INSTITUTE FOR DEFENSE ANALYSES

Thoughts on Applying Design of Experiments (DOE) to Cyber Testing

James M. Gilmore
Kelly M. Avery
Matthew R. Girardi
Rebecca M. Medlin

February 2022
IDA Publication NS D-33023
Log: H 2022-000110

INSTITUTE FOR DEFENSE ANALYSES 730
East Glebe Road
Alexandria, Virginia 22305

Approved for public release; distribution is unlimited.

About This Publication

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-19-D-0001, Project AX-1-3100, "Technical Analysis for the Director, Developmental Test, Evaluation, and Assessments," for the AX / Dir, DTE&A / Director, Developmental Test Evaluation and Assessments. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Approved for public release; distribution is unlimited. Cleared for public release by the DoD Office of Prepublication Review, Case 22-S-1540

Acknowledgments

The authors would like to thank IDA committee, Dr. Stephen Ouellette (chair), Dr. John S. Hong, and Dr. Rachel Kuzio de Naray for providing technical review of this effort.

For More Information

John S. Hong, Project Leader
jhong@ida.org, (703) 845-2564

Stephen M. Ouellette Director, SED
souellet@ida.org, (703) 845-2443

Approved for public release; distribution is unlimited.

I N S T I T U T E F O R D E F E N S E A N A L Y S E S

IDA Document NS D-33023

Thoughts on Applying Design of Experiments (DOE) to Cyber Testing

James M. Gilmore
Kelly M. Avery
Matthew R. Girardi
Rebecca M. Medlin

Approved for public release; distribution is unlimited.

Executive Summary

This presentation for Dataworks 2022 provides ideas for how Design of Experiments (DOE) could be applied to Cybersecurity testing. Hypothetical examples of systems are used to illustrate two potential Cyber applications of DOE: (1) Using DOE to plan Mission-Based Cyber Risk Assessments (MBCRAs) conducted by Subject Matter Experts (SMEs) comprehensively covering a system's potential vulnerabilities without assessing every one of an often very large number of such vulnerabilities; and (2) Using DOE to generate a more detailed Cyber test plan using the results of the MBCRA (or other analogous assessments).

Thoughts on Applying Design of Experiments (DOE) to Cyber Testing

Mike Gilmore, Kelly Avery, Matt Girardi, Rebecca Medlin

April 2022

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Approved for public release; distribution is unlimited.

Can/Should DOE be Applied to Cyber Testing?

The DoD Cybersecurity T&E Guidebook **"promotes data-driven mission-impact-based analysis and assessment methods for cybersecurity test and evaluation..."**

In that regard, Design of Experiments offers:

- Efficient coverage of operational space and potential vulnerabilities consistent with limited resources and time
  - Objective and quantitative determination of how much testing is enough and risks of insufficient testing
- Identification and statistical quantification of significant factors/vulnerabilities
- Quantitative evaluation of what is lost if rules of engagement (ROE) are too constraining and/or time is too short
- Addition of structure to previously ad hoc test events, thereby aiding comprehensive evaluation, while not eliminating free play

Framework for Applying DOE (or for Planning any Test and Evaluation)

Determine scope of test
- Questions you can ask about the system

Identify appropriate metrics
- How you should measure system performance

Identify factors that affect performance
- Types of data to collect, operational envelope

Develop Test Design
- Quantity of data necessary, best resource allocation, objective plans

Conduct the test
- Adjust test execution if necessary

Analyze the data
- Structured mathematical data analysis plan appropriate for the design

Draw conclusions
- Defensible risk assessments based on test results

Test & Evaluation requires collaboration

Subject Matter Expertise

Analytical Expertise

DOE tools can be applied at each step

Determine scope of test

Where/what are the potential vulnerabilities?

Example 1 – Using DOE to Help Structure a Systematic Cyber Assessment of a Hypothetical Processing System (PS)

Approved for public release; distribution is unlimited.

Hypothetical PS—Comprises 15 Subsystems; 2 Operations Consoles

How can DOE help?

DOE can be used to---

- Initially guide systematic assessments in narrowing the number of subsystems to be tested*
- Aid structuring the "final" tests
- Aid analysis of test results

*Potential venues include Cyber Table Tops (CTTs) and other Mission-Based Cyber Risk Assessments (MBCRAs)

1 Subsystem 1
2 Subsystem 2
3 Subsystem 3
4 Subsystem 4
5 Subsystem 5
6 Subsystem 6
7 Subsystem 7
8 Subsystem 8
9 Subsystem 9
10 Subsystem 10
11 Subsystem 11
12 Subsystem 12
13 Subsystem 13
14 Subsystem 14
15 Subsystem 15
16 Operations Console 1
17 Operations Console 2

Structuring a Systematic Cyber Assessment of a Hypothetical Processing System (PS)

–Attacks on Single Subsystems—

Narrow the Number of Potential Vulnerabilities

–Attacks Spanning Multiple Subsystems—

Options for Design of PS Cyber Assessment—Single Subsystem Attacks

Consider entry using Operations Consoles---2-level factor (Entry)

Remaining subsystems are targets---15-level factor (Target)

PS Option 1: Operations Console 1, Operations Console 2 for Entry (2)
Remaining Subsystems are Targets (15)
Nearsider and Insider Attack Postures (2)
Native, Foreign Tools (2)

