INSTITUTE FOR DEFENSE ANALYSES

A Review of Sequential Analysis

Heather Wojton, Project Leader

Rebecca Medlin
John Dennis
Keyla Pagan-Rivera
Leonard Wilkins

December 2020

Approved for Public Release.

Distribution Unlimited.

IDA Document NS D-20487

Log: H 2020-000515

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BD-9-2299(90), "Test Science Applications," 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. Allison Goodman, Dr. Curtis Miller, Dr. Elliot Bartis, Dr. John Haman, and Dr. Stephen DeVito from the Operational Evaluation Division.

For more information:
Heather Wojton, Project Leader
hwojton@ida.org • 703-845-6811

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

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-20487

A Review of Sequential Analysis

Heather Wojton, Project Leader

Rebecca Medlin
John Dennis
Keyla Pagan-Rivera
Leonard Wilkins

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Executive Summary

Introduction

Sequential analysis concerns statistical evaluation in 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 course of the investigation. Although the formal genesis of sequential analysis originated in ballistics testing during World War II for the Department of Defense (DoD) (Wald, 1945; Wallis, 1980), we find that it has been underused in recent DoD testing. Expanding the use of sequential analysis has the potential to save a lot of money and reduce test time (National Research Council, 1998). This paper summarizes the literature on sequential analysis and offers fundamental information for providing recommendations for its use in DoD test and evaluation.

A first step in establishing a roadmap for the applied use of sequential analysis in DoD test and evaluation may be to consider its use as it applies to clinical trials, and to review the guidance of the FDA as a means of setting standards. Sequential procedures are frequently used in clinical trials. Periodically, the FDA releases guidance that represents their current thinking on the topic. In their most recent guidance,[1] the FDA lists the following advantages of using sequential procedures: statistical efficiency, ethical considerations, improved understanding, and acceptability to stakeholders. DoD testing and clinical trials share many similarities. Both require careful planning, and practitioners in each express great interest in reducing the number of test events, speeding up testing, and saving money without sacrificing information needed to support a decision.

This literature review is organized by broad categories within sequential analysis, and concludes with a list of annotated references. To summarize the topic, we subdivide the field of sequential analysis into three broad functional categories: sequential testing, sequential design, and sequential estimation. Because this categorization does not imply mutual exclusivity, it is possible for a citation to appear under more than one category.

Sequential Testing

Sequential testing involves a collection of hypothesis tests performed in a sequential manner in which one must decide whether more data need to be collected after each hypothesis test. This may involve repeated testing of the same hypothesis or testing multiple hypotheses. In particular, sequential testing procedures allow the number of observations to depend upon information

[1] "Adaptive Designs for Clinical Trials of Drugs and Biologics, Guidance for Industry," November 2019.

acquired during a testing procedure instead of being predetermined at the start of an investigation. A key benefit of sequential testing is the expected reduction in the sample size required to reach a conclusion regarding the hypothesis, as compared to a non-sequential or fixed sample size testing procedure.

MIL-HDBK-781A and the STAT COE (2017) recommend the use of sequential testing specifically for reliability testing. The intent of reliability testing is to determine the distribution of failure times; it uses top-level metrics such as the mean time between failures (MTBF), or a probability of failure. The size or length of a reliability sampling plan is determined by the reliability requirement and desired statistical metrics. Often a fixed-duration test plan is selected to estimate reliability because the length of a test must be known in advance. MIL-HDBK-781A presents the use of a sequential probability ratio test (SPRT) plan, based on Wald's (1945) SPRT, for determining compliance with a specific reliability requirement. An SPRT plan will save test time as compared to fixed-duration test plans that have similar risks when the demonstrated MTBF is high or very low. With respect to determining an initial test length when using a sequential test plan, MIL-HDBK-781A notes, "for sequential test plans, test duration should be planned on the basis of maximum allowable test time (truncations), rather than the expected decision point, to avoid the probability of unplanned test cost and schedule overruns."

Sequential Design

Sequential design refers to a class of problems and procedures concerned with the design of experiments (DOE) for which the pattern and composition of the resulting data, as well as the number of observations, are not predetermined at the start of an investigation but instead depend upon the information acquired throughout the course of the investigation. In addition to the number of observations, the conditions on which those observations are collected depend on acquired information from previous experiments.

The T&E community has embraced the use of non-sequential DOE for planning developmental and operational testing (Freeman et al., 2017). 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. DOE is not a sequential technique in nature, 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.

Sequential Estimation

Sequential estimation describes a point or interval estimation procedure that allows the number of observations to depend upon the information acquired during the course of the investigation. While some sequential estimation procedures seem to offer little benefit over fixed sample procedures, others can solve problems that fixed sample procedures cannot solve. In general, there are a variety of sequential estimation methods, each constructed with a specific purpose in mind. Such procedures may involve stopping criteria that indicate when the number of

observations is sufficient. Other sequential estimation procedures simply seek to recursively update the estimate as new data arrive, without consideration of stopping.

Johnson et al., (2014) illustrate several methods of sequential estimation in the application of ballistic resistance testing. Ballistic resistance testing is conducted in DoD to estimate the probability that a projectile will perforate the armor of a system under test. Ballistic resistance testing routinely employs sensitivity experiment techniques in which sequential methods are used to estimate a particular quantile of the probability of perforation.

Special Topics

The final category of our review is special topics, and include references to relevant guidelines, policies, and best practices, as well as specific challenges in implementing sequential analysis. For example, Avery and Simpson (2020) note that sequential procedures are challenging to use in DoD because the number of test runs, conditions for those runs, and resources required to execute those runs are often decided early on and codified in the Test and Evaluation Master Plan (TEMP) and Test Plan. Furthermore, sequential procedures may prove challenging to implement when the time required to score individual test events and perform analysis is longer than the scheduled time between tests, and when stakeholders have divergent assessments of test runs.

Conclusion

Sequential procedures may prove to be more challenging than non-sequential procedures to implement in DoD T&E. In cases where they can be applied, however, we find from our review that sequential procedures offer further opportunities to gain efficiencies in testing – such as with autonomous defense systems, which have gained much attention over the last several years. The use of sequential methods has been highlighted as a critical tool that can help testers execute the testing adaptively, efficiently, and effectively (Ahner and Parson, 2016; Porter, et al., 2020).

Approved for public release; distribution is unlimited.

Introduction

The majority of statistical analyses involve observing a fixed set of data, then analyzing those data after the final observation has been collected to draw some inference about the population from which they came (Ghosh, 2014). Unlike these traditional methods, 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 course of the investigation. Although the formal genesis of sequential analysis was in ballistics testing for the Department of Defense (DoD) (Wald, 1945; Wallis, 1980), we find sequential analysis is underused in recent DoD testing.[2] Expanding the use of sequential analysis in DoD testing has the potential to save substantial test dollars and decrease test time (National Research Council, 1998). Barnard (1946) points to two ways in which sequential procedures save time: "First, they make us stop work just as soon as we have enough evidence on the points in question. And, second, they enable us to use judgment in arranging the experiment, and to use any plausible guesses we can make about the true results to minimize the work involved; these guesses may be wrong, but if they are, they affect only the amount of work we do, not the validity of our results." In this paper, we review the sequential analysis literature as a first step in providing a recommendation for its use in DoD Test and Evaluation (T&E).

Perhaps the best-known applications of sequential analysis involve 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 the development of his Sequential Probability Ratio Test (SPRT).[3] In particular, the SPRT involves taking observations one at a time; each addition of an 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 non-sequential tests, while controlling for type I and type II errors[4] to exactly the same extent.

The statistical community quickly realized the broad implications of sequential procedures, and researchers began to investigate the properties of sequential procedures for examining multiple hypotheses, estimation, and allocation of effort in experimentation. Robbins (1952) uses the phrase sequential analysis to describe the "design and analysis of sampling experiments in which the size and composition of the samples are [not] completely determined before the

[2] DoD ballistic resistance testing continues to use sequential test designs to estimate a particular probability of perforation (Johnson et al., 2014).

[3] The SPRT is sometimes referred to as the Probability Sequential Ratio Test (PSRT).

[4] A type I error is the rejection of a true null hypothesis; one minus the probability of a type I error is referred to as the confidence of the test. A type II error is the failure to reject a false null hypothesis. One minus the probability of a type II error is the probability of correctly rejecting a false null hypothesis; we call this the power of a test.

experimentation begins." Johnson (1961) defines the term sequential analysis to apply to any statistical procedure in which the final pattern (including the number) of observations is not determined before the start of the investigation, but depends in some way on the values observed in the course of the investigation. Similarly, Ghosh (2014) describes sequential analysis as those techniques in which the final size and composition of the data need not be predetermined but may depend, in some specified way, on the data themselves as they become available during the course of an investigation. It is important to note that these definitions agree that the key defining feature of sequential analysis is that the composition, pattern, and/or number of observations used in the statistical procedure are not predetermined at the start of the analysis but instead depend on the information acquired as the analysis progresses.

