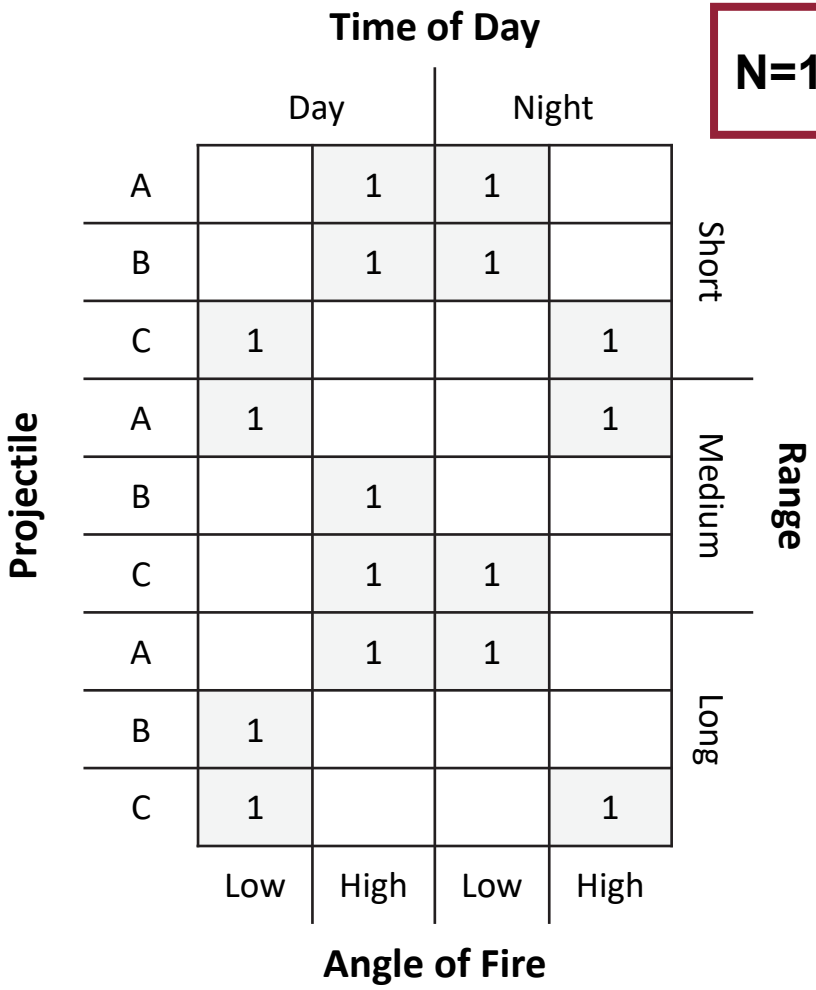
Includes center points to check for curvature.

Response Surface Designs

- Support the test goals of characterize, optimize, predict, improve.
- Helpful in locating where in the design space response are optimal.

TestScience
Data . Driven . Defense

Optimal Designs Are Most Useful When The Number Of Test Points Is Constrained To Preclude A Factorial Design



N=16

Places test points at the right conditions to estimate specified model terms

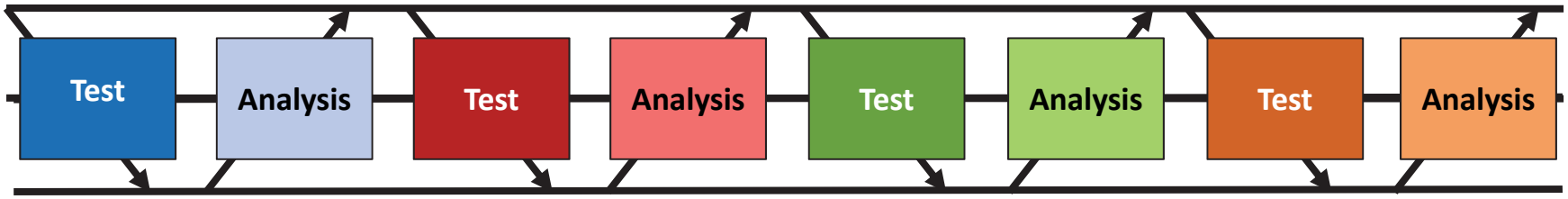
Optimal Designs

- Support the test goals of characterize, optimize, predict.
- Useful when the number of test points is constrained to preclude a factorial design.
- Requires researcher-specified model and a fixed sample size (a subset of the runs required for a full factorial).

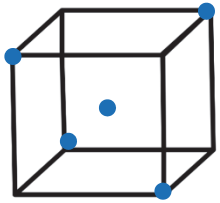
TestScience
Data . Driven . Defense

Plan for a Sequential Experimental Campaign

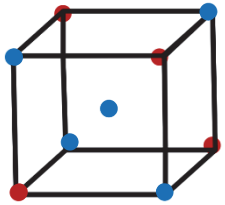
Experiments, Data, and Observations



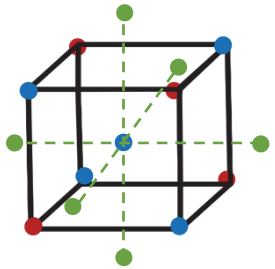
Theories, Hypotheses, Expert Opinion



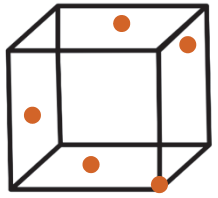
Screening Designs



Interaction Designs



Higher-Order Designs



Validation Designs

TestScience
Data . Driven . Defense

Rarely is the Test Design Process Simple

Multidimensional Evaluation

Many Categorical Factors with More than Two Levels

Restrictions on Randomization

Disallowed Factor Combinations

Limited Resources for Replication

Our test of the new cannon includes factors that are Difficult to Completely Randomize and have Disallowed Combinations

Factor	Level	Difficult to Randomize	Disallowed Combinations
Time of Day	Day, Night	Hard	Night and Short Range
Range to Target	Short, Medium, Long	Hard	
Projectile Type	A, B, C	Easy	A and Long Range
Angle of Fire	Low, High	Easy	

Split-plot designs allow your DOE to reflect your test execution

Time of Day	Range	Angle of Fire	Projectile Type
Day	Long	Low	B
		Low	C
		High	B
		High	C
Tactical Move			
Day	Short	Low	A
		Low	B
		High	C
		High	B
Tactical Move			
Night	Medium	Low	C
		High	B
		Low	B
		Low	A
Tactical Move			
Day	Medium	High	A
		High	B
		Low	C
		Low	A
Tactical Move			
Night	Medium	High	A
		Low	B
		Low	A
		High	C
Tactical Move			
Night	Long	High	B
		Low	B
		High	C
		Low	C
Tactical Move			
Night	Long	Low	B
		Low	C
		High	C
		High	C
Tactical Move			
Day	Medium	High	C
		High	B
		High	A
		Low	B

Whole-Plot

Sub-Plots

Failing to correctly design and account for execution order could lead to wrong conclusions.

Lighting and **Range** are ***hard*** to change between Tactical Moves

Angle of Fire and **Projectile Type** are ***easy*** to vary from one fire mission to the next

TestScience
Data . Driven . Defense

Which Points? Testing for Problem Cases in a Software System Application

System required to pull and integrate information from multiple sources and display via applications.

Response Variable

Works as Expected /

Does not Work as Expected

There are 22 applications total; personnel typically would not need to open more than 8 at a time.

