



INSTITUTE FOR DEFENSE ANALYSES

Why are Statistical Engineers needed for Test & Evaluation?

Rebecca Medlin
Kayla Pagan-Riveria
Monica Ahrens

July 2021

Approved for Public Release.

Distribution Unlimited.

IDA Document NS-D-22722

Log: H 2021-000273/2

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," 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 Kelly Avery, John Haman, Curtis Miller, and Han Yi from the Operational Evaluation Division.

For more information:

Rebecca Medlin, Project Leader
rmedlin@ida.org • (703)-845-6731

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

Copyright Notice

© 2021 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 [Feb. 2014].

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS-D-22722

Why are Statistical Engineers needed for Test & Evaluation?

Rebecca Medlin
Kayla Pagan-Riveria
Monica Ahrens

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Executive Summary

The Department of Defense (DOD) develops and acquires some of the world's most advanced and sophisticated systems. As new technologies emerge and are incorporated into systems, OSD/DOT&E faces the challenge of ensuring that these systems undergo adequate and efficient test and evaluation (T&E) before operational use.

Statistical engineering is a collaborative, analytical approach to problem solving that integrates statistical thinking, methods, and tools with other relevant disciplines. This process provides better solutions to large, unstructured, real-world problems and supports rigorous decision-making.

This briefing, developed for a presentation at the 2021 Quality and Productivity Research Conference, includes two case studies that highlight why statistical engineers are necessary for successful T&E. These case studies center on the important theme of improving methods to integrate testing and data collection across the full system life cycle – a large, unstructured, real-world problem. Integrated testing supports efficient test execution, potentially reducing cost.

1. Case Study 1: Sequential Test and Evaluation

Design of experiments (DOE) is an approach that allows for systematic variation of controllable input factors in the process of determining the effect these factors have on an output. The T&E community has embraced the use of non-sequential DOE for planning developmental and operational testing.

In this case study, we advocate for and illustrate the use of sequential DOE as a method to support integrated effectiveness testing. We demonstrate the ability to learn the same information with less testing using sequential DOE as compared to non-sequential DOE. Sequential DOE is useful in the planning of a series of tests, leveraging the information obtained from one sequence of testing to help plan the next. Sequential DOE is a critical tool in helping testers execute testing adaptively, efficiently, and effectively.

2. Case Study 2: Incorporating Legacy Data into Test Planning

In testing and evaluating a system's reliability requirements, DOD traditionally uses a demonstration test. This is a classical hypothesis test, which uses data from only the current test to assess whether reliability requirements have been met. Demonstration tests often require a lot of testing.

In this case study, we advocate for and illustrate the use Bayesian assurance testing as a method to support integrated reliability testing. Assurance testing leverages information from various sources in an attempt to reduce the amount of testing required to meet a requirement. We demonstrate an efficient way to assess reliability with less testing than is necessary with traditional methods.



Approved for public release; distribution is unlimited.

Why are Statistical Engineers Needed for Test & Evaluation?

Rebecca Medlin

Keyla Pagan-Rivera

Monica Ahrens

Institute for Defense Analyses

2021 Quality and Productivity Research Conference

July, 28, 2021

Institute for Defense Analyses

4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

Outline

Overview of IDA, OED, and Test Science

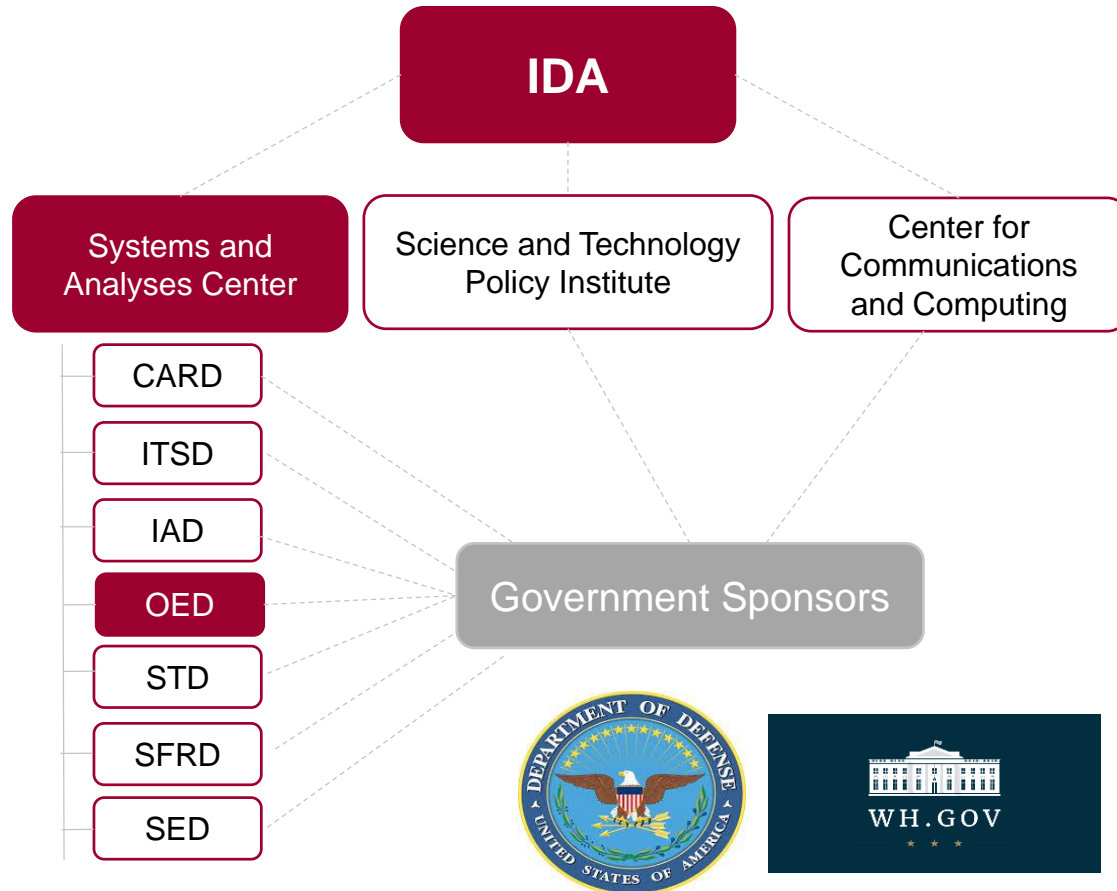
Case Study 1: Sequential Test & Evaluation

Case Study 2: Incorporating Legacy Data into Test Planning

Lessons Learned

Institute for Defense Analyses (IDA)

Answer the most challenging U.S. security and science policy questions with objective analysis leveraging extraordinary scientific, technical, and analytic expertise