Sequential procedures are frequently used in clinical trials. The National Institutes of Health defines a clinical trial as a research study performed in people that is aimed at evaluating a medical, surgical, or behavioral intervention. The U.S. Food and Drug Administration (FDA) uses the term "adaptive design" to describe sequential procedures, which it defines as "a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in trial" (2019). Periodically, the FDA releases guidance that represents their current thinking on the topic. In their most recent guidance,[5] the FDA lists the following advantages of using an adaptive design: statistical efficiency, ethical considerations, improved understanding, and acceptability to stakeholders.

A first step in establishing a roadmap for the applied use of Sequential Analysis in Test and Evaluation may be to consider the use of sequential analysis as it applies to clinical trials and to review the guidance of the FDA as a means of setting standards. DoD testing and clinical trials share many similarities. Both require careful planning, and practitioners in each express great interest in reducing the number of test events, speeding up testing, and saving money without sacrificing information needed to support a decision. For this reason, we include several citations to clinical trial papers in our literature review.

This literature review is far from exhaustive. A Google Scholar search with the keyword "sequential analysis" lists over four million results; we looked at only a subset of these for this review. Sequential analysis was first formalized during World War II, with some ad hoc sequential procedures existing before that time (Lai, 2001). Researchers have extensively studied the subject from both theoretical and practical perspectives for the last 80 years, and have achieved many developments, breakthroughs, and advances. However, although the subject is considered mature, many challenges remain in its application.

We document some of the challenges associated with sequential procedures. One of the most frequently encountered challenges occurs when a series of tests is conducted sequentially on a dataset. In many situations, such a procedure can be considered data snooping, as it involves

[5] "Adaptive Designs for Clinical Trials of Drugs and Biologics, Guidance for Industry," November 2019.

repeatedly looking at the data. Failing to account for the influence of these repeated looks can result in incorrect test results.

A second type of challenge we highlight concerns the testing of autonomous systems. Autonomous defense systems have gained much attention over the last several years. The use of sequential methods has been highlighted as a critical tool that can help testers adaptively, efficiently, and effectively execute the testing (Ahner and Parson, 2016; Porter, et al., 2020). To this end, we include several references regarding the challenge of testing autonomous weapon systems. The papers referenced specifically call for the use of sequential analysis.

This literature review is organized by broad categories within sequential analysis, and concludes with a list of annotated references. Each section ends with a table of references applicable to that area. First, we discuss the formal genesis of sequential analysis, highlighting several of the seminal papers contributing to its development and this field of study. Next, we subdivide the field of sequential analysis into three broad functional categories: sequential testing, sequential design, and sequential estimation. Our final category lists special topics, and includes references to relevant guidelines, policies, and best practices for sequential analysis procedures; challenges to sequential analysis; and other fields of study closely related to sequential analysis.

Our categorization of sequential testing, sequential design, and sequential estimation does not imply mutual exclusivity. Sequential design procedures often include an objective related to testing or estimation. We categorize methods as either sequential testing or sequential estimation when only the number of observations depends upon information acquired throughout the investigation. We categorize methods as sequential design when elements affecting the pattern and composition of the observations depend upon acquired information as well (Govindarajulu, 1975). Because our categories are not mutually exclusive, it is possible for a citation to appear under more than one category.

Approved for public release; distribution is unlimited.

Historical Note

While some ad hoc sequential methods existed earlier, 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). The problem that gave rise to sequential analysis was a specific question posed to the Statistical Research Group (SRG)[6] 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. Captain Schulyer was interested in determining a rule, specified in advance, that stated the conditions under which the experiment might be terminated earlier than planned. Captain Schulyer's problem was brought to the attention of Abraham Wald, a member of the SRG at the time. As a result, Wald devised the SPRT, a statistical test that takes advantage of the sequential nature of the data to reduce the required number of observations. Wald (1945) notes that "Because of the substantial savings in the expected number of observations effected by the SPRT, and because of the simplicity of this test procedure in practical applications, the National Defense Research Committee considered these developments sufficiently useful for the war effort to make it desirable to keep the results out of the reach of the enemy at least for a certain period of time."

Fisher (1952) notes that "Wald introduced the sequential test, but the sequential idea is much older." The origin of sequential analysis is closely related to sampling methods, in which classical sampling methods and theory were based on fixed sample sizes (Johnson, 1961). Many, including Wald, attribute the first idea of a sequential test to Dodge and Romig (1929), who first proposed a double sampling inspection procedure based on the idea that only a relatively small sample is needed to determine whether the quality of a batch of manufactured goods is very good or very bad, and that a larger sample is needed only in borderline cases. A different two-sample procedure allows pre-assigned accuracy of an estimator by estimating the variance and mean on different samples (Stein, 1945). Haldane (1945) proposes an inverse sampling procedure for estimating a binomial probability by sampling until a certain number of successes is encountered, instead of using a fixed number of observations. Another procedure iteratively updates an estimator based on its previous value without concern for when to stop sampling (Robbins and Munro, 1952).

Wald (1945) also indicates Mahalanobis's (1940) introduction of chain experiments, which involve the design of a large-scale experiment in successive stages, as being an early example of sequential analysis. Fisher (1952) himself describes the essence of sequential experimentation as "a series of experiments each of which depends on what has gone before." Johnson (1961) notes Fisher's lack of statistical formalism in referring to his ideas as "sequential thinking."

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

The principle of design was invented in an agricultural context, and while these principles would be useful to industry, Box (1999) notes two key differences between industrial experimentation and agricultural experimentation that he termed immediacy and sequentiality. Immediacy refers to the quick turn-around of results, and sequentiality refers to the notion that "experimental runs were made in sequence and information obtained from each run or small groups of runs was known and could be acted upon quickly and used to plan the next set of runs." Box goes on to note that the chief quarrel among experimenters with whom he interacted involved the fear that "statistics" would mean giving up enormous advantages offered by immediacy and sequentiality. To address the industry concerns of sacrificing immediacy and sequentiality for statistics, Box and Wilson introduced Response Surface Methodology (RSM) in 1951. RSM was a first attempt to provide suitable adaptation of statistical methods "to catalyze a process of investigation that was not static, but dynamic" (Box, 1999). See Myers, et al., (2016) for a thorough text on RSM.

Table 1 lists the sequential analysis foundational works we reviewed, with Appendix B providing full annotations for each reference.

Table 1. Foundational Works

Topic	Reference
Sequential Testing	Dodge, H. F. and Romig, G. (1929); Wald, A. (1945); Stein, C. (1945); Wald, A. (1947a); Johnson, N.L. (1961); Wallis, W. A. (1980); Lai, T.L. (2001)
Sequential Estimation	Stein, C. (1945); Haldane, J.B.S. (1945); Robbins, H. (1952); Johnson, N.L. (1961); Lai, T.L. (2001)
Sequential Design	Mahalanobis, P. C. (1940); Fisher, R.A. (1952); Robbins, H. (1952); Box, G.E.P and Wilson, K. B. (1951); Chernoff, H. (1959); Box, G. E. P. (1999); Myers, R. H. et al., (2016)
Notes: Appendix B includes annotations for each reference listed.	

Sequential Testing

Sequential testing involves a collection of hypothesis tests performed in a sequential manner in which one must decide whether more data need to be collected after each hypothesis test. This may involve repeated testing of the same hypothesis or testing multiple hypotheses. In particular, sequential testing procedures allow the number of observations to depend upon information acquired during a testing procedure instead of being predetermined at the start of an investigation. A key benefit of sequential testing is the expected reduction in the sample size required to reach a conclusion regarding the hypothesis, as compared to a non-sequential or fixed sample size testing procedure.

A sequential testing procedure is implemented in stages, where each stage involves collecting additional data and conducting the statistical test of a hypothesis. Each stage of the sequential testing procedure is accompanied by a decision rule indicating whether to continue collecting data or stop experimentation in favor of accepting or rejecting a hypothesis. Group sequential methods and Wald's (1945) SPRT are examples of sequential testing procedures.

Table 2 includes textbook treatments of sequential testing and reviews that provide a broad overview of topics. We also list existing guidebooks and software applications that can aid the practitioner in implementing some sequential testing procedures. Additionally, we include a broad variety of theoretical and methodological papers representing the breadth of research regarding sequential testing procedures that we categorize into one of the four detailed topics: hypothesis testing, SPRT and extensions,[7] group sequential testing, and Bayesian sequential decision analysis. Finally, we include case studies and papers focusing on applications of sequential testing procedures illustrating the breadth of use that has been found for sequential testing across many fields, including ballistic resistance testing, reliability testing, clinical trials, health policy, and A/B testing.

Table 2. Sequential Testing References by Type and Topic

Type	Topic	Reference
Books		Govindarajulu, Z. (1975); Wald, A. (1947a); Ghosh, B. K. and Sen, P. K. (1991)
Review Papers		Johnson, N. L. (1961); Lai, T. L. (2001)
Guidebooks	Reliability Testing	MIL-HDBK-781A; STAT COE (2019)
Theoretical / Methodological Papers	Hypothesis Testing	Wald, A. (1945); Chernoff, H. (1959); Wald, A. and Wolfowitz, J. (1948); Armitage, P. (1950); Tartakovsky, A. (2014), Albert, A. (1961); Malek, A. et al., (2017); Stein, C. (1945); Sobel, M. and Wald, A. (1949);

[7] The SPRT is a type of hypothesis test. We include the SPRT as a separate category because of the large focus on related tests in the literature.