Are there any combinations of applications that do not work together?

TestScience
Data . Driven . Defense

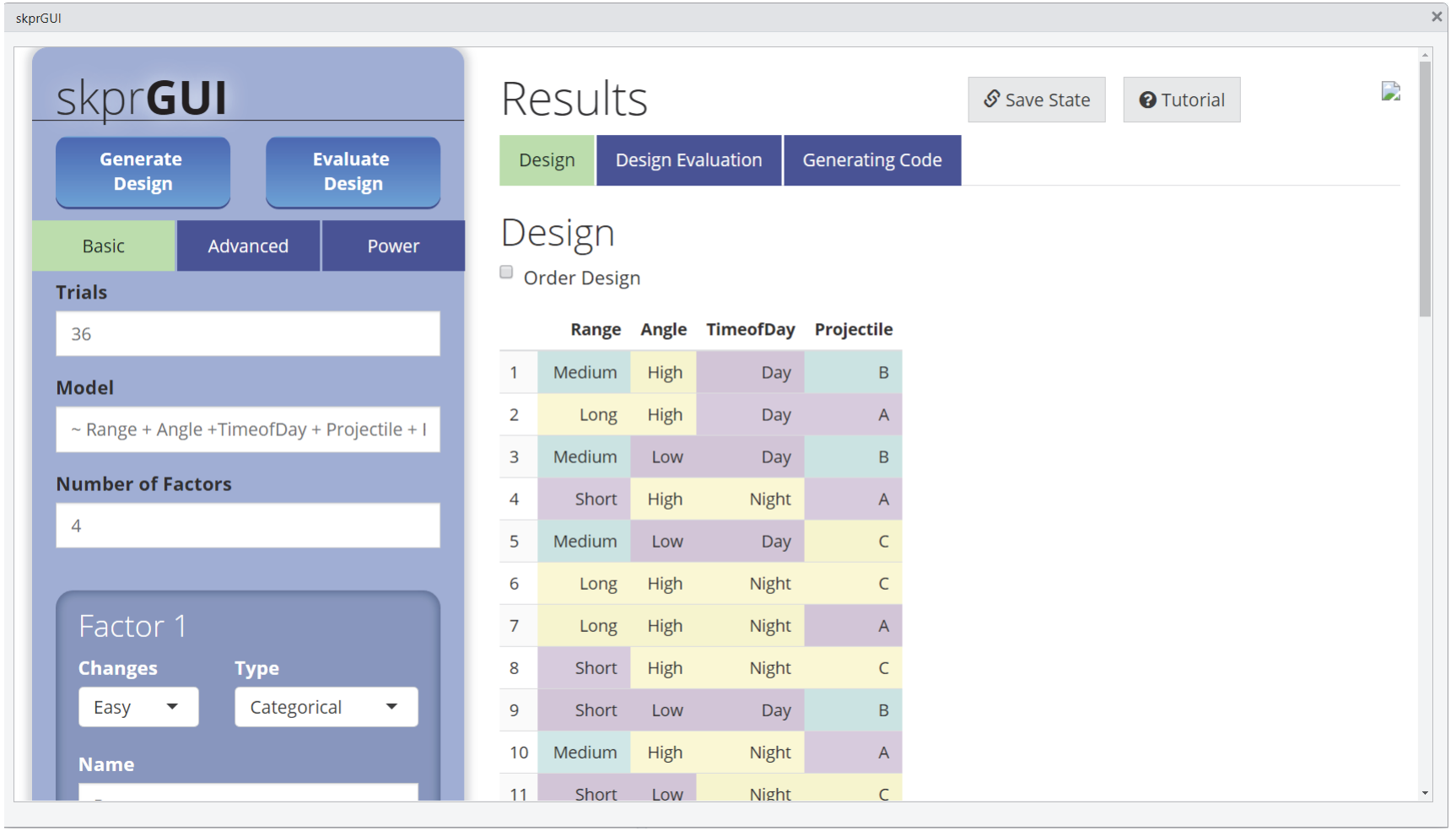
Combinatorial Designs are useful for finding bugs and conditions where the system breaks.

C#	W 1	W 2	W 3	W 4	W 5	W 6	W 7	W 8	W 9	W 10	W 11	W 12	W 13	W 14	W 15	W 16	W 17	W 18	W 19	W 20	W 21	W 22
1	0	0	0	1	1	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0	0	1
2	0	0	1	0	0	0	0	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1
3	0	1	0	0	0	0	0	0	1	0	1	1	0	0	1	1	1	1	0	0	0	1
4	0	1	1	0	0	0	1	1	0	0	1	0	1	0	1	0	1	0	0	0	0	1
5	1	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	1	1	1	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
70	1	1	0	0	0	0	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	1
71	1	1	1	0	0	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	1
72	0	0	0	1	0	1	0	1	0	1	1	0	0	0	0	1	1	0	1	0	0	1
73	0	0	0	0	1	0	0	1	1	1	0	0	1	0	1	1	1	0	0	0	0	1
74	0	0	0	1	0	0	1	1	1	0	0	1	1	0	1	1	0	0	0	0	0	1
75	0	0	0	1	1	0	0	0	0	1	1	1	1	1	0	0	1	0	0	0	0	1

BUT because of their inability to determine cause and effect,
Combinatorial Designs are not the best choice for Operational Tests.

TestScience
Data . Driven . Defense

Software programs assist testers in selecting and constructing the test designs



skpr: Design of Experiments Suite: Generate and Evaluate Optimal Design

TestScience
Data . Driven . Defense

Summary of Test Designs

Type of DOE	Size	Size Scales With …	Uses	Robustness
Factorial Designs (Full factorial, fractional factorial, factorial with center point, replicated factorial)	Large to medium	Number of factors	Estimate performance at extremes/"corners" of the design space	Very robust to lost data points
Optimal Designs (D-optimal, I-optimal)	Small	Nothing	Design minimal tests when there is minimal risk to lost data, and factors affecting performance are well understood	Fragile design; lost data can cause major problems
Response Surface Design	Large	Number of factors, flexibility of response surface	Estimate performance with non-linear factor effects and interactions	Somewhat robust to lost data points
Hierarchical Designs (blocking, split-plot designs)	Medium to small	Number of groups/blocks, group/block size	Properly randomize, while accounting for natural grouping in your data	Somewhat robust to lost data points; less robust to losing full groups of data points
Combinatorial Designs	Large	Number of factors, size of combinations	Find problems through systematically testing all possible cases	Not robust; must check all combinations

TestScience
Data . Driven . Defense

IDA

Test Design

Lesson 4: Evaluate the Design

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

When evaluating a test design ask:

1. Does this design ensure that **adequate data** are collected?

2. If the system doesn't perform well, will I be able to **determine why**?

3. Does the DOE reflect **the way the test will be conducted**?

4. Will the data collected let me do the **analysis** I want to do?

When evaluating a test design ask:

1. Does this design ensure that **adequate data** are collected?

Power and Confidence

A primary reason to use DOE is to identify the number of observations required to satisfactorily address the test objective