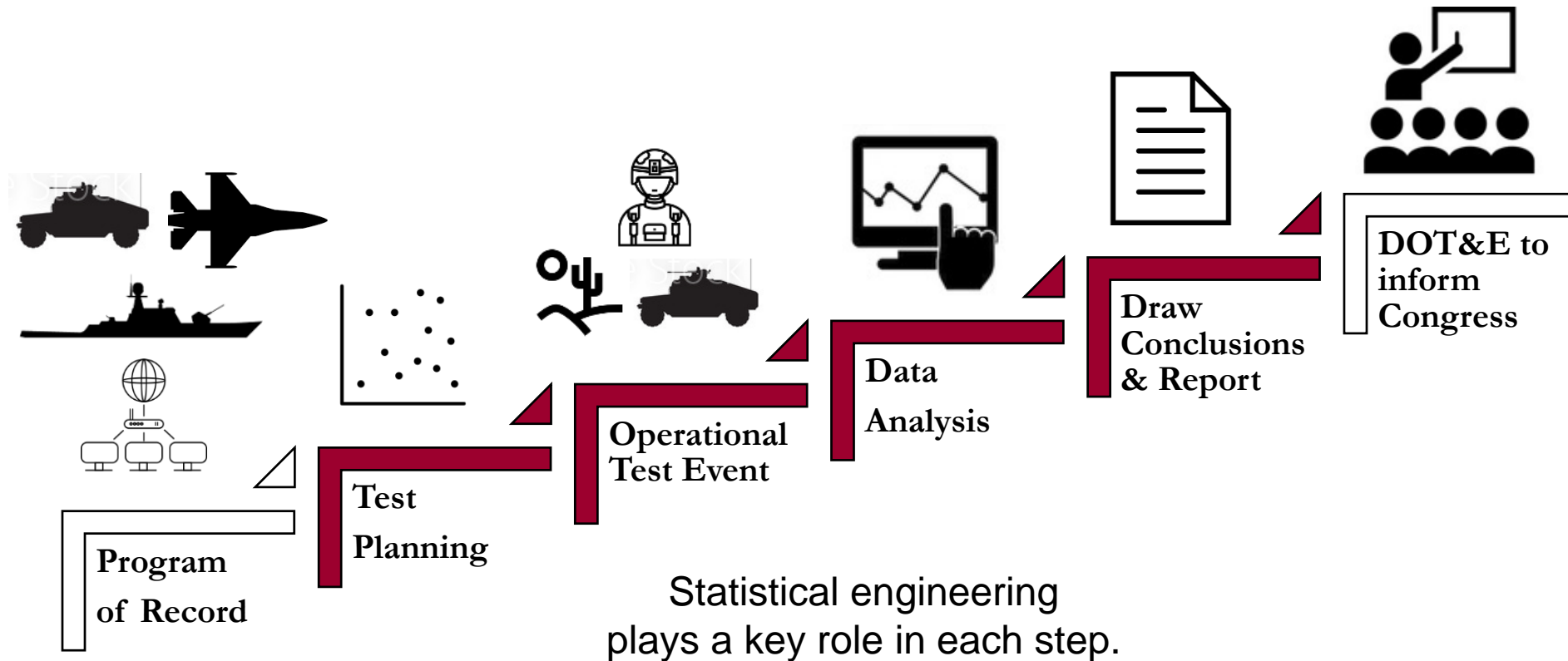
IDA, headquartered in Alexandria, Virginia, is a private nonprofit corporation that operates three Federally Funded Research and Development Centers

CARD – Cost Analysis and Research Division; ITSD – Information Technology and Systems Division; IAD – Intelligence Analyses Division; **OED – Operational Evaluation Division**; STD – Science and Technology Division; SFRD – Strategy, Forces and Resources Division; SED – System Evaluation Division

All military systems undergo operational testing before fielding or full-rate production



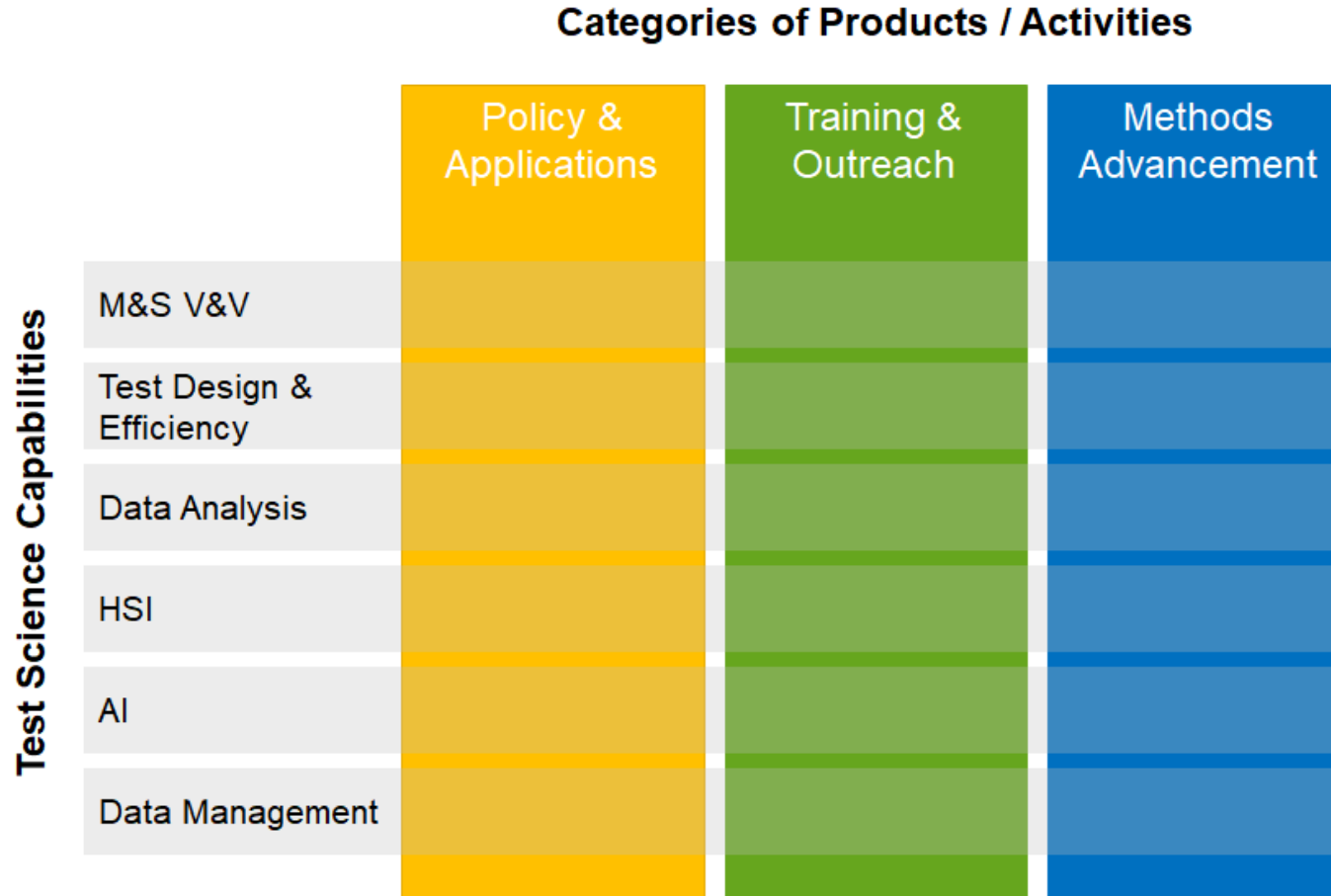
One role of the Operational Evaluation Division is to provide support to the Director, Operational Test & Evaluation (DOT&E)



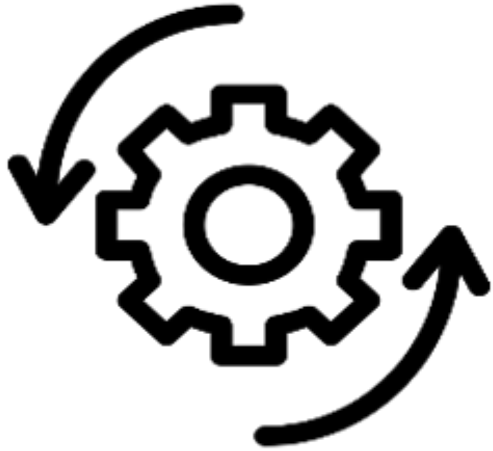
TestScience

Data . Driven . Defense

We are a diverse team of statisticians, behavioral and data scientists, mathematicians, and engineers



Method advancement includes ways to better integrate testing and data collection across the full system life cycle



Test Design &
Efficiency

Case Study 1: Sequential Test & Evaluation

Case Study 2: Incorporating Legacy Data into
Test Planning

How does Statistical Engineering factor in?