Type	Topic	Reference
		Golz, M. et al., (2014); Tal, O. et al., (2001); Telford, J.K. (1992)
	SPRT and Extensions	Wald, A. (1945); Wald, A. and Wolfowitz, J. (1948); Armitage, P. (1950); Tartakovsky, A. (2014); Robbins, H. and Siegmund, D. O. (1974); Johari, R., et al., (2019); Sobel, M. and Wald, A. (1949); Golz, M. et al., (2014); Tal, O. et al., (2001); Telford, J.K. (1992)
	Group Sequential Testing	Whitehead and Stratton (1983); Schuler, S. et al., (2017); Pocock (1977); Daldal, R., et al., (2017)
	Bayesian Sequential Decision Analysis	Carlin, B. et al., (1998)
Application / Case Study Papers	Ballistic Resistance Testing	Johnson, T. H., et al., (2014)
	A/B Testing	Johari, R. et al., (2019)
	Clinical Trials	Gerke, O. et al., (2017); Lu C. Y. et al., (2018); Wetterslev, J. et al., (2017)
	Health Policy	Ryan, E. G. et al., (2019)
	Reliability Testing	Telford, J.K. (1992)
Software	SPRT and Extensions	Silva, I.R. and Kuldorff, M. (2019)
	Group Sequential Testing	Silva, I.R. and Kuldorff, M. (2019); Hack, N. et al., (2019); Yuan, Y. (2016); Wason. J. M. S. (2015); Kirk, J. L. and Fray M. P. (2014)
Notes: Appendix A includes notes for each topic listed. Appendix B includes annotations for each reference listed.		

Sequential Design

Sequential design refers to a class of problems and procedures concerned with the design of experiments (DOE) for which the pattern and composition of the resulting data, as well as the number of observations, are not predetermined at the start of an investigation but instead depend upon the information acquired throughout the course of the investigation. In addition to the number of observations, the conditions on which those observations are collected depend on acquired information from previous experiments. Govindarajulu (1975) notes that the theory of sequential design problems is more complicated than the theory of sequential testing or estimation; this is because more than just the number of observations in an experiment is sequentially dependent.

The T&E community has embraced the use of non-sequential DOE for planning developmental and operational testing. 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. DOE is not a sequential technique in nature, but many, including Montgomery (2017), recommend planning and executing a DOE based on the results of previous experiments. Because experiments are usually iterative in nature, Montgomery notes that it would be unwise to design too comprehensive of an experiment at the start of a study.[8] For this reason, Montgomery recommends that no more than 25 percent of the available test resources be invested in the first experiment. Regarding ways to implement a sequential experimental design strategy, Montgomery suggests starting with a screening design in which many factors can be tested in order to assess their importance. Using the results from the screening design, testers can augment the design by adding additional experimental runs. An augmented design can help determine, for example, whether higher-order terms are needed in the statistical model. There are many ways to plan for sequential data collection in the context of executing a DOE. The STAT COE (2018) provides a discussion of ways in which one might applying sequential DOE in the context of T&E.

The T&E community may be more familiar with the sequential DOE planning approach described by Montgomery (2017) above, but sequential design problems are more generally those that involve a sequential search for informative experiments (Chernoff, 1959). In particular, a key feature of sequential design problems is that the researcher may choose future experiments based on the results from existing observations, allowing the pattern and composition of experiments and observations to depend upon results of previous experiments. While this is a common feature of sequential design procedures, various procedures may differ in terms of the goal of the design. For example, one common objective of design procedures involves discrimination among hypotheses,

[8] Specifically, Montgomery (2017) notes, "A single comprehensive experiment requires the experimenters to know the answers to a lot of questions, and if they are wrong, the results will be disappointing. This leads to wasting time, materials, and other resources and may result in never answering the original research questions satisfactorily."

which is closely related to sequential testing. The difference is that the design problem allows the experimenter to choose from various experiments to gather each subsequent collection of observations (Chernoff, 1959). Such design procedures will typically require a stopping rule to determine when to stop testing, in addition to an allocation rule to determine the choice of next experiment. This objective differs from those of sequential design procedures, in which the objective of experimentation is to maximize or minimize some function by choosing from a set of experiments, as with bandit processes (Robbins, 1952)[9] and Response Surface Methodology (Box and Wilson, 1951).

Response surface methodology (RSM) is a collection of techniques used to optimize processes and product designs. It is based on the philosophy of sequential experimentation, with the objective of approximating the response with a lower-order polynomial in a relatively small region of interest that contains the optimum solution (Box and Wilson, 1951; Myers et al., 2016). RSM leverages experimental design fundamentals by using optimally informative values of covariates. It does this sequentially by allowing the most informative values of covariates and the inclusion of higher-order terms to depend upon sequentially acquired information.

Table 3 includes textbook treatments of sequential design and reviews that provide a broad overview of topics for the interested reader. We also list existing guidebooks that can aid the practitioner in implementing some sequential testing procedures. Additionally, we include a broad variety of theoretical and methodological papers representing the breadth of research regarding sequential design. Finally, we include case studies and papers focusing on applications of sequential design, specifically focusing on literature related to DoD testing.

Table 3. Sequential Design References by Type and Topic

Type	Topic	Reference
Books		Myers, R. H., et al., (2016); Montgomery, D. C. (2017); Hamada, M.S., et al., (2014); Ghosh, B. K. and Sen, P. K. (1991)
Review Papers		Box, G. E. P. (1999); Lai, T. L. (2001)
Guidebooks	Design of Experiments	STAT COE (2019)
	Testing via Sequential Experimentation	STAT COE (2018)
Theoretical / Methodological Papers	Response Surface Methodology	Box, G. E. P. and Wilson, K. B. (1951)

[9] Experimentation is often targeted toward and constrained by certain control objectives, and the procedures often describe a trade-off between exploration and exploitation, sometimes referred to as a trade-off between information and control. In particular, bandit processes seek to maximize a total reward by sequentially choosing lines of effort, each of which has its own random reward. The optimal allocation procedure involves considering the trade-offs associated with continuing to pursue a perceived high reward effort versus gaining information regarding efforts with unknown reward.

Type	Topic	Reference
	Bayesian Experimental Designs	Cuturi, Marco et al., (2020); Terejanu G. et al., (2012); Drovandi, C.C et al., (2014); Chaloner, K and Verdinelli, I. (1995); Huan, X and Marzough Y.M. (2013); Pauwels, E. et al., (2014); Albert, C. et al., (2012); Huan, X. and Marzouk, Y. (2014); Sieck, V. R. C., and Christensen, F. G. W. (In Press).
	Simulation Experimentation	Kim, J. H. et al., (2018); Chen, P. A., et al., (2018a)
	Design of Experiments	Chernoff, H. (1959); Albert, A. (1961)
	Reliability Assurance Tests	Smith, C., et al., (2010); Gilman, J. F., et al., (2018)
	Sensitivity Testing	Johnson, T.H., et al., (2014); Ray, D. M. and Roediger, P. A. (2018)
	Bandit Process	Whittle, P. (1988); Weitzman, M. L. (1979); Gittins, J. C. (1979); Robbins, H. (1952); Chen, P. A., et al., (2018a)
	Allocation of Effort	Anderson-Cook, C.M., et al., (2008); Chernoff, H. (1959); Albert, A. (1961); Robbins, H. (1952); Gittins, J. C. (1979); Weitzman, M. L. (1979); Whittle, P. (1988)
Application / Case Study Papers	Design of Experiments for Defense Testing	Johnson, R. T., et al., (2012); Simpson, J., et al., (2013); Simpson J. (2020); Freeman, L.J., et al., (2018); Johnson, T.H., et al., (2014); Kim, J. H. et al., (2018)
	Reliability of Complex Systems	Anderson-Cook, C.M. et al., (2008)
	Reliability Assurance Testing for Defense Testing	Gilman, James F. et al., (2018); Hamada, M.S., et al., (2014);
	Simulation Experimentation	Chen, A. P. et al., (2018a); Chen, A. P., et al., (2018b)
Software		
Notes: Appendix A includes notes for each topic listed. Appendix B includes annotations for each reference listed.		

Approved for public release; distribution is unlimited.

Sequential Estimation

Sequential estimation describes a point or interval estimation procedure that allows the number of observations to depend upon the information acquired during the course of the investigation. While some sequential estimation procedures seem to offer little benefit over fixed sample procedures, others can solve problems that fixed sample procedures cannot solve. In general, there are a variety of sequential estimation methods, each constructed with a specific purpose in mind. Examples of sequential estimation procedures include inverse binomial sampling (Haldane, 1945) to estimate a frequency with a lower variance than is possible in fixed sample procedures, and double sampling for estimation of a confidence interval with specified length (Stein, 1945).[10] Such procedures may involve stopping criteria that indicate when the number of observations is sufficient. Other sequential estimation procedures simply seek to recursively update the estimate as new data arrive, without consideration of stopping, as with the Kalman Filter (Kalman, 1960).

