INSTITUTE FOR DEFENSE ANALYSES

Improving Operational Test Efficiency: Sequential Methods in Operational Testing

Rebecca M. Medlin, Project Leader

Dr. Keyla Pagan-Rivera

August 2023

Public release approved. Distribution is unlimited.

IDA Document NS 3000050

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

IDA

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, "CRP Statistics Work Group". 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 Dr. V. Bram Lillard and consisted of Dr. John T. Haman from the Operational Evaluation Division.

For more information:
Dr. Rebecca M. Medlin, Project Leader
rmedlin@ida.org • (703) 845-6731

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS 3000050

Improving Operational Test Efficiency: Sequential Methods in Operational Testing

Rebecca M. Medlin, Project Leader

Dr. Keyla Pagan-Rivera

Executive Summary

Sequential methods is a type of statistical evaluation in which the number, pattern, or composition of the data is not determined at the start of the investigation, but instead depends on the information acquired during the investigation. Although sequential methods originated in ballistics testing for the Department of Defense (DoD), it is underutilized in the DoD. Expanding the use of sequential methods may save money and reduce test time.

In this presentation, we introduce sequential methods, describe its potential uses in operational test and evaluation (OT&E), and present a method for applying it to the test and evaluation of defense systems. We evaluate the proposed method by performing simulation studies and applying the method to a case study. Additionally, we discuss some of the challenges we might encounter when using sequential analysis in OT&E.

Improving Operational Test Efficiency: Sequential Methods in Operational Testing

Keyla Pagán-Rivera

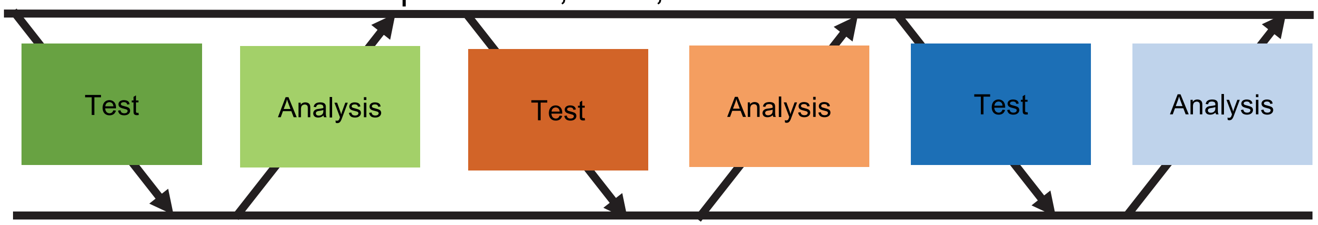
August 2023

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

Outline

- Introduction
- Motivating Example
- Sequential Probability Ratio Test
- Sequential Design Of Experiments
- Challenges

Use information gained at each stage of experimentation when considering how to continue the investigation



Examples of

Sequential Testing*	$H_0 \; vs \; H_a$ $S_1 = log\Delta_1$	$H_0 \; vs \; H_a$ $S_2 = S_1 + log\Delta_2$	$H_0 \; vs \; H_a$ $S_3 = S_2 + log\Delta_3$	$H_0 \; vs \; H_a$ $S_n = S_{n-1} + log\Delta_n$
Sequential Estimation	$\widehat{\mu_1}, \widehat{\sigma_1}$	$\hat{\mu}_2, \widehat{\sigma_2}$	$\widehat{\mu_3}, \widehat{\sigma_3}$	$\widehat{\mu_n}, \widehat{\sigma_n}$
Sequential Design				

* H_0 is the null hypothesis; $\mathrm{H_a}$ is the alternative hypothesis; S is the test statistic; $log\Delta$ is the log-likelihood ratio; $\hat{\mu}$ is the recursive estimate of the mean; $\hat{\sigma}$ is the estimate of the recursive standard deviation

Sequential methods support integrated testing

DT (experimental testing)	Integrated Testing*	IOT&E (evaluation testing)

Statistical Methods to support Test Efficiency:

Sequential Methods
(Frequentist and Bayesian)

Requirements for Successful Implementation:

- Collaborative Planning
- Early Planning
- Shared Data
- “How to” Trainings
- Applications

DT – developmental test; IOT&E – initial operational test and evaluation

*“Integrated testing requires the collaborative planning and execution of test phases and events to provide shared data in support of independent analysis, evaluation, and reporting by all stakeholders.” – Department of Defense Instruction 5000.89, “Test and Evaluation,” November 19, 2020.