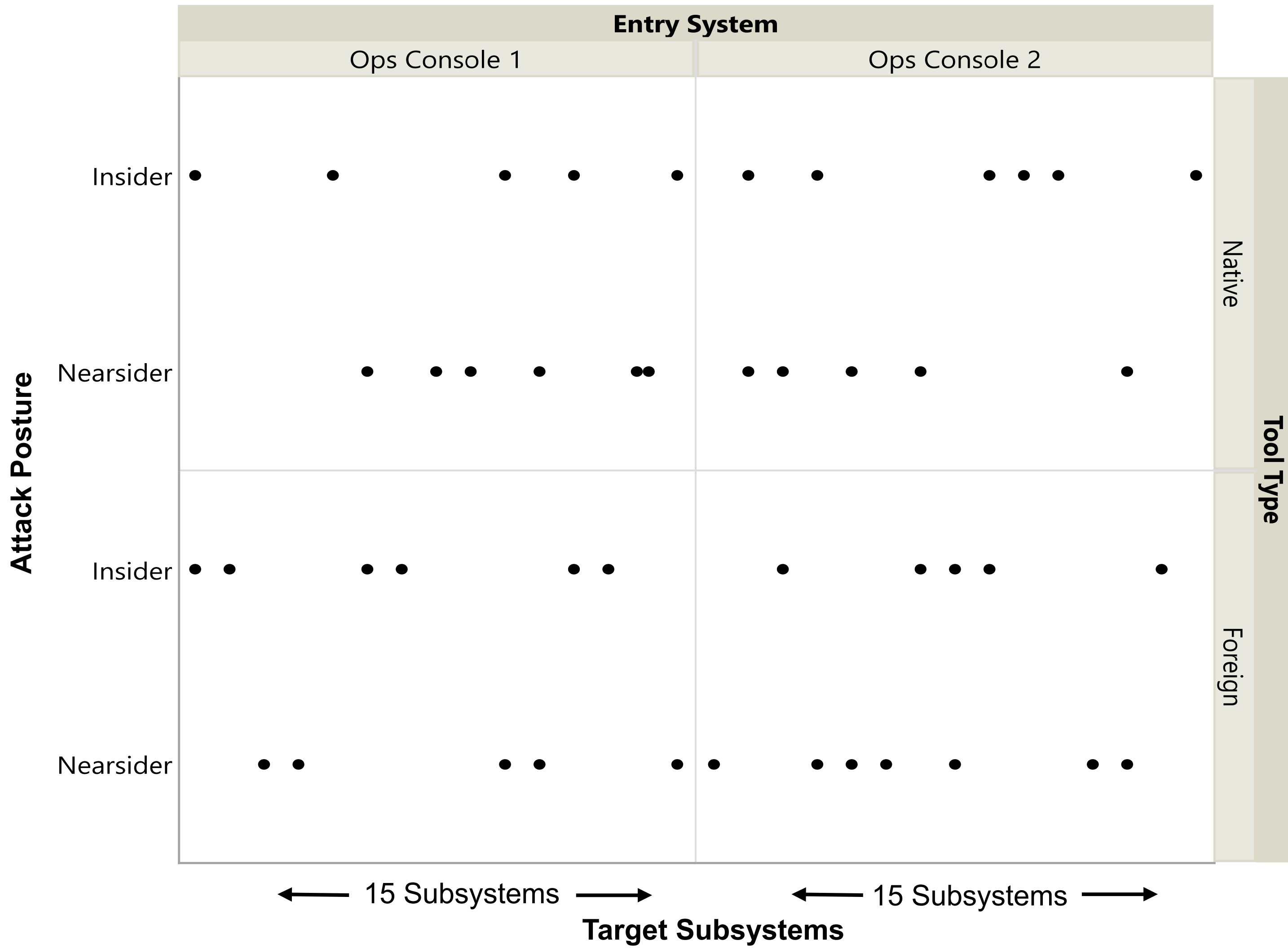
120 Total Combinations

Consider 68 percent (minimal) and 80 percent power to correctly assess/identify vulnerabilities to subsystems (true positive)

Consider 80 percent confidence of correctly excluding vulnerabilities (true negative)

1 Subsystem 1
2 Subsystem 2
3 Subsystem 3
4 Subsystem 4
5 Subsystem 5
6 Subsystem 6
7 Subsystem 7
8 Subsystem 8
9 Subsystem 9
10 Subsystem 10
11 Subsystem 11
12 Subsystem 12
13 Subsystem 13
14 Subsystem 14
15 Subsystem 15
16 Operations Console 1
17 Operations Console 2

PS Design Options for Assessment—Single Subsystem Attacks



Assessing 45 potential vulnerabilities covers 120 combinations with 68% power and 80% confidence; 65 assessments required for 80% power

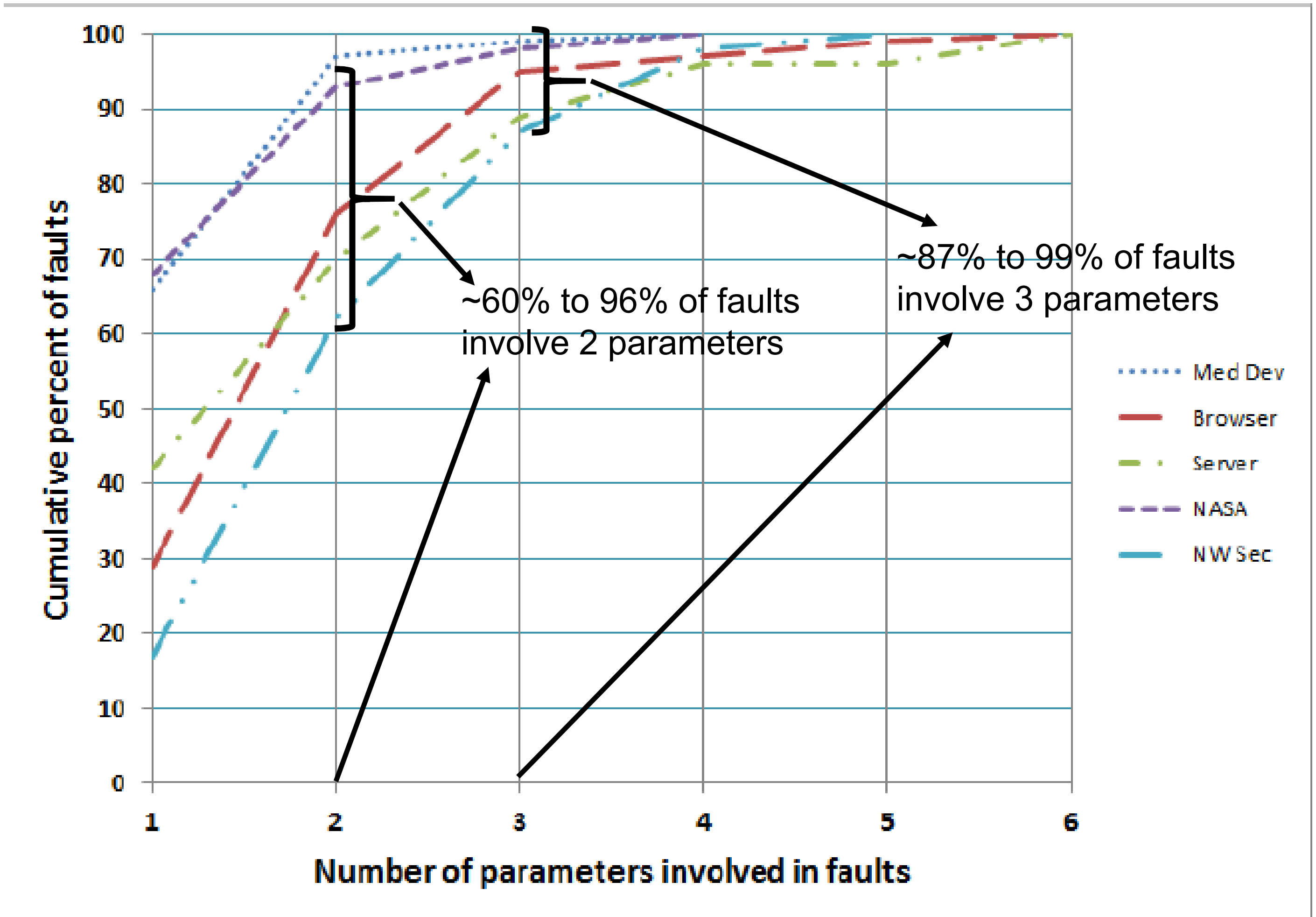
Structuring a Systematic Cyber Assessment of a Hypothetical Processing System (PS)

–Attacks on Single Subsystems—

Narrow the Number of Potential Vulnerabilities

–Attacks Spanning Multiple Subsystems—

Software Faults versus Number of Interacting Parameters



Source: Kuhn, D., et al, Practical Combinatorial Testing, October 2010,
available at https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-142.pdf, accessed January 14, 2022.

PARAMETER = Input Data OR Configuration

Treat Subsystems spanned as a Configuration

Options for Design of PS Cyber Assessment—Attacks Spanning Two Subsystems

Suppose: Assessment of single subsystems described previously narrows focus to 8 subsystems for initial insider (only) penetration/attack through Operations Console 1 or 2; but---