Table 4 includes textbook treatments of sequential estimation and reviews that provide a broad overview of topics for the interested reader. We include a broad variety of theoretical and methodological papers representing the breadth of research regarding sequential estimation. We include one case study paper focusing on applications of sequential estimation to ballistic resistance testing.

Table 4. Sequential Estimation References by Type and Topic

Type	Topic	Reference
Books		Wald, A. (1947a); Govindarajulu, Z. (1975); Ghosh, B. K. and P. K. Sen (1991); Ghosh, M., et al., (1997)
Review		Lai, T. L. (2001); Johnson, N. L. (1961); Anscombe, F. J. (1953)
Theoretical / Methodological	Point Estimation	Haldane, J. B. S. (1945); Kalman, R. E. (1960)
	Interval Estimation	Hall, P. (1983); Stein, C. (1945)
	Inverse Binomial Sampling	Haldane, J. B. S. (1945)
	Double Sampling	Hall, P. (1983); Stein, C. (1945)

[10] Note that the double sampling procedure of Stein (1945) is different from the double sampling inspection procedure described by Dodge and Romig (1929).

Type	Topic	Reference
	Recursive Estimation / Stochastic Approximation	Kalman, R. E. (1960); Robbins, H. (1952)
Application / Case Study	Ballistic Resistance Testing	Johnson, T.H., et al., (2014)
Software		
Notes: Appendix A includes notes for each topic listed. Appendix B includes annotations for each reference listed.		

Special Topics

We close our review with citations of several special topic papers in Table 5 that do not fit into the sequential testing, sequential design, or sequential estimation categories. We include papers regarding guidelines, policies, and best practices relevant to the application of or call for sequential procedures in DoD. We also highlight the FDA's most recent policy on sequential analyses, an organization that has been employing sequential methods for decades in clinical trials.

One best practice to highlight is the use of a Data Monitoring Committee (DMC), which is standard practice in clinical trials. The primary purpose of a DMC is to ensure that the continuing of a trial according to its protocol is ethical. The DMC judges the evidence and makes recommendations on whether to continue or terminate a trial.

Our review also emphasizes a primary challenge related to the use of sequential analysis procedures: data snooping. Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection, and is perhaps one of the most commonly encountered challenges. Data snooping can result in incorrect model selection by causing the researcher to believe that some factors are relevant when they are unrelated to the outcome. One must account for repeated views of a dataset during analysis to avoid this pitfall.

A second type of challenge we highlight concerns the testing of autonomous systems. Autonomous defense systems have gained much attention over the last several years. To test them, testers must understand how these systems make decisions across different operating conditions if they are to provide stakeholders with appropriate levels of trust in autonomous or AI-enabled capabilities (Porter, et al., 2020). The use of sequential methods has been highlighted as a critical tool that can help testers adaptively, efficiently, and effectively execute the testing (Ahner and Parson, 2016; Porter, et al., 2020). To this end, we include several references regarding the challenge of testing autonomous weapon systems. The papers referenced specifically call for the use of sequential analysis.

Last, we mention the closely related topics of Decision Theory and Quality Control, two fields that are associated with the application of sequential analysis.

Table 5. Special Topic References by Type and Topic

Type	Topic	Reference
Guidelines, Policy, and Best Practice Papers	Test & Evaluation in the DoD	National Research Council, (1998); STAT COE (2019); STAT COE (2018); Avery, M. R. and Simpson, J., (2020)
	Data Monitoring Committee	Armitage, P. (1991); Grant, A.M. et al., (2005)
	Clinical Trials	U.S. Food and Drug Administration (2019)

Type	Topic	Reference
Challenges	Data Snooping	White, H. (2000); Lovell, M.C. (1983); Romano, J.R. and Wolf, M. (2005); Hansen, P.R. (2005); Lo, A.W. and Mackinlay, C. (1990); Johari, R., et al., (2019); Malek, A. et al., (2017)
	Testing Autonomous Weapon Systems	Simpson, J. (2020); Ahner, D. and Parson, C. (2016); OSD R&E (2015); Deonandan, I., et al., (2010); Hess, J. and Valerdi, R. (2010); Porter, D., et al., (2020); Hess, J. and Valerdi, R. (2010)
Related Topics	Decision Theory	Ghosh, B. K. and Sen, P. K. (1991); Wald, A. (1947b)
	Quality Control	Lai, T. L. (2001)
Notes: Appendix A includes notes for each topic listed. Appendix B includes annotations for each reference listed.		

Appendix A – Reference Topic Notes

A/B Testing – A/B Testing is the industry terminology for a completely randomized controlled trial where two versions of something (usually a control and a treatment) are compared to figure out which performs better. In A/B testing practice, analysts are not constrained to simply analyze the output of an experiment; they can also adjust the experimental design in response to the data observed. A pervasive form of this technique is users continuously monitoring the p-values and confidence intervals reported in order to set the sample size of an experiment dynamically.

Allocation of Effort – Allocation of effort is the determination of the best way to collect new data in order to maximally improve understanding.

Bandit Process – Bandit processes are those for which experimentation is targeted toward and constrained by certain control objectives, such as those allowing the researcher to choose among experiments in order to maximize some reward function. This is different from other more traditional goals of experimentation such as those in which the researcher chooses among a collection of experiments with the goal of discerning among a collection of hypotheses. The terminology originated with the idea of choosing which arm to pull on an imagined slot machine (bandit) with two or more arms in such a way that one maximizes the payout. In particular, Lai and Robbins (1985) note "Ordinary slot machines with one arm are one-armed bandits, since in the long run they are as effective as human bandits in separating the victim from his money."

Bayesian Experimental Design – Bayesian experimental designs use Bayesian methods to determine the conditions in which the next set of data points will be collected. The data collection decision is usually based on a utility function that maximizes the expected information gained from the next set of runs. The utility function allows the tester to rank the outcome of an action (e.g., deciding which data point to next collect).

Bayesian Sequential Decision Analysis – A Bayesian approach to sequential analysis that involves making decisions to minimize the expected value of a loss function associated with a particular outcome during an interim analysis. Bayesian methods enable the formal combination of expert opinion and objective information into interim and final analyses of clinical trial data. Approaches to find decision rules include the backward induction method (see, e.g., DeGroot 1970; Berger 1985), forward sampling (see e.g., Carlin et al., 1998), and the gridding method (see, e.g., Brockwell and Kadane, 2003)

Clinical Trials – Clinical trials are research studies performed on people that are aimed at evaluating a medical, surgical, or behavioral intervention. Clinical trials are the primary way researchers determine whether a new treatment, like a new drug or diet or medical device, is safe and effective in people. Often a clinical trial is used to learn if a new treatment is more effective and/or has less harmful side effects than the standard treatment. Clinical trials frequently leverage sequential procedures, which allow for prospectively planned modifications to one or more aspects of the test based on accumulating data from subjects in the trial.

Data Monitoring Committee (DMC) – A DMC is a committee convened to judge the evidence and make recommendations on whether to continue or terminate a trial. Their primary purpose is to ensure that continuing a trial according to its protocol is ethical. Data Monitoring Committees are common in clinical trials.

Data Snooping – Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection. Such data reuse decreases the likelihood that satisfactory results are due to any merit inherent in the method yielding the results.

Decision Theory (Statistical) – Statistical decision theory is the study of constructing a statistical decision function which associates each sample point with a decision such that the decision is made when the sample point is

observed. A sequential decision function is a decision function for which the number of observations needed to reach a decision depends on the outcome of the observations.

Design of Experiments (DOE) – DOE is the process of planning an experiment so that the appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions (Montgomery, 2017). Specifically, DOE is an approach to systematically varying the controllable input factors in the process and determining the effects these factors have on the output parameters, while controlling for type I and II errors. A statistical DOE allows for multiple input factors to be manipulated, determining their effect on a desired output (response). By manipulating multiple inputs at the same time, DOE can identify important interactions that may be missed when experimenting with one factor at a time.

Double Sampling - Double sampling typically refers to the method in which a first sample is taken and used to estimate the variance of a population, and then a second sample (whose number of observations depends upon the results of the first sample) is taken and used to estimate the mean of the population (Stein, 1945). This procedure allows calculation of a confidence interval of fixed width that is independent of the mean. More generally, double sampling can also refer to splitting a sample in such a way that the number of observations in the second sample depends upon results obtained from the first sample.

Group Sequential Testing – Group sequential testing refers to a trial, which allows for one or more prospectively planned interim analyses of comparative data with prespecified criteria for stopping the trial. At the beginning of the design process, testers carefully plan how many data points they need for the entire study, and divide them across all interim analyses. After a certain amount of data points have been collected, data are analyzed for the first time. The results from this subset of data are compared to pre-defined boundaries and the decision is made to either stop for futility or efficacy or collect the next set of data points. This process is repeated for subsequent interim analyses until the trial is stopped or the complete set of data have been collected.