Sequential methods have been around since World War II and are still being used

SRG – Statistical Research Group

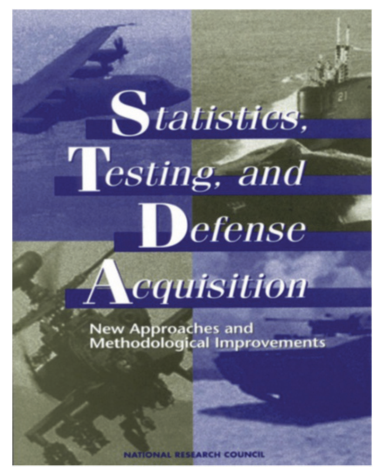
The NAS report (1998) included some of the benefits and challenges of using sequential methods in T&E

"In the application of operational testing, wide use of sequential testing could result in substantial savings of test dollars and a decrease in test time."

Why hasn't sequential testing become the norm in defense testing?

"The need to obtain expedited analysis of test results and the scheduling of soldiers and test facilities makes sequential designs difficult to apply in some circumstances... The panel is concerned that the technical demands of these tests contributes to their infrequent use. "

But not trying would be unfortunate given the advantage of reducing test cost and time.



NAS – National Academy of Sciences

Sequential methods could be applied to operational testing and evaluation

- 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?

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.

We would like to know if the radar's failure rate* is higher than what is expected

Current Test Procedure:
One Sample Proportion

$$H_0: p \leq p_0$$
$$H_1: p \geq p_1 = p_0 + \Delta$$

$$p - value = P(X > x | p_0)$$

Decision Rule:

- Reject if $p - value < \alpha$

Sequential Test Procedure:
Sequential Probability Ratio Test

$$H_0: p \leq p_0$$
$$H_1: p \geq p_1 = p_0 + \Delta$$

$$S_i = S_{i-1} + \log(\Delta_i)\,,\ \ i = 1,2, \ldots ??$$

Decision Rule:

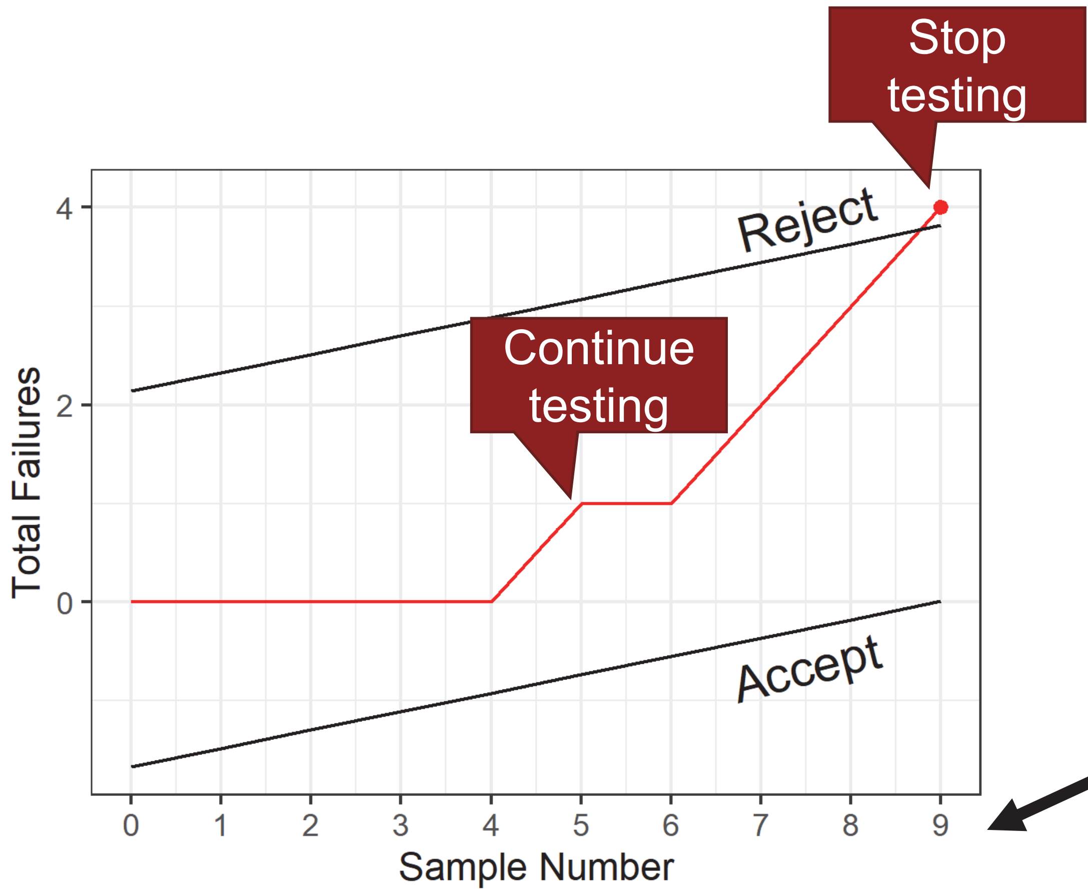
- $S_i \geq R$ reject H_0 and accept H_1
- $A \leq S_i \leq R$, continue sampling
- $S_i \leq A$ accept H_0

*The failure rate is the proportion of times the radar fails to detect the incoming projectile.