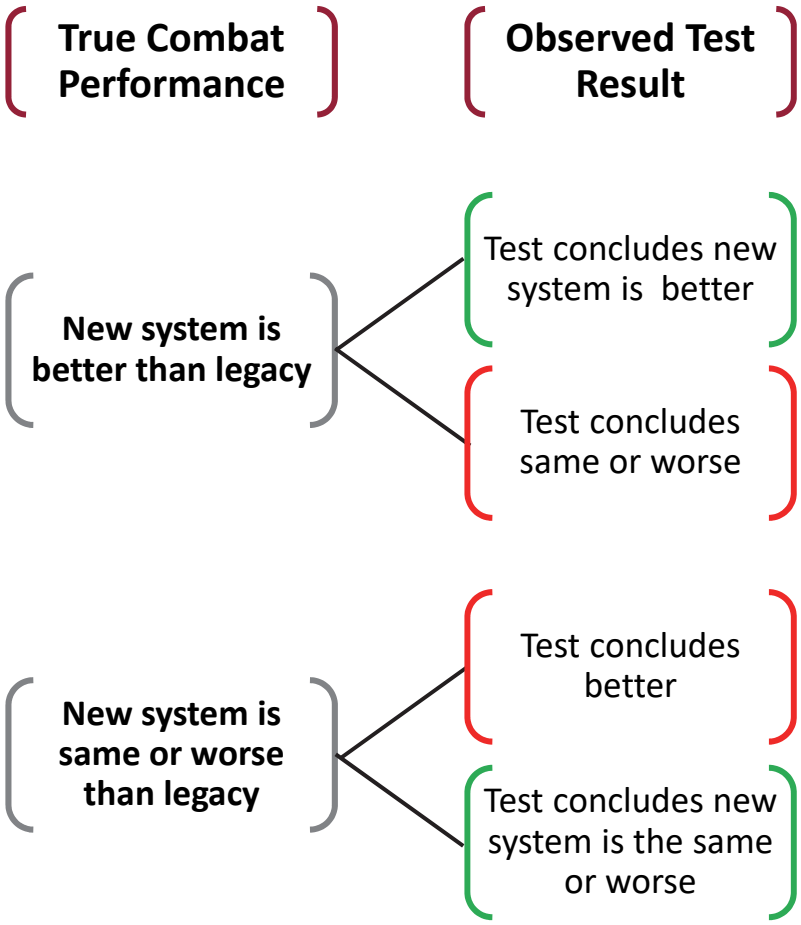
A primary reason to use DOE is to identify the number of observations required to satisfactorily address the test objective

True Combat Performance

New system is better than legacy

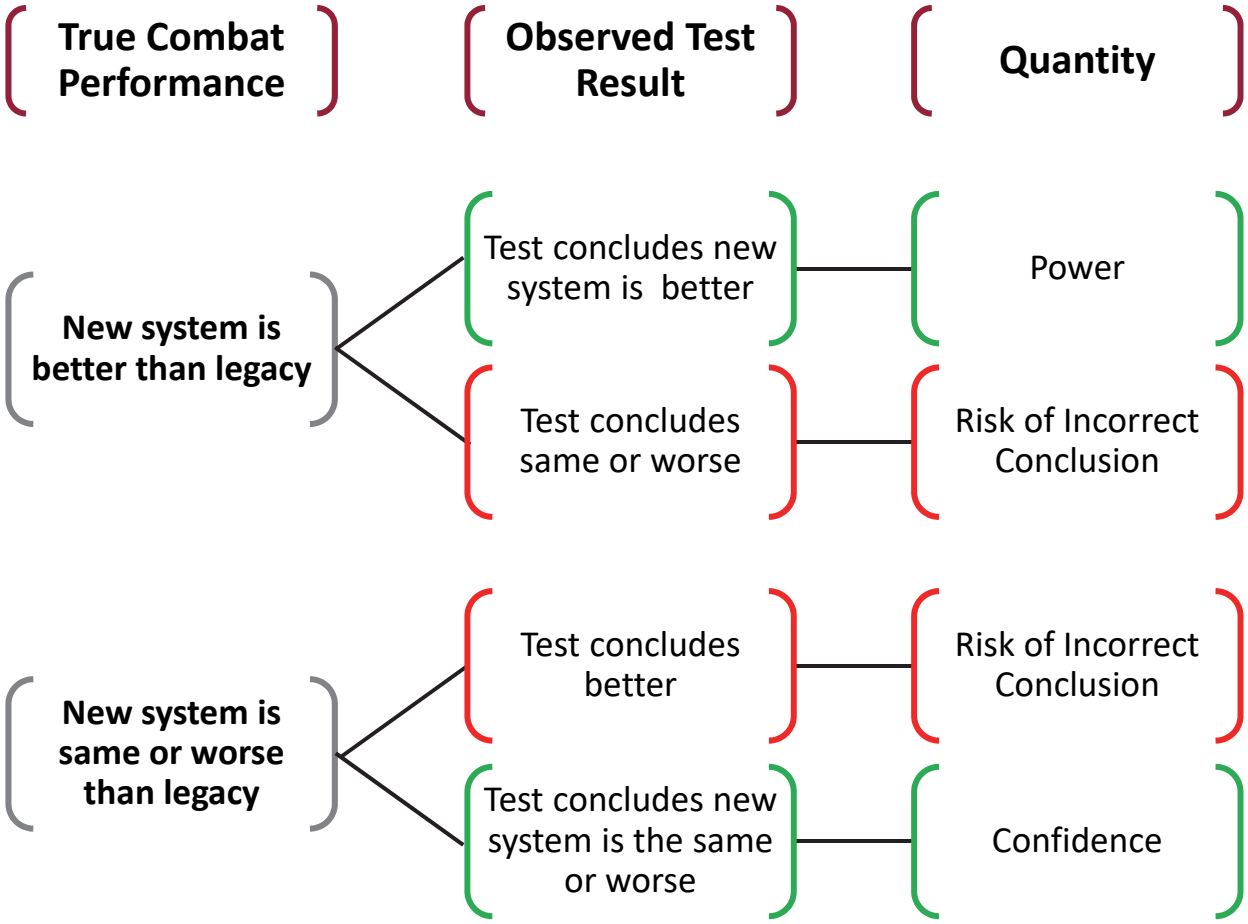
New system is same or worse than legacy

A primary reason to use DOE is to identify the number of observations required to satisfactorily address the test objective