Hypothesis Testing – Hypothesis testing provides a formal procedure for testing whether the results of an experiment are meaningful. Specifically, it is a rule that specifies i) for which sample values the decision is made to accept the Null Hypothesis and ii) for which sample values the Null Hypothesis is rejected and the Alternative Hypothesis is accepted as true. Sequential hypothesis testing requires taking observations sequentially until some pre-defined stopping criterion is satisfied. The Sequential Probability Ratio Test (SPRT) is a specific case of sequential hypothesis testing.

Interval Estimation – Interval estimation refers to the estimation of an interval for a parameter with a specified coverage probability.

Inverse Binomial Sampling – Inverse binomial sampling is a method for estimating the frequency of an attribute by counting a sample until a predetermined number of members possessing this attribute are found.

Point Estimation – Point estimation refers to the estimation of a parameter.

Quality Control – Quality control is the monitoring and evaluation of the quality of products from a continuous production process. Shewhart introduced this fundamental concept in 1931, and it is now widely applied in industry today. Shewhart's control chart is a "single-sample" scheme whose decision depends solely on the current sample although the results of previous samples are available from the chart. To improve the sensitivity of the Shewhart control chart, Page (1954, not included in annotations) modified Wald's theory of sequential hypothesis testing to develop the cumulative sum control chart (CUSUM) control chart. Lai (2001) provides a review of some major developments in sequential change-point detection and diagnosis.

Recursive Estimation – Recursive estimation is a type of sequential estimation procedure in which the analyst corrects an estimate after every trial, with the correction determined from the results of all previous trials. (Pugachev, 1984)

Reliability Assurance Tests – A reliability assurance test uses additional supplementary data and information to reduce the required amount of testing The supplementary data and information may include appropriate reliability models, earlier test results on the same or similar devices, expert judgement regarding performance, knowledge of the environmental conditions under which the devices are used, benchmark design information on similar devices, or prior knowledge of possible failure modes (Hamada et al, 2014).

Reliability Testing – Reliability testing intent is on determining the distribution of failure times, and it uses top-level metrics like the mean time between failures (MTBF), or a probability of failure. The size or length of a sampling plan is determined by the reliability requirement and desired statistical metrics. Often a fixed-duration test plan is selected to estimate reliability because the length of a test must be known in advance. However, MIL-HDBK-781A and the STAT COE (2017) also recommend the use of a sequential test plan, based on Wald's SPRT, for determining compliance with a specific reliability requirement. With respect to determining an initial test length when using a sequential test plan, MIL-HDBK-781A notes, "for sequential test plans, test duration should be planned on the basis of maximum allowable test time (truncations), rather than the expected decision point, to avoid the probability of unplanned test cost and schedule overruns."

Response Surface Methodology (RSM) – RSM is a collection of statistical and mathematical techniques used to optimize processes and product designs. This methodology identifies and fits an appropriate response surface model from experimental data requires some knowledge of statistical experimental design fundamentals, regression modeling techniques, and elementary optimization methods. RSM was introduced as a first attempt to provide suitable adaptation of statistical methods to catalyze a process of investigation that was not static, but dynamic (Box and Wilson 1951). The RSM approach is based on a philosophy of sequential experimentation, with the objective of approximating the response with lower-order polynomial in a relatively small region of interest that contains the optimum solution (Myers et al., 2016).

Sensitivity Testing – Sensitivity testing refers to a testing procedure used to quickly determine the stimulus level at which a specified fraction of test items fail. Specifically, with respect to Test & Evaluation in the DoD, the sensitivity testing problem considered can be described as follows: a binary response is observed when a stimulus level was applied to an experimental unit. For example, suppose a projectile penetrates an armored plate, it is called a response or success. If it does not, it is called a nonresponse or failure. The main goal is to find the stimulus level at which the experimental unit has a success with a given probability. Sensitivity testing seeks to generate a set of data that supports a predictive regression model for the probability or quantile associated with various stimulus levels. Adaptive sensitivity testing "adapts" to the results as each data point is generated and analyzed, thereby significantly reducing the risk of generating useless or unbalanced data sets.

Sequential Analysis – Sequential analysis is concerned with 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. Sequential analysis procedures differ from the usual statistical procedures in that properties of the sample, such as size, are not fixed in advance. In sequential analysis procedures, the results of the observations themselves influence the number, pattern, or composition of future observations. For standard sequential testing and estimation problems, only the number of observations is treated as random. Design problems increase the complexity of the sequential procedure by allowing the pattern and composition of the observations to depend upon acquired information as well.

Sequential Design – Sequential design refers to a class of experimental design problems and procedures for which the pattern and composition of the data, as well as the number of observations, are not predetermined at the start of an investigation, but instead depend upon the information acquired throughout the course of the investigation. Compared to standard sequential testing and estimation procedures, sequential design procedures allow the experimenter to choose among experiments to perform at each stage or to vary the treatments sequentially.

Sequential Estimation – Sequential estimation describes an estimation procedure performed in such a way that the number of observations depends upon the information acquired during the course of the investigation. Estimation can refer to point or interval estimation, and while some sequential estimation procedures seem to offer little benefit over fixed sample procedures, others do not have a fixed sample analogue.

Sequential Probability Ratio Test (SPRT) and Extensions – The (SPRT) is a specific sequential hypothesis test used during sequential testing to determine with the smallest number of observations possible whether the null hypothesis should be accepted or rejected. As in classical hypothesis testing, the SPRT starts with a pair of hypothesis: the Null and the Alternative. Next, the cumulative sum of the log-likelihood ratio is calculated as new data are collected. The log-likelihood ratio is compared to bounds provided by a stopping rule, which defines the acceptance and rejection regions based on the desired type 1 and type 2 errors. Wald and Wolfowitz (1948) later prove the SPRT to be optimal in the sense that it requires the smallest expected number of observations to reach a conclusion among comparable tests. Since Wald's formalization of the SPRT, many extensions of the test have been added to the literature, including the Generalized SPRT (GSPRT), multihypothesis generalized sequential likelihood ratio test (MGSLRT), the multihypothesis adaptive sequential likelihood ratio test with one-stage delayed estimators (MASLRT), and many others.

Sequential Testing – Sequential testing describes a set of testing procedures for which the total number of observations is not pre-defined, but rather is determined throughout the course of testing based on the collected results. A sequential testing procedure is implemented in stages, where each stage involves collecting additional data, conducting the statistical hypothesis testing, then deciding to either continue collecting data or stopping experimentation based on a pre-defined decision rule. Sequential testing, analysts perform a collection of hypothesis tests in a sequential manner to decide if more data needs to be collected. This may involve repeated testing of the same hypothesis or testing multiple hypotheses. Wald's (1945) Sequential Probability Ratio Test and Group Sequential Methods are famous examples of sequential testing procedures.

Simulation Experimentation — Simulation Experimentation can be defined as a test or series of tests in which purposeful changes are made according to inputs of a target system, that uses a simulation model in place of the real-world system.

Stochastic Approximation – Stochastic approximation is the process of recursive estimation in which the correction after every trial depends only on the result of this trial and the previous estimate. The general theory of stochastic approximation was initiated by Robbins and Monro (1951) and developed in the subsequent works by Kiefer and Wolfowitz (1952), Blum (1954a, b), and many others. (Pugachev, 1984)

[Developmental and Operational] Test & Evaluation in the DoD – Developmental testing covers a wide range that includes component testing, modeling and simulation, and engineering systems testing. Developmental testing and evaluation (DT&E) presents the first opportunity to measure the performance and effectiveness of the system. Operational testing and evaluation examine the performance of a fully integrated set of systems, including the subject system, under realistic operating environments. Operational testing and evaluation (OT&E) occur in an operationally representative environment with operational uses to determine whether a system is operationally effective, operationally suitable, and survivable for its intended use. DT&E and OT&E are key parts of military system development.

[Challenges of] Testing Autonomous Weapon Systems – There are many challenges associated with testing autonomous systems. Some of these challenges fall under sequential analysis. While sequential test design is a critical existing technology, the current use of sequential testing is not adequate for autonomous systems. In particular, there is a need for near real time or autonomous sequential testing and a method to determine when to stop testing when the next test depends upon the results of the previous test. A primary shortcoming is the need for a user in the loop to identify new design points and update the test.

Appendix B – References & Annotations

Ahner, D. K., & Parson, C. (2016). Workshop Report: Test and Evaluation of Autonomous Systems. Wright-Patterson AFB, OH.

This paper highlights the challenges associated with testing autonomous systems. Some of these challenges fall under sequential analysis. The authors indicate that, while sequential test design is a critical existing technology, the current use of sequential design and testing is not adequate for autonomous systems. In particular, they note the need for near real time or autonomous sequential testing and a method to determine when to stop testing when the next test depends upon the results of the previous test. Further, they note, “a primary shortcoming is the need for a user in the loop to identify new design points and update the test.”

Albert, A. E. (1961). The Sequential Design of Experiments for Infinitely Many States of Nature. The Annals of Mathematical Statistics, 32(3), 774-799.

Similar to Chernoff (1959), this paper discusses two important issues: when to stop experimentation, and which experiment to choose if experimentation continues. The optimal stopping rule is characterized by a bound on the sequential likelihood ratio test. The optimal selection of the "next experiment" can be characterized by a probability distribution over experiments.