Notional probabilities; not based on data.

A – acceptance region; R – rejection region; S – test statistic; x – number of successes; α – significance level; $\log(\Delta)$ -- log-likelihood ratio

The sequential probability ratio test in action



In this example, we could reject our null hypothesis after nine shots!

Source: https://test-science.shinyapps.io/Shiny-SPRT/

Does it really work? Can we control for error?

$$H_0 : p = p_0 \quad H_A : p = p_1$$

Method	Error Rate		Average Sample Size (Standard Deviation)	
	Type I	Type II	Under $\boldsymbol{H_0}$	Under $\boldsymbol{H_1}$
SPRT	0.151	0.203	26.5 (18.5)	23.7 (18.7)
Exact Binomial Test	0.122	0.192	50	50

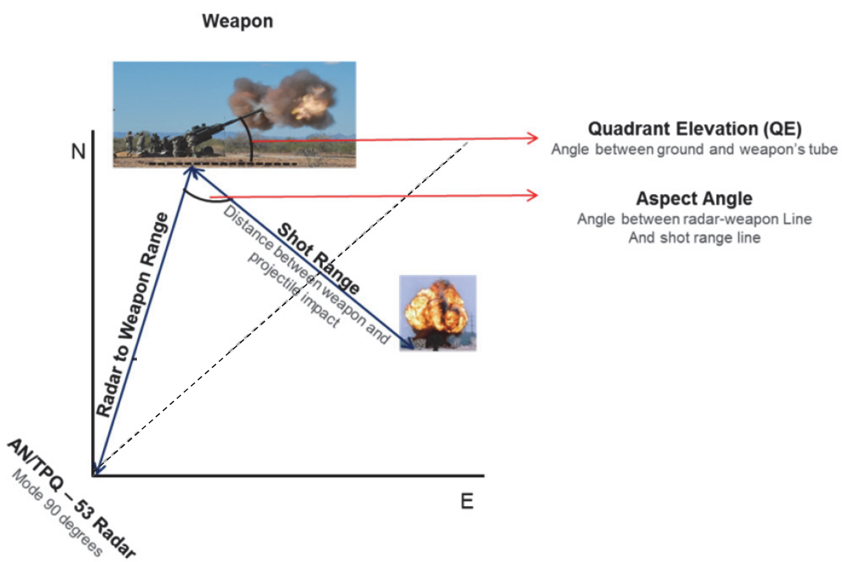
Simulation Study Settings: N = 1000, $\alpha = 0.2$, $\beta = 0.2$

Traditional Method Settings: obtained using JMP One Sample Proportion Sample Size Estimate

p_0 and p_1 are the failure rates under the null and alternative hypotheses, respectively

The performance of combat systems is likely affected by a variety of physical factors