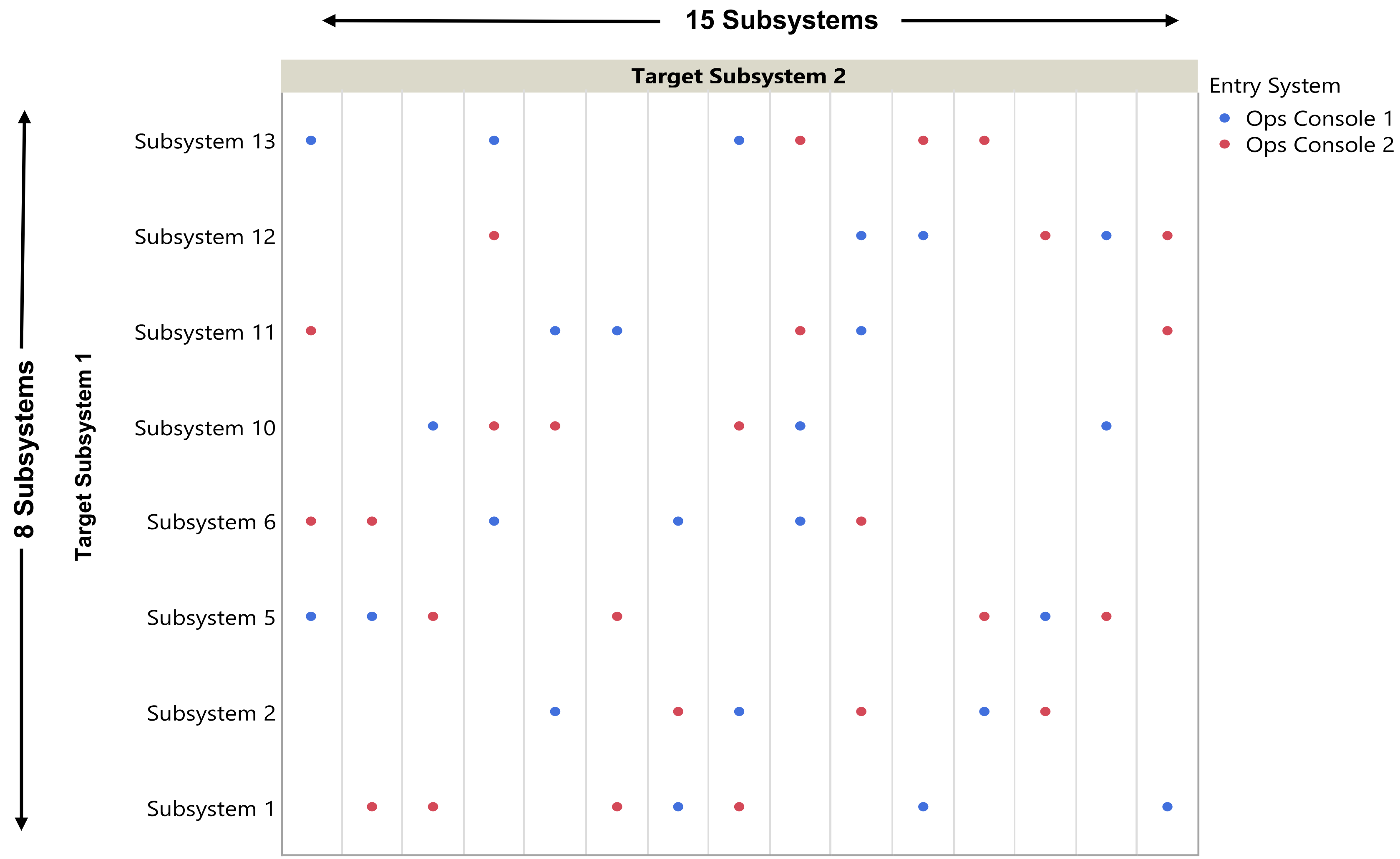
Concern exists regarding attacks spanning more than one subsystem

Consider attacks spanning those 8 subsystems and any one of the other 15-1 with the tool(s) used unspecified, but assumed to be those most applicable in each case as determined by prior assessment (e.g., specific native or foreign)

<u>PS Option 2</u>: Operations Console 1, Operations Console 2 for Entry
8 Subsystems are first Targets (Target Subsystem 1)
14 Subsystems are second targets (Target Subsystem 2)
Insider Attack Posture
Most Applicable Tool

224 Total Combinations (2x8x14)

PS Design Options for Assessment—Attacks Spanning Two Subsystems



Assessing 50 potential vulnerabilities covers 224 combinations with 68% power and 80% confidence; 65 assessments for 80% power

PS Design Options for Assessment—Attacks Spanning Three Subsystems

Suppose Further: Assessment of two-subsystem combinations narrows focus to 6 subsystems as second targets; but---

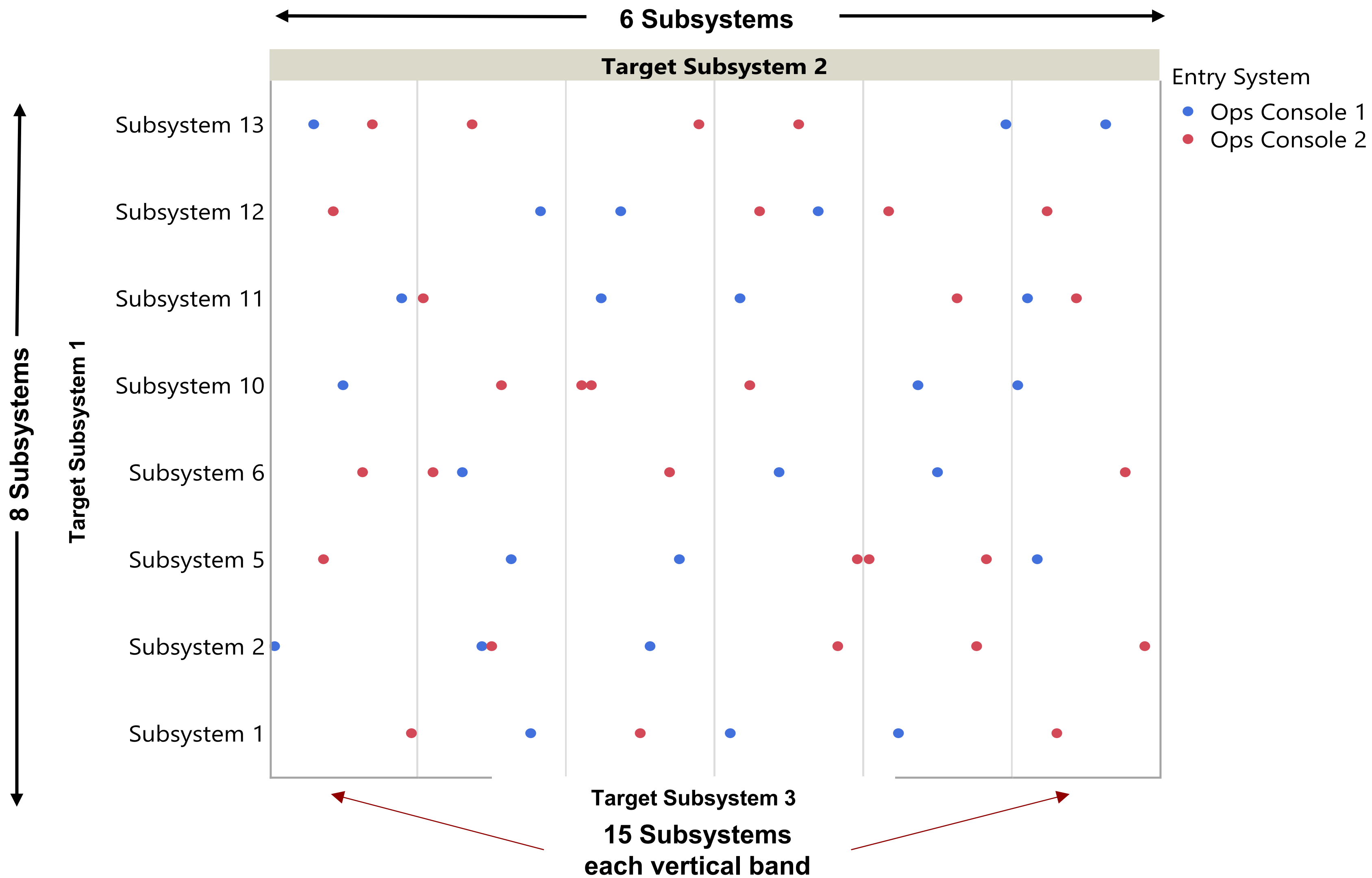
Concern exists regarding attacks spanning up to three subsystems

Consider attacks spanning the identified 8 first targets, 6 second targets, and any one of the remaining 15-2 subsystems

<u>PS Option 3</u>: Operations Console 1, Operations Console 2 for Entry
8 Subsystems as first Targets (Target Subsystem 1)
6 Subsystems as second targets (Target Subsystem 2)
13 Subsystems as third targets (Target Subsystem 3)
Insider Attack Posture
Most Applicable Tool

1248 Total Combinations (2x8x6x13)

PS Design Options for Assessment—Attacks Spanning Three Subsystems



Assessing 55 potential vulnerabilities covers 1248 combinations with 68% power and 80% confidence; 70 assessments for 80% power

Framework for Applying DOE (or for Planning any Test and Evaluation)

Determine scope of test
- Questions you can ask about the system

Identify appropriate metrics
- How you should measure system performance

Identify factors that affect performance
- Types of data to collect, operational envelope

Develop Test Design
- Quantity of data necessary, best resource allocation, objective plans

Conduct the test
- Adjust test execution if necessary

Analyze the data
- Structured mathematical data analysis plan appropriate for the design

Draw conclusions
- Defensible risk assessments based on test results

Demonstrated

How might this work?

