INSTITUTE FOR DEFENSE ANALYSES

Case Study on Applying Sequential Analyses in Operational Testing

Rebecca M. Medlin, Project Leader

Monica L. Ahrens
Rebecca M. Medlin
Keyla Pagán-Rivera
John W. Dennis

December 2021

Public release approved. Distribution is unlimited.

IDA Document NS D-32904

Log: H 2021-000468

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

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 Dr. John T. Haman and Dr. Matthew R. Avery from the Operational Evaluation Division.

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

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

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-32904

Case Study on Applying Sequential Analyses in Operational Testing

Rebecca M. Medlin, Project Leader

Monica L. Ahrens
Rebecca M. Medlin
Keyla Pagán-Rivera
John W. Dennis

Case Study on Applying Sequential Analyses in Operational Testing

Monica L. Ahrens

Rebecca M. Medlin

Keyla Pagán-Rivera

John W. Dennis

Abstract

Sequential analysis concerns 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 analysis originated in ballistics testing for the Department of Defense (DoD) (Wald 1945; Wallis 1980), it is underutilized in the DoD. Expanding the use of sequential analysis may save money and reduce test time. In this paper, we introduce sequential analysis, describe its current and 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 challenges to address when using sequential analysis in OT&E.

Executive Summary

Most statistical analyses involve observing a fixed set of data and analyzing those data after the final observation has been collected to draw some inference about the population from which they came (Ghosh 2014). In contrast, sequential analysis concerns situations in which the number, pattern, or composition of the data is not determined at the start of the investigation but instead depends upon the information acquired throughout the investigation.

Despite the fact that sequential analysis originated in ballistics testing for the Department of Defense (DoD) (Wald 1945; Wallis 1980), we find that it is underutilized in the DoD. Expanding the use of sequential analysis potentially can reduce costs and test time.

In this paper, we introduce sequential analysis, describe its current and potential uses in operational test and evaluation (OT&E), and present methods for applying it to the test and evaluation (T&E) of defense systems. We present a case study involving the AN/TPQ-53 Counterfire Radar and perform simulation studies demonstrating the potential value in applying sequential methods. Finally, we discuss challenges to address when using sequential analysis for OT&E.

Motivating Example

We present a case study regarding the AN/TPQ-53 Counterfire Radar (Q-53) to motivate the use of sequential methods. The Q-53 is a ground-based radar designed to detect incoming mortar, rocket, and artillery projectiles; predict impact locations; and locate the threats geographically. The following two questions are key to understanding and evaluating the Q-53's performance:

1. Can the Q-53 detect shots with high probability?
2. Can the Q-53 locate a shot's origin with sufficient accuracy to provide an actionable counterfire grid location?

We use the Q-53 to demonstrate how one might consider conducting the test using a sequential T&E strategy, and how by doing so one – on average – could save test resources.

Sequential Methods

Sequential Probability Ratio Test

Often, when planning an operational test, the number of observations is treated as fixed. The traditional approach is to first calculate the number of test points needed to achieve a specified power and confidence for a given effect size (for example, 80 percent power and 80 percent confidence), and then execute that exact number of test points. For example, to test whether the detection rate of the Q-53 radar is below a certain probability, p, we would calculate the number of shots we need to fire to have a desired power of finding a true increase or decrease from p.

Alternatively, one could consider conducting the test using a sequential approach. One of the best-known applications of sequential analysis involves testing a hypothesis when the final sample size is not fixed at the start of the analysis; instead, it depends on the information obtained as the data are collected. The sequential probability ratio test (SPRT) involves taking observations one at a time; each additional observation is used to decide whether to stop sampling and accept or reject the null hypothesis in question.

Using a simulation study, we compare the results of a fixed-sample-size hypothesis test to the SPRT. Our results show that the sample size required for our SPRT analysis is, on average, approximately half the sample size required by a fixed-sample-size hypothesis test (for example, the binomial test).

It's important to note that neither the fix-sample-size hypothesis test nor the SPRT characterize the system's performance across the operational test space, which is often a goal of operational testing. For this reason, we also describe a sequential design of experiments method, which allows the experimenter to adaptively, efficiently, and effectively characterize the operational test space and determine how factors affect an output.

Sequential Design of Experiments

The performance of the Q-53 is likely affected by a wide variety of factors, including: operating conditions, threat types, system operating modes, and other physical factors. To study the impact of each factor as well as the impact of interactions among factors, we often use design of experiments to plan a test event and determine a fixed set of test points to execute.

A sequentially planned design allows testers to learn from one test and use that knowledge to modify subsequent tests. Modifications might include adding or removing factors (or levels of a factor) and adding or modifying the response variables in order to capture more precise information. We implement a phased approach to collecting and analyzing the data in an operational test.

Using a simulation study, we compare the use of a sequential design of experiment to a traditional non-sequential design of experiment. We find that the average sample size required under each set of test conditions is always smaller for the sequential design when we assume that we will need to fit a model with main effects, quadratic effects, and all two-way interactions. If, however, the true model is known in advance (for example, a main-effects-only model), a traditional non-sequential design (for example, D-optimal design) is, on average, more efficient than the sequential design. This is due to keeping the sample size fixed within phases and only allowing the number of testing phases to vary according to the sequential design.[1]

Conclusion

We introduced sequential analysis, described its current and potential uses in T&E, and presented two methods for applying a sequential analysis to the T&E of defense systems. Additionally, we demonstrated the methods' potential reduction to test costs and time.

However, sequential procedures may prove to be more challenging to implement in DoD T&E than non-sequential procedures. Sequential procedures are challenging to use in DoD because the number of test points, conditions for those test points, and resources required to execute those test points are often decided early on and codified in the T&E strategy and test plans. Furthermore, sequential procedures may prove challenging to implement when the time required to score individual test events and perform analysis takes longer than the scheduled time between tests, and when stakeholders have divergent assessments of test points. When sequential methods can be applied, however, we find from our review and our examples that sequential procedures offer opportunities to make testing more efficient. Therefore, we recommend that test teams consider sequential methods when planning T&E.

---

1 A sequential design allowing the sample size to vary within phases may fare better against a non-sequential design with a known true model, but this is beyond the scope of this work.

Contents

1. Introduction

Most statistical analyses involve observing a fixed set of data and analyzing those data after the final observation has been collected to draw some inference about the population from which they came (Ghosh 2014). In contrast, sequential analysis concerns situations in which the number, pattern, or composition of the data is not determined at the start of the investigation; instead, these depend upon the information acquired throughout the investigation (Robbins 1952; N. L. Johnson 1961; Ghosh 2014).