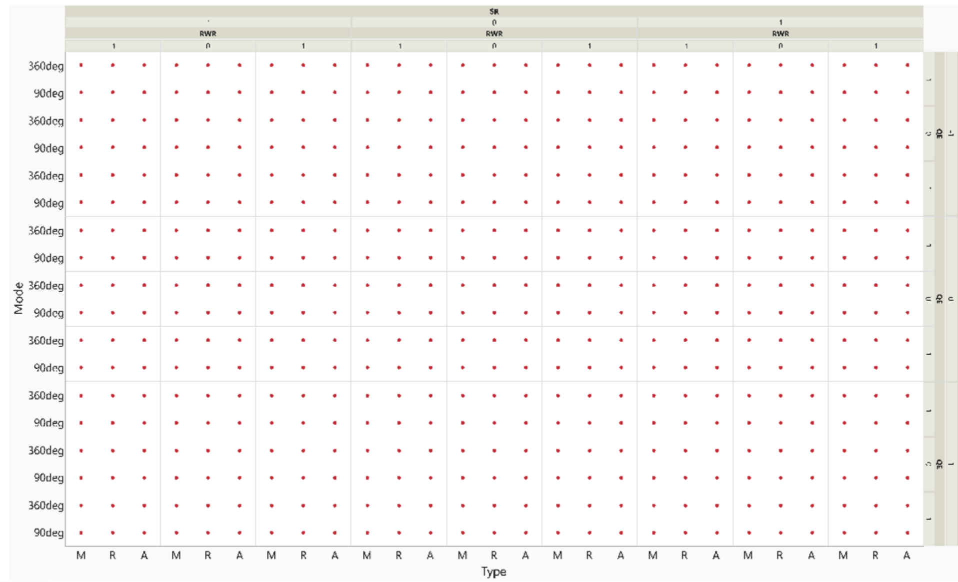
Example: Q-53 Counterfire Radar Mission



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

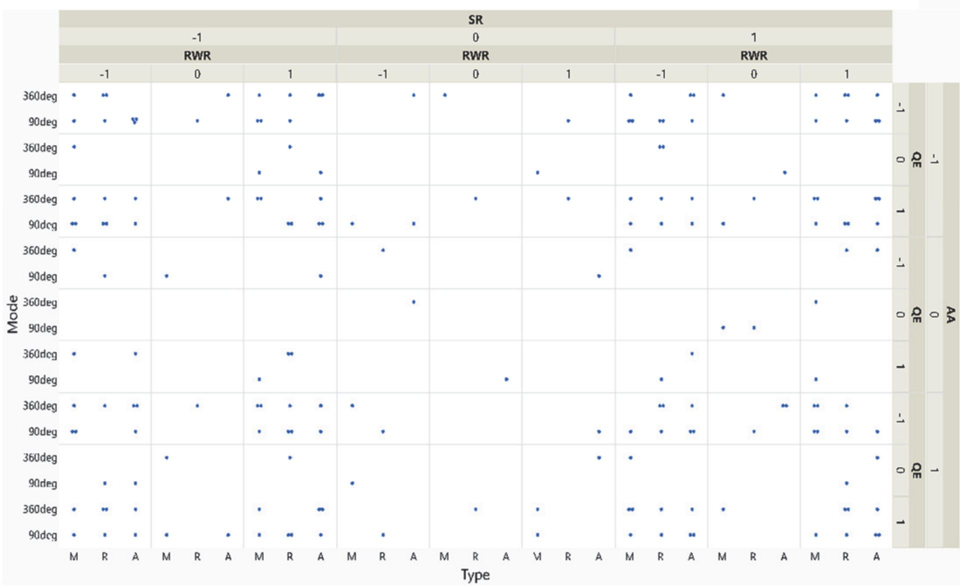
Figure Source: Freeman, Laura J., et al. "Testing defense systems." Analytic Methods in Systems and Software Testing (2018): 441.

