Taylor & Francis
Taylor & Francis Group

Quality Engineering

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lqen20

Case study on applying sequential analyses in operational testing

Monica Ahrens, Rebecca Medlin, Keyla Pagán-Rivera & John W. Dennis

To cite this article: Monica Ahrens, Rebecca Medlin, Keyla Pagán-Rivera & John W. Dennis (2023) Case study on applying sequential analyses in operational testing, Quality Engineering, 35:3, 534-545, DOI: 10.1080/08982112.2022.2146510

To link to this article: https://doi.org/10.1080/08982112.2022.2146510

Published online: 12 Dec 2022.

Submit your article to this journal

Article views: 308

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lqen20

QUALITY ENGINEERING
2023, VOL. 35, NO. 3, 534–545
https://doi.org/10.1080/08982112.2022.2146510

CASE STUDY

Case study on applying sequential analyses in operational testing

Monica Ahrens[a], Rebecca Medlin[b], Keyla Pagán-Rivera[b], and John W. Dennis[c]

[a]Virginia Tech Center for Biostatistics and Health Data Science, Roanoke, Virginia; [b]Operational Evaluation Division, Institute for Defense Analyses, Alexandria, Virginia; [c]Strategy, Forces and Resources Division, Institute for Defense Analyses, Alexandria, Virginia

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)and it is widely used in other disciplines, 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 for sequential analysis in OT&E. Lastly, while operational testing is the focus in this paper, the methodology presented is applicable to campaigns of experimentation and general testing across numerous disciplines.

KEYWORDS
Design of experiments; sample size determination; sequential probability ratio test; statistical power

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; Johnson 1961; Ghosh 2014).

Although Sequential Analysis has its formal origin in Department of Defense (DoD) testing (Wald 1945; Wallis 1980), it has been underused in recent test and evaluation of military systems. Exceptions to this include the myriad of sequential testing techniques[1] used in DoD ballistic resistance sensitivity testing to estimate a particular probability of perforation (Johnson et al. 2014). Expanding the use of sequential analysis to other areas of DoD testing may save money and reduce test time (National Research Council 1998).

Medlin et al. (2021) subdivides the field of Sequential Analysis into three broad functional categories: sequential testing, sequential design, and sequential estimation. These categories are not mutually exclusive in the sense that design procedures often include an objective related to testing or estimation. The problems categorized as sequential testing and estimation only allow the number of observations to depend upon information acquired throughout the investigation. Design problems increase the complexity of the sequential procedure by allowing elements affecting the pattern and composition of the observations to depend upon acquired information as well (Govindarajulu 1975). In this paper, we focus on the methods for sequential testing and sequential design. Even though these methods can be used to benefit a variety of disciplines, 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 of 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

CONTACT Rebecca Medlin ✉ rmedlin@ida.org Operational Evaluation Division, Institute for Defense Analyses, Alexandria, VA, USA.

[1] Up and Down Method (UD), Langlie Method (LM), Delayed Robbins Monroe Method (DRM), Wu's three-phased approach (3POD), Neyer's Method (NM), the Robbins Monroe Joseph Method (RMJ), and K-in-a-row (KR).

© 2022 Institute for Defense Analyses

4 and 5 include a simulation study to compare 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. Historical context & 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 posted to the Statistical Research Group (SRG) by Captain Garret L. Schulyer of the Bureau of Ordnance, Navy Department, who was interested in calculating the probability of a hit by an anti-aircraft fire on a directly approaching dive bomber.[2] 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).[3] 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; however, many 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), Box and Liu (1999), Box (1999), Myers, Montgomery, and Anderson-Cook (2016), and Montgomery (2020) all allude to the fact 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; in this paper, we will refer to this type of experimental campaign as sequential DOE (SDOE).

Regarding how to implement a SDOE strategy, Montgomery (2020) 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 (2018) 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. We propose and illustrate a method for planing SDOE, similar to the recommendations of Montgomery (2020) and Simpson (2018). We expand on the work of Simpson (2018) by illustrating how one might size each phase of test for a specified confidence and power. Additionally, we briefly discuss some details regarding the sizing of such a test.