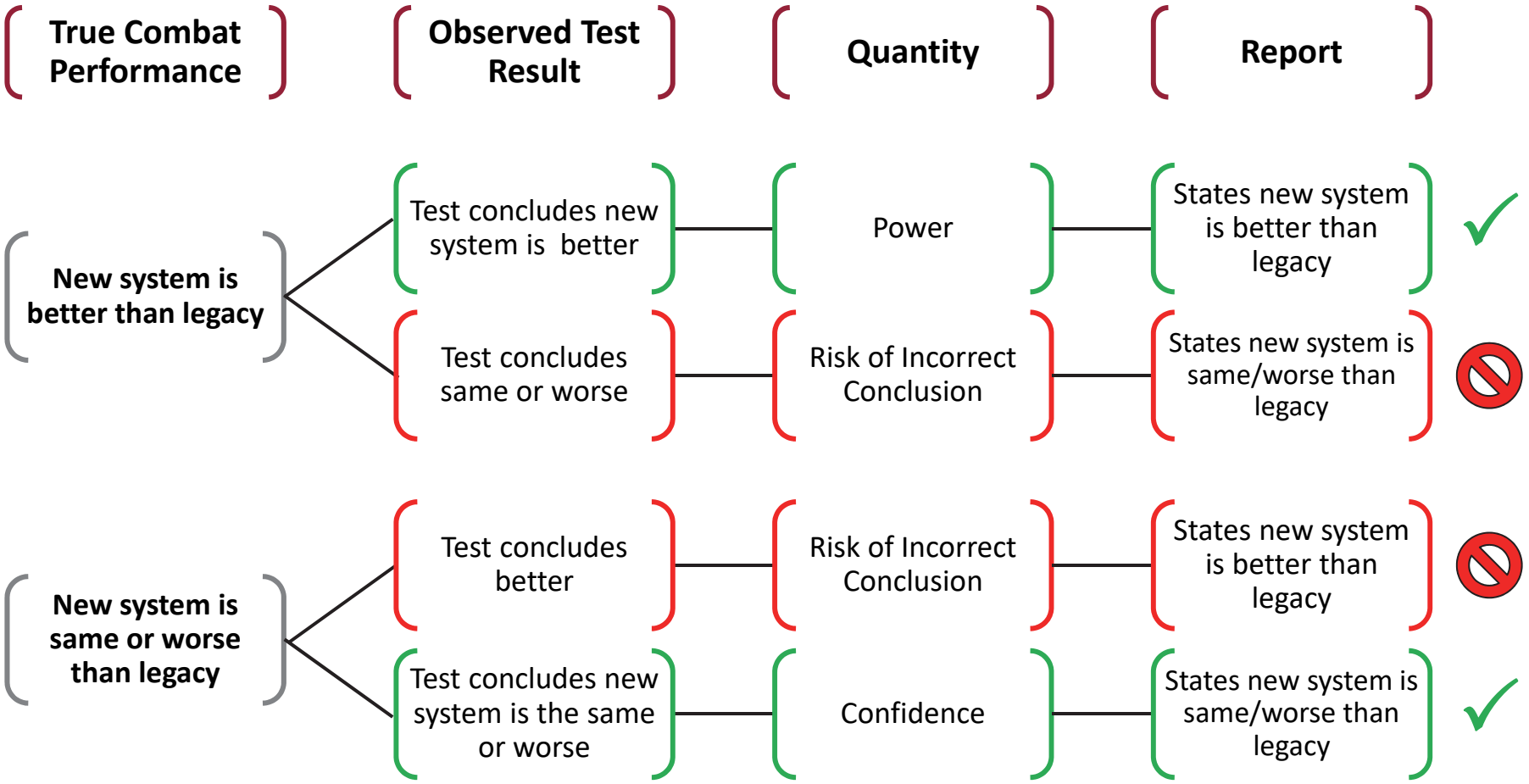
TestScience
Data . Driven . Defense

A primary reason to use DOE is to identify the number of observations required to satisfactorily address the test objective



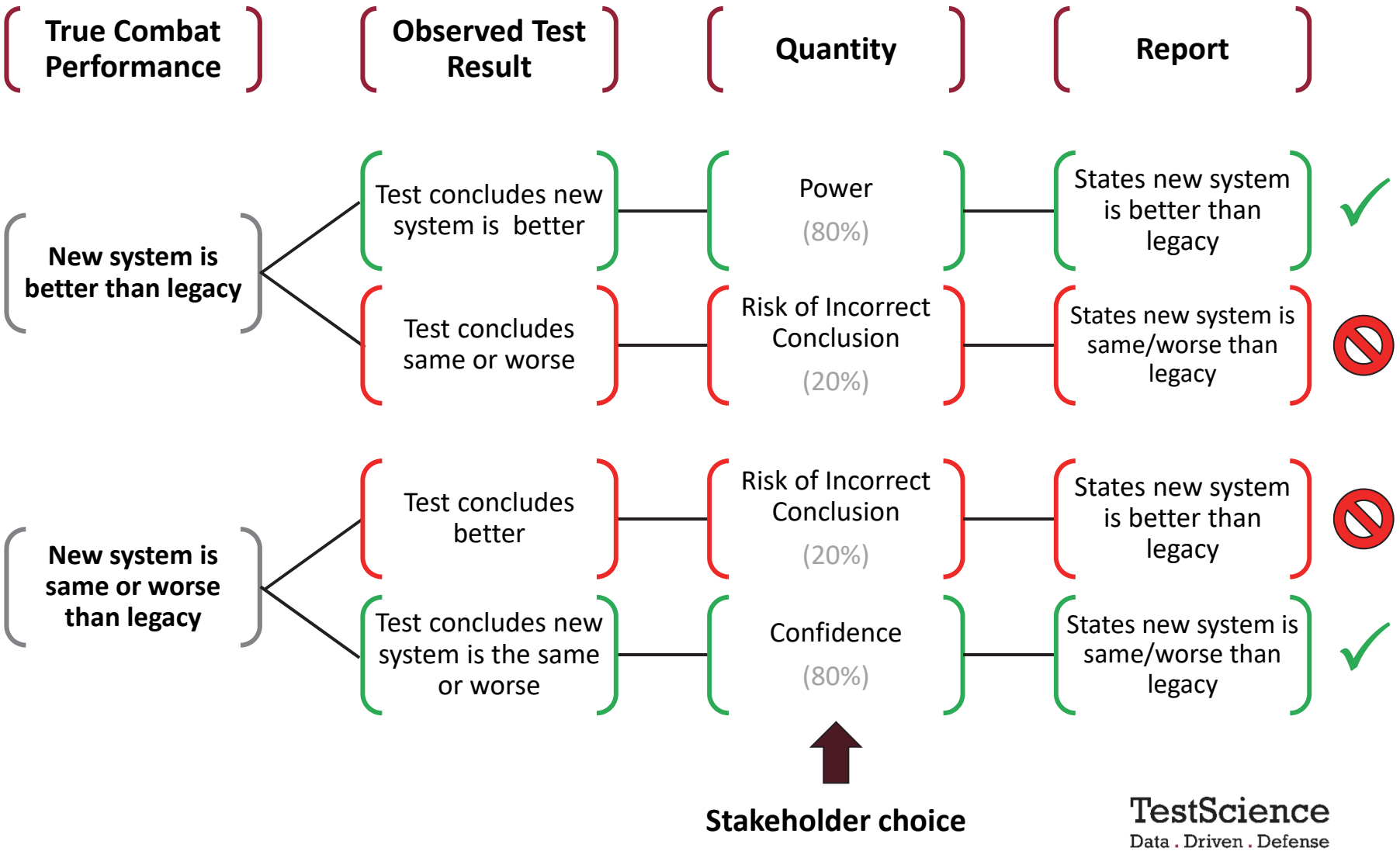
TestScience
Data . Driven . Defense

A primary reason to use DOE is to identify the number of observations required to satisfactorily address the test objective