We could plan a test using the "traditional" DOE approach…



Full Factorial – 486 test points

- Characterize
- Main Effects + Interactions + Quadratic Terms

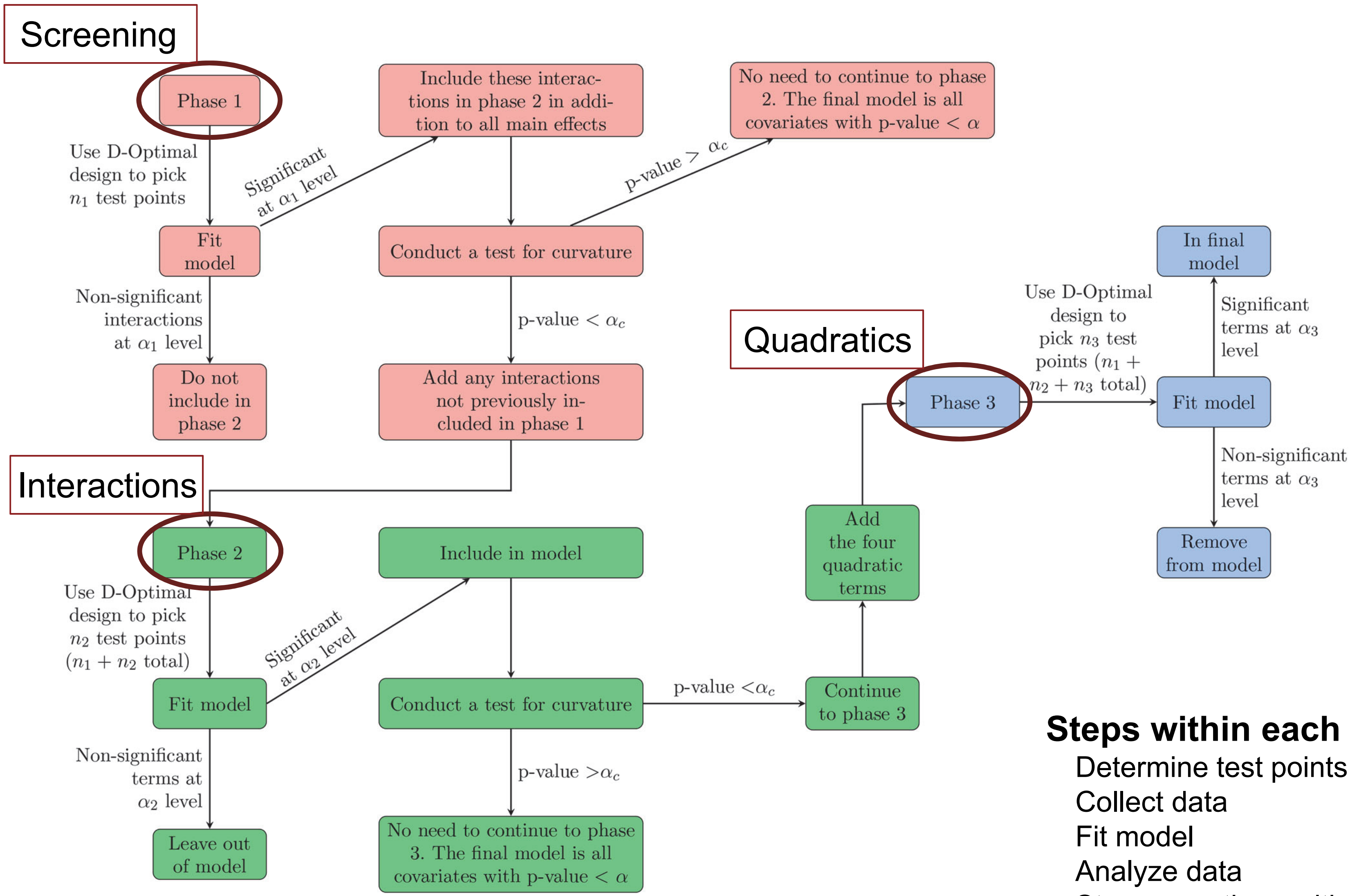


D-optimal design – 184 test points

- Characterize
- Main Effects + Interactions + Quadratic Terms
- Requires research-specified model

DOE – Design of Experiments

...or we could use a sequential DOE approach



Steps within each test phase
- Determine test points
- Collect data
- Fit model
- Analyze data
- Stop or continue with next phase

Each phase of data collection informs the collection of the next set of test points.

Does it work?

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

Simulation Settings:

- D-Optimal design
- 1,000 data sets
- $\sigma = \{1,3,4\}$
- α
    - Phase 1 = 0.30
    - Phase 2 = 0.15
    - Phase 3 = 0.15

*Models are notional and not based on true system performance.

The simulation stops at the test phase we expect it to

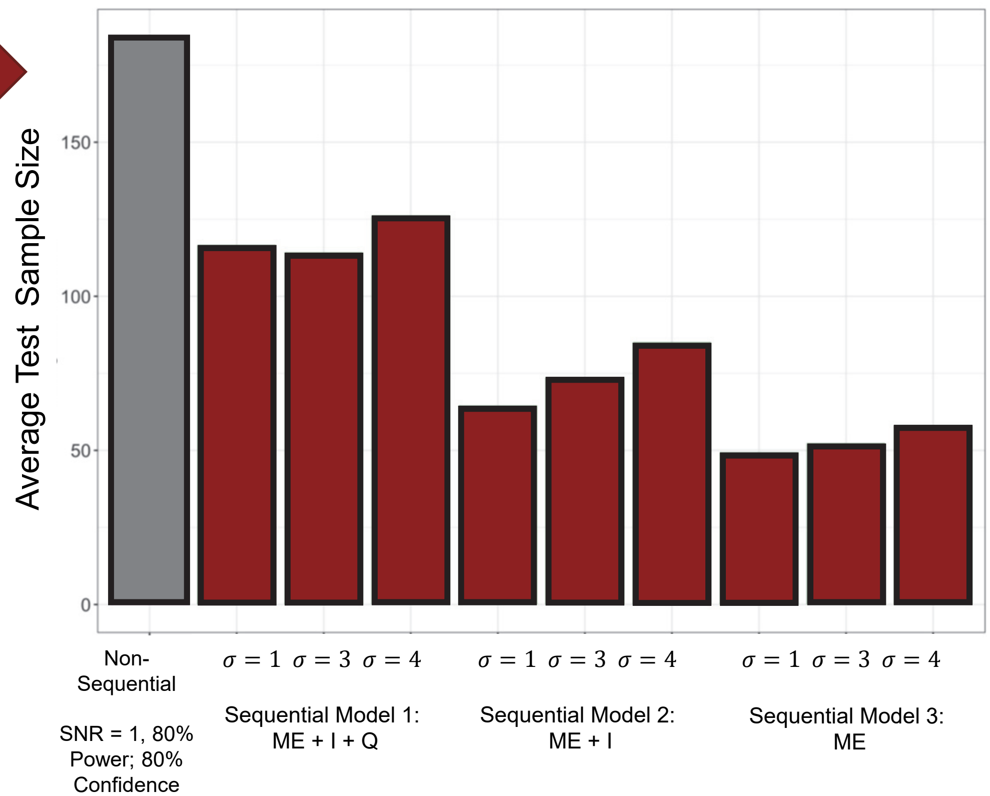
True Model*
Model 1: Main Effects + Interactions + Quadratic
Model 2: Main Effects + Interactions
Model 3: Main Effects

	Model 1			Model 2			Model 3		
Phase / σ	1	2	3	1	2	3	1	2	3
1	0%	9.10%	91%	0%	91%	9.40%	90%	1.80%	7.80%
3	0.30%	22%	78%	1.70%	91%	7.20%	91%	5.90%	2.70%
4	1.80%	23%	76%	7.10%	85%	8.30%	90%	6.80%	3.70%

*Models are notional and not based on true system performance.