Statistical Engineering: a collaborative, analytical approach that integrates statistical thinking, methods, and tools with other relevant disciplines to generate better solutions to large, unstructured, real-world problems sustainably and supports rigorous decision-making.

Case Study 1: Collecting operational test data using sequential planning



Soldiers Emplacing the AN/TPQ-53 (Q-53) Counterfire Radar

Figure Source: Freeman, Laura J., et al. "Testing defense systems." *Analytic Methods in Systems and Software Testing* (2018): 441.

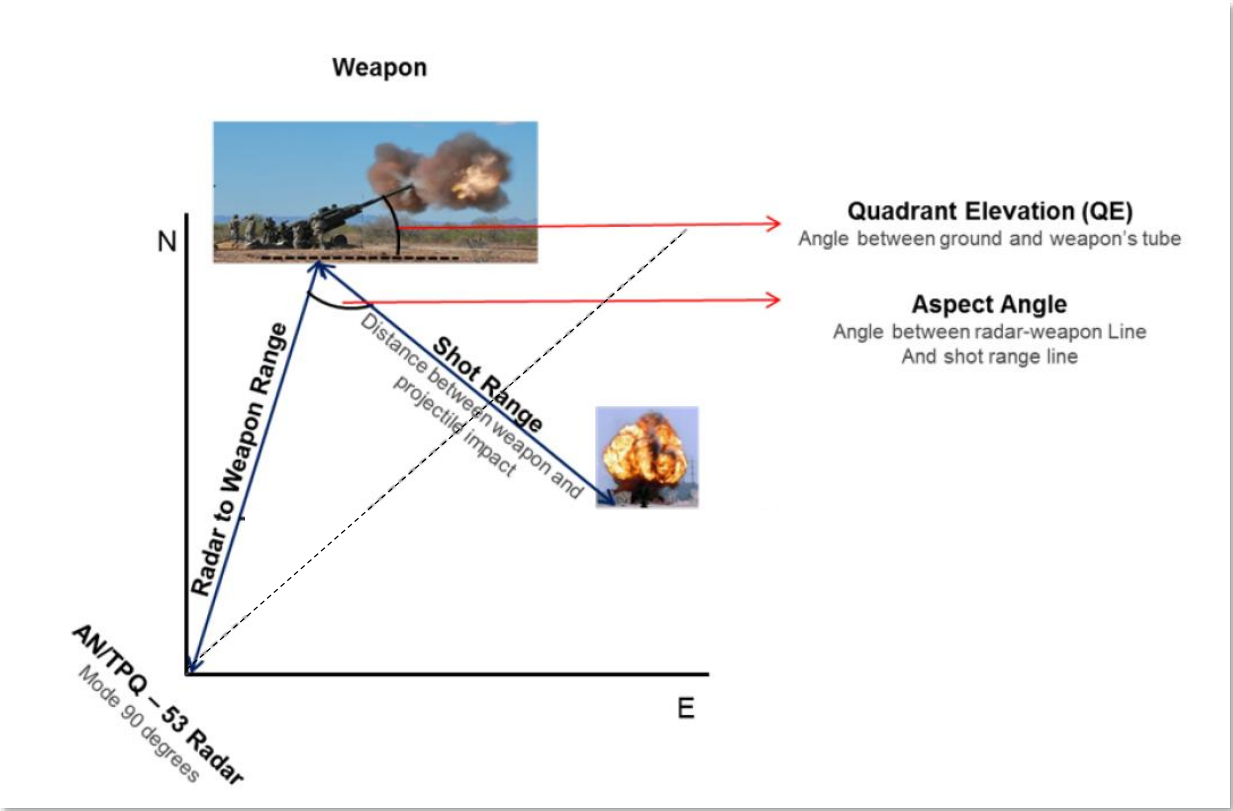
Approved for public release; distribution is unlimited.

Can the Q-53 detect shots with high probability?

Can the Q-53 locate the origin of a shot with sufficient accuracy to provide an actionable counterfire location?

The performance of combat systems is likely affected by a wide variety of operating conditions, threat types, system operating modes, and other physical factors

Example Fire Mission



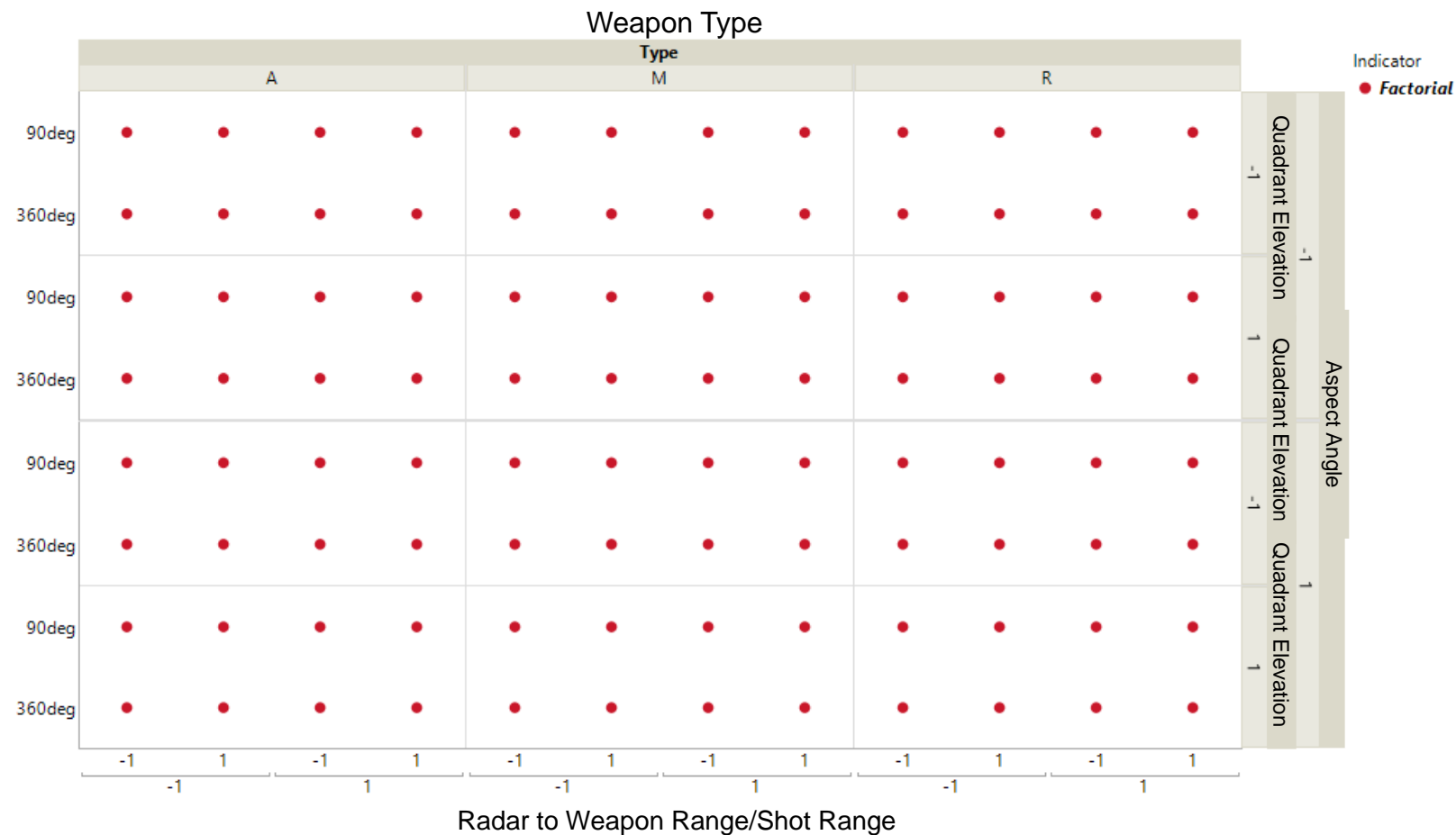
Design Factor	Level
Quadrant Elevation (QE)	Low, High
Aspect Angle (AA)	Low, High
Munition Type	Mortar, Rockets, Artillery
Shot Range (SR)	Low, High
Radar Operating Mode	90 deg, 360 deg
Radar to Weapon Range (RWR)	Low, High