Although sequential analysis has its formal origin in ballistics testing for the Department of Defense (DoD) (Wald 1945; Wallis 1980), it has been underused in recent DoD testing.[2] Expanding the use of sequential analysis in DoD testing may save money and reduce test time (National Research Council 1998).

Medlin et al. (2021) subdivide the field of sequential analysis into three broad functional categories: sequential testing, sequential design, and sequential estimation. In this paper, we focus on the methods for sequential testing and sequential design. We use sequential analysis techniques to demonstrate an efficient and effective way for the DoD to maximize system understanding in test.

The remainder of the paper is organized as follows. We first briefly review the history and literature in Section 2. In Section 3, we describe a case study involving the AN/TPQ-53 Counterfire Radar. Using this system, we demonstrate in Section 4 an application of Wald's (1945) Sequential Probability Ratio Test (SPRT). In Section 5, we present a sequential design of experiments (SDOE) method. Both Sections 4 and 5 include a simulation study comparing sequential methods to non-sequential methods of testing. We conclude the paper in Section 6 by discussing the challenges and potential benefits of using sequential procedures.

[2] Sequential analysis is used in some areas; for example, DoD ballistic resistance testing continues to use sequential test designs to estimate a particular probability of perforation (T. H. Johnson et al. 2014).

2. Historical Context and Literature Review

The field of sequential analysis "was born in response to demands for more efficient testing of anti-aircraft gunnery during World War II, culminating in Wald's development of the SPRT in 1943" (Lai 2001). In particular, Wald's (1945) solution to the problems underlying sequential analysis arose in connection with a specific question posed to the Statistical Research Group (SRG)[3] by Captain Garret L. Schulyer of the Bureau of Ordnance, Navy Department, who was interested in calculating the probability of a hit by anti-aircraft fire on a directly approaching dive bomber. Captain Schulyer wanted to determine a rule, specified in advance, for stating the conditions under which the experiment might be terminated earlier than planned.

The problem came to the attention of Abraham Wald, a member of the SRG. Wald devised the SPRT, a statistical test that takes advantage of the sequential nature of the data to reduce the required number of observations. This test is the most efficient one, in terms of sample size, for testing a simple hypothesis H_0 against a single alternative H_1. In this paper, we apply the SPRT to an operational test and evaluation (OT&E) case study.

Sequential techniques may also be applied to design of experiments (DOE).[4] 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 test and evaluation (T&E) community has embraced the use of non-sequential DOE for planning developmental and operational testing (Freeman et al. 2018). DOE is not by nature a sequential technique, but many people, including Montgomery (2017), strongly recommend planning and executing a DOE based on the results of previous experiments to either augment or inform later testing.

Because experiments are usually iterative in nature, Box and Wilson (1951) and Montgomery (2017) note that it is unwise to design too comprehensive of an experiment, a "one-shot" experiment, at the start of a study. Rather, one should plan for a series of tests that leverage the information obtained from one sequence of the experiment to help plan the next; this defines a sequential DOE (SDOE). Regarding how to implement a sequential

---

[3] The Statistical Research Group was an Office of Scientific Research and Development activity at Columbia University during the Second World War.

[4] The test and evaluation community may be more familiar with the SDOE planning approach described by Box and Wilson (1951) and Montgomery (2017), which is the focus of this paper, but sequential design problems are more generally those that involve a sequential search for informative experiments (Chernoff 1959).

experimental design strategy, Montgomery (2017) suggests starting with a screening design in which many factors are tested to assess their importance. Using the results from the screening design, testers can augment the test design matrix by adding additional experimental test points. An augmented design can help determine whether higher-order terms are needed in the statistical model. In this paper, we illustrate how to plan a DOE in phases, where each phase is an augmentation of the previous phase.

Simpson, Listak, and Hutto (2013) discuss the ways one might apply SDOE in the context of T&E – encouraging the use of a staged or phased process of testing that allows for analysis pauses. In this paper, we propose and illustrate a method for planning a DOE sequentially, similar to the recommendations of Montgomery (2017) and Simpson, Listak, and Hutto (2013). We expand on the work of Simpson, Listak, and Hutto (2013) by illustrating how one might size each phase of test for a specified confidence and power. We additionally briefly discuss what such a test is and is not sized for.

Sequential methods are a critical tool in helping testers adaptively, efficiently, and effectively execute testing. Recently, the testing community for artificial intelligence and autonomous systems (AI&AS) cited the need for sequential methods. Ahner and Parson (2016) and Porter et al. (2019) mention some of the reasons sequential analysis methods could be useful in testing and analyzing data from AI&AS: understanding the decision-making process of the system, sequentially covering the AI&AS operational space, and using DOE as an efficient tool for test planning. Learning to analyze the data sequentially will allow the T&E community to learn and adapt as new data arrive and, if necessary, to inform and modify the data collection plan. We believe this paper serves as a first step in helping the T&E community think about and apply sequential methods to their programs.

3. Motivating Example: AN/TPQ-53 Counterfire Radar

To motivate the use of sequential methods, we use the AN/TPQ-53 Counterfire Radar (Q-53) as a case study. Mortar, rocket, and artillery fire posed a significant threat to U.S. forces in Afghanistan and Iraq and will likely remain a significant threat to ground troops in future conflicts. The Q-53 (Figure 1) is a ground-based radar designed to detect incoming mortar, rocket, and artillery projectiles; predict impact locations; and locate the threats geographically. Threat location information allows U.S. forces to return fire, and impact location information also can be used to warn U.S. troops. The Army conducted the initial operational test and evaluation (IOT&E) of the Q-53 in June 2015.

Figure 1. Soldiers Emplacing the AN/TPQ-53 Counterfire Radar during Operational Testing

Freeman et al. (2018) used the 2015 IOT&E data to illustrate how a statistical analysis can be used to summarize complex system behavior. The authors suggest that the following two questions are key to understanding and evaluating the Q-53's performance:

3. Can the Q-53 detect shots with high probability?
4. Can the Q-53 locate a shot's origin with sufficient accuracy to provide an actionable counterfire grid location?

In this paper, we revisit the Q-53 example but from the perspective of test planning. We show how one might have conducted the test using a sequential T&E strategy, and how by doing so one – on average – could have saved test resources. In Section 4, we illustrate how one might apply the SPRT to evaluate the Q-53's ability to detect a shot. In Section 5, we illustrate how one might apply SDOE to strategically select test points for evaluating the Q-53's ability to accurately locate a shot's origin.