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 the use of 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 arrives 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.

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

[3] The T&E community may be more familiar with the SDOE planning approach described by Box and Wilson (1951) and Montgomery (2020), 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).

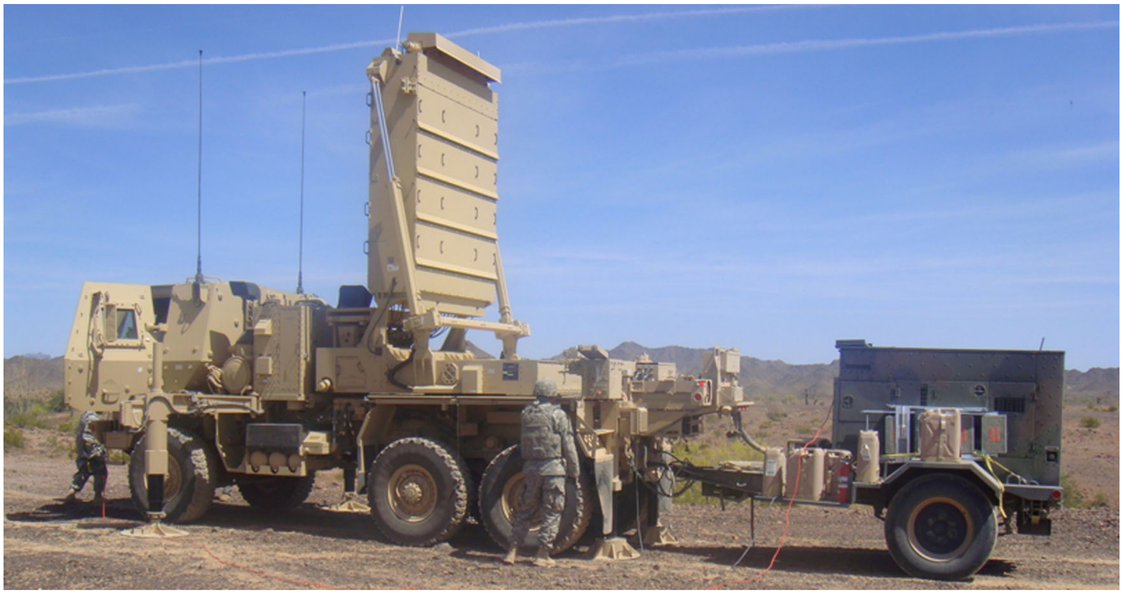

Figure 1. Soldiers emplacing the AN/TPQ-53 Counterfire Radar during operational testing.

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.

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:

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?

A response variable related to (1) could be whether the Q-53 detected the shot or not (a binary response), and the miss distance, a continuous response, could be a response related to (2).

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 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 only after collection was complete, and the results of the hypothesis test reported. While this method proves reliable for providing power and confidence to detect the desired difference from 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.

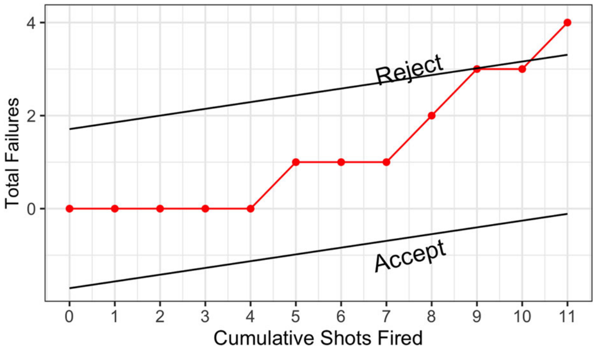

Figure 2. Illustration of SPRT process for a true value $p = p_1$. After the 11th observation, the test indicates we have collected enough information to reject the null hypothesis. For comparison, the required sample size for a non-sequential approach is 50 test points.