Test & Evaluation requires collaboration

Subject Matter Expertise

Analytical Expertise

DOE tools can be applied at each step

Applying the Framework to Cyber T&E (Steps 2 - 3)

Objectives---

Cooperative test – attempt to comprehensively identify vulnerabilities and validate exposures in system

Adversarial test – using the results of the cooperative test in as realistic setting as appropriate, assess system/users to protect, mitigate, and restore when faced with various types of cyber threats

Potential response variables---

Attack thread length/number of steps

Level of threat capability required to achieve action (Nascent, Limited, Moderate, Advanced)

Severity of mission effects (None, Low, Med, High) (*AA only*)

Time to detect / mitigate / restore

Time to penetrate / achieve effect

Potential factors---

Protocol or objective (Web application, servers, interfaces with other systems, etc.)

Type of cyber effect (Confidentiality, Integrity, Availability)

Starting posture (Outsider, Near-sider, Insider)

Tool Type (Native, Foreign)

System load/Number of users (Low, High)

Level of defender participation (Users only, Users + local defenders, Users + local + CSSP)

Examples of many possibilities

Applying the Framework to Cyber T&E (Steps 2 – 3)

- Consider a sequential approach –
  - First stage -- screen for potential vulnerabilities
  - Second stage – refine test, characterize significance of factors and interactions in greater detail
- Cyber/system SMEs should determine which interaction effects are likely/interesting, which specific response variables are most meaningful
- Create design first, then update based on specifics, such as rules of engagement (ROE) and disallowed combinations, while considering tradeoffs
  - Enables effects/constraints of ROE to be understood
- Could include ability to control for learning effects over time
  - Would need to randomize to the extent possible and collect enough data to be able to include coefficients for time and person in the model

Approved for public release; distribution is unlimited.

Applying the Framework to Cyber T&E (Steps 2 – 3)

A model is fit to data to form an empirical relationship between the response variable and factor settings for the purposes of:

--Determining which factors have a large effect on the response

--Making predictions across the factor space (including combinations that were not explicitly tested)

--Quantifying uncertainty in test results

One such model could be:

Responses: Time to get in/achieve effect, Thread length, Level of threat required, Time to detect/mitigate/restore, Severity of mission effects

$$y = \beta_0 + \beta_1(Protocol) + \beta_2(Starting\ Posture) + \beta_3(Tool\ Type) + \beta_4(Network\ Load) + \beta_5(Defenders) + \varepsilon$$

Normally-distributed error

Estimated model coefficients

While the model is linear in its parameters, the factors/responses are not necessarily linear or normal:

Time-based responses are likely right-skewed, so lognormal regression or a survival model may be appropriate

The mission effects response is categorical so a multinomial logistic regression is one appropriate modeling choice

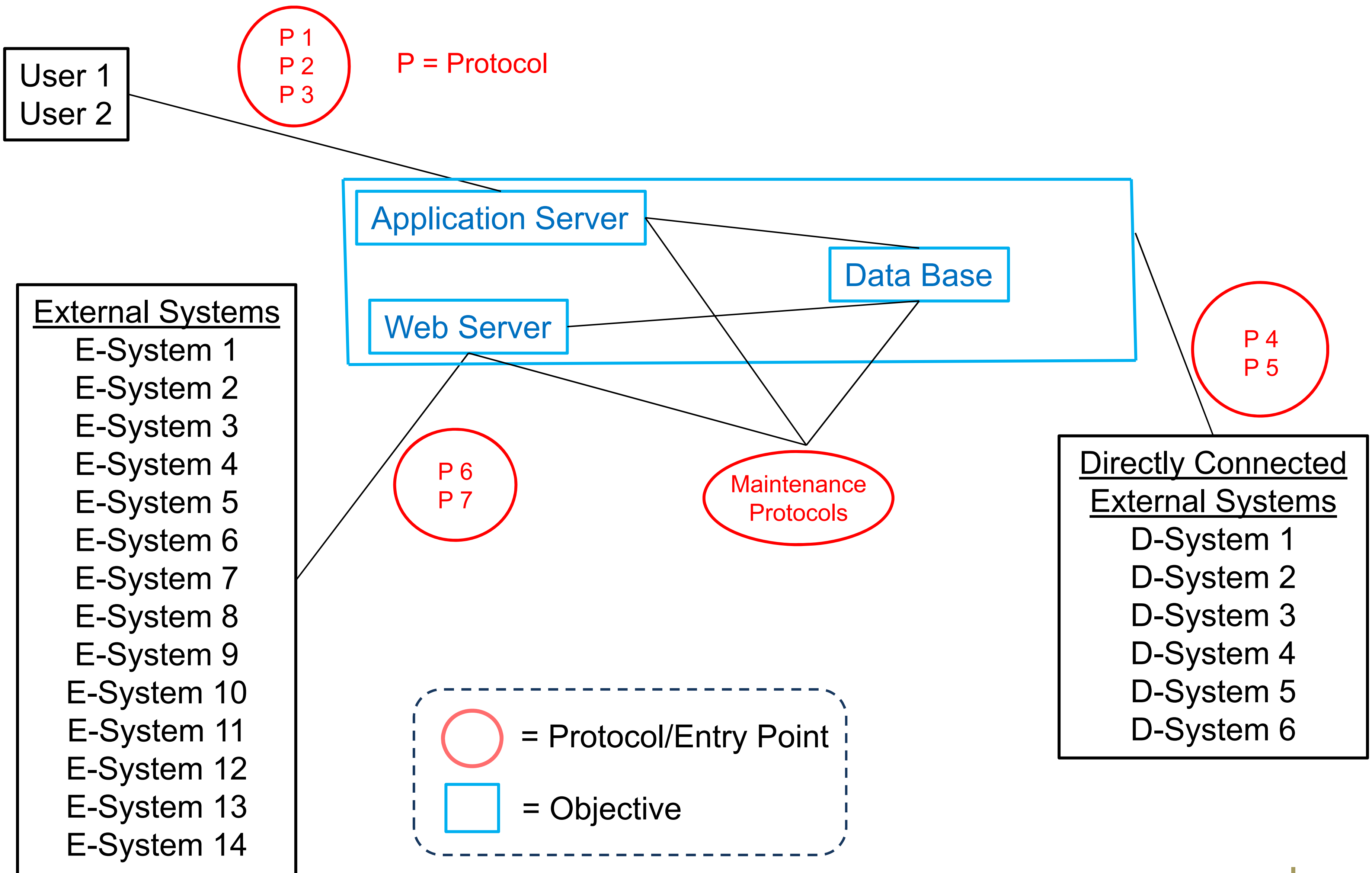
The test could be designed to allow the ability to include additional recorded factors (e.g. tool/method, time) in the model and estimate their effects

Develop Test Design

Example 2 – Hypothetical Command and Control (C^2) System

Approved for public release; distribution is unlimited.