4. Sequential Probability Ratio Test

Often, when planning an operational test the number of observations is treated as fixed. The traditional approach is to first calculate the number of test points needed to achieve a specified power and confidence for a given effect size (for example, 80 percent power and 80 percent confidence), and then execute that exact number of test points. For example, to test whether the detection rate of the Q-53 radar is below a certain probability, p, we would calculate the number of projectiles or shots we need to fire to have a desired power of finding a true increase or decrease from p. In testing, all shots would be fired, the radar data collected and analyzed, and the results of the hypothesis test reported. While this method proves reliable for providing power and confidence to detect the desired difference in p, it may not be the most efficient or the best use of test resources.

Alternatively, one could consider conducting the test using a sequential approach. One of the best-known applications of sequential analysis involves testing a hypothesis when the final sample size is not fixed at the start of the analysis; instead, it depends on the information obtained as the data are collected. This procedure underlies the genesis of sequential analysis as formalized by Wald (1945) in his SPRT. The SPRT involves taking observations one at a time; each additional observation is used to decide whether to stop sampling and accept or reject the null hypothesis in question. Wald (1945) notes that the SPRT requires, in general, a considerably smaller expected number of observations than the fixed number of observations required by a corresponding non-sequential test, while controlling the type I and type II errors.

In testing the Q-53 radar, of primary interest is knowing whether the failure rate[5] differs from what is expected. For instance, a scenario might be that we expect the failure rate to be p_0, and we want to test whether it is actually higher by a $\Delta\ p_1$, say .10; such a hypothesis statement would be written as

$$H_0: p = p_0$$
$$H_1: p = p_1$$

Figure 2 illustrates the application of the SPRT to this scenario. The figure presents a cumulative count of the shots fired (x-axis), a cumulative count of the detection failures that have occurred (y-axis), and the upper and lower SPRT-calculated stopping boundaries. For each shot fired, the Q-53 radar's ability to detect or fail to detect is recorded, the SPRT test statistic (the sum of failures) is calculated, and the test statistic is compared to the upper

[5] The failure rate is the proportion of times the radar fails to detect the incoming projectile.

and lower bounds to determine whether to stop the testing or continue the testing. In this example, no failures are observed in the first four shots; the fifth shot results in the first failure. Testing continues, however, because the calculated SPRT test statistic is still within the predetermined stopping boundaries. In fact, testing continues in this example until after we observe the 11th shot, at which point the test statistic crosses the rejection boundary set by the SPRT. Our conclusion, after 11 shots, is to reject the null hypothesis ($H_0 = p_0$) in favor of the alternative hypothesis ($H_1 = p_1$) and to state the failure rate is higher than expected. We provide the SPRT mathematical details in Appendix A.

Figure 2. Illustration of SPRT process for a true value $p = p_2 > p_1$. After the 11th observation, the test indicates we have collected enough information to reject the null hypothesis. For comparison, the required sample size for a non-sequential approach is 50 test points.

A. Simulation Study

To further motivate the use of Wald's SPRT, we conducted a simulation study[6] to compare the results of the SPRT to a fixed-sample-size hypothesis test, which in this case is the exact binomial test. The results in Table 1 show that the sample size required for our SPRT analysis is, on average, approximately half the sample size required by the exact

---

[6] We used the following simulation settings: true failure proportions of $p = p_0$ and $p = p_1$ and type I and type II error rate set to 20 percent.

binomial test.[7] Our results also show that we are still able to maintain the proper type I and type II error rates when using the SPRT.

Table 1. Observed error rates and sample sizes for the SPRT compared to the exact binomial test. The true failure proportions are $p = p_0$ and $p = p_1$ and the type I and type II error rate are set to 20 percent.

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

In this section, we demonstrate the benefits in test efficiency from using the SPRT compared to a traditional hypothesis test, namely, the exact binomial test.[8] However, neither test approach characterizes the system's performance across the operational test space, which is often a goal of operational testing. In the next section, we describe an SDOE method, which allows the experimenter to adaptively, efficiently, and effectively characterize the operational test space and determine how factors affect an output.

It is worth noting that one area in which the SPRT may be most useful to OT&E is reliability testing. The DoD (1996) recommends the use of sequential testing for reliability testing. Recall the intent of reliability testing is to determine the distribution of failure times; reliability is based on top-level metrics, such as the mean time between failures (MTBF), or a probability of failure. The size or length of a reliability test plan is determined by the reliability requirement and desired statistical metrics. Often in OT&E, fixed-duration test plans are selected to estimate reliability because the length of a test must be known in advance. The DoD (1996) presents the use of an SPRT plan, based on Wald's (1945) SPRT, for determining compliance with a specific reliability requirement. When the demonstrated MTBF is high enough or low enough, an SPRT plan will save test time compared to a fixed-duration test plan that has similar risks. With respect to determining an initial test length when using a sequential test plan, the DoD (1996, pp. 18–19, sec. 5.4.2.3) notes, "for sequential test plans, test duration should be planned on the basis of maximum allowable test time (truncation), rather than the expected decision point, to avoid the probability of unplanned test cost and schedule overruns."

---

7 Because the SPRT does not have a fixed sample size, we present the average sample size and standard deviation (SD) for each simulation scenario in Table 1.

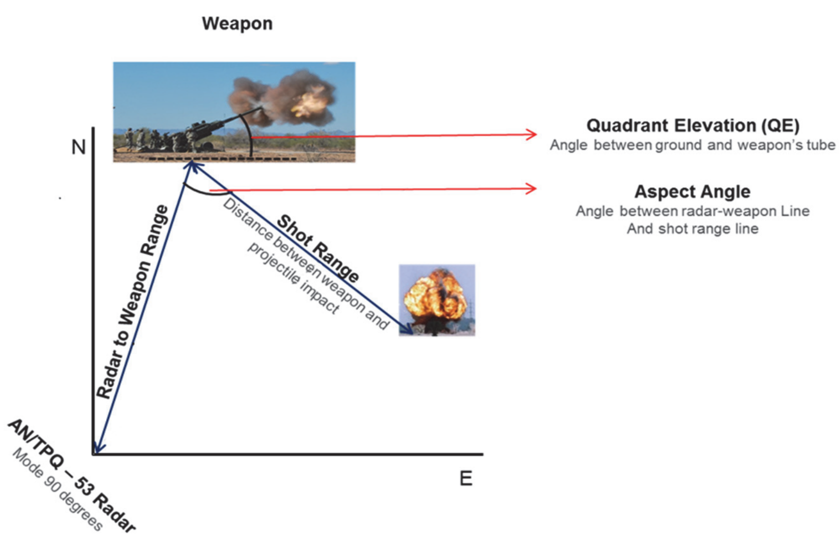
8 In this example we used the exact binomial test, which is commonly used in T&E for analyzing binomial data. However, a more natural comparison would be to use the Neyman-Pearson Likelihood Ratio Test, because the hypotheses are identical to the SPRT. The exact binomial is mathematically identical to the Neyman-Pearson Likelihood Ratio Test (see Appendix A).