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 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, a primary interest is to know if the detection failure rate[4] differs from what is expected. For example, we might expect the failure rate to be p_0, and we want to test if the failure rate is larger by some Δ,[5] say 10%. In this case, our alternative hypothesis would be specified by $p_1 = p_0 + \Delta$:

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

Figure 2 illustrates the application of the SPRT for a binary response 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 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 $(p \leq p_0)$ in favor of the alternative hypothesis $(p \geq p_1)$ and to state the failure rate is higher than expected. We provide the SPRT mathematical details in Appendix A.

Table 1. Observed error rates and sample sizes for the SPRT compared to the exact binomial test.

	Error Rates		Average Sample Sizes (SD)	
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

True failure proportions $p = p_0$ and $p = p_1$, and the type I and type II error rate set to 20%.

4.1. Simulation study

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

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 a set of experimental factors that span 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

[4] Failure rate = proportion of times radar fails to detect the incoming projectile.

[5] $\Delta > 0$ is the effect size, which can be thought of as a statement regarding an acceptable level of risk; experimental setups are optimized for detection of deviations from the null hypothesis of at least Δ. Detecting large deviations from the null hypothesis is easier than detecting small deviations, so smaller effect sizes are associated with larger sample sizes needed to make a decision. Deviations less than Δ will still be detected, but the error rate for detecting small deviations will be larger than our chosen error rate thresholds. This is a necessary trade off in test planning, and testers should utilize subject matter expertise to determine an acceptable value of Δ.

[6] Simulation settings: true failure proportions $p = p_0$ and $p = p_1$, and type I and type II error rate set to 20%.

[7] Note, 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 Neyman-Pearson LRT (see Appendix A).

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 MIL-HDBK-781A. (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 MIL-HDBK-781A presents the use of a 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 MIL-HDBK-781A (1996, 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."

Table 2. Q-53 test design factors.

Design Factors	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

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 the detection and accuracy performance of the Q-53 (Freeman 2020). The response variable of interest is miss distance, which is a continuous response. Figure 3 depicts a standard fire mission for the Q-53. During a threat fire mission, the threat will fire projectiles at a target inside the search area of the Q-53. Figure 3 shows the Q-53 operating in a 90-degree mode, so its search sector is limited to the area withing the black bars (in this example, between North and East). The specific geometry of the scenario may impact the ability of the Q-53 to track the projectile and estimate the threat's position.

In this section, we present an SDOE strategy for the test and evaluation of the Q-53. The strategy we present allows testers to fully characterize the system's performance across the set of experimental factors that span the operational test space and to potentially do so with fewer test points than if a non-SDOE approach had been employed.

5.1. Design of experiments

Full-factorial designs are often paired with the test goal of characterization, which is a common goal

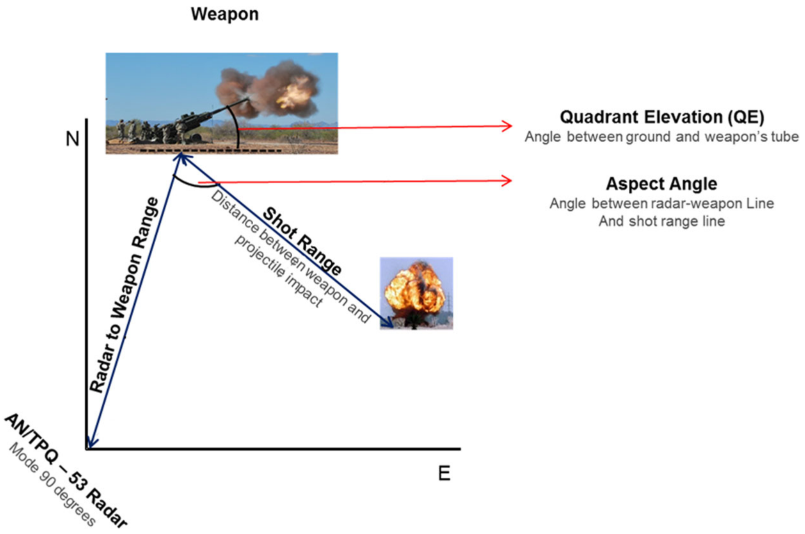

Figure 3. Example of fire mission including relevant geometric factors impacting Q-53 system performance. During a threat fire mission, the threat will fire projectiles at a target inside the search area of the Q-53. In this figure, the Q-53 radar is operating in a 90-degree mode, and so its search sector is limited to the area within the black bars. Source: Freeman et al. (2018).

when designing an experiment for an operational test.[9] 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 2020), though potentially prohibitively costly when many factors are involved. One could also consider using a fractional factorial design. However, these designs are more frequently used in applications with two-level factors. Our example contains a mixed-level factor set, with (typically categorical) factors having more than 2 levels. Optimal designs are the appropriate design technique for this situation, which is often encountered when planning for an operational test.