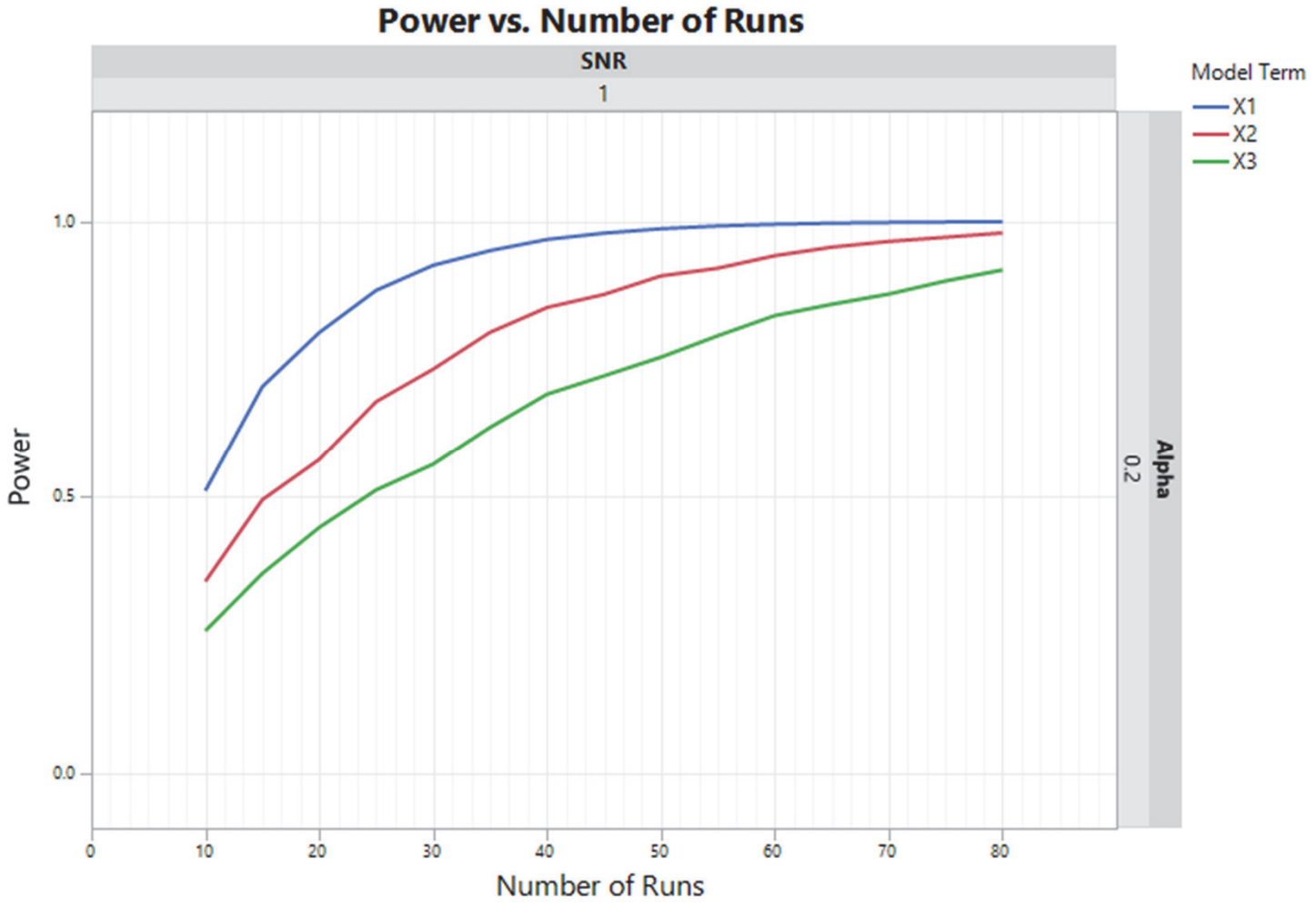
TestScience
Data . Driven . Defense

A primary reason to use DOE is to identify the number of observations required to satisfactorily address the test objective



TestScience
Data . Driven . Defense

Power curves are good ways to visualize and compare power for different test designs



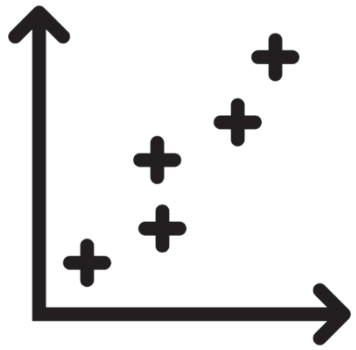
Power estimated by JEDIS v3.0

TestScience
Data . Driven . Defense

When evaluating a test design ask:

2. If the system doesn't perform well, will I be able to determine why?

Correlation and Identifiability



Tests should be designed so that we can identify the situations or missions where performance drops off

Poor performance during the day?

Poor performance against wheeled targets?

Poor performance against wheeled targets during the day?

Narrow Radius Explosive* Test Design

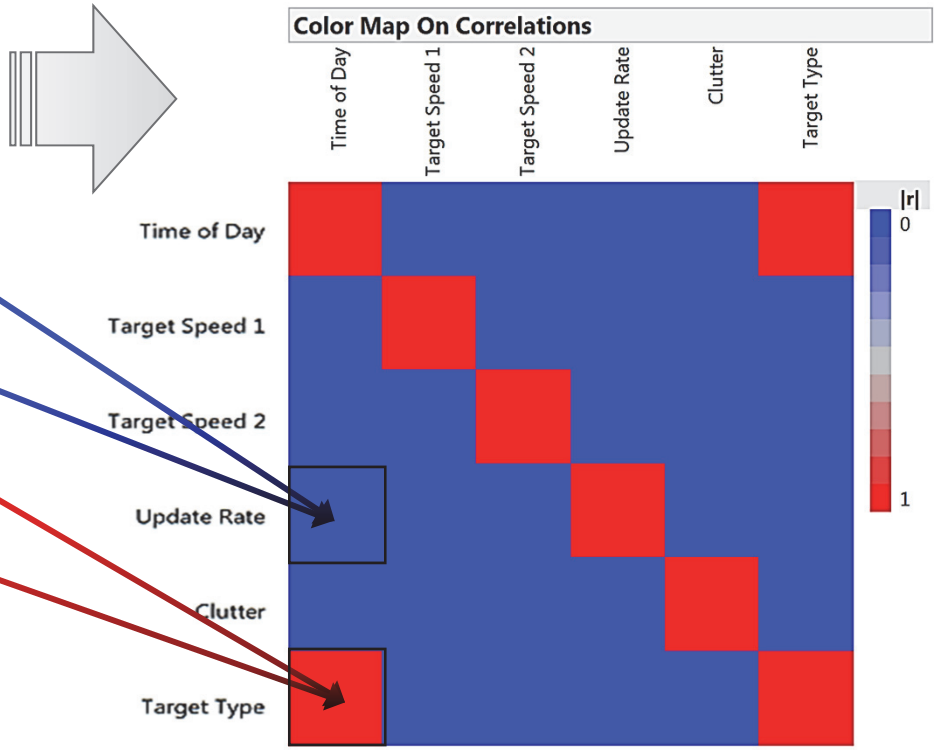
Time of Day	Target Speed	Target Type	Update Rate	Clutter
Day	Fast	Wheeled	30	Y
Day	Fast	Wheeled	12	N
Day	Fixed	Wheeled	12	N
Day	Fixed	Wheeled	30	N
Day	Slow	Wheeled	12	Y
Day	Slow	Wheeled	30	Y
Night	Fast	Tracked	30	N
Night	Fast	Tracked	12	Y
Night	Fixed	Tracked	30	Y
Night	Fixed	Tracked	12	Y
Night	Slow	Tracked	12	N
Night	Slow	Tracked	30	N

*Notional

TestScience
Data . Driven . Defense

Use correlation maps to quickly visualize how hard it will be to distinguish between factors