Figure Source: Freeman, Laura J., et al. "Testing defense systems." Analytic Methods in Systems and Software Testing (2018): 441.

Approved for public release; distribution is unlimited.

Planning a test: “traditional” DOE approach

- D-optimal design – 184 test points
- Characterize
 - Requires research-specified model



Planning a test: sequential DOE approach

Test Phase 1:

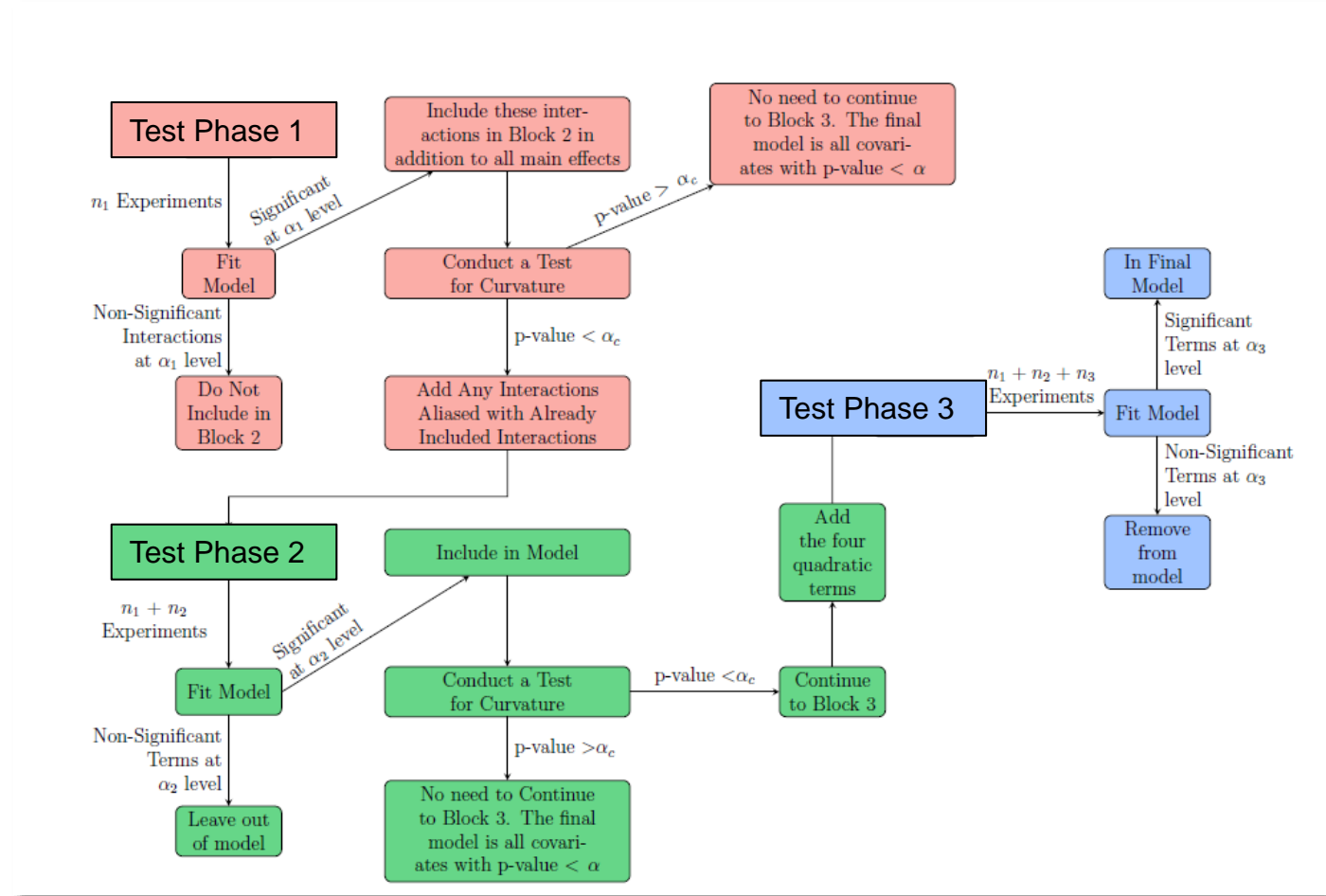
- Select test points to screen for main effects and some interactions
- Collect data
- Fit model
- Test for curvature

Test Phase 2:

- Select test points to fit model for remaining interactions
- Collect data
- Fit model
- Test for curvature

Test Phase 3:

- Select test points to fit model with quadratic effects
- Collect data
- Fit final model



Does it work?

True Model
Model 1: Main Effects + Interactions + Quadratic
Model 2: Main Effects + Interactions
Model 3: Main Effects

Simulation Settings	Design	N	σ	Phase 1 α_1	Phase 2 α_2	Phase 3 α_3	α_c
	D-optimal	1,000 data sets	1	.30	.15	.15	.20
			3				
			4				

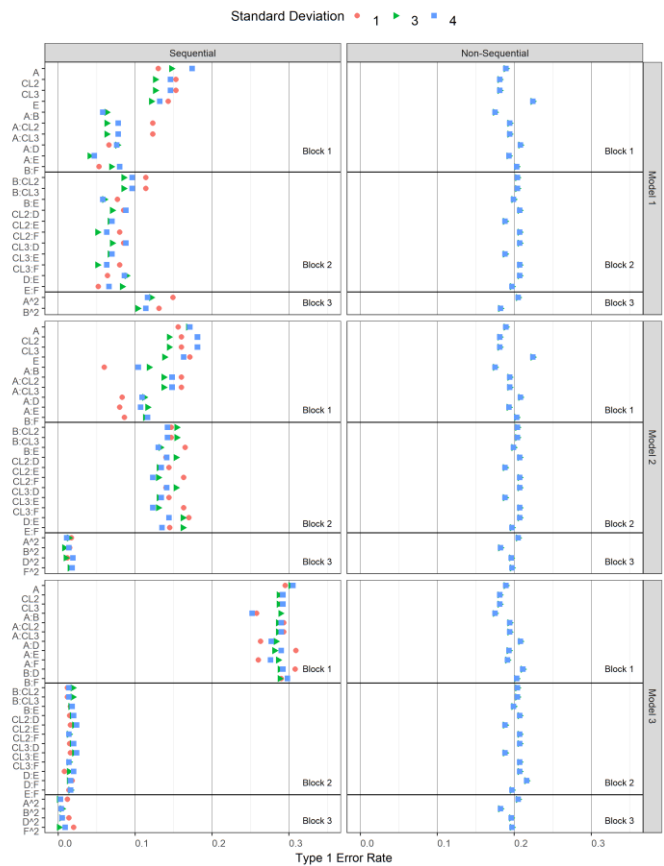
* Models are notional and not based on true system performance

Method performs as expected

Approved for public release; distribution is unlimited.