Optimal designs are frequently used when planning for an operational test. They are especially useful when the factor space includes categorical factors and the number of test points is constrained to preclude a full-factorial design. Myers, Montgomery, and Anderson-Cook (2016) note three reasons for the appropriateness of leveraging an optimal design over a classical design like a factorial or fractional factorial: 1) an irregular experimental region, 2) a nonstandard model to include categorical design factors, and 3) unusual sample or block size requirements. Because operational testing often involves at least one of the three conditions we center our example around the use of an optimal 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 2020).

Using an optimal design can result in a more efficient DOE approach relative to classical designs. However, there are tradeoffs 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 planning for main effects, two-way interactions, and quadratic effects may be necessary. In our example, to account for all possible effects and to achieve adequate power and confidence (for example, a signal-to-noise ratio of 1, 80 percent confidence, and 80 percent power), the design requires 184 test points.[10]

5.2. 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.

Common sequential planning approaches include point-by-point sequential designs (Johnson et al. 2014) and experimental campaign sequential designs. The specific sequential approach will depend on the research goals and test limitations. We implement an experimental campaign by application of 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, like Simpson (2018), we use three phases to follow the natural progression of screening, decoupling aliased interactions, and augmenting for higher order terms. A different number of phases may be appropriate depending on the logistics and goals of the test. Additionally, in practice, one might consider incorporating blocking in the design and analysis if variation between test phases is expected (Myers, Montgomery, and Anderson-Cook 2016; Montgomery 2020). 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. Because the response, miss distance, is continuous, we fit our statistical model in each

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

[10] We encourage using subject matter experts (SMEs) in the test planning and design process. For example, a SME may be able to eliminate, based on their knowledge of the system, the need to plan for certain two-way interactions and quadratic effects.