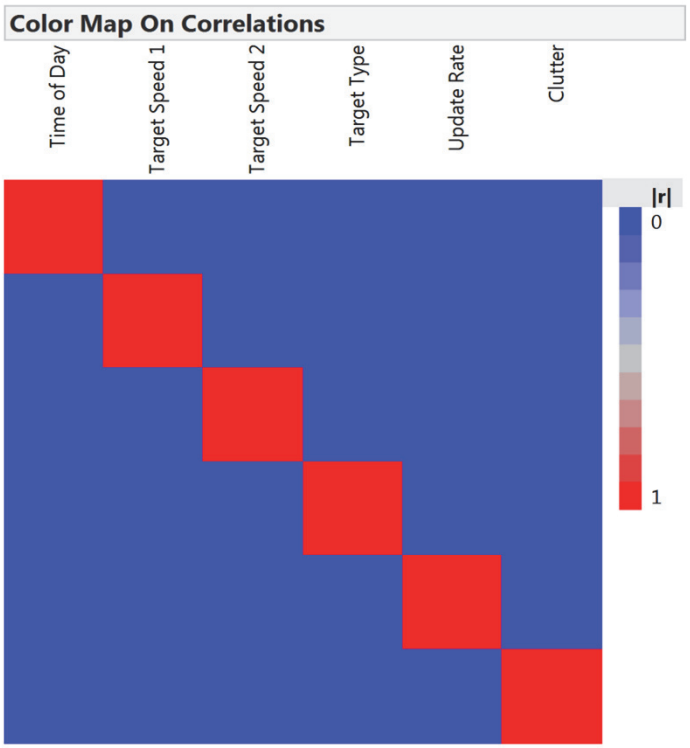
Time of Day	Target Speed	Target Type	Update Rate	Clutter
Day	Fast	Wheeled	30	Y
Day	Fast	Wheeled	12	N
Day	Fixed	Wheeled	12	N
Day	Fixed	Wheeled	30	N
Day	Slow	Wheeled	12	Y
Day	Slow	Wheeled	30	Y
Night	Fast	Tracked	30	N
Night	Fast	Tracked	12	Y
Night	Fixed	Tracked	30	Y
Night	Fixed	Tracked	12	Y
Night	Slow	Tracked	12	N
Night	Slow	Tracked	30	N



- Blue means perfectly uncorrelated (No problems!)
- Red is perfectly (100%) correlated (Cannot differentiate)
- Colors in between indicate some degree of correlation (Can differentiate, but estimates may be biased/incorrect)

TestScience
Data . Driven . Defense

Ideally, your test design will have factors completely independent to ensure identifiability



Time of Day	Target Speed	Target Type	Update Rate	Clutter
Day	Fast	Tracked	30	Y
Day	Fast	Wheeled	12	N
Day	Fixed	Tracked	12	N
Day	Fixed	Wheeled	30	N
Day	Slow	Tracked	12	Y
Day	Slow	Wheeled	30	Y
Night	Fast	Tracked	30	N
Night	Fast	Wheeled	12	Y
Night	Fixed	Tracked	30	Y
Night	Fixed	Wheeled	12	Y
Night	Slow	Tracked	12	N
Night	Slow	Wheeled	30	N

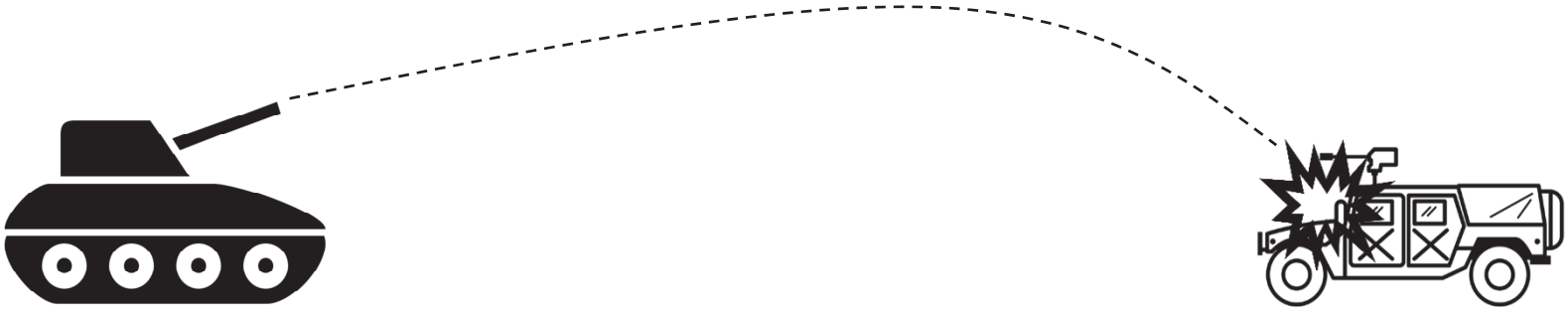
When evaluating a test design ask:

3. Does the DOE reflect the way the test will be conducted?

Restricted Randomization and Disallowed Combinations

RECALL: Our test of the new cannon includes factors that are Difficult to Completely Randomize and have Disallowed Combinations

Factor	Level	Difficult to Randomize	Disallowed Combinations
Time of Day	Day, Night	Hard	Night and Short Range
Range to Target	Short, Medium, Long	Hard	
Projectile Type	A, B, C	Easy	A and Long Range
Angle of Fire	Low, High	Easy	
Traverse Angle	0-15, 16-30, 31-45	Easy	



TestScience
Data . Driven . Defense

Time of Day	Range	Angle of Fire	Projectile Type
Day	Long	Low	B
		Low	C
		High	B
		High	C
Tactical Move			
Day	Short	Low	A
		Low	B
		High	C
		High	B
Tactical Move			
Night	Medium	Low	C
		High	B
		Low	B
		Low	A
Tactical Move			
Day	Medium	High	A
		High	B
		Low	C
		Low	A
Tactical Move			
Night	Medium	High	A
		Low	B
		Low	A
		High	C
Tactical Move			
Night	Long	High	B
		Low	B
		High	C
		Low	C
Tactical Move			
Night	Long	Low	B
		Low	C
		High	C
		High	C
Tactical Move			
Day	Medium	High	C
		High	B
		High	A
		Low	B

Split-plot designs allow your DOE to reflect your test execution

Failing to correctly design and account for execution order could lead to wrong conclusions.

Whole-Plot

Sub-Plots

Lighting and **Range** are ***hard*** to change between Tactical Moves

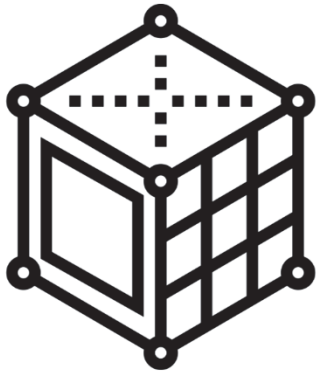
Angle of Fire and **Projectile Type** are ***easy*** to vary from one fire mission to the next

TestScience
Data . Driven . Defense

When evaluating a test design ask:

4. Will the data collected let me do the **analysis** I want to do?

Statistical Model Supported



TestScience
Data . Driven . Defense

Determine what you want to learn from the test and build the design accordingly

Factors?