5. Sequential Design of Experiments

The performance of combat systems may be affected by a wide variety of operating conditions, threat types, system operating modes, and other physical factors. Table 2 lists planning factors that could impact a radar's performance, and Figure 3 illustrates those factors. In this section, we present an SDOE strategy. The strategy allows testers to fully characterize the system's performance across the operational test space and to potentially do so with fewer test points than if a non-SDOE approach had been employed.

Table 2. Q-53 Test Design Factors

Design Factor	Label	Type	Levels
Quadrant Elevation	A	Continuous	Low, High
Aspect Angle	B	Continuous	Low, High
Munition Type	C	Categorical	Mortar, Rockets, Artillery
Shot Range	D	Continuous	Low, High
Radar Operating Mode	E	Categorical	90-degree, 360-degree
Radar to Weapon Range	F	Continuous	Low, High



Source: Freeman et al. 2018

Figure 3. Example of fire mission including relevant geometric factors impacting Q-53 system performance. During a threat fire mission, the threat will fire projectiles at a target inside the search area of the Q-53. In this figure, the Q-53 radar is operating in a 90-degree mode, and so its search sector is limited to the area within the black bars.

A. Design of Experiments

Full-factorial designs are often paired with the test goal of characterization,[9] which is a common goal when designing an experiment for an operational test. A full-factorial design includes at least two factors and examines all possible combinations of each factor's levels. These designs allow evaluators to determine the impact of each factor as well as the impact of interactions among factors. Full-factorial designs are informative for studying the effect of more than one factor (Montgomery 2017), though potentially prohibitively costly when many factors are involved. For our example, to test each possible combination once, not accounting for quadratic terms, at least 96 test points are required. If we include quadratic terms, treating continuous factors as a three-level factor, at least 486 test points are required.

Optimal designs are also frequently used when planning an operational test. They are especially useful when the number of test points is constrained to preclude a full-factorial design. An optimal design approach starts with the factors and response variables of interest and then creates a corresponding tailor-made design that matches the needs of the experimenter. An optimal design requires a researcher-specified model and a fixed sample size. A D-optimal design is a common optimal design choice when the test is intended to characterize, as the design seeks to minimize the overall variance of the parameter estimates (Montgomery 2017). While using an optimal design can result in a more efficient DOE approach relative to classical designs like factorials, there are trade-offs to using these types of designs compared to a more classical design. For example, consider model robustness – if the researcher-specified model is incorrect, will the design selected provide valuable information? To avoid a mis-specified model and to achieve adequate power and confidence (for example, a signal-to-noise ratio of 1, 80 percent confidence, and 80 percent power), 184 test points are required to fit a model with main effects, quadratic effects, and all two-way interactions.

B. Planning a Sequential Design of Experiments

SDOE is not commonly used in the testing of military systems; however, many, including Freeman et al. (2018) and Simpson (2018), encourage its use for T&E. In general, a sequentially planned design allows testers to learn from one test and use that knowledge to modify subsequent tests. Modifications might include adding or removing factors (or levels of a factor) and adding or modifying the response variables in order to capture more precise information.

There is more than one way to implement a sequential planning approach. The approach will depend on the research goals and test limitations. In this paper, we

---

[9] By *characterization* we mean describing the system's performance under different operational conditions.

implement a phased approach to collecting the data within an operational test. We use the term *phase* to define a set of test points to be executed and evaluated before the next set of test points is identified, executed, and evaluated. In our example, we use three phases, but different numbers may be appropriate depending on the logistics and goals of the tests. We design each phase to have enough power to detect certain prescribed effect sizes in the presence of noise.

The objective of our test is to characterize the operational space. Each phase of the test has a specific goal. In our example, the first phase screens for all main effects and some two-way interactions. In the second phase, we augment the test design to screen for any remaining two-way interactions. In the third phase, we augment the design to verify that there are not any quadratic effects in our remaining continuous factors. The flow chart in Figure 4 communicates the general process we employed. Steps 1–11 below outline the specific steps we followed for the Q-53 radar sequential test design example.