Sequential methods might provide opportunities for test efficiency (substantial time & cost savings)

Results from a Simulation Study*



Important Caveat: With fixed sample sizes within each phase, it is possible to design smaller tests using non-sequential DOE, but we would be inherently making an assumption about the true model and potentially missing out on important factors.

Sequential methods do not necessarily guarantee the smallest test design – assumptions about the true model matter!

I – interactions; ME – main effects; Q – quadratic; SNR – signal-to-noise ratio

*Models are notional and not based on true system performance.

Implementation of sequential analysis has potential challenges

Scheduling and budgeting

- How can we plan for an unknown test length?
- Real time updates

Data collection, management, and analysis

- Quick decisions that influence the next phase

Misuse

- How could we avoid this?
- This could result in incorrect conclusions

Competing objectives

- Test phases might not share the same test goal

Modifications

- Did the system change much since the last test phase?

Challenges do not mean this is not doable, but will take time to think through the appropriate applications

Closing thoughts

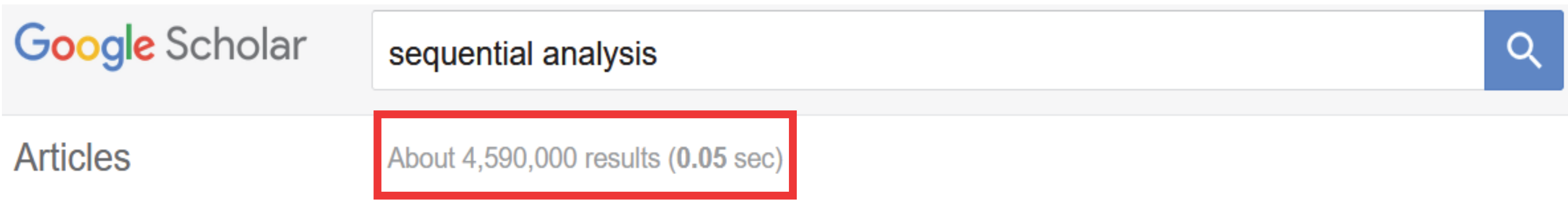
Applying a sequential strategy to T&E may help to speed up testing and save in cost by reducing the number of test runs required on average

BUT

Methods require:

- Access to data during testing to perform real-time analyses and decision making
- Flexibility in the manner in which the experiments are completed
- Close collaboration between testers, subject matter experts, and analysts

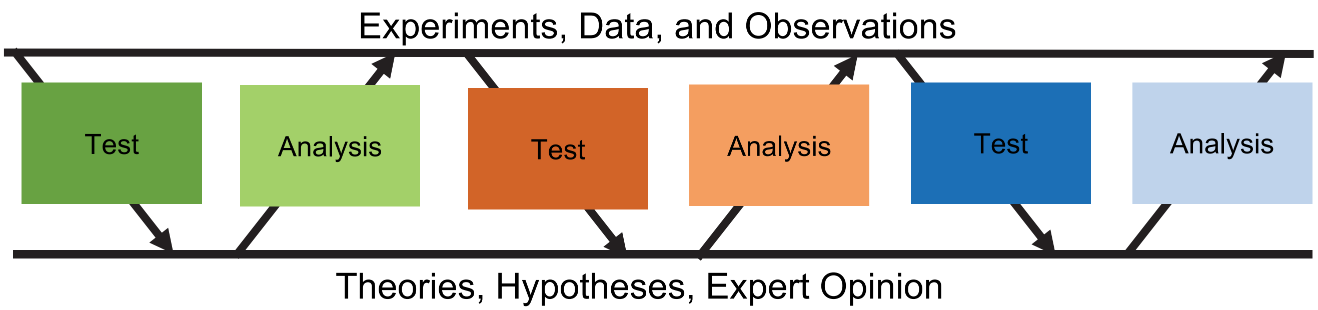
Backup



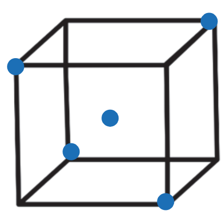
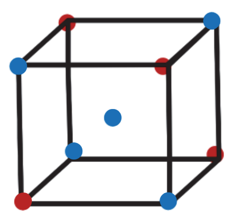
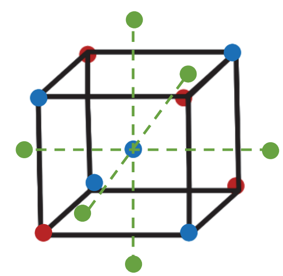
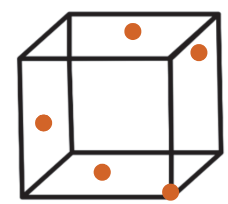
Sequential Testing: A collection of hypotheses tests performed in a sequential manner for which one must decide if more data needs to be collected.