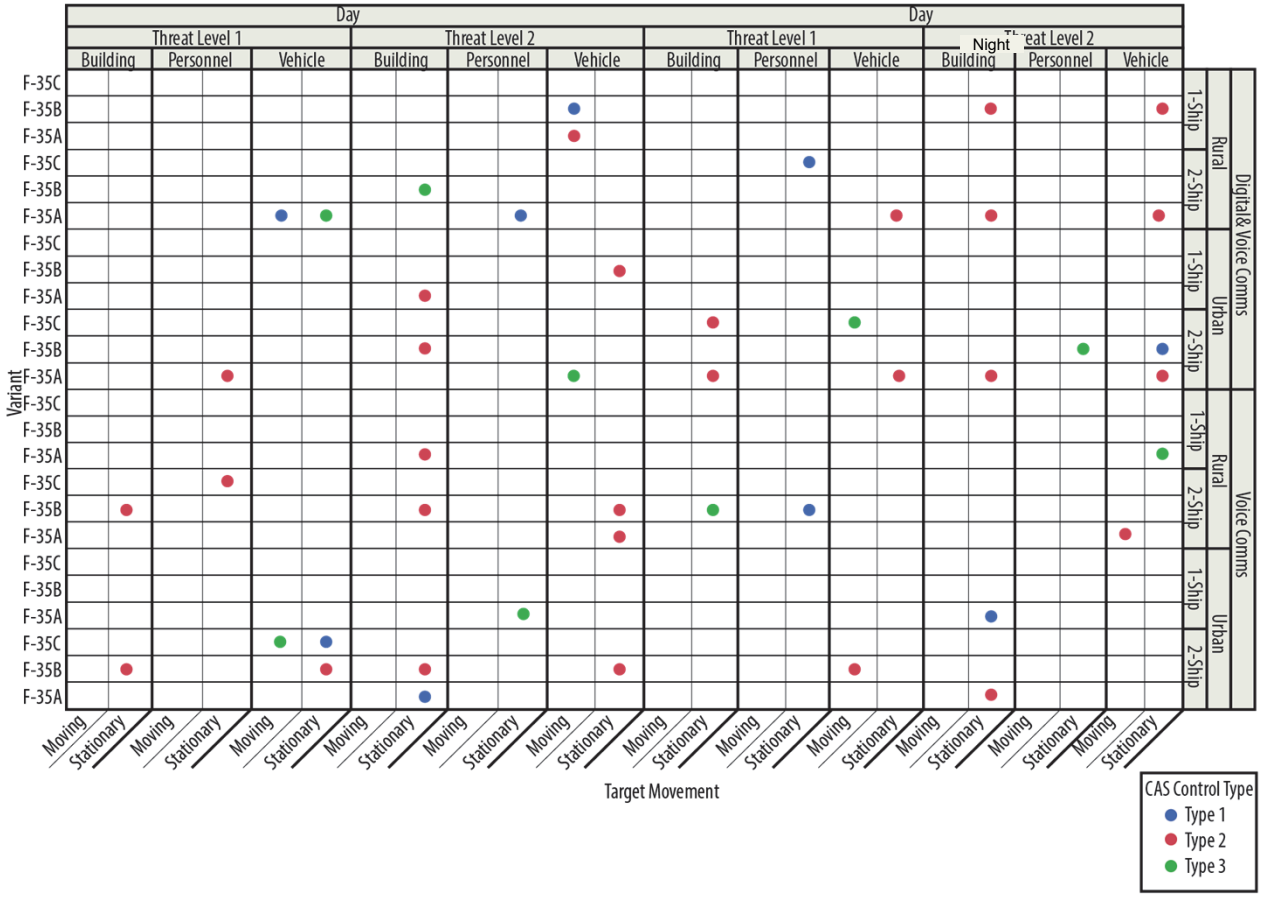
When resources are constrained:

1. Carefully describe the operational conditions in which the system will be employed, and
2. Make sure to get the data points required to fully describe that space.

Interactions?

Curvature?

DOE supports getting the information you need in the most efficient way possible



TestScience
Data . Driven . Defense

TestScience
Data . Driven . Defense

IDA

Test Design

Lesson 5: Manage and Store the Test Data

Institute for Defense Analyses
4850 Mark Center Drive ● Alexandria, Virginia 22311-1882

What is Data Management?

Data Management is the development and execution of policies and practices that organize, protect, and control data and information throughout their lifecycle*

- Access restrictions
- Naming conventions
- Software and hardware systems for storage, search, and discovery
- Roles and responsibilities for data management
- Etc.

Data Management plays a key role throughout the T&E Lifecycle

*Adapted from the Data Management Association' (DAMA) Data Management Body of Knowledge (DMBOK) v2

TestScience
Data . Driven . Defense

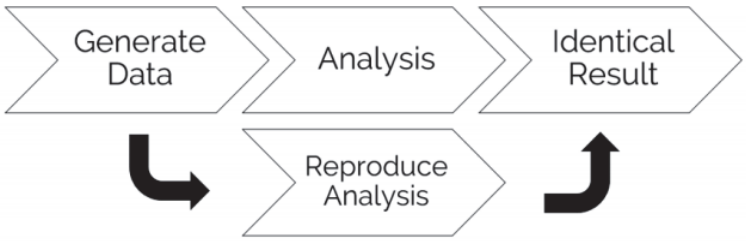
Each organization must define its own goals for data management

FAIR – *Nature* article identifies four properties (Findable, Accessible, Interoperable, Reusable) as goals for research data


SCIENTIFIC DATA

OPEN

Comment: The FAIR Guiding Principles for scientific data management and stewardship



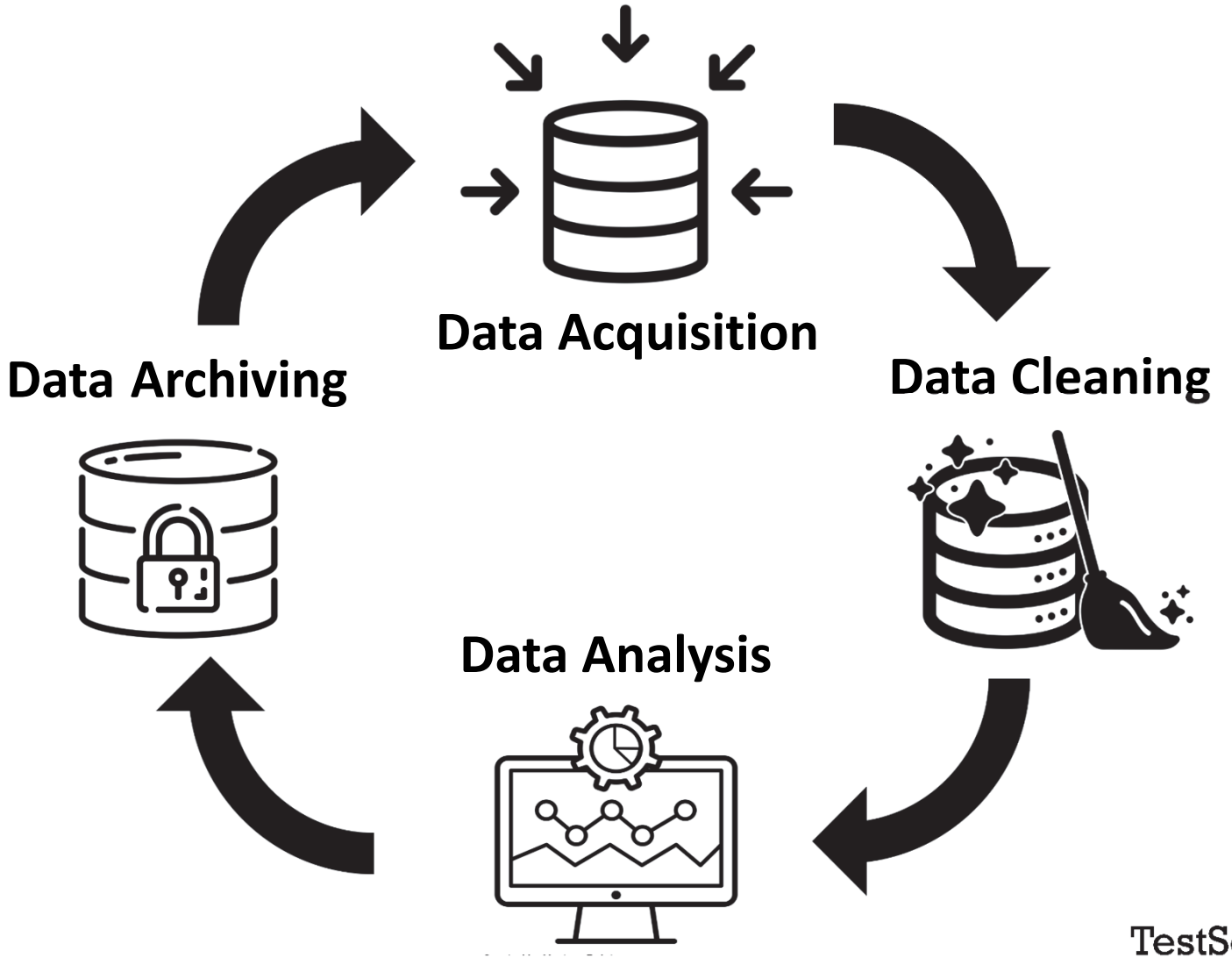
Reproducibility – Manage data to facilitate Reproducible Research*