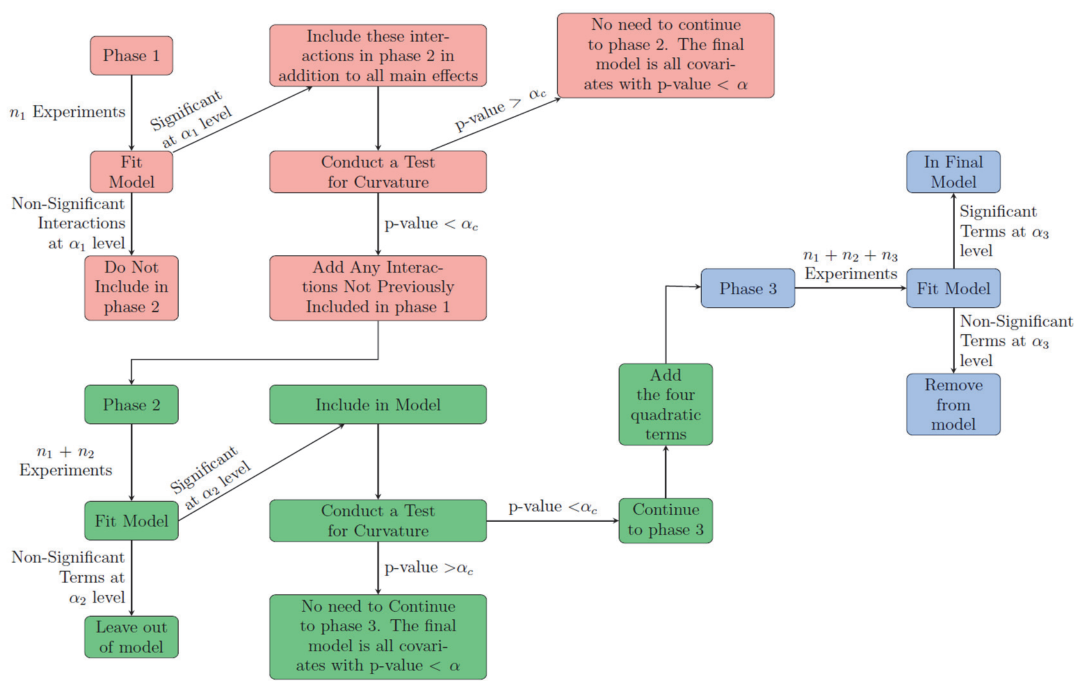


Figure 4. The SDOE Planning Process Used in This Paper

We used a D-optimal design to select our design matrix for each phase of test points and used the skpr library in R to generate the design matrices and calculate power. In planning, we allowed for different significance levels when sizing each phase, labeled α_1, α_2, α_3. For example, in the simulation study in Section 5.D, we used $\alpha_1 = 30\%$ significance for the first phase, $\alpha_2 = 15\%$ for the second phase, and $\alpha_3 = 15\%$ for the third phase. We allowed "moderately significant" coefficients to be retained early in testing, with further examination as testing progressed. In the steps below, α_c represents the significance level for a test for curvature. Since we conducted two different tests for curvature, a Bonferroni correction is needed, and we used a significance level of $\alpha_c/2$ for each test for curvature. In the simulation, we used $\alpha_c = 20\%$. The test team should discuss the pros and cons associated with determining desired significance levels as part of the planning process.

1. **Phase 1: Screen for Main Effects and Some Interactions.** Size the first phase of test points to find one standard deviation effect sizes under the model, including all main effects and some interactions with $1-\beta$ power and $1-\alpha_1$ confidence. Our design allows us to look at the following model terms:[10] $A,\ B,\ C,\ D,\ E,\ F,\ A{:}B,\ A{:}C,\ A{:}D,\ A{:}E,\ A{:}F,\ B{:}D,\ B{:}F.$ Include replication test points to test for a lack of fit. Details on properly picking the number of replications are in Appendix B.
2. **Phase 1: Collect Data.** Collect the data for the test design generated in Step 1.
3. **Phase 1: Fit Model with Phase 1 Data.** Fit the model based on the data collected in Step 2. Remove factors that are not statistically significant at the α_1 level. Maintain model hierarchy by keeping main effects that are included in any statistically significant interaction term.
4. **Phase 1: Test for Curvature.** Test for curvature by using a test for pure error, also known as a lack of fit test, as detailed in Appendix B. If the test is statistically significant at the $\alpha_c/2$ level, continue to Phase 2. Otherwise, the model chosen in Step 3 is the final model.
5. **Phase 2: Size to Fit Model for Rest of the Interactions.** Fit a model that includes all covariates in Step 3 plus the rest of the interactions. Use a two-step process to size Phase 2. First, size for power of $1-\beta$ assuming a standard deviation effect size. Then, make a second power calculation to size the number of replications needed to test for curvature for $1-\beta$ power and $\alpha_c/2$ level. Details on this power calculation are in Appendix B.
6. **Phase 2: Collect Data.** Run experiments for test points chosen in Step 5.
7. **Phase 2: Fit Model with Phase 2 Data.** Using the data collected in Step 6, fit the Phase 2 model and retain any statistically significant coefficients at the α_2 level while maintaining hierarchy.
8. **Phase 2: Test for Curvature.** Conduct a second test for curvature. If this test is statistically significant at the $\alpha_c/2$ level, then use a third phase to test for

[10] The letters correspond to the factor labels in Table 2, and the colon (:) between letters represents a two-way interaction between those factors.

quadratic effects. If it is not statistically significant, then the model picked in Step 7 is the final model.

9. **Phase 3: Size for a Model with Quadratic Effects.** Find the design matrix for the third phase using a one standard deviation effect size with power of $1 - \beta$ and α_3 significance. The model that we plan to fit in Phase 3 contains all statistically significant terms from Step 7, all quadratic effects for numeric factors, and any main effects needed to maintain hierarchy.
10. **Phase 3: Collect Data.** Collect the data for the test design generated in Step 9.
11. **Phase 3: Fit Final Model.** After selecting a design and fitting a model, maintain any statistically significant variables at the α_3 level.

C. Characterization versus Model Selection

The goal of our implementation of the DOE and SDOE is not to select exactly the correct model but to obtain a better idea of the design space while allowing some type I errors to be made in order to conservatively capture the "true" model as best as possible. Note that the planning and analysis procedure we describe is not nominally sized[11] for the null hypothesis associated with selecting the correct model. Consider the testing procedure outlined within Phase 1. This testing procedure is nominally sized for the null hypothesis of a single parameter being zero. However, we are using the same data set to test multiple parameters individually to determine whether they are different from zero. This necessarily results in a multiple hypothesis problem, producing larger than nominal type I error rates for the joint null hypothesis that all parameters are zero (Lovell 1983).[12] The distorted type I error implies that at least one variable will be included spuriously with a much higher probability than the significance level indicates. For this reason, it is important to note that this design is not intended to be used as a true model selection procedure.

Modification of this procedure for use as a true model selection procedure with nominal type I error for the null hypothesis that all parameters are jointly zero follows by instituting a Bonferroni correction within each phase.[13] This will necessarily increase the required sample size to maintain power for detecting parameters that are actually different from zero. In this sense, the procedure outlined here is likely to use a smaller sample size

---

[11] The type I error rate does not equal α.

[12] For example, a test based on a planned significance level $\alpha = 0.3$ for the null hypothesis that a single parameter is different from zero will produce an actual type I error rate $\tilde{\alpha} \approx 0.97$ associated with the null hypothesis that all parameters are jointly zero when separately testing 10 different parameters for inclusion in the model.

[13] Discussion of this and related considerations are beyond the scope of this paper. White (2000), Hansen (2005), and Romano and Wolf (2005) describe procedures that account for the multiple hypothesis problem in model selection.

than a true model selection procedure at the expense of including irrelevant covariates with a higher probability than was intended by the experimental design.

D. Simulation Study

To further motivate use of a sequential design strategy, we conducted a simulation study to compare models developed under a sequential planning approach to those developed under the traditional non-sequential planning approach.

We conducted the simulation for three true models (see Table 3) with random error standard deviations of 1, 3, and 4. In the traditional planning approach, we selected a design matrix that provided 80 percent power and 80 percent confidence for each coefficient value to fit a full model with all quadratic terms and two-way factor interactions. We used D-optimality criteria to select all design matrices in the simulation study. Each scenario was run on 1,000 data sets.