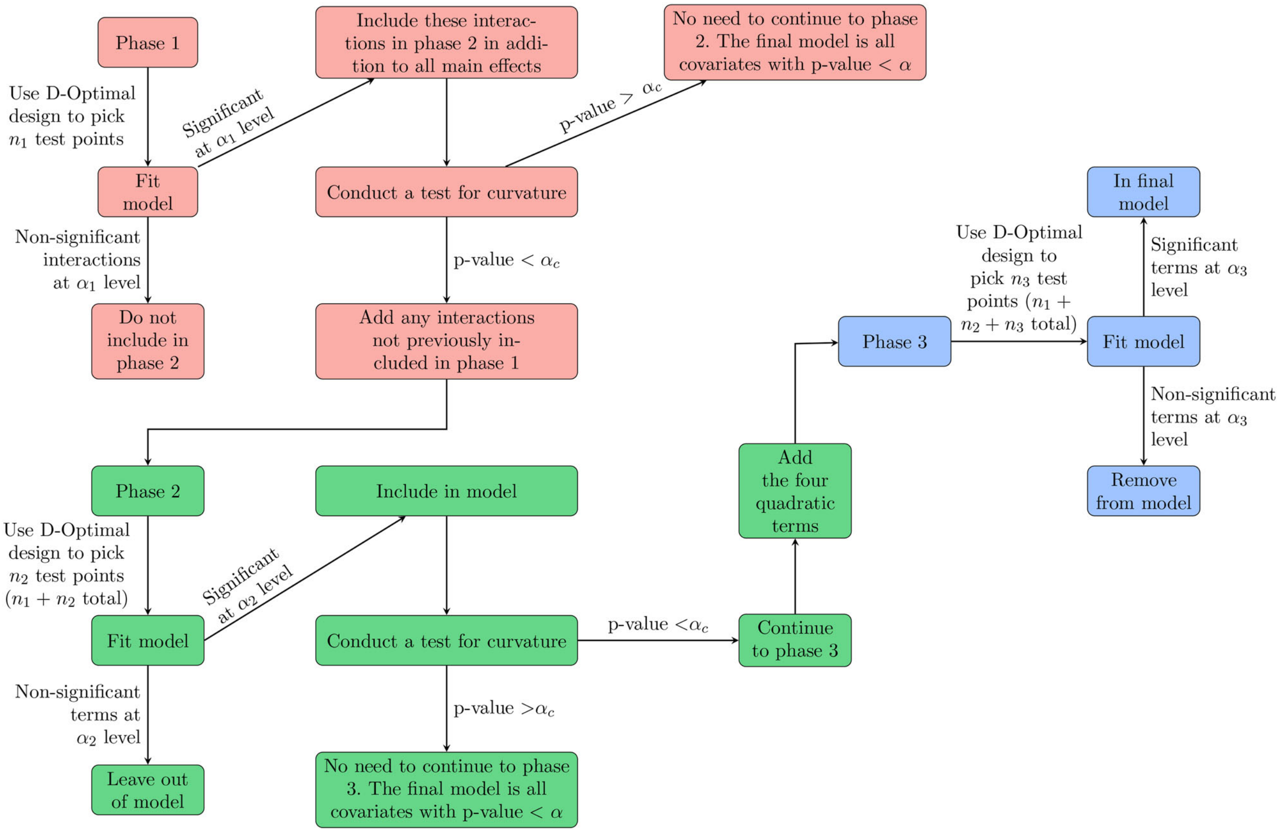
Phase 1
Include these interactions in phase 2 in addition to all main effects
No need to continue to phase 2. The final model is all covariates with p-value < α
Use D-Optimal design to pick n_1 test points
Significant at α_1 level
p-value > α_c
Fit model
Conduct a test for curvature
In final model
Use D-Optimal design to pick n_3 test points ($n_1 + n_2 + n_3$ total)
Significant terms at α_3 level
Non-significant interactions at α_1 level
p-value < α_c
Do not include in phase 2
Add any interactions not previously included in phase 1
Phase 3
Fit model
Non-significant terms at α_3 level
Add the four quadratic terms
Phase 2
Include in model
Remove from model
Use D-Optimal design to pick n_2 test points ($n_1 + n_2$ total)
Significant at α_2 level
Fit model
Conduct a test for curvature
p-value <α_c
Continue to phase 3
Non-significant terms at α_2 level
p-value >α_c
Leave out of model
No need to continue to phase 3. The final model is all covariates with p-value < α

Figure 4. The SDOE planning process used in this paper.

phase using least squares estimation, but the same process would apply for a generalized linear model. Steps 1–11 below outline the specific steps we followed for the Q-53 radar sequential test design example.

We use a D-optimal design to select our design matrix for each phase of test points and use the skpr library in R to generate the design matrices and calculate power. In planning, we allowed for different significance levels to size each phase, labeled α_1, α_2, α_3. For example, in the simulation study in Section 5.4, we use $\alpha_1 = 30\%$ significance for the first phase, $\alpha_2 = 15\%$ for the second phase, and $\alpha_3 = 15\%$ for the third phase. We allow for "moderately significant" coefficients to be retained early in testing with further examination as testing progresses. In the steps below, α_c represents the significance level for a test for curvature. Since two different tests for curvature are conducted, a Bonferoni correction is needed and significance level of $\alpha_c/2$ is used for each test for curvature. In the simulation, we use $\alpha_c = 20\%$. The test team should discuss pros and cons associated with determination of the 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: A, B, C, D, E, F, $A:B$, $A:C$, $A:D$, $A:E$, $A:F$, $B:D$, $B:F$.[11] Include replication test points in order to test for a lack of fit. Details for 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 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.

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

(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 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 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 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.

5.3. 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 for the null hypothesis associated with selecting the correct model.[12] 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).[13] 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.[14] 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 than a true model selection procedure at the expense of including irrelevant covariates with a higher probability than was intended by the experimental design.