Albert, C., Ashauer, R., Künsch, H., & Reichert, P. (2012). Bayesian experimental design for a toxicokinetic–toxicodynamic model. Journal of statistical planning and inference, 142(1), 263-275.

The authors apply a Bayesian method of identifying optimal experimental designs, specifically to a toxicokinetic-toxicodynamic model. A Bayesian optimal experimental design is determined by comparing the knowledge gained from various sets of design points. Note, a design of experiment is called optimal within this set of designs if it maximizes the expected gain of knowledge about the parameters. They use Bayesian methods to calculate the expected gain of knowledge by comparing the variances of the prior and posterior distributions for each set of points. The authors determine the next set of design points will be the ones that maximize the expected gain of knowledge given the prior.

Anderson-Cook, C. M., Graves, T. L., & Hamada, M. S. (2008). Resource allocation: Sequential design for analyses involving several types of data. Paper presented at the IEEE International Conference on Industrial Engineering and Engineering Management, Singapore.

In analyzing the reliability for complex system, several types of data from full-system tests to component level tests are commonly available and used. Following preliminary analysis, additional resources may be available and one might wish to identify the best new data to collect to improve the prediction of system reliability. This paper presents a methodology for determining what new data to collect to improve understanding. The goal of resource allocation is to determine the best collection of new data that will maximally improve our understanding of the system. The authors' approach to resource allocation is to simultaneously consider cost and reduction of uncertainty. By using current understanding of system reliability, the authors point out that we can leverage this knowledge into making a more informed decision about where future resources should be spent. Using estimates based on available data can help us spend our resources most effectively to maximally reduce the uncertainty of our system reliability estimates.

Anscombe, F. J. (1953). Sequential estimation. Journal of the Royal Statistical Society: Series B (Methodological), 15(1), 1-21.

In this review paper, Anscombe discusses early literature regarding sequential estimation. Initially, researchers were focused on the design of sequential sampling procedures on testing hypotheses with estimation being an afterthought, but "Wald in his Sequential Analysis (1947) devoted a short section to discussing sequential procedures whose primary object was to provide, with as few observations as possible, a confidence interval for an unknown parameter having preassigned confidence coefficient and preassigned width, or satisfying some other similar condition. He studied, in particular, the problem of estimating the mean of a normal population, of which the variance was unknown, by a confidence interval of prescribed width and confidence coefficient. He did not then attempt to obtain useful solutions to such problems. Meanwhile, two particular procedures for sequential estimation, of

considerable practical interest, had been published, though not under the label "sequential". These were (i) Haldane's inverse binomial sampling and (ii) Stein's double-sampling method for estimating the mean of a normal population (Wald's problem just mentioned). [Anscombe begins] this survey with an outline of these two procedures, together with some later developments, and then [reviews] the other work that has appeared on sequential estimation, which is mostly rather more abstract and theoretical."

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. Journal of the Royal Statistical Society. Series B, Methodological, 12(1), 137-144.

This paper extends Wald's SPRT to handle sequential decision problems with three or more hypotheses, and is related to the test proposed by Sobel and Wald (1949). In particular, "in a review of Wald's book (1947), Barnard (1947) pointed out that, as a generalization of Wald's method, one could formulate a procedure for deciding between more than two simple hypotheses. The procedure would be based on likelihood ratios, and the risks involved could be controlled by suitable choice of the acceptance conditions. The present note contains an outline of the theory of such sequential procedures, which are closely related to some recent developments in the theory of discriminant functions. The methods should prove useful for two-sided tests of statistical hypotheses. In particular, [Armitage considers] a two-sided test for the value of a binomial probability, and, as a development of this, a two-sided test for comparative trials." As a further note, Armitage published several well cited papers on the topic of sequential analysis in clinical trials.

Armitage, P. (1991). Interim analysis in clinical trials. Statistics in medicine, 10(6), 925-937.

The history and functions of data monitoring committees (DMCs) are reviewed in this paper. Armitage notes that DMCs come in many shapes and sizes. DMCs need to consider many aspects of the data before making recommendations to the investigators, who have ultimate responsibility for early termination or protocol changes. This paper is mainly concerned with interim analysis for clinical trials, but the discussion itself is relevant for defense testing. Amongst other topics, Armitage covers early stopping. Two broad approaches to early stopping are (i) demonstration of strong evidence that a treatment effect falls above or below some critical value) and (ii) stochastic curtailment based on prediction of discrete time points. Armitage references a number of papers describing the statistical consequences of various stopping rules. A primary purpose of the paper is to indicate that decisions about early stopping must depend on a variety of factors (for clinical trials: management data, safety data, efficacy data, effects of repeated looks), and that it will usually be inappropriate to think of an inviolable decision rule based on the results of interim analysis. It is usually unwise to define and publicize a stopping-rule at the outset of a trial. Armitage further states that if repeated hypothesis tests are performed, then the Type I error probability (or overall significance level) is affected and increases with the number of tests; much literature exists to address this issue.

Avery, M., & Simpson, J. (2019). How Much Testing is Enough? 25 Years Later. Institute for Defense Analyses.

The question of “How much testing is enough?” is persistent across Department of Defense (DoD) test and evaluation (T&E) endeavors. In 1994, the Military Operations Research Society (MORS) and the International Test and Evaluation Association (ITEA) looked to answer this question, or at least focus future inquiry, with a three-day mini-symposium. Content from the 1994 symposium sessions and keynotes were summarized in a report, which captured the major points of discussion and provided some general recommendations. It has been 25 years since that symposium and, in some respects, the T&E community has experienced great progress in answering the question that inspired the symposium. Avery and Simpson summarize progress since 1994 in answering the question, “How much testing is enough?” Sequential Testing is addressed in the context of Discussion Area 3: Using early T&E data to influence budgeting and acquisition strategy. The report defines sequential testing as using multiple test phases, arranged such that the results from earlier test phases influence runs or test points in subsequent test phases. A multi-phase test strategy that schedules the most critical test events and runs early can be advantageous, since it enables the test team to use those early results to inform decisions about later testing. If critical hurdles are passed early on, less testing may be required later on. If the system struggles in those critical events, a more thorough set of tests than originally planned may be necessary. The report notes that sequential designs are challenging to use in DoD, since the number of test runs, conditions for those runs, and resources required to execute those runs are often decided early on and codified in the Test and Evaluation Master Plan (TEMP). Issues with sequential designs can arise if system performance is assessed by many measures, each of which is affected differently by multiple factors (i.e., they vary over different conditions). Sequential design can also prove challenging to implement when the time required to score individual test events takes longer than the scheduled time between tests, and when OT&E stakeholders have divergent assessments of test runs. Although

sequential designs can prove challenging to implement, the authors conclude by noting, in cases where they can be applied, sequential test designs offer further opportunities to gain efficiencies in testing.

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 1-21.

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.

This is a companion paper to Part I, in which Box illustrates a number of concepts which together embody what he understands to be response surface methodology (RSM). RSM is a group of statistical techniques specifically designed to catalyze scientific learning of this kind. Box notes that if statistical methods are to act as a catalyst to investigation, they must be robust to likely deviations from assumption (all models are wrong, but some are useful). This paper, Part II, is concerned with the implications raised when RSM is used as a statistical technique for the catalysis of learning. It might also be considered an overview paper. The paper is written from an industrial experimentation perspective. Box notes that industrial success requires efficient experimentation both for the improvement of existing products and processes and for development of new ones. Because results are usually known quickly (immediacy), Box argues the natural way to experiment is to use information from each group of runs to plan the next (sequentially). Such investigation employs a scientific paradigm in which data drives an alternation of induction and deduction. This process can suggest how to resolve questions at each stage of the experiment. Results from previous experiments, combined with subject matter knowledge, motivate the next step.

Historical Note: Box notes the foundations of RSM date back to World War II, during which Box, having had no formal statistical training, served at a research station concerned with defense against chemical warfare and was assigned the job of designing and analyzing many statistically planned experiments. This experience led him to pursue a theoretical statistics degree at the University College in London. It was later, while at the Imperial Chemical Industries in England, Box watched what experimenters did and tried to find ways to help them do it better. Box found that most of the principles of design originally developed for agricultural experimentation would be of great value in industry, but that usual industrial experimentation differed in two major respects: immediacy and sequentially. Results were available, if not within hours, within days. Furthermore, runs were usually made in sequence, and the information obtained from each run or small group of runs was known and could be acted upon quickly to plan the next set of runs. Box found the chief quarrel that experimenters had with using "statistics" was that they thought it would mean giving up enormous advantages offered by immediacy and sequentially. For them, statistical design, meant planning an all-encompassing "one-shot" factorial experiment. Box noticed that the "one-shot" experiments often quickly petered out. Because, in light of their knowledge of chemistry and engineering, after a few runs the experimenters might say, "Now that we see these early results, we realize that we should be using much higher pressures and temperature. Also, the data suggested that some of the factors first thought important were not..." Box defined their need to be: find ways of using statistics to catalyze a process of investigation that was not static, but dynamic. RSM was introduced by Box and Wilson (1951) as a first attempt to provide a suitable adaptation of statistical methods to meet these needs. By chance, experimental design was invented in an agricultural context. For example, Fisher's earlier interest in aerodynamics could have resulted in a career in aircraft design, perhaps producing a somewhat different emphasis in the "design of experiments." However, the circumstances of agricultural experimentation are very unusual (certain industrial life testing experiments are an exception) and should certainly not be perceived as sanctifying methods in which all assumptions are fixed a priori and lead to a one-shot procedure. Iterative learning, of course, goes on in agricultural trials as elsewhere. The results from each year's trials are used in planning the next.