Table 3. True Models

Model	True Model Specification
1	$79 - 6B + 4D - 7.5F + 5A * F - 5.5B * D + 4.5D * F$ $+4D^2 - 9F^2$
2	$79 - 6B + 4D - 7.5F + 5A * F - 5.5B * D + 4.5D * F$
3	$79 - 6B + 4D + 5E - 7.5F$

Note. Variables are defined in Table 2.

Since the goal is to capture the true model while allowing for the selection of a few extra irrelevant factors, we are interested in the proportion of models that contain the true model and the average number of extraneous factors that are included in the final model. Table 4 shows these two measures for both the SDOE method and the traditional D-optimal design.

Table 4. 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.

Table 4 shows that the SDOE method produces models that include fewer extraneous factors than the traditional D-optimal method. The consequence of having fewer extraneous factors, holding all else constant, is that we leave out true factors more often in the SDOE. The SDOE method only contains the true model between 75 percent and 100 percent of the time, whereas the traditional D-optimal method always contains the true model. We are overpowered to find the coefficient effects in the traditional D-optimal method and at each phase of the SDOE. The lower power in the SDOE seems to originate in the test for curvature. The number of data sets where the correct model was contained in the final model closely aligns with the percentage of times we reached the correct phase, recorded in Table 5.

Table 5. Proportion of Data Sets Stopped After Each Phase for Each Model

	Model 1			Model 2			Model 3		
Phase	1	2	3	1	2	3	1	2	3
$\sigma = 1$	0.0%	9.1%	90.9%	0.0%	90.6%	9.4%	90.4%	1.8%	7.8%
$\sigma = 3$	0.3%	22.2%	77.5%	1.7%	91.1%	7.2%	91.4%	5.9%	2.7%
$\sigma = 4$	1.8%	22.7%	75.5%	7.1%	84.6%	8.3%	89.5%	6.8%	3.7%

The average sample size required under each set of conditions is always smaller for the SDOE than for the traditional D-optimal design when we assume that we will need to fit a model with main effects, quadratic effects, and all two-way interactions (see Figure 5). This is the model we assumed for the sequential design. If, however, the true model is known (for example, a main-effects-only model), a traditional D-optimal design is, on average, more efficient than an SDOE.

For each of the settings, the design size of the traditional D-optimal approach remains fixed at 184. For a sequential method, the sample size varies by the true model. Model 1 requires the largest sample size, but this is due to the fact that the true model includes quadratic effects, necessitating the use of all three phases of test points. To fit Model 2, the first two phases of test points are required, causing the sample size to be slightly larger on average than for Model 3. Model 3 only includes main effects and therefore was expected to stop after Phase 1 most of the time, leading to the smallest average sample size of the three models.

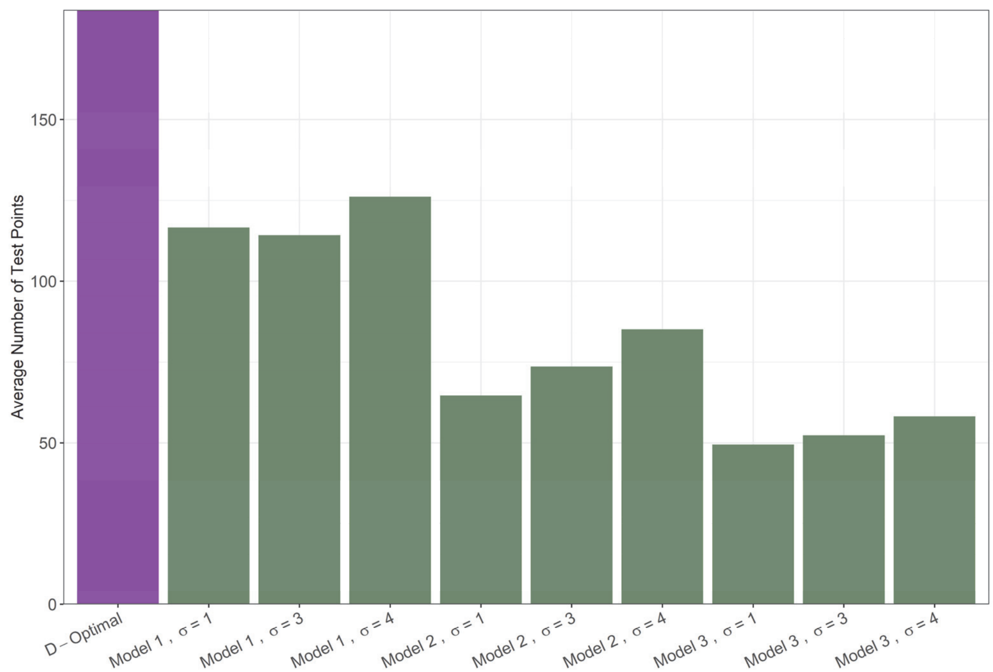


Figure 5. Average Total Size for Each Model

Finally, we see that including the test for curvature – that is, determining whether to continue to the next phase of test points – works as expected and that across all models our simulation study stops at the correct phase between 75 percent and 90 percent of the time. Table 5 reports the proportion of times the simulation study stopped at each phase for each model. Model 1 contains quadratic effects, so we expected it to stop after Phase 3. For Model 2, 85 percent of the experiments stopped after Phase 2, which is expected since Model 2 contains only main effects and interactions. For Model 3, 90 percent of the experiments stopped after Phase 1, which is expected because Model 3 contains only main effects. As the standard deviation increases, some error is introduced in the stopping phase.

6. Conclusions

We have introduced sequential analysis, described its current and potential uses in OT&E, and presented two methods for applying a sequential analysis to the T&E of defense systems. However, sequential procedures may prove to be more challenging to implement in DoD T&E than non-sequential procedures. For example, Avery and Simpson (2020) note that sequential procedures are challenging to use in DoD because the number of test points, conditions for those test points, and resources required to execute those test points are often decided early on and codified in the T&E strategy and test plans. Furthermore, sequential procedures may prove challenging to implement when the time required to score individual test events and perform analysis takes longer than the scheduled time between tests, and when stakeholders have divergent assessments of test points. When sequential methods can be applied, however, we find from our review and our examples that sequential procedures offer opportunities to make testing more efficient. Therefore, we recommend that test teams consider sequential methods when planning T&E.

References

Ahner, Darryl K., and Carl R. Parson. 2016. *Workshop Report: Test and Evaluation of Autonomous Systems* (STAT COE-Report-01-2016). Wright-Patterson Air Force Base, OH: Scientific Test and Analysis Techniques Center of Excellence.