5.4. 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. Our simulation results reveal the following properties: the SDOE models include fewer extraneous factors, the SDOE usually reaches the correct phase of testing, and the SDOE requires on average a smaller sample size.

The simulation study considers three true models (see Table 3) with random error standard deviations of 1, 3, and 4. Model 1 matches the truth model used by Simpson (2018), where care was taken to make a truth model characteristic of a "standard" 2nd order model encountered in practice. Simpson (2018) notes that aspects of the truth model that are consistent with many historical defense systems are effect sparsity, model term heredity, and quadratic behavior. For the traditional planning approach, we selected a design

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

[13] 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.

[14] Discussion of this and other related considerations are beyond the scope of this paper; we will instead discuss these considerations in more detail in a separate paper. White (2000), Hansen (2005), and Romano and Wolf (2005) each provide procedures that account for the multiple hypothesis problem in model selection.

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.

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 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 and a smaller sample size on average, holding all else constant, is that we leave out true factors more often in the SDOE. However, simulations indicate that our SDOE method selects a model containing the correct model often,-with the frequency of selecting the correct model being affected by the noise in the experiment. 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 3. True models.

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

Variables are defined in Table 2.

Table 4. Average number of extra factors included in final model and number of times the correct model was contained in the final model across all settings. “CCM” is “Contained Correct Model” and denotes the number of data sets where the correct model was contained in the final selected model.

		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

Table 5 reports the proportion of times the simulation study stopped at each phase for each model. From the results in this table, we see that including the test for curvature – that is, determining whether to continue to the next phase of test points or to stop collecting data – 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. As the standard deviation increases, some error is introduced in the stopping phase. The highlighted cells in this table represent the target phase for each model. For example, consider Model 1 with $\sigma = 4$: the model contains quadratic effects, and our simulation reveals that our SDOE procedure correctly proceeds to Phase 3 in 76% of our simulation trials. Similarly, Model 2 contains only main effects and interactions, and our SDOE procedure correctly stops at Phase 2 in 90.6% of our simulation trials when $\sigma = 1$.

Finally, the average and median sample size required under each set of conditions is always smaller for the

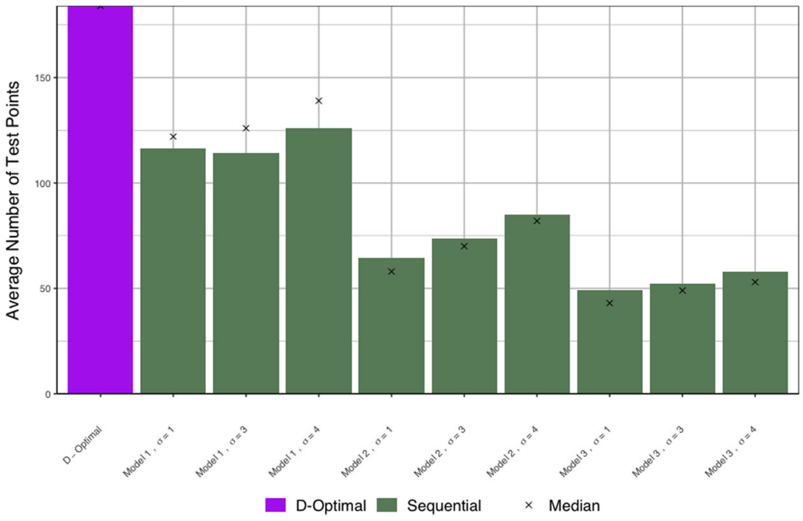

Figure 5. Mean and median total size for each model.

Table 5. Reported 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 highlighted area represents the target phase for each model.

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. 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, but knowing the true model seems unlikely in practice.

Figure 5 presents a comparison of the average and median sample size of the SDOE to the traditional D-optimal design. For each simulation setting, the design size of the traditional D-optimal approach remains fixed at 184. For the SDOE, the sample size varies depending on 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 achieve 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 expect the SDOE to stop after Phase 1 most of the time, leading to the smallest average sample size of the three models.