Security – Ensure that classified, proprietary, and PII data are handled appropriately

* See, e.g., Flack, Kirshenbaum, and Haman tutorial from 2019 DATAWorks, available at www.testcience.org/archive

TestScience
Data . Driven . Defense

Data Management Plays a Key Role throughout the T&E Lifecycle



TestScience
Data . Driven . Defense

Data Acquisition

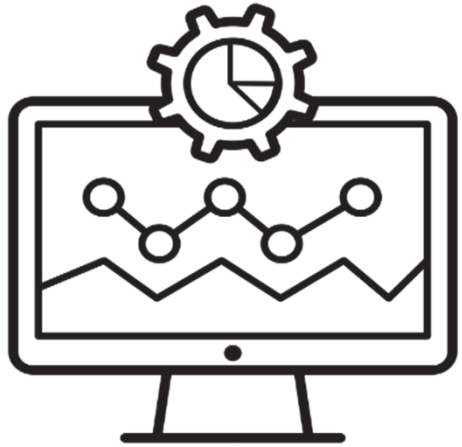
- Storage volume
- Backup options
- Preserving pedigree
- Metadata
- Security

Data Cleaning

- Version control
- Documentation
- Intermediate data products
- Naming conventions
- Automation and repetition

Data Analysis

- Version control
- Documentation
- Organization
- Reproducibility



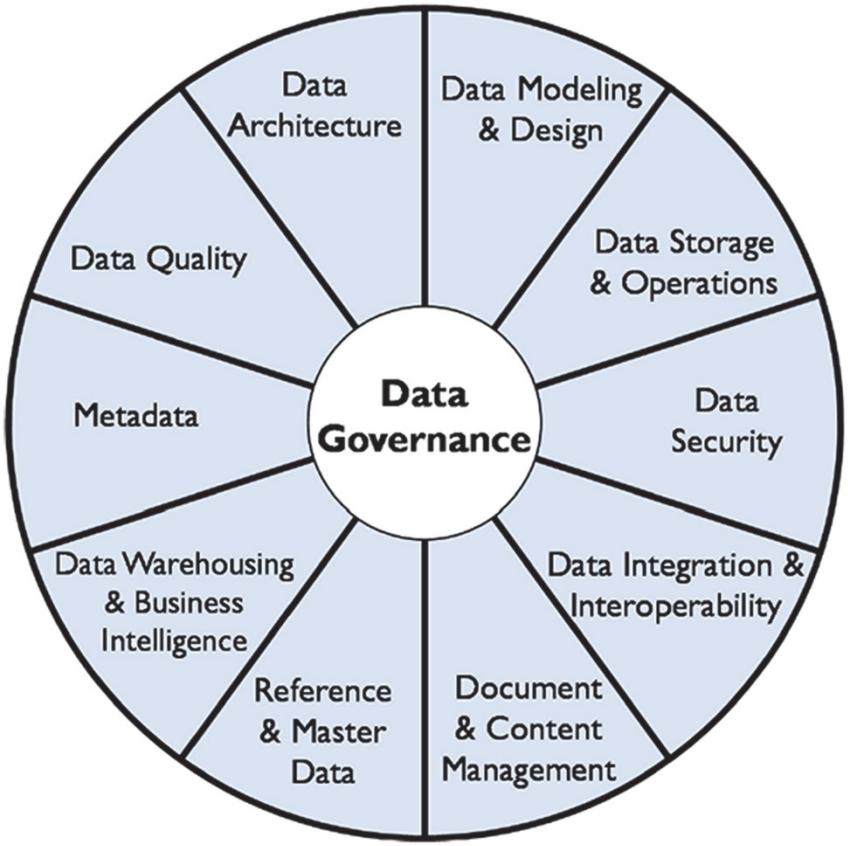
Data Archiving

- Future users
- Findability
- Searchability
- Metadata
- Stewardship

*Adapted from DAMA-DMBOKv2

Just the beginning…

DMBOK identifies 10 key data management functions:



Data are managed – whether through conscious, deliberate process or not.

By building a data management strategy and executing on it, organizations can ensure that their people, practices, and systems are helping them to get the most out of limited data resources and do the best analysis possible.

TestScience
Data . Driven . Defense

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
04-2021	IDA Publication	

4. TITLE ANDSUBTITLE

Test Science Website Videos - Test Planning and Test Design

5a. CONTRACT NUMBER
Separate Contract

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kelly M. Avery (OED); Caitlan A. Fealing (OED); Rebecca M. Medlin (OED); Rachel A. Haga (OED); Matthew R. Avery (OED);

5d. PROJECT NUMBER
C9082

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
NS-D-16393
H 2020-000425

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

10. SPONSOR/MONITOR'S ACRONYM(S)
IDA

11. SPONSOR/MONITOR'S REPORT NUMBER

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Understanding the test design process is fundamental to conducting efficient and effective tests of systems. Planning is a critical element of this process and requires collaboration and iteration among subject matter experts. Test planning involves identifying the test objectives, determining the appropriate outcome measure(s), and identifying what factors may influence those outcomes. The first set of website videos introduces the overarching test design process and provides best practices and examples for each of the three planning steps.

Test design is the process of actually creating and evaluating a set of experimental runs. It entails careful consideration of the test goals to ensure objectives can be met by the data collected, building the corresponding type of design, and then evaluating it across multiple criteria. The test design should have a large enough sample size to be useful, be executable in the real world, and support the desired analysis. The second set of videos describes and provides recommendations for each element of the test design process, and also introduces considerations for how to manage and store data collected from tests.

15. SUBJECT TERMS
Design of Experiments (DOE); OT&E; Statistics; Test Planning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Rebecca Medlin (OED)
Unclassified	Unclassified	Unclassified	Unlimited	171	**19b. TELEPHONE NUMBER** *(include area code)* (703) 845-6731