Box, G. E., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society: Series B (Methodological), 13(1), 1-38.

This paper is the first to introduce Response Surface Methodology (RSM), a method commonly used with experimental design. The goal of RSM is to find the values of variables that maximize a dependent response variable; RSM provides a method to find these values by searching the domain space in an iterative, sequential fashion. Box and Wilson explain their procedure: "The general problem is discussed of finding experimentally the levels of a number of quantitative variables at which some dependent response has a maximum value. The problem can be solved by exploring the whole experimental region; but the number of experiments necessary to do this would usually be prohibitively large. When the experimental error is small and experiments are conducted sequentially,

the derivatives determined in a given sub-region may be used to locate another in which the response is greater. By repetition of this process a near-stationary region is found. This region may then be explored by determining derivatives of higher order. The derivatives are deduced by performing experiments in the sub-region and fitting an equation of suitable degree to the experimental points. Two possible sources of error arise, that due to errors of observation, and that due to bias which might occur if the response function were not capable of representation by an equation of the type assumed."

Burke, S., Divis, E., Guldin, S., Harman, M., Kolsti, K., McBride, A., . . . Welker, T. (2019). Guide to Developing an Effective STAT Test Strategy V7. 0. Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

The Scientific Test and Analysis Techniques Center of Excellence (STAT COE) guide provides an outline for how one might develop an effective STAT Test Strategy for testing DoD weapon systems. Scientific Test and Analysis Techniques are deliberate, methodical processes and procedures that seek to relate requirements to analysis in order to inform better decision making. Applying design of experiments to testing is the primary focus of the guidebook. The guidebook recommends the implementation of a sequential testing approach, which the authors describe as a progressive test strategy that starts with low level test design and culminates in top-level operational testing. Appendix A includes an example application of a sequential testing. The example is limited though in that it only illustrates how one might build an initial screening design. In addition to recommending a DOE approach, the guidebook lists several other statistical methods for various applications including reliability. The guide recommends using the Sequential Probability Ratio Test (SPRT) as a modified sampling plan for reliability test planning as it could enable testing to stop early if certain conditions are met.

Carlin, B. P., Kadane, J. B., & Gelfand, A. E. (1998). Approaches for optimal sequential decision analysis in clinical trials. Biometrics, 964-975.

This paper is written in the context of clinical trials. Unlike traditional approaches, Bayesian methods enable formal combination of expert opinion and objective information into interim and final analyses. The authors of this paper describe a fully Bayesian sequential decision-making method for numerous interim looks. The authors illustrate their methodology using a set of published data for an AIDS clinical trial. The trial includes two covariates (treatment - active drug or placebo and baseline CD4 cell count) and response variable time from start of trial until development of toxoplasmic encephalitis (TE), a major cause of morbidity and mortality among AIDS patients, or death.

Chaloner, K., & Verdinelli, I. (1995). Bayesian Experimental Design: A Review. Statistical Science, 10(3), 273-304.

This paper provides a thorough review of Bayesian experimental design through 1995. Bayesian experimental designs can be viewed as a form of sequential analysis, because for each experiment (or set of experiments) we can incorporate prior information. The authors present Bayesian experimental design from a decision theory point of view that includes linear and nonlinear design problems.

Chen, P.-H. A., Santner, T. J., & Dean, A. M. (2018a). Sequential Pareto minimization of physical systems using calibrated computer simulators. Statistica Sinica, 28(2), 671-692.

This paper proposes a sequential design methodology for a combined physical system-computer simulation experiment, in a scenario where the goal is to optimize the means of the physical system outputs. The authors implement a minimax fitness function for guiding the sequential search for new vectors of simulator control settings when additional observations on the physical system are to be taken. They use a Bayesian calibrated model to maximize the posterior expected fitness function. When additional runs of the simulator are to be taken, the method chooses the input settings that minimize the sum of the squared prediction errors. The design is empirically verified on simulated data.

Chen, P.-H. A., Villarreal-Marroquín, M. G., Dean, A. M., Santner, T. J., Mulyana, R., & Castro, J. M. (2018b). Sequential design of an injection molding process using a calibrated predictor. Journal of Quality Technology, 50(3), 309-326.

This paper features a case study on sequential planning of experiments to optimize an injection molding process. The authors identified factors and outcomes to design physical and simulated experiments using the sequential design method by Chen et al. (2018a). This method uses Bayesian calibrated predictors and an expected improvement function to choose the next design.

Chernoff, H. (1959). Sequential Design of Experiments. The Annals of Mathematical Statistics, 30(3), 755-770.

This early work on the sequential design of experiments presents two particularly important conclusions. The first is the characterization of a stopping rule based on bounds of the likelihood ratio test, similar to Wald (1943). The second is the characterization of the optimal choice of experiments, which can be thought of as a probability distribution over the set of available experiments. For the special case where there are only a finite number of states of nature and a finite number of available experiments this procedure is shown to be "asymptotically optimal" as the cost of sampling approaches zero. The procedure can be partially described by saying that at each stage the experimenter acts as though he is almost convinced that the current maximum likelihood estimate (MLE) of the state of nature is very close to the true state of nature. In problems where the cost of sampling is not small, this procedure may leave something to be desired. More specifically, until enough data are accumulated, the procedure may suggest very poor experiments because it does not sufficiently distinguish between the cases where the MLE is a poor estimate and where the MLE is a good estimate. When the cost of experimentation is small, this is relatively unimportant as more experimentation can be done more easily. Two important issues are discussed: when to stop experimentation, and which experiment to choose if it is decided that experimentation is to continue. The optimal stopping rule is characterized by bound on the sequential likelihood ratio test. The optimal choice of experimentation can be characterized by a potentially non-unique mixed strategy; that is, a (possibly non-unique) probability distribution over experiments.

Chernoff, H. (1968). Sequential Designs. Technical Report No. 35. Department of Statistics, Stanford University.

A sequential design problem is a design problem for which the choice of optimal setup itself may vary with acquired information. To clarify this, consider the example from Chernoff (1968) in which we wish to test the reliability of a type of component. These components have a probability of success, p, and can be arranged in series, so the probability of success of such a setup is p^r, where r is the number connected components. An experiment, e_r, can be thought of as a setup with r components connected in series; hence, experiments can be indexed by r. Suppose the cost of a trial or run (that is, the collection of one observation from an experiment) is 1+r. An optimal design would account for both the information gained from and the cost of each trial. For this component scenario, and Chernoff (1968) illustrates that an optimal design for this scenario calls for r=3. However, in situations for which the cost of a trial is not known, an analytic solution for the optimal choice of experiment may not be available; hence, the researcher will want to update her choice of experiment as she gains information regarding the costs and benefits through experimentation. For example, if the researcher does not know the cost of an experiment e_r, she will start by choosing an r, say r=1. As she conducts trials, she gains information regarding both the probability of success of a component and the cost of experiment e_1. She must decide between continuing to collect data via experiment e_1or choose another experiment e_r with r different from 1. As she varies experiments, she learns the costs of other experiments and eventually will have enough information to determine that the optimal choice of experiment involves r=3.

Chernoff, H. (1972). Sequential Analysis and Optimal Design. Society for Industrial and Applied Mathematics. Philadelphia, PA.

The addition of covariates is neither necessary nor sufficient to characterize a design problem. However, allowing the choice of covariates for observation n+1 to depend upon results obtained using the first n observations does provide an example of a sequential design problem as does allowing the choice of the most informative values of the covariates to depend upon acquired information. According to Chernoff (1959), sequential design problems are those which involve “sequential searching for relevant and informative experiments.” An early illustration of a non-sequential design problem without covariates due to Hotelling and discussed in Chernoff (1972) involves estimating the weights of eight objects in the presence of measurement error by use of a chemist’s scale, selecting the best setup for an experiment based on the expected information gained among all experiments, and the cost of each experiments. While a naïvely planned experiment requires examination of 64 different cases, a carefully constructed optimal design requires only eight. The design procedure is characterized by the search for this optimal experimental setup.

Chernoff, H. (1981). Lectures on Optimal Design and Sequential Analyses. Institute of Statistics Mimeo Series 1349.

Cuturi, M., Teboul, O., & Vert, J.-P. (2020). Noisy Adaptive Group Testing using Bayesian Sequential Experimental Design. arXiv preprint arXiv:2004.12508.

This article proposes a method that could be used when trying to determine the prevalence of a disease in a population. The goal of the method is to find the next group of people to test for an infection based on some criteria

and results from observed data. These criteria could be the mutual information or the area under the curve of a receiver operating characteristic curve, and the observed data are past infection detection test results. The method is capable of handling a noisy environment where tests for infection could be wrong. The authors show that this sequential Bayesian optimal method is more efficient (fewer tests needed) than testing each person for infection when the disease prevalence is low.