6. Conclusions

We introduced Sequential Analysis, described its current and potential uses in DoD T&E, and presented two methods for applying a sequential analysis to the OT&E of defense systems. While our focus was on OT&E, the methodology presented is applicable to campaigns of experimentation and general testing across numerous disciplines.

Sequential procedures may prove to be more challenging to implement in DoD T&E than non-sequential procedures. For example, Avery, Simpson, and Wojton (2021) 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.

About the authors

Dr. Monica Ahrens, PhD is a Research Scientist at Virginia Tech in the Center for Biostatistics and Health Data Science. She completed a summer internship with the Institute for Defense Analyses in 2022 and received her PhD from the University of Iowa Department of Biostatistics in 2022.

Dr. Rebecca Medlin, PhD is a research staff member in the Operational Evaluation Division at the Institute for Defense Analyses. She supports the Director, Operational Test and Evaluation on training, research and applications of statistical methods for the planning and evaluation of military systems. She received her PhD in Statistics from Virginia Tech in 2014.

Dr. Keyla Pagán-Rivera has a Ph.D. in Biostatistics from The University of Iowa and serves as a Research Staff Member in the Operational Evaluation Division at the Institute for Defense Analyses. She supports the Director, Operational Test and Evaluation on training, research and applications of statistical methods for the planning and evaluation of military systems.

Dr. John W. Dennis, PhD is a research staff member focusing on Econometrics, Statistics, and Data Science with the Institute for Defense Analyses' Human Capital Group. He received his PhD in Economics from the University of North Carolina at Chapel Hill in 2019.

References

Ahner, D. K., and C. R. Parson. 2016. Workshop report: Test and evaluation of autonomous systems. https://www.afit.edu/stat/statcoe_files/TE/%20Auto/%20Syst/%20Workshop/%20DASDDTE/%20Memo.pdf.

Avery, M. R., J. R. Simpson, and H. M. Wojton. 2021. Determining how much testing is enough: An exploration of progress in the department of defense test and evaluation community. *The ITEA Journal of Test and Evaluation* 42 (1):39–47.

Box, G. E. 1999. Statistics as a catalyst to learning by scientific method part II–a discussion. *Journal of Quality Technology* 31 (1):16–29. doi:10.1080/00224065.1999.11979890.

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):1–38. doi:10.1111/j.2517-6161.1951.tb00067.x.

Box, G. E., and P. Y. Liu. 1999. Statistics as a catalyst to learning by scientific method part I–an example. *Journal of Quality Technology* 31 (1):1–15. doi:10.1080/00224065.1999.11979889.

Chernoff, H. 1959. Sequential design of experiments. *The Annals of Mathematical Statistics* 30 (3):755–70. doi:10.1214/aoms/1177706205.

Freeman, L. 2020. Test and evaluation for artificial intelligence. *Insight* 23 (1):27–30. doi:10.1002/inst.12281.

Freeman, L. J., T. Johnson, M. Avery, V. B. Lillard, and J. Clutter. 2018. Testing defense systems. In *Analytic methods in systems and software testing* eds. R. S. Kenett, F. Ruggeri and F. W. Faltin. Wiley. doi:10.1002/9781119357056.ch18.

Ghosh, B. 2014. Sequential analysis-historical. In *Wiley StatsRef: Statistics Reference Online*, eds. N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels. Wiley. doi:10.1002/9781118445112.stat02062.

Govindarajulu, Z. 1975. *Sequential statistical procedures*. New York, NY: Academic Press.

Hansen, P. R. 2005. A test for superior predictive ability. *Journal of Business & Economic Statistics* 23 (4):365–80. doi:10.1198/073500105000000063.

Johnson, N. L. 1961. Sequential analysis: A survey. *Journal of the Royal Statistical Society: Series A (General)* 124 (3): 372–411. doi:10.2307/2343243.

Johnson, T. H., L. Freeman, J. Hester, and J. L. Bell. 2014. A comparison of ballistic resistance testing techniques in the department of defense. *IEEE Access* 2:1442–55. doi:10.1109/ACCESS.2014.2377633.