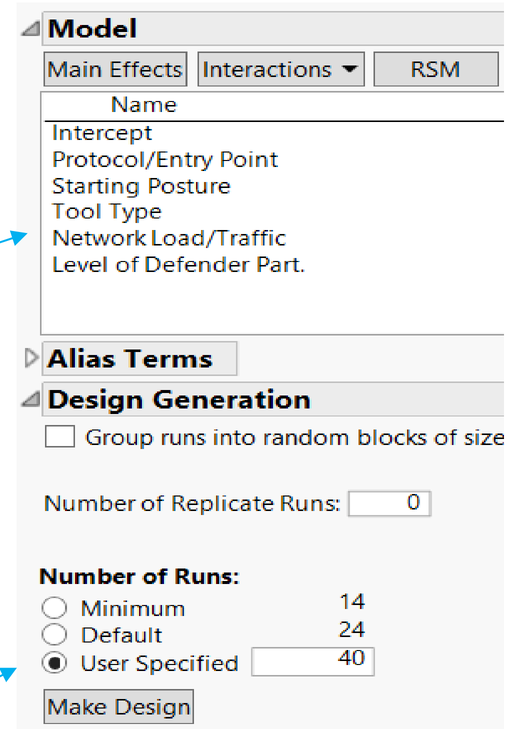
Hypothetical C2 System



Design for Cooperative Test (1 of 2)

- Create a design using the 5 varied factors presented earlier
- For the cooperative test, cover the space of all entry point/protocol combinations (an 8-level factor)

Protocol/Entry Point	Starting Posture	Tool Type	Network Load/Traffic	Level of Defender Part.
P1	Outsider	Foreign	Low	Users only
P2	Near-sider	Native	High	Users + Local Defenders
P3	Insider			Users + Local + CSSP
P4				
P5				
P6				
P7				
Maintenance Protocol				

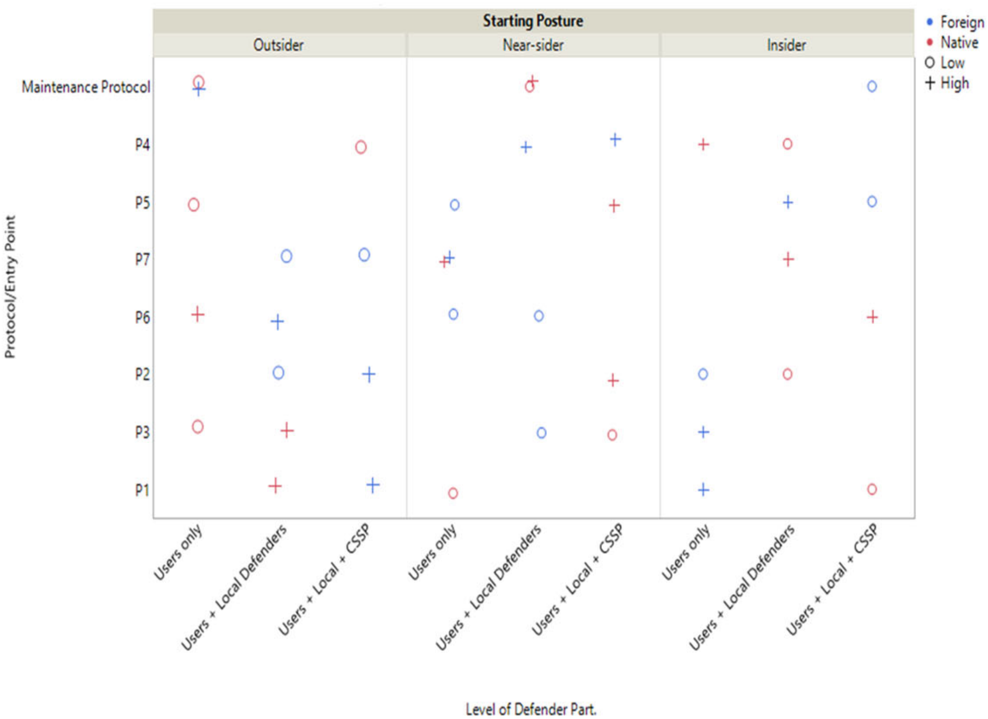


- Focus on main effects
- Can choose more than the minimum number of runs enabling additional covariates to be included in the statistical model during analysis
- Forty runs (attempted penetrations) chosen as an example, but more usually better

Design for Cooperative Test (2 of 2)

- The resulting 40 run design provides coverage (albeit sparse) of the 8 X 3 X 3 X 4 = 288 factor space

Run	Protocol/Entry Point	Starting Posture	Tool Type	Network Load/Traffic	Level of Defender Part.
1	P1	Outsider	Native	High	Users + Local Defenders
2	P6	Outsider	Foreign	High	Users + Local Defenders
3	P7	Near-sider	Native	High	Users only
4	Maintenance Protocol	Near-sider	Native	Low	Users + Local Defenders
5	P3	Outsider	Native	High	Users + Local Defenders
6	P5	Near-sider	Foreign	Low	Users only
7	P1	Insider	Foreign	High	Users only
8	P6	Outsider	Native	High	Users only
9	P3	Near-sider	Foreign	Low	Users + Local Defenders
10	P4	Near-sider	Foreign	High	Users + Local + CSSP
11	P5	Outsider	Native	Low	Users only
12	P5	Insider	Foreign	High	Users + Local Defenders
13	P1	Insider	Native	Low	Users + Local + CSSP
14	P7	Outsider	Foreign	Low	Users + Local + CSSP
15	P2	Near-sider	Native	High	Users + Local + CSSP
16	P6	Near-sider	Foreign	Low	Users only
17	P7	Near-sider	Foreign	High	Users only
18	P6	Insider	Native	High	Users + Local + CSSP
19	P3	Near-sider	Native	Low	Users + Local + CSSP
20	P1	Near-sider	Native	Low	Users only
21	P4	Outsider	Native	Low	Users + Local + CSSP
22	P5	Near-sider	Native	High	Users + Local + CSSP
23	P5	Insider	Foreign	Low	Users + Local + CSSP
24	P4	Insider	Native	Low	Users + Local Defenders
25	P7	Insider	Native	High	Users + Local Defenders
26	P4	Near-sider	Foreign	High	Users + Local Defenders
27	P3	Outsider	Native	Low	Users only
28	P6	Near-sider	Foreign	Low	Users + Local Defenders
29	Maintenance Protocol	Near-sider	Native	High	Users + Local Defenders
30	P3	Insider	Foreign	High	Users only
31	P4	Insider	Native	High	Users only
32	Maintenance Protocol	Outsider	Native	Low	Users only
33	Maintenance Protocol	Outsider	Foreign	High	Users only
34	P2	Outsider	Foreign	Low	Users + Local Defenders
35	P1	Outsider	Foreign	High	Users + Local + CSSP
36	P7	Outsider	Foreign	Low	Users + Local Defenders
37	Maintenance Protocol	Insider	Foreign	Low	Users + Local + CSSP
38	P2	Insider	Foreign	Low	Users only
39	P2	Insider	Native	Low	Users + Local Defenders
40	P2	Outsider	Foreign	High	Users + Local + CSSP



IDA | 21

Cooperative Test Measures of Merit

- The design is sufficient to provide high power to detect large differences (SNR=2) in main effects with 80% confidence
- There is necessarily some aliasing in the design, but it is mostly among higher order terms. Correlations between main effects are very low and not a concern

Term	Power
Protocol/Entry Point	0.77
Starting Posture	0.99
Level of Defender Participation	0.99
Tool Type	1.00
Network Load/Traffic	1.00