Avery, Matthew R., and James R. Simpson. 2020. *How Much Testing Is Enough? 25 Years Later* (IDA Paper P-10994). Alexandria, VA: Institute for Defense Analyses.

Box, G. E. P., and K. B. Wilson. 1951. "On the Experimental Attainment of Optimum Conditions." *Journal of the Royal Statistical Society. Series B (Methodological)* 13 (1): 1–45.

Chernoff, Herman. 1959. "Sequential Design of Experiments." *Annals of Mathematical Statistics* 29: 755–70.

DoD (Department of Defense). 1996. *Handbook for Reliability Test Methods, Plans, and Environments for Engineering Development, Qualification, and Production* (MIL-HDBK-781A). Arlington, VA: DoD.

Freeman, Laura J., Thomas Johnson, Matthew Avery, V. Bram Lillard, and Justace Clutter. 2018. "Testing Defense Systems." In *Analytic Methods in Systems and Software Testing*, edited by Ron S. Kenett, Fabrizio Ruggeri, and Frederick W. Faltin, 439–487. Hoboken, NJ: John Wiley & Sons.

Ghosh, B. K. 2014. "Sequential Analysis – Historical." *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat02062

Hansen, Peter Reinhard. 2005. "A Test for Superior Predictive Ability." *Journal of Business & Economic Statistics* 23 (4): 365–80.

Johnson, N. L. 1961. "Sequential Analysis: A Survey." *Journal of the Royal Statistical Society: Series A (General)* 124 (3): 372–411.

Johnson, Thomas H., Laura Freeman, Janice Hester, and Jonathan L. Bell. 2014. "A Comparison of Ballistic Resistance Testing Techniques in the Department of Defense." *IEEE Access* 2: 1442–55.

Lai, Tze Leung. 2001. "Sequential Analysis: Some Classical Problems and New Challenges with Rejoinder." *Statistica Sinica* 11: 303–408.

Lovell, Michael C. 1983. "Data Mining." *The Review of Economics and Statistics* 65 (1): 1–12.

Medlin, Rebecca, John Dennis, Keyla Pagán-Rivera, and Leonard Wilkins. 2021. *A Review of Sequential Analysis* (IDA Document D-20487). Alexandria, VA: Institute for Defense Analyses.

Montgomery, Douglas C. 2017. *Design and Analysis of Experiments*. Hoboken, NJ: John Wiley & Sons.

National Research Council. 1998. *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements*. Washington, DC: National Academies Press.

Porter, Daniel J., Yevgeniya K. Pinelis, Chad M. Bieber, Heather M. Wojton, Michael O. McAnally, and Laura J. Freeman. 2019. "Operational Testing of Systems with Autonomy." *JSTOR*. http://www.jstor.com/stable/resrep22754.

Robbins, Herbert. 1952. "Some Aspects of the Sequential Design of Experiments." *Bulletin of the American Mathematical Society* 58 (5): 527–35.

Romano, Joseph R., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–82.

Simpson, James. 2018. *Testing via Sequential Experiments: Best Practice and Tutorial* (STAT COE-Report-01-2014). Wright-Patterson Air Force Base, OH: Scientific Test and Analysis Techniques Center of Excellence.

Simpson, James R., Charles M. Listak, and Gregory T. Hutto. 2013. "Guidelines for Planning and Evidence for Assessing a Well-Designed Experiment." *Quality Engineering* 25 (4): 333–55.

Wald, Abraham. 1945. "Sequential Tests of Statistical Hypotheses." *The Annals of Mathematical Statistics* 16 (2): 117–86.

Wallis, W. Allen. 1980. "The Statistical Research Group, 1942–1945." *Journal of the American Statistical Association* 75 (370): 320–30.

White, Hal. 2000. "A Reality Check for Data Snooping." *Econometrica* 68 (5): 1097–1126.

Appendix A.
Details of SPRT for Binomial Data

Wald's (1945) Sequential Probability Ratio Test (SPRT) is based on the likelihood ratio under the alternative versus the null hypotheses, and the rejection bounds are estimated as a function of the desired type I and type II error rates (α and β, respectively). Suppose the hypotheses one wishes to test are

$$H_0: p = p_0$$
$$H_1: p = p_1$$

where p_0 is the proportion under the null hypothesis and p_1 is the proportion under the alternative hypothesis. Assuming each experiment is an independent and identically distributed Bernoulli trial and the aggregated data come from a binomial distribution with x successes out of n experiments, the likelihood takes the form $L_i(x) = \binom{n}{x} p_i^x (1 - p_i)^{n-x},\ i = 0{,}1$. This gives the following likelihood ratio:

$$\frac{L_1(x)}{L_0(x)} = \frac{\binom{n}{x} p_1^x (1 - p_1)^{n-x}}{\binom{n}{x} p_0^x (1 - p_0)^{n-x}}$$
$$= \frac{p_1^x (1 - p_1)^{n-x}}{p_0^x (1 - p_0)^{n-x}}$$

Wald's method requires calculating the likelihood ratio after each observation is collected. The likelihood ratio after m experiments are conducted is

$$\frac{L_{1m}}{L_{0m}} = \frac{p_1^{d_m} (1 - p_1)^{m-d_m}}{p_0^{d_m} (1 - p_0)^{m-d_m}}$$
$$= \left(\frac{p_1}{p_0}\right)^{d_m} \left(\frac{1 - p_1}{1 - p_0}\right)^{m-d_m}$$

where $d_m = \sum_{i=1}^{m} x_i$, the total observed successes in the first m experiments. Wald's SPRT method establishes that when $\frac{L_{1m}}{L_{0m}} \geq \frac{1-\beta}{\alpha}$, one rejects H_0, and when $\frac{L_{1m}}{L_{0m}} \leq \frac{\beta}{1-\alpha}$, one fails to reject H_0. However, if $\frac{\beta}{1-\alpha} \leq \frac{L_{1m}}{L_{0m}} \leq \frac{1-\beta}{\alpha}$, then there is not yet enough evidence to make a decision. In these situations, one continues running experiments until the likelihood ratio falls outside of the two bounds. One can use the log-likelihood to obtain the bounds for the SPRT and then compare the total number of successes after each experiment to these bounds. These bounds are a function of $\alpha,\ \beta,\ p_0$ and p_1 and can be expressed as

$$\frac{\log\frac{\beta}{1-\alpha}+m\log\frac{1-p_0}{1-p_1}}{\log\frac{p_1}{p_0}-\log\frac{1-p1}{1-p0}} \leq d_m \leq \frac{\log\frac{1-\beta}{\alpha}+m\log\frac{1-p_0}{1-p_1}}{\log\frac{p_1}{p_0}-\log\frac{1-p1}{1-p0}}$$