Sequential Estimation: An estimation procedure (e.g., point or interval estimation) performed in a sequential manner

Sequential Design: A procedure that allows the experimenter to choose among experiments to perform at each stage or to vary the treatments sequentially. Each experiment builds on information gained from the previous experiment in considering how to continue the investigation.

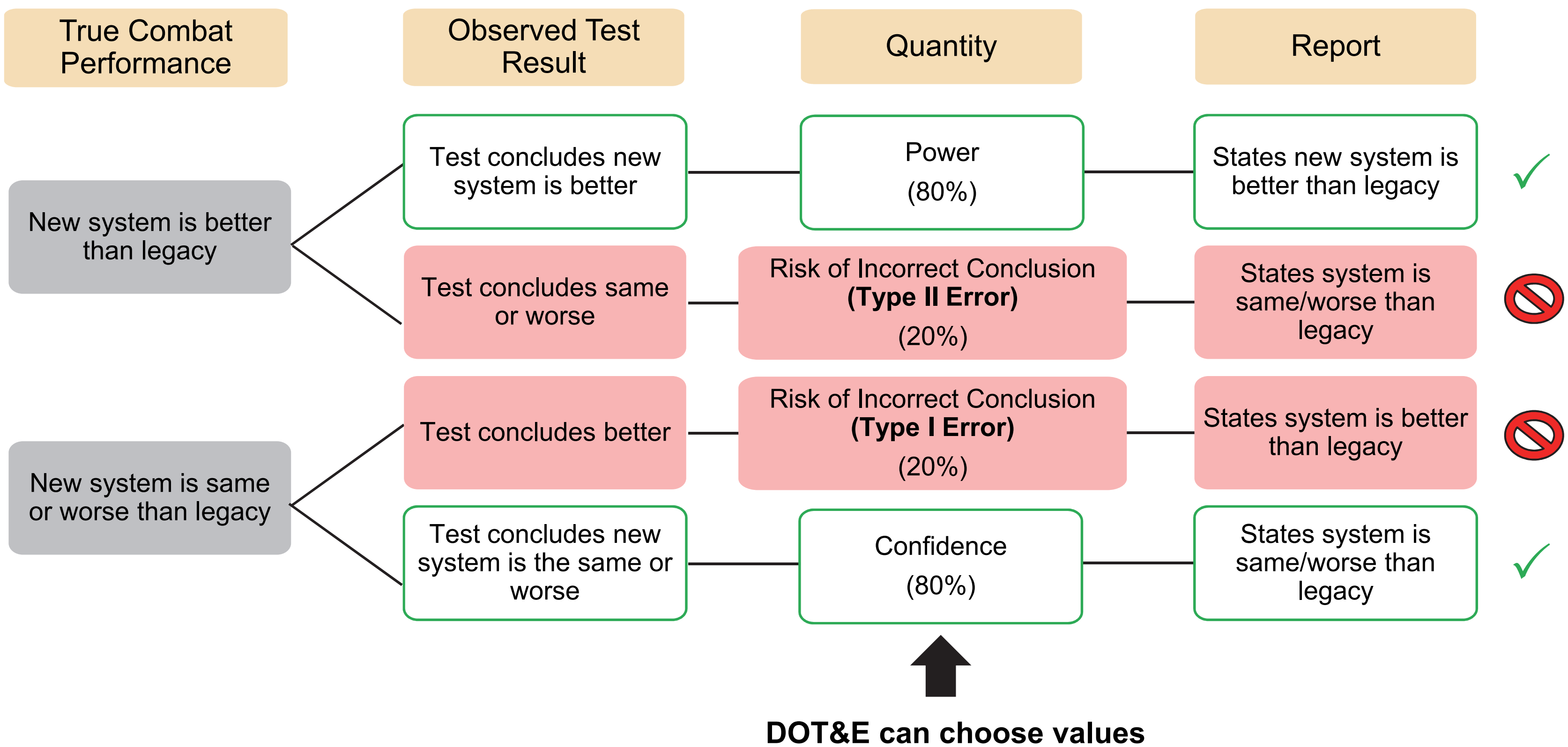


Examples of

Sequential Testing	$H_0\ vs\ H_a$ $S_1 = log\Delta_1$	$H_0\ vs\ H_a$ $S_2 = S_1 + log\Delta_2$	$H_0\ vs\ H_a$ $S_3 = S_2 + log\Delta_3$	$H_0\ vs\ H_a$ $S_n = S_{n-1} + log\Delta_n$
Sequential Estimation	$\widehat{\mu_1}, \widehat{\sigma_1}$	$\hat{\mu}_2, \widehat{\sigma_2}$	$\widehat{\mu_3}, \widehat{\sigma_3}$	$\widehat{\mu_n}, \widehat{\sigma_n}$
Sequential Design	 Screening Designs	 Interaction Designs	 Higher-Order Designs	 Validation Designs

H_0 is the null hypothesis; H_a is the alternative hypothesis; S is the test statistic; $log\Delta$ is the log-likelihood ratio; $\hat{\mu}$ is the recursive estimate of the mean; $\hat{\sigma}$ is the estimate of the recursive standard deviation

In sequential tests, as well as in current test procedures, we have to protect against committing two kinds of errors



What do we mean by statistical modeling?

$$Location\ Accuracy = \beta_0$$
$$+\beta_1 RWR + \beta_2 Projectile + \beta_3 Radar\ Mode$$
$$+\beta_4 (RWR * Projectile) + \beta_5 RWR^2 + \epsilon$$

- Outcome / Response
- Intercept
- Main Effect
- Interaction Effect
- Quadratic Effect
- Error (unexplained)
- β is a coefficient that we estimate from the collected data

Main Effect: The change in the response produced by changing the level of a factor.

➢ Ex: A difference in the mean location accuracy when we change radar mode.

Interaction effect: Occurs when the change in the response between the levels of one factor is not the same at all levels of the other factors (e.g., factors work in a synergistic fashion)

➢ Ex: The artillery projectile had larger miss distances for longer radar weapon to range distances. This same change was not observed for the mortar projectile.

SDOE method produces models that include fewer extraneous factors than the traditional D-optimal method

Average Number of Extra Factors Included in the Final Model and Number of Times the Correct Model Was Contained in the Final Model Across All Settings