Lai, T. L. 2001. Sequential analysis: Some classical problems and new challenges with rejoinder. *Statistica Sinica* 11: 303–408.

Lovell, M. C. 1983. Data mining. *The Review of Economics and Statistics* 65 (1):1–12. doi:10.2307/1924403.

Medlin, R., J. Dennis, K. Pagán-Rivera, and L. Wilkins. 2021. A review of sequential analysis. Institute for Defense Analyses D-20487. https://idalink.org/D-20487.

MIL-HDBK-781A. 1996. Reliability test methods; plans, and environments for engineering development, qualification, and production. Washington, DC.

Montgomery, D. C. 2020. *Design and analysis of experiments*. 10th ed. Hoboken, NJ: John Wiley & Sons.

Myers, R. H., D. C. Montgomery, and C. M. Anderson-Cook. 2016. *Response surface methodology: Process and product optimization using designed experiments*. Hoboken, NJ: John Wiley & Sons.

National Research Council. 1998. Statistics, testing, and defense acquisition: New approaches and methodological improvements. Washington, DC: The National Academies Press.

Porter, D. J., Y. K. Pinelis, C. M. Bieber, H. M. Wojton, M. O. McAnally, and L. J. Freeman. 2019. Operational testing of systems with autonomy. Institute for Defense Analyses. http://www.jstor.org/stable/resrep22754.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58 (5):527–35. doi:10.1090/S0002-9904-1952-09620-8.

Romano, J. R., and M. Wolf. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica* 73 (4): 1237–82. doi:10.1111/j.1468-0262.2005.00615.x.

Simpson, J. 2018. Testing via sequential experiments best practice and tutorial. https://www.afit.edu/stat/statcoe_files/Testing_via_Sequential_Experiments_Best_Practice.pdf.

Wald, A. 1945. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* 16 (2):117–86. doi:10.1214/aoms/1177731118.

Wallis, W. A. 1980. The statistical research group, 1942–1945. *Journal of the American Statistical Association* 75 (370):320–30.

White, H. 2000. A reality check for data snooping. *Econometrica* 68 (5):1097–126. doi:10.1111/1468-0262.00152.

Appendix A. Details of SPRT for binomial data

Wald's SPRT (Wald 1945) is based on the likelihood ratio under the alternative versus the null, 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 we wish 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, the aggregated data comes from a binomial distribution with x successes out of n experiments, the likelihood is of 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 $d_m = \sum_{i=1}^m x_i$, is the total observed successes in the first m experiments. Wald's SPRT method establishes that when $\frac{L_{1m}}{L_{0m}} \geq \frac{1-\beta}{\alpha}$ we reject H_0, but when $\frac{L_{1m}}{L_{0m}} \leq \frac{\beta}{1-\alpha}$ we fail 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 us 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 α, β, p_0 and p_1, and could be expressed as

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

A.1. *neyman pearson likelihood ratio test vs. 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,

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

Where $p_0 < p_1$. The likelihood ratio is of 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 in x. One conducts the 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 success, $P(X > x|p_0)$, 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 level of significance with the critical value $c = 3$ and the critical value $c = 4$ is too conservative.

Appendix B. Details for test for curvature and its power analysis

The test for curvature or lack of fit, conducted using 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 number of points used for the regression, and n_i the number of observations under the factor vector x_i. The sum of squares of lack of fit (SSLF) and sum of squares pure error (SSPE) are defined as follows:

$$SSLF = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(\bar{y}_i - \hat{y}_i\right)^2 \tag{B1}$$

$$SSPE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 \tag{B2}$$

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

SSLF evaluates the average vertical 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 vertical 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 there is no unexplained curvature from the model. Under the null hypotheses, the test statistic, $F = \frac{MSLF}{MSPE} = \frac{SSLF/(c-p-1)}{SSPE/(n-c)}$ follows an F distribution with $c - 2$ numerator degrees of freedom and $n - c$ denominator degrees of freedom

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

An assumption regarding λ must be made to calculate power for this F test, where 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 get 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 where drawn from a random sample of center points.