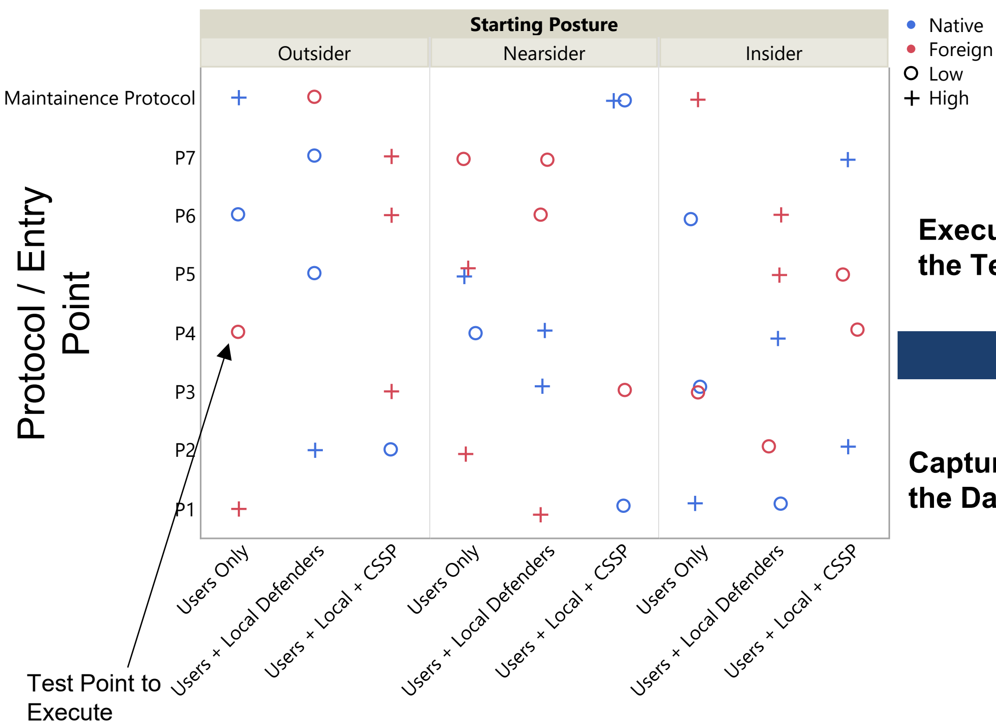
No major confounding between factors

Approved for public release; distribution is unlimited.

Analyze the data

Analysis—How it Might Work

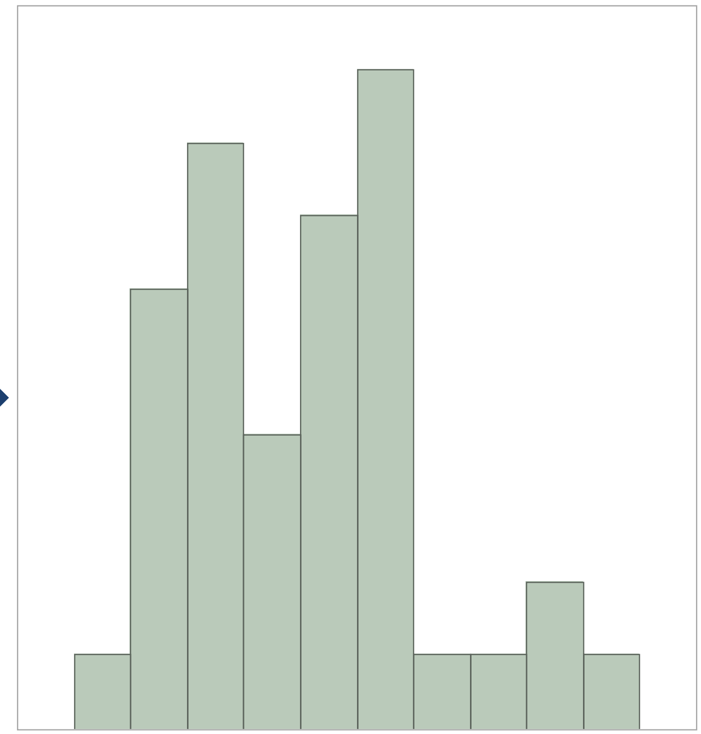
Example Analysis of a Continuous Response Variable



Level of Defender Participation

Execute the Test

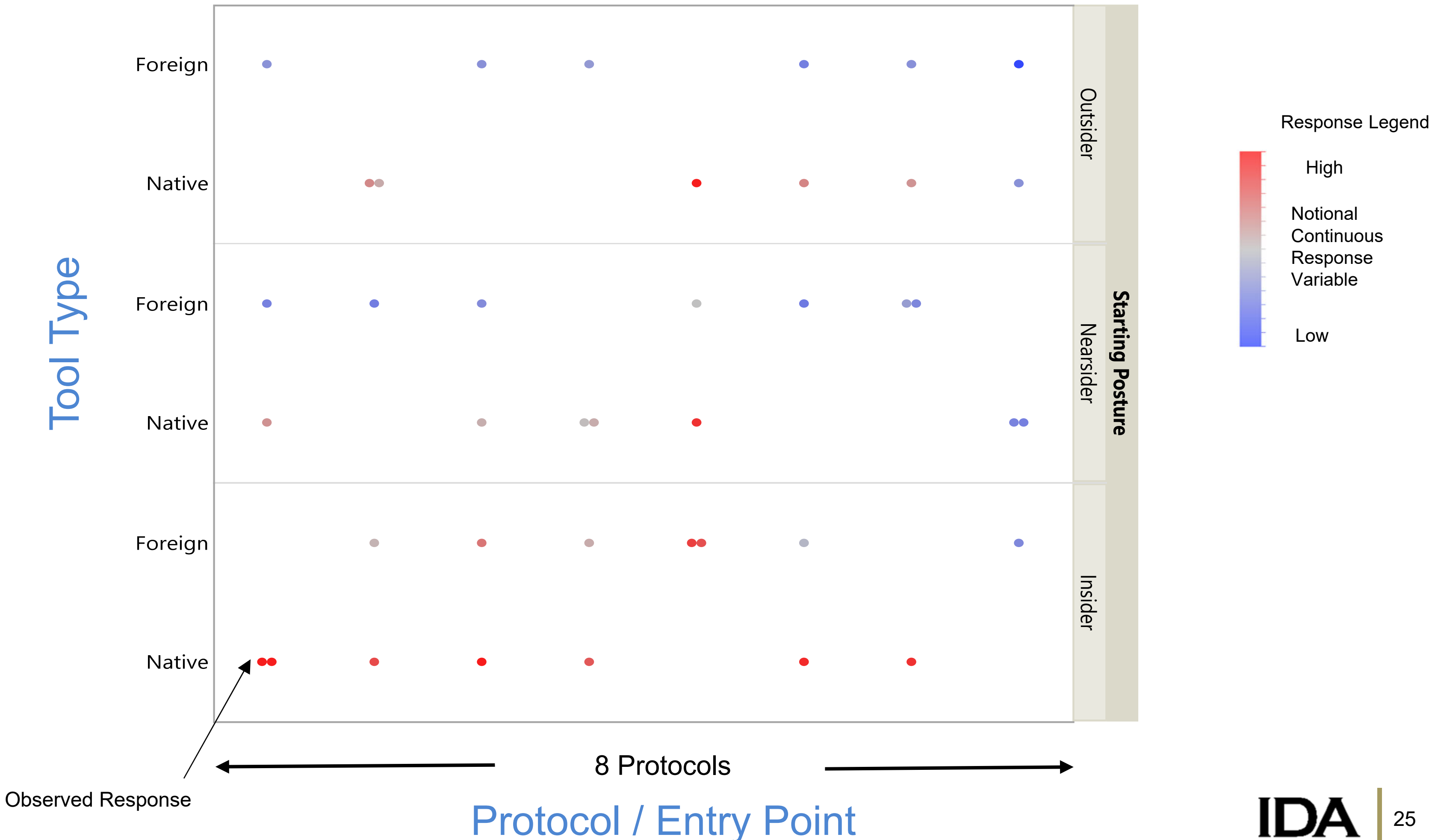
Capture the Data



Notional distribution of the continuous response variable collected from the 40 test points

Example Analysis of a Continuous Response Variable

<u>After executing the test, we can perform an exploratory analysis.</u> Observations considering three of the factors include Native Tools appear to have higher responses than Foreign Tools, as do Insider Attacks. There also appear to be some differences in responses across the Protocols.