		SDOE		Traditional D-Optimal	
Model	**σ**	**Extra Factors (SD)**	**CCM**	**Extra Factors (SD)**	**CCM***
1	1	2.22 (2.00)	909	4.54 (2.78)	1000
	3	2.09 (1.93)	775	4.54 (2.78)	1000
	4	1.90 (1.77)	753	4.54 (2.78)	1000
2	1	3.00 (2.42)	1000	4.94 (2.85)	1000
	3	2.97 (2.42)	975	4.94 (2.85)	1000
	4	2.98 (2.33)	876	4.94 (2.85)	1000
3	1	3.37 (2.06)	1000	5.33 (2.86)	1000
	3	3.38 (2.10)	998	5.33 (2.86)	1000
	4	3.39 (2.06)	997	5.33 (2.86)	1000

*CCM is “contained correct model” and denotes the number of data sets in which the correct model was contained in the final selected model.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	XX-08-2023
2. REPORT TYPE	Final
3. DATES COVERED *(From - To)*	
4. TITLE AND SUBTITLE	Improving Operational Test Efficiency: Sequential Methods in Operational Testing
5a. CONTRACT NUMBER	HQ0034-19-D-0001
5b. GRANT NUMBER	
5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)	Keyla Pagan-Rivera
5d. PROJECT NUMBER	C9082
5e. TASK NUMBER	C9082
5f. WORK UNIT NUMBER	
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 3000050
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	Central Research Project
10. SPONSOR/MONITOR'S ACRONYM(S)	CRP
11. SPONSOR/MONITOR'S REPORT NUMBER(S)	

12. DISTRIBUTION/AVAILABILITY STATEMENT

This OED Draft has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

13. SUPPLEMENTARY NOTES

Project Leader: Rebecca M. Medlin

14. ABSTRACT

The Department of Defense develops and acquires some of the world's most advanced, sophisticated, and expensive systems. As new technologies emerge and are incorporated into systems, Director, Operational Test and Evaluation faces the challenge of ensuring that these systems undergo adequate and efficient test and evaluation (T&E) prior to operational use. In this talk, I will highlight ways in which relatively well-known statistical methods can help the T&E community in taking steps toward increasing test efficiency. In particular, a motivating example will be used to demonstrate the value in applying sequential test planning approaches to operational effectiveness testing.

15. SUBJECT TERMS

Design of Experiments (DOE); Sequential Design; Sequential Probability Ratio Test; Test and Evaluation

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON	19b. TELEPHONE NUMBER *(Include area code)*
Unclassified	Unclassified	Unclassified	SAR	28	Rebeca M. Medlin	703-845-6731