Does the method control the Type I and Type II error rates?

Does the simulation stop at the test phase we expect it to?

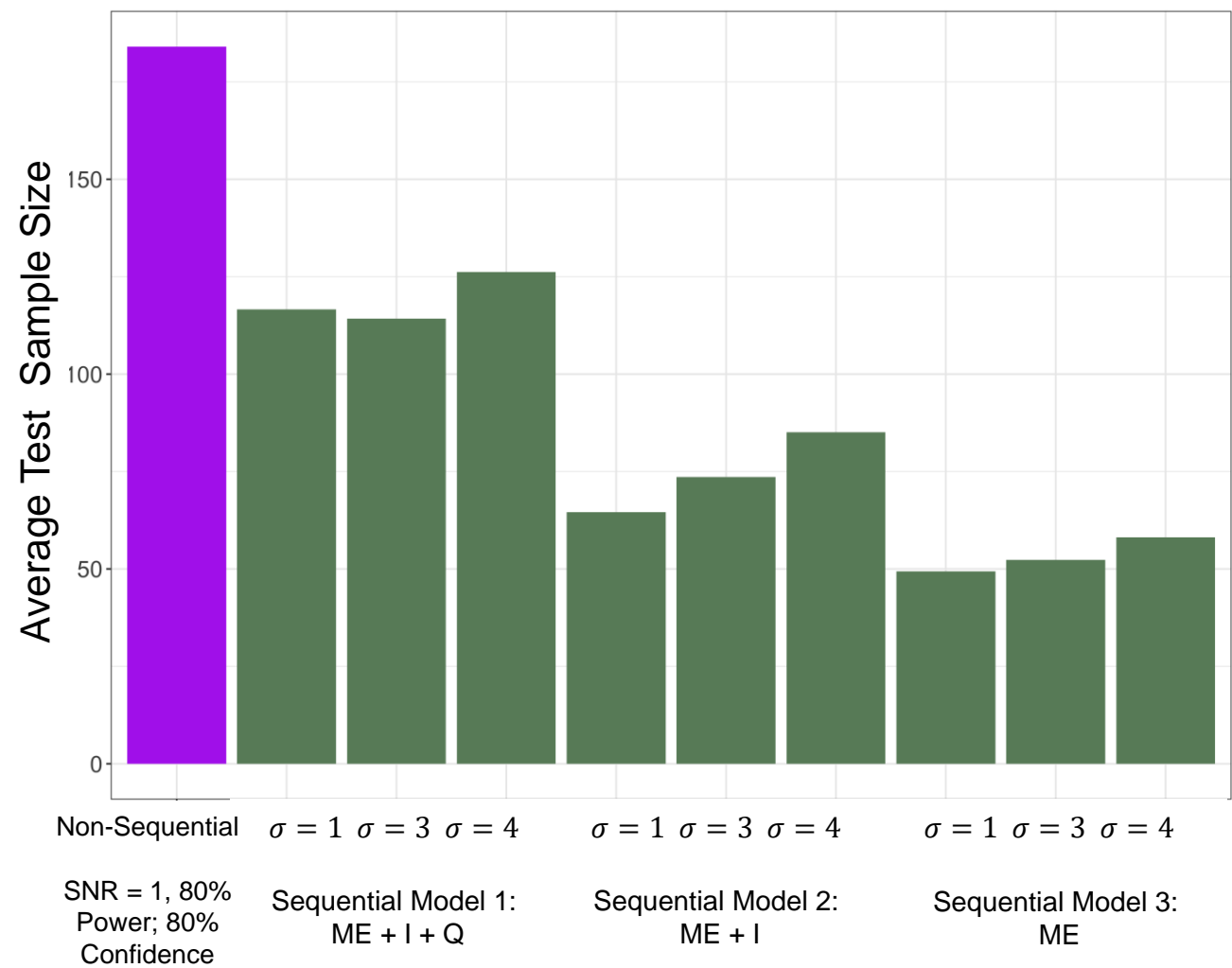


	Model 1			Model 2			Model 3		
σ \ Phase	1	2	3	1	2	3	1	2	3
1	0%	9.1%	91%	0%	91%	9.4%	90%	1.8%	7.8%
3	0.3%	22%	78%	1.7%	91%	7.2%	91%	5.9%	2.7%
4	1.8%	23%	76%	7.1%	85%	8.3%	90%	6.8%	3.7%

Approved for public release; distribution is unlimited.

Opportunities for substantial cost savings!

Approved for public release; distribution is unlimited.



Approved for public release; distribution is unlimited.

Challenges

Within a
Test Event

Is real-time test planning possible?

Is real-time data analysis possible?

How might the approach impact the test schedule and logistics?

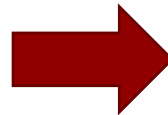
Between
Test Events

Does the amount of time between test events matter?

What if the system experiences modification between events?

Does the maturity of the system matter?

When should we try to use
sequential methods?



Develop a roadmap for the applied use
of sequential analysis in T&E.

Demonstrate method using case
studies.

Case Study 2: Incorporating legacy data into reliability test planning



Paladin Integrated Management (PIM) self-propelled howitzer

How many hours of reliability testing do we need to determine whether the system meets its requirement?

We can incorporate legacy data into a reliability test plan

We need to determine the test size (hours) for operational testing and the number of failures allowed before we declare whether the PIM howitzer meets or does not meet its 64-hour MTBF reliability threshold. There are two errors we can make:

- Test is passed when the vehicle reliability is actually below threshold (consumer risk)
- Test is failed when the reliability is actually above threshold (producer risk)

Traditional DoD demonstration tests

- ✓ Fix the risk of passing the test given that the true reliability is less than the threshold (consumer risk)
- ✓ Effectively ignore the risk of failing the test given that the true reliability is greater than a threshold (producer risk)
- ✓ Find the minimum test size around a fixed number of failures
- ✓ Often require an exorbitant amount of testing

Bayesian assurance testing¹

- ✓ Fixes the risk that the true reliability is less than the threshold given that the test is passed (consumer risk)
- ✓ Fixes the risk that the true reliability is greater than a threshold given that the test is failed (producer risk)
- ✓ Finds the minimum test size around a fixed number of failures
- ✓ By incorporating all available information, typically requires less testing

1. If the information we are looking to incorporate is poor – say, for example, developmental testing suggests poor system reliability – then incorporating this information will buy us no advantages and will not shorten the length of a test.

What is the maximum number of failures, c , permitted for a successful test of length T ?

Traditional Risk Criteria

$$\begin{aligned} & \text{Consumer's Risk} \\ &= P(\text{Test is Passed} | \lambda = T/MTBF_{Req}) \\ &= P(y \leq c | \lambda) = \sum_{y=0}^c \frac{\lambda^y e^{-\lambda}}{y!} \leq \alpha \end{aligned}$$

We choose c to be the largest non-negative integer that satisfies this inequality.