Neyman-Pearson Likelihood Ratio Test versus Exact Binomial

The Neyman-Pearson Likelihood Ratio Test tests the following set of hypotheses:

$$H_0: X_1, \ldots, X_n \sim f_{\theta_0}(x)$$

$$H_1: X_1, \ldots, X_n \sim f_{\theta_1}(x)$$

To conduct this test, the likelihood ratio, $L(x) = \frac{f_0(x)}{f_1(x)}$, is compared to some critical constant, c^*. For a binomially distributed random variable, $X = \sum X_i \sim \text{Binomial}(n, p)$, these hypotheses reduce to

$$H_0: p = p_0$$

$$H_1: p = p_1$$

The likelihood ratio takes the following form:

$$L(x) = \left(\frac{p_0/(1-p_0)}{p_1/(1-p_1)}\right)^x \left(\frac{1-p_0}{1-p_1}\right)^n$$

where x is the observed number of successes. If $p_0 < p_1$, the likelihood ratio is a decreasing function of x. One conducts the Neyman-Pearson Likelihood Ratio Test by comparing the likelihood to a chosen threshold value: $L(x) < c^*$. Equivalently, since $L(x)$ is monotonic in x, we can choose a critical value, c, to compare to x. The critical value c is chosen by first setting an α value and then selecting c so that $P(X > c|p_0) \leq \alpha$. One rejects the null hypothesis if the count of successes in the observed data is larger than the critical value; that is, if $x > c$.

For comparison, the most common method for conducting the exact binomial test involves comparing the p-value associated with the count of successes, $P(X > x|p_0)$, to the significance level, α. The null hypothesis is rejected when $P(X > x|p_0) < \alpha$. For a chosen critical value or significance level, the two approaches are equivalent.

Note that the discrete support of a binomial random variable often prevents us from testing at exactly the α level of significance. For example, let $n = 20$ and $p_0 = 0.1$, then $P(X > 3|p_0 = 0.1) = 0.133$ and $P(X > 4|p_0 = 0.1) = 0.043$. If we test at the $\alpha = 0.1$ level of significance, the test cannot achieve the desired significance level with the critical value $c = 3$ and the critical value $c = 4$ is too conservative.

Appendix B.
Details of Test for Curvature and Its Power Analysis

The test for curvature or lack of fit, conducted via an F test, requires replicated factor points to test whether unexplained variation can be accounted for by including additional terms in the model. Lack of fit refers to the terms that we could have fit to the model but chose not to fit. Let c represent the number of distinct values of x, n the total sample size used for the regression, and n_i the number of observations under the factor vector x_i. The sum of squares due to lack of fit (SSLF) and sum of squares pure error (SSPE) are defined as follows:

$$SSLF = \sum_{i=1}^{c} n_i(\bar{y}_i - \hat{y}_i)^2$$

$$SSPE = \sum_{i=1}^{c}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2$$

Note that $SSE = SSLF + SSPE$, where SSE is the residual sum of squares.

$SSLF$ evaluates the average squared distance from the predicted values to the average of the y_i values for the particular value of the vector x_i. $SSPE$ evaluates the average squared distance from the observed data to the average of the y_i values for the particular value of the vector x_i.

The null hypothesis for this test is that the group of terms being considered for inclusion in the model do not improve the model's fit to the data. Under the null hypothesis, the test statistic, $F = \frac{SSLF/(c-(\mathrm{p}))}{SSPE/(n-c)}$, follows an F distribution with $c - p$ numerator degrees of freedom and $n - c$ denominator degrees of freedom. In this equation, p represents the number of parameters in the model.

Under the alternative, the test statistic follows an F distribution with $c - p$ numerator degrees of freedom, $n - c$ denomenator degrees of freedom, and a non-centrality parameter $\lambda = \frac{\sum_i (E(y_i|x_i) - \beta\prime x_i)^2}{\sigma^2}$.

An assumption regarding λ must be made to calculate power for this F test. λ is the ratio of $E(SSLF)$ and the true variance. One must assume the smallest value of $E(SSLF)$ that can result in a meaningful amount of unexplained variation. Based on this assumption, a root-finding algorithm is used to find the number of replicated center points, r, needed to achieve the desired power. The final sample size is calculated as $c + r$. Choosing which values to replicate is left to the researcher. For the simulation in this paper, the replicates were drawn from a random sample of center points.

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**	
12-2021	Final	**START DATE**	**END DATE** Dec 2021

4. TITLE AND SUBTITLE
Case Study on Applying Sequential Analyses in Operational 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**
C9082	C9082	

6. AUTHOR(S)
Monica L. Ahrens (OED); Rebecca M. Medlin (OED); Keyla Pagan-Rivera (OED); John W. Dennis (SFRD)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	**8. PERFORMING ORGANIZATION REPORT NUMBER**
Institute for Defense Analyses 730 E. Glebe Road Alexandria, Virginia 22305	NS D-32904 H 2021-000468

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**	**11. SPONSOR/MONITOR'S REPORT NUMBER**
Central Research Project	AE/CRP	

12. DISTRIBUTION/AVAILABILITY STATEMENT
Public release approved. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Sequential Analysis is concerned with statistical evaluation in situations for which the number, pattern, or composition of the data is not determined at the start of the investigation, but instead depends upon the information acquired throughout the course of the investigation. The formal start of Sequential Analysis originates in ballistics testing for the Department of Defense (DOD) (Wald 1945; Wallis 1980), but we find that the use of sequential analysis is underutilized in the DOD. Expanding the use of sequential analysis has the potential to produce cost savings and reduce test time. In this paper, we provide a brief introduction to Sequential Analysis, describe its current and potential uses in operational test and evaluation (OT&E), and present a method for applying a sequential analysis to the test and evaluation of defense systems. We perform simulation studies and apply the method to a case study to evaluate the proposed method. Additionally, we discuss challenges that will need to be addressed for sequential analysis that are relevant for OT&E.

15. SUBJECT TERMS
Design of Experiments (DOE); Sequential Design; Sequential Probability Ratio Test; Test and Evaluation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	SAR	39

19a. NAME OF RESPONSIBLE PERSON	**19b. PHONE NUMBER**
Rebecca Medlin	703-845-6731

PREVIOUS EDITION IS OBSOLETE. **STANDARD FORM 298 (REV. 5/2020)**
Prescribed by ANSI Std. Z39.18