Example Analysis of a Continuous Response Variable

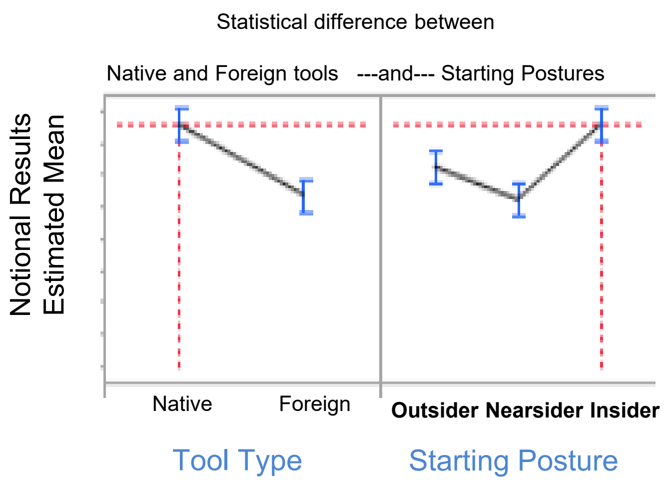
<u>Our test design enables us fitting the statistical model</u> as a function of the design factors

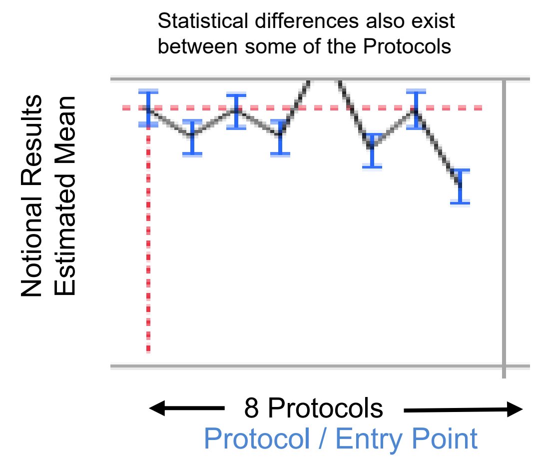
$$y = \beta_0 + \beta_1(Protocol) + \beta_2(Starting\ Posture) + \beta_3(Tool\ Type) + \beta_4(Network\ Load) + \beta_5(Defenders) + \varepsilon$$

Observed Response

From the model fit, we see that **some factors have an effect on the Notional Continuous Response Variable**

We can **<u>summarize the results using the point estimate and confidence intervals</u>**