Bayesian Posterior Risk Criteria

$$\begin{aligned} & \text{Consumer's Risk} = P(\lambda \geq \lambda_1 | \text{Test is Passed}, x) \\ & \approx \frac{\sum_{j=1}^N \left[\sum_{y=0}^c \frac{(\lambda^{(j)} T)^y \exp(-\lambda^{(j)} T)}{y!} \right] I(\lambda^{(j)} \geq \lambda_1)}{\sum_{j=1}^N \left[\sum_{y=0}^c \frac{(\lambda^{(j)} T)^y \exp(-\lambda^{(j)} T)}{y!} \right]} \leq \alpha \end{aligned}$$

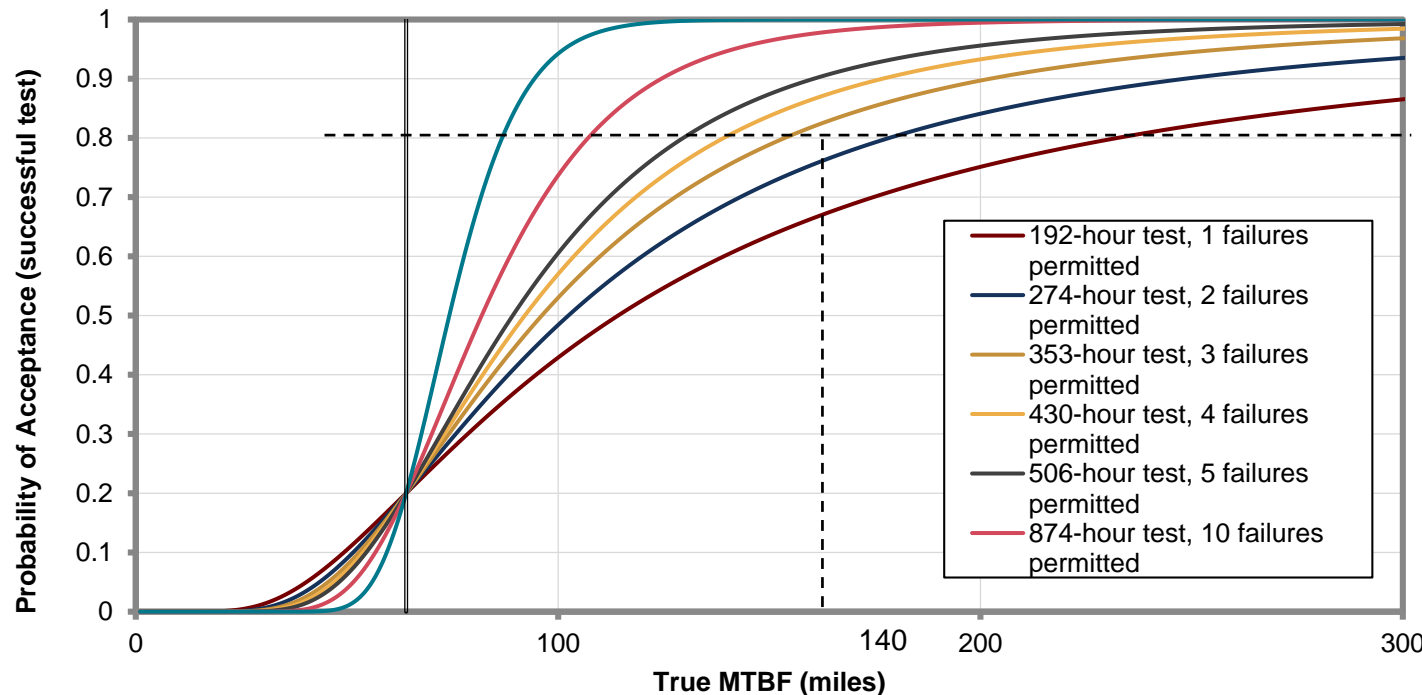
$$\begin{aligned} & \text{Producer's Risk} = P(\lambda \leq \lambda_0 | \text{Test is Failed}, x) \\ & \approx \frac{\sum_{j=1}^N \left[1 - \sum_{y=0}^c \frac{(\lambda^{(j)} T)^y \exp(-\lambda^{(j)} T)}{y!} \right] I(\lambda^{(j)} \leq \lambda_0)}{\sum_{j=1}^N \left[1 - \sum_{y=0}^c \frac{(\lambda^{(j)} T)^y \exp(-\lambda^{(j)} T)}{y!} \right]} \leq \beta \end{aligned}$$

Where x is available data, $\lambda^{(j)}$ are the posterior predictive draws, and $\lambda_0 < \lambda_1$

Evaluating a mission-based threshold with traditional methods requires a very long test

Traditional Operating Characteristic Curves are based on demonstrating 64-hour MTBF.

- Using optimistic assumptions about true **howitzer** reliability, a minimum of **430 hours** of operational testing are required to evaluate **the howitzer's** reliability with $\alpha = 0.2$ and probability of acceptance = 0.80



We need to find the combination of test length, T , and maximum allowed failures, c , that satisfies

$$P(y \leq c | \lambda) = \sum_{y=0}^c \frac{\left(\frac{T}{64}\right)^y \exp\left(-\left(\frac{T}{64}\right)\right)}{y!} \leq .2$$

We can use available data to construct our test plan

Likelihood Distribution

T = 400 hours of testing
N = 2 failures

Available data, x, from
previous test event

$$P(x|\lambda) \sim \text{Exponential}(\lambda)$$

Prior Distribution Non-Informative Prior

$$P(\lambda) \sim \text{Gamma}(\alpha = .001, \beta = .001)$$

Posterior Distribution

$$P(\lambda|x) \sim \text{Gamma}(\alpha' = \alpha + N, \beta' = \beta + T)$$

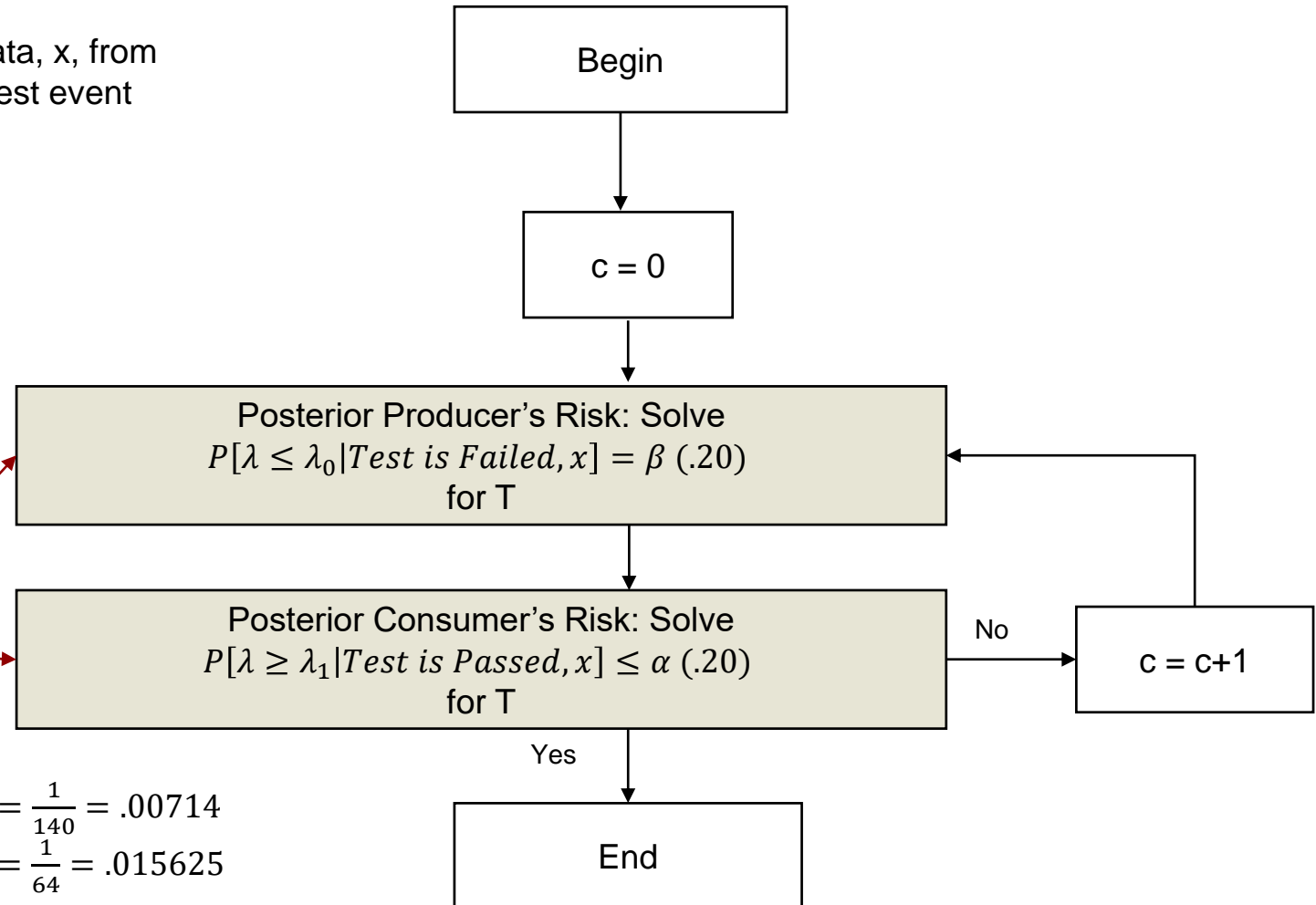
Acceptable and rejectable levels

Inputs based on growth curve test objective and requirement

$$\lambda_0 = \frac{1}{140} = .00714$$

$$\lambda_1 = \frac{1}{64} = .015625$$

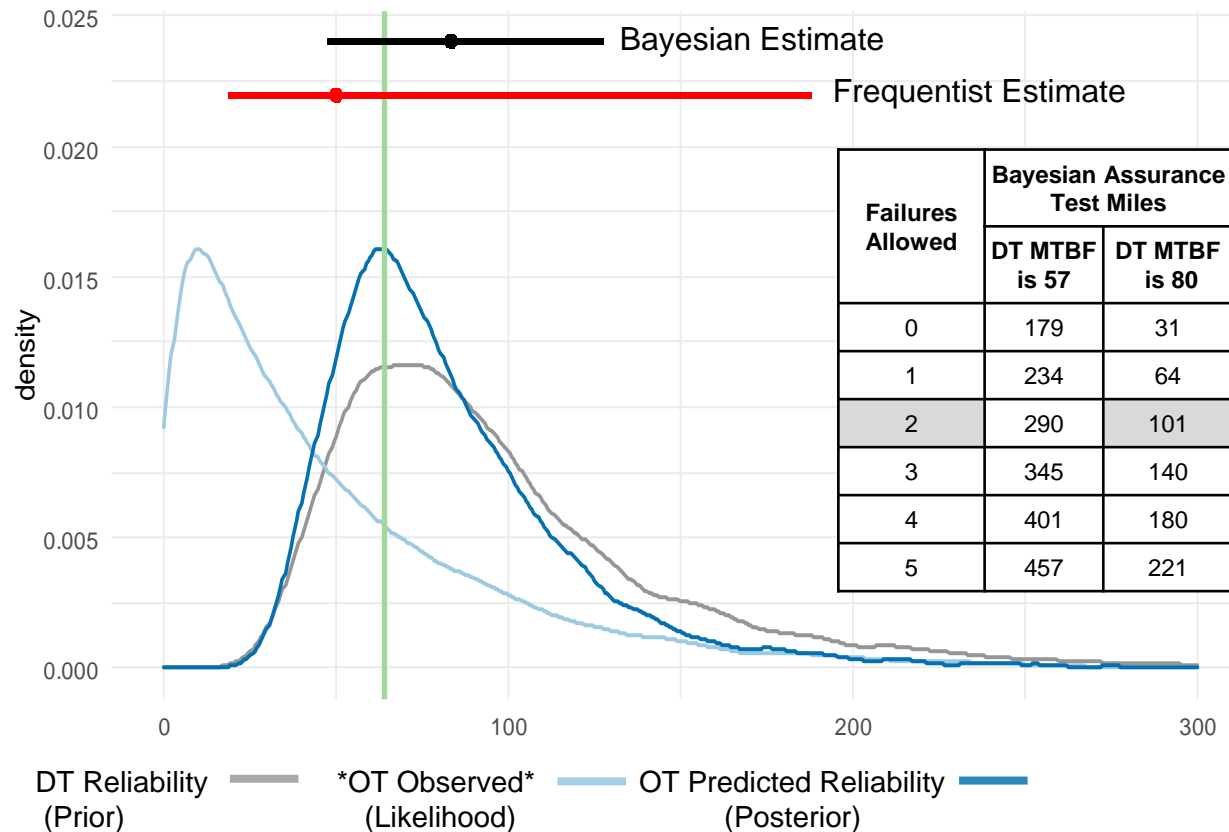
Bayesian Test Plan Algorithm



Note: The expression *Test is failed* means that the number of observed failures is larger than the maximum number of allowed failures, that is $y > c$. Similarly, *Test is Passed* means that $y \leq c$.

Bayesian assurance testing offers a more powerful and efficient way to assess reliability compared to traditional methods

Example: OT analysis for 101 test hours with 2 observed failures



When **DT and OT testing** are carried out under **similar conditions**, incorporating DT data into the OT reliability assessment may be reasonable.

- Bayesian assurance methods provide a structured way to leverage DT
- Similar to Operational Characteristic Curve assumptions, chart on the left assumes the howitzer attains a true reliability of 80 hours
- Permits a more powerful assessment of reliability with fewer miles compared to traditional methods that consider only OT data

Challenges

When is a Bayesian Assurance Test appropriate?

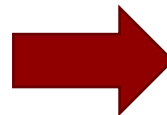
Does the analysis you plan to conduct after the test is executed matter?

How do we convince the OTA that this might be a reasonable path forward?

What data will you need to collect?

How much past test data should we use/include?

When should we try to use
Assurance Testing methods?