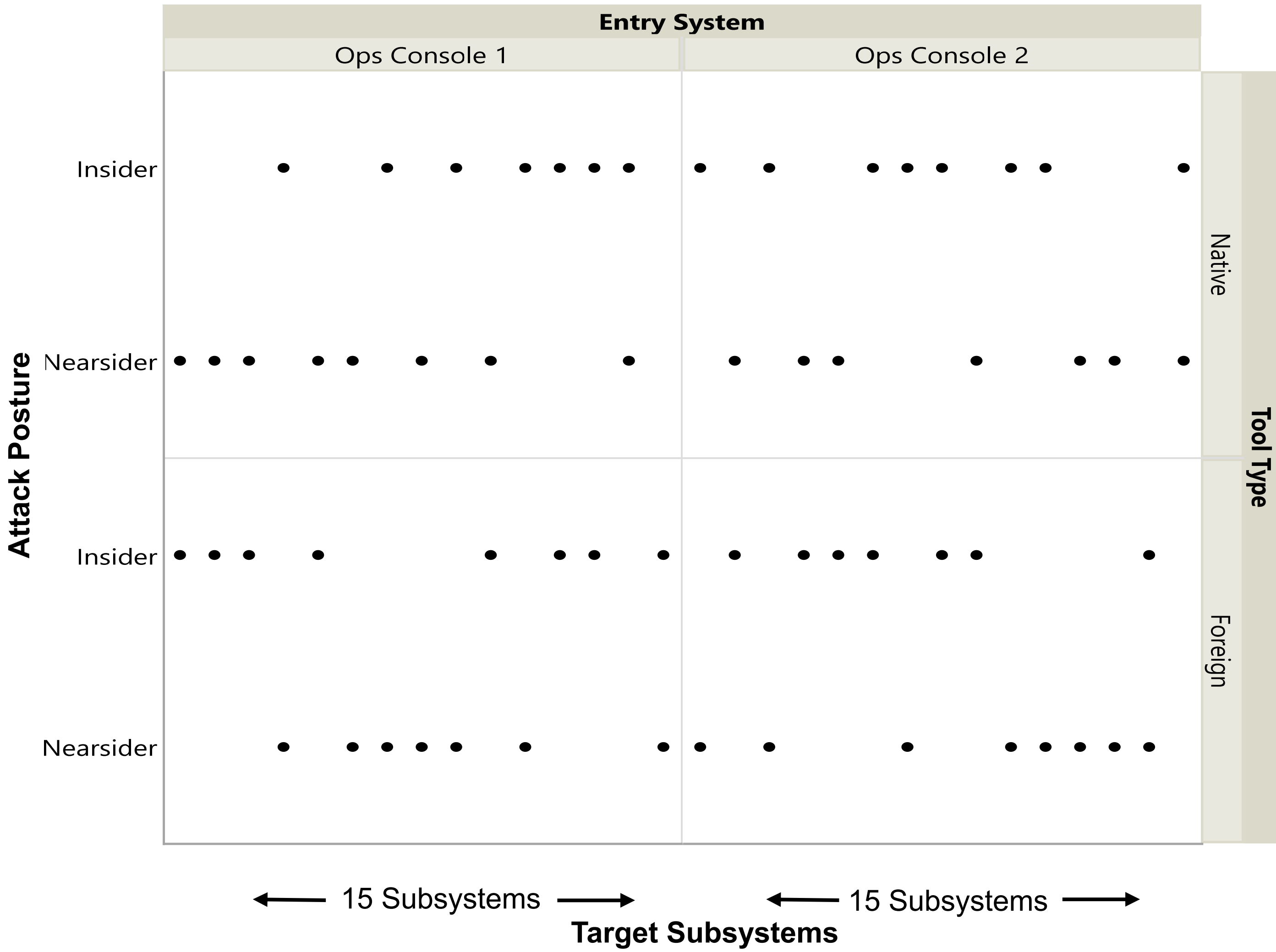




Back-up

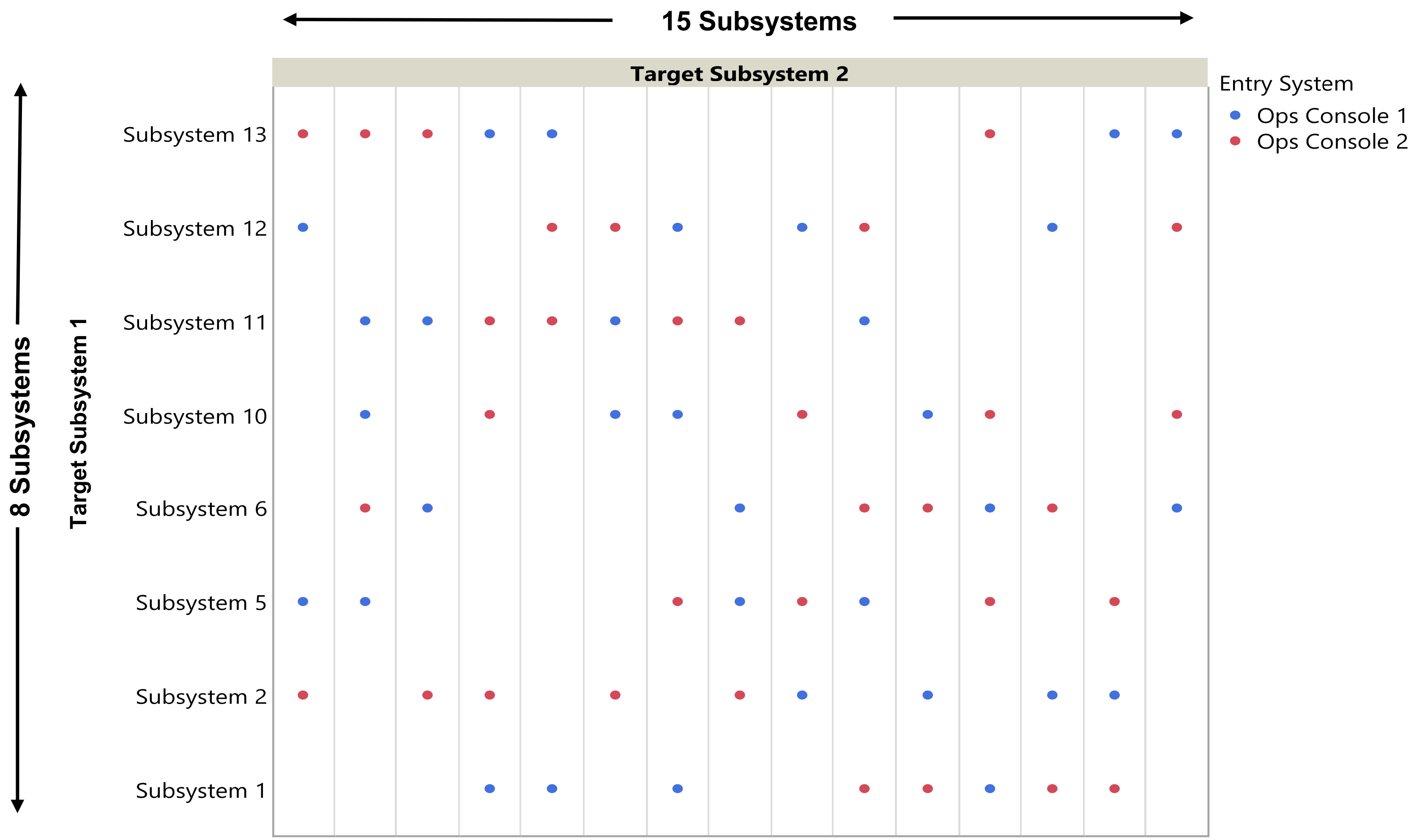
Approved for public release; distribution is unlimited.

PS Design Options for Assessment—Single Subsystem Attacks



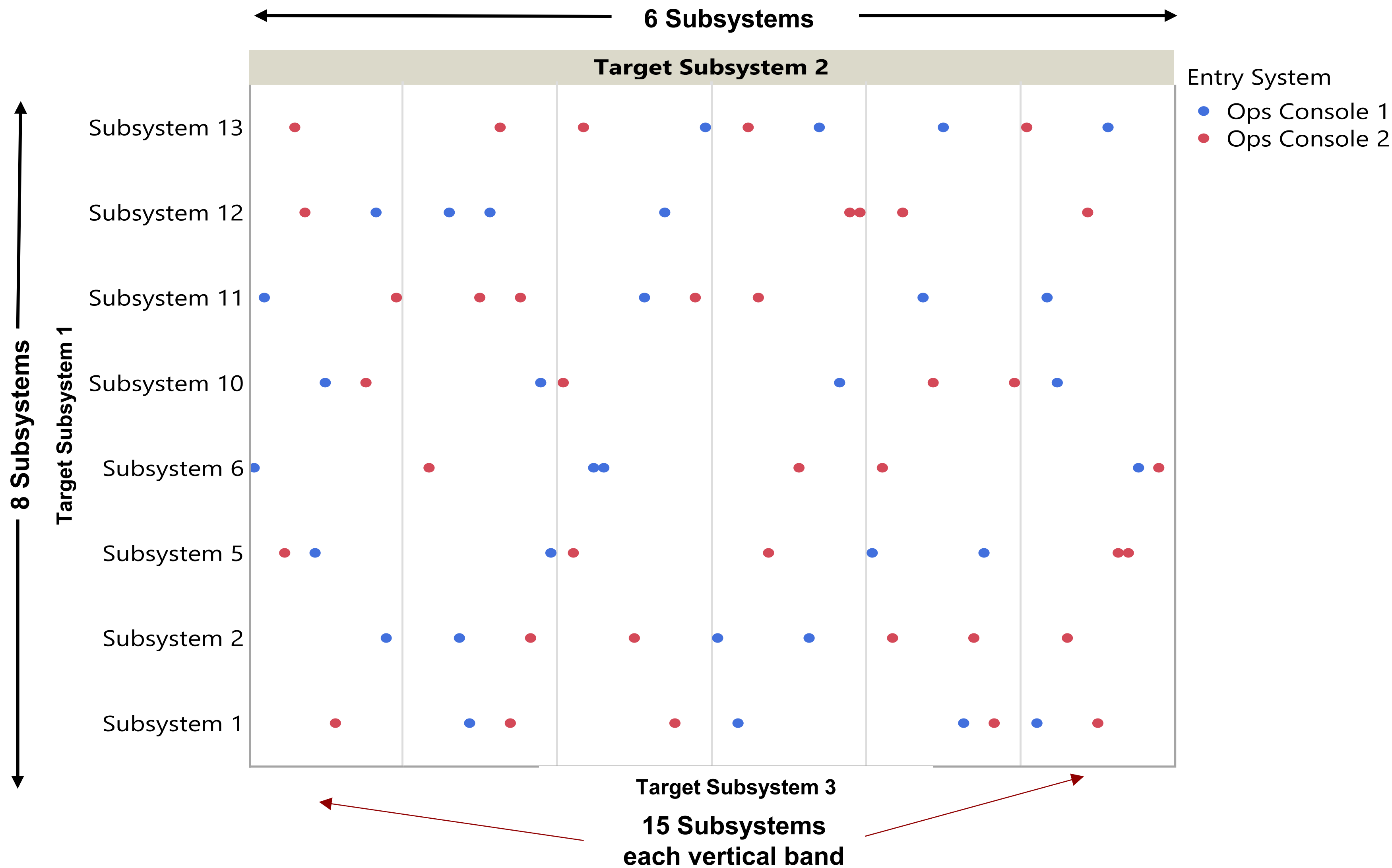
Assessing 65 potential vulnerabilities covers 120 combinations with 80% power and 80% confidence

PS Design Options for Assessment—Attacks Spanning Two Subsystems



Assessing 65 potential vulnerabilities covers 224 combinations with 80% power and 80% confidence

PS Design Options for Assessment—Attacks Spanning Three Subsystems



Assessing 70 potential vulnerabilities covers 1248 combinations with 80% power and 80% confidence

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED** START DATE	END DATE
02-2022	Final		

4. TITLE AND SUBTITLE
Thoughts on Applying Design of Experiments (DOE) to Cyber Testing

5a. CONTRACT NUMBER	**5b. GRANT NUMBER**	**5c. PROGRAM ELEMENT NUMBER**
HQ0034-19-D-0001		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
AX-1-3100		

6. AUTHOR(S)
Gilmore, James, M.; Avery, Kelly, M.; Girardi, Matthew, R.; Medlin, Rebecca, M.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER
NS D-33023
H 2022-000110

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Ms. Sarah Standard
Cybersecurity/Interoperability Technical Director
OUSD(R&E)/DTE&A

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.
Cleared for public release by the DoD Office of Prepublication Review, Case 22-S-1540

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This briefing presented at Dataworks 2022 provides examples of potential ways in which Design of Experiments (DOE) could be applied to initially scope cyber assessments and, based on the results of those assessments, subsequently design in greater detail cyber tests.

15. SUBJECT TERMS
cyber assessments; cyber testing; Design of Experiments

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
Unclassified	Unclassified	Unclassified	SAR	

19a. NAME OF RESPONSIBLE PERSON	**19b. PHONE NUMBER**
John Hong	703-845-2564

Approved for public release; distribution is unlimited.