Develop a best practices guide

Build an R library

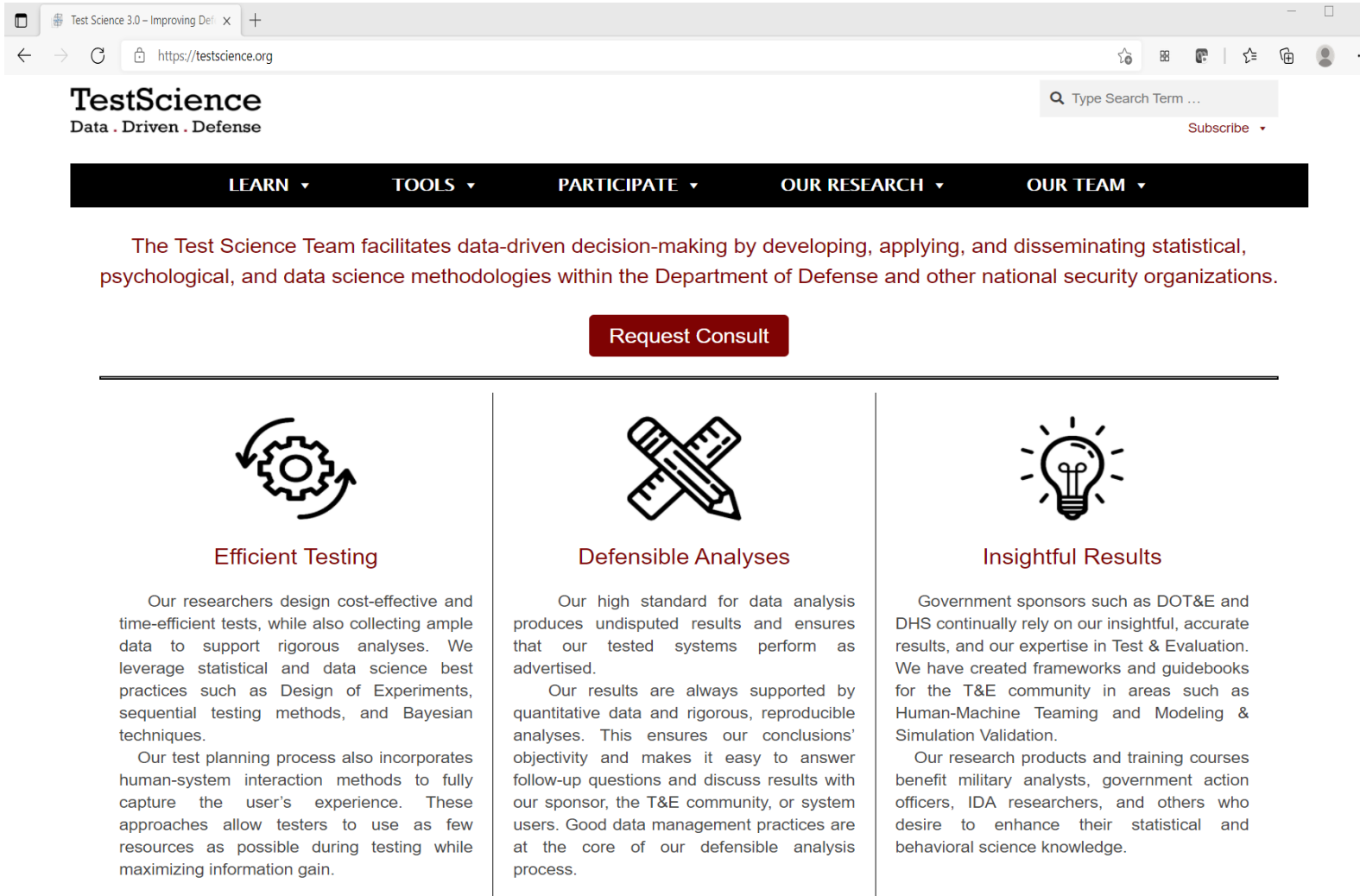
Illustrate using case studies

Lessons Learned

1. Successful test planning requires frequent communication and collaboration
2. It's okay to compromise
3. Most people do not speak statistics. Learn to speak their language!
4. Observe the test event; test execution rarely goes as planned
5. The 80% solution is usually good enough
6. Don't be afraid to ask questions
7. Statistics isn't everything

The Test Science team strives to ensure that our methodologies and best practices are widely disseminated and accessible to everyone in the T&E community

Approved for public release; distribution is unlimited.




The screenshot shows the Test Science website homepage. At the top, the Test Science logo is displayed with the tagline "Data . Driven . Defense". A navigation bar contains links for LEARN, TOOLS, PARTICIPATE, OUR RESEARCH, and OUR TEAM. A search bar and a "Subscribe" button are also present. Below the navigation bar, a paragraph describes the team's mission: "The Test Science Team facilitates data-driven decision-making by developing, applying, and disseminating statistical, psychological, and data science methodologies within the Department of Defense and other national security organizations." A red "Request Consult" button is centered below this text. The main content area is divided into three columns, each with an icon and a title: "Efficient Testing" (gears icon), "Defensible Analyses" (ruler and pencil icon), and "Insightful Results" (lightbulb icon). Each column contains a paragraph of text describing the team's work in that area.

TestScience
Data . Driven . Defense

LEARN ▾ TOOLS ▾ PARTICIPATE ▾ OUR RESEARCH ▾ OUR TEAM ▾

The Test Science Team facilitates data-driven decision-making by developing, applying, and disseminating statistical, psychological, and data science methodologies within the Department of Defense and other national security organizations.


[Request Consult](#)



Efficient Testing

Our researchers design cost-effective and time-efficient tests, while also collecting ample data to support rigorous analyses. We leverage statistical and data science best practices such as Design of Experiments, sequential testing methods, and Bayesian techniques.


Our test planning process also incorporates human-system interaction methods to fully capture the user's experience. These approaches allow testers to use as few resources as possible during testing while maximizing information gain.



Defensible Analyses

Our high standard for data analysis produces undisputed results and ensures that our tested systems perform as advertised.

Our results are always supported by quantitative data and rigorous, reproducible analyses. This ensures our conclusions' objectivity and makes it easy to answer follow-up questions and discuss results with our sponsor, the T&E community, or system users. Good data management practices are at the core of our defensible analysis process.



Insightful Results

Government sponsors such as DOT&E and DHS continually rely on our insightful, accurate results, and our expertise in Test & Evaluation. We have created frameworks and guidebooks for the T&E community in areas such as Human-Machine Teaming and Modeling & Simulation Validation.

Our research products and training courses benefit military analysts, government action officers, IDA researchers, and others who desire to enhance their statistical and behavioral science knowledge.

Join Us at DATAWorks 2022! A Statistical Engineering Workshop in Disguise

Approved for public release; distribution is unlimited.



DATAWorks

Defense and Aerospace Test and Analysis (DATA) Workshop

SAVE THE DATE

APRIL 26-28

INSTITUTE FOR DEFENSE ANALYSES, ALEXANDRIA, VA

Stay up to date at dataworks.testscience.org



Organized for Defense and Aerospace Communities by



No endorsement of non-NASA and non-DOT&E organizations intended.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 07-2021		2. REPORT TYPE IDA Publication		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Why are Statistical Engineers needed for Test & Evaluation?				5a. CONTRACT NUMBER Separate Contract	
				5b. GRANT NUMBER _____	
				5c. PROGRAM ELEMENT NUMBER _____	
6. AUTHOR(S) Rebecca M. Medlin (OED); Monica L. Ahrens (OED); Keyla Pagan-Rivera (OED);				5d. PROJECT NUMBER	
				5e. TASK NUMBER C9082	
				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-22722 H 2021-000239	
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 The Department of Defense (DoD) develops and acquires some of the world's most advanced and sophisticated systems. As new technologies emerge and are incorporated into systems, OSD/DOT&E faces the challenge of ensuring that these systems undergo adequate and efficient test and evaluation (T&E) prior to operational use. Statistical engineering is a collaborative, analytical approach to problem solving that integrates statistical thinking, methods, and tools with other relevant disciplines. The statistical engineering process provides better solutions to large, unstructured, real-world problems and supports rigorous decision-making. In this talk, we provide two case study examples related to looking at ways to improve approaches to integrate testing and data collection across the full system lifecycle. These case studies highlight why we believe statistical engineers are necessary for successful T&E.					
15. SUBJECT TERMS Assurance Testing, Design of Experiments, Integrated Testing and Evaluation					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			Rebecca Medlin (OED)
			Unlimited	35	19b. TELEPHONE NUMBER (include area code) (703) 845-6731