Daldal, R., Özlük, Ö., Selçuk, B., Shahmoradi, Z., & Ünlüyurt, T. (2017). Sequential testing in batches. Annals of Operations Research, 253(1), 97-116.

The authors study an extension of the sequential testing problem where - instead of performing one costly test at a time to identify the state of a system - it is possible to perform multiple tests simultaneously, or in batches, in order to gain a cost advantage through reduced total fixed costs. This approach is oriented towards medical clinical trials but described in sufficient generality to be applicable to many different scenarios. To model a batched approach to sequential testing, the authors assume that the total cost of each test consists of a fixed and a variable component, with the variable cost portion for each batch equal to the sum of attribute costs of individual tests, while the fixed cost - corresponding to set up, ordering, transportation, administration, etc. - is incurred only once per batch, regardless of the number or type of individual tests involved. The main decision is to find out how the tests should be batched together and optimally sequenced to yield the minimum expected total cost. The authors' approach is more mathematically rigorous than is usually found in sequential testing and clinical trials literature, while still retaining flexibility. They assume a priori knowledge of which groups of tests are allowed to be executed together, thus allowing that not every subset of the tests to be performed can be batched. They also determine the complexity of the model, and develop efficient heuristic algorithms and compare their performances on simulated examples.

Deonandan, I., Valerdi, R., Lane, J. A., & Macias, F. (2010). Cost and risk considerations for test and evaluation of unmanned and autonomous systems of systems. Paper presented at the 2010 5th International Conference on System of Systems Engineering.

This article addresses the challenges of testing Artificial Intelligence (AI)/ Autonomous Systems (AS) on budgeting and proposes a risk-based, parametric cost estimation procedure to help with choosing, scoping, and planning test events. The authors mention allocation of testing effort (specifically field vs simulation), testing throughout the life cycle of the Unmanned Autonomous System of Systems (UASoS), prioritizing the testing process based on risks, and the optimal stopping point of a test based on trade-offs between marginal effort and marginal risk. While they do not appear to be aware of sequential analysis, they follow a “seven step modeling methodology” that could be described as sequential thinking. The seven-step methodology follows:

Step 1: Analyze existing literature

Step 2: Perform Behavioral analysis

Step 3: Identify relative significance

Step 4: Perform Expert Judgment, Delphi Analysis

Step 5: Gather Project Data

Step 6: Determine Bayesian A-Posteriori Update

Step 7: Gather more data: Refine model

Dodge H. F. & H. G. Romig (1929). A method of sampling inspection. The Bell System Technical Journal. Vol. 8, pp. 613-631.

Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2014). A Sequential Monte Carlo Algorithm to Incorporate Model Uncertainty in Bayesian Sequential Design. Journal of Computational and Graphical Statistics, 23(1), 3.

The authors introduce a Bayesian method to choose the next design point. They use a utility function to discriminate between experimental designs and also incorporate model uncertainty when designing the experiment. The best design is the one that maximizes the mutual information (or dependence) between variables in a model and a future observation. The authors explain the method and present a few examples using datasets that include one factor and one outcome.

Fisher, R. A. (1952). Sequential experimentation. Biometrics, 8(3), 183-187.

Fisher states, "This willingness to learn from [the experiment] how to proceed is the essential quality of sequential procedures." Some might refer to this as Sequential Thinking. Specifically, Fisher mentions that "Wald introduced the sequential test, but the sequential idea is much older"; however, he does not provide any citations or references. The paper is Fisher's notes on a lecture he delivered on June 18, 1952 at Institute of Statistics Conferences, Blue Ridge, N. C. The paper attempts to discuss sequential experimentation with two examples from the field of genetics, but the examples given are only loosely, and colloquially rather than formally, related to sequential experimentation. Overall, this paper does not have much content; it only contains two examples from the field of genetics regarding experimentation with brief comments on how they can be interpreted as sequential experimentation. For example, in describing the test of milk yield from cows, the relation to sequential experimentation is the quote "There should be some such method of using animal reactions to guide experiments in milk yield. I would call this a sequential test of milk yield."

Freeman, L. J., Johnson, T., Avery, M., Lillard, V. B., & Clutter, J. (2018). Testing defense systems. Analytic Methods in Systems and Software Testing, 441.

This paper highlights core statistical strategies that have proven useful in the testing of defense systems. The authors include case studies to illustrate the use of statistical techniques to the design of test and analysis of resulting data. The authors note that all of the case study examples included in the paper are "one-shot" experiments as opposed to sequential experiments. A "one-shot" experiment is one in which there is no current plan for immediate follow-up experimentation. The authors note that it has long been a goal of DoD testing to conduct integrated testing, where data from earlier test phases could be used to either augment or inform later testing, but that it is often unachievable due to competing test objectives and limited resources. This fact does not stop the authors from encouraging the use of "experimental campaign sequential designs", where the idea is that a defense test plan employ a series or sequence of smaller tests with the goal being to learn from one test and modify subsequent tests based on this information. Results from preliminary tests may lead the evaluator to drop or add factors, modify the levels of a factor, or to the creation of more precise response variable measures. One area where sequential designs are commonly employed is in ballistic resistance testing (Johnson et al. (2014)).

Gerke, O., Vilstrup, M. H., Halekoh, U., Hildebrandt, M. G., & Høilund-carlsen, P. F. (2017). Group-sequential analysis may allow for early trial termination: illustration by an intra-observer repeatability study. EJNMMI Research, 7(1), 1-6.

This case study presents the implementation of interim analyses by using alpha-spending functions and sequential one-sided hypothesis tests. An alpha-spending function is a function that specifies the distribution of the Type I error across interim analyses (Lan and DeMets, 1983). The authors use data from an intra-observer ovarian cancer study to show that a trial could be stopped early which will lead to less testing. Testers should consider a few considerations when following this method: the plan should be defined at the beginning and the sample size at each interim analysis must be large (for example, one third of the total sample size) to minimize the effect of any outliers.

Ghosh, B. K., & Sen, P. K. (1991). Handbook of sequential analysis. New York: M. Dekker.

This comprehensive text covers many developments in sequential analysis. Ghosh and Sen provide a historical and comprehensive analytical overview of many developments in sequential analysis, including the Sequential Probability Ratio Test (SPRT) and Generalized SPRT (GSPRT), stopping times, testing with simple and multiple hypotheses, point and interval estimation, group sequential testing, methods for finite populations, parametric and non-parametric methods, decision theory, hierarchical and empirical Bayes sequential estimation, adaptive control, sequential design and allocation, and many more topics.

Ghosh, M., Mukhopadhyay, N., & Sen, P. K. (1997). Sequential estimation. Hoboken: New York: Wiley.

The book, written for an audience of researchers and advanced graduate students, focuses on sequential estimation, a sub-topic of sequential analysis that the authors indicate has not received much attention (from a text book perspective). In particular, they note that significant advances have been made that are not captured in any text, and there is a need to tie up the diversities in sequential estimation in a unified manner. This text covers classical and modern techniques, and many topics not covered elsewhere including shrinkage, empirical and hierarchical Bayes procedures, time-sequential estimation, and others.

Gilman, J. F., Fronczyk, K. M., & Wilson, A. G. (2019). Bayesian modeling and test planning for multiphase reliability assessment. Quality and Reliability Engineering International, 35(3), 750-760.

The authors propose a method for combining time to failure data to assess the reliability of a system and to plan for future reliability testing by way of an assurance test. The method combines the failure rate data for multiple failure modes by using a Bayesian hierarchical model. The authors show how to apply their method through theoretical explanations and by using an example. This method is valuable to the testing of military systems and should be considered.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society: Series B (Methodological), 41(2), 148-164.

This paper introduces the Dynamic Allocation Index, sometimes called the Gittins Index, the use of which provides an optimal solution to a large class of sequential design problems. According to Gittins, "Forwards induction policies are optimal for a class of problems [described in this paper], in which effort is allocated in a sequential manner between a number of competing candidates for that effort. ... These candidates will be described as alternative bandit processes. From the optimality of forwards induction policies it follows that a dynamic allocation index (DAI) may be defined ... with the property that an optimal policy must at each stage allocate effort to one of those bandit processes with the largest DAI value. ... The paper aims to give a unified account of the central concepts in recent work on bandit processes and dynamic allocation indices; to show how these reduce some previously intractable problems to the problem of calculating such indices; and to describe how these calculations may be carried out. Applications to stochastic scheduling, sequential clinical trials and a class of search problems are discussed."

Golz, M., Fauss, M., & Zoubir, A. (2017). A bootstrapped sequential probability ratio test for signal processing applications. In (pp. 1-5): IEEE, Piscataway, NJ.

The bootstrapped Generalized Sequential Probability Ratio Test (GSPRT) algorithm can reduce the incidence of erroneous early termination in the GSPRT. While there is a computation cost associated with the algorithm, this is mitigated by the idea that it only need be used when the regular GSPRT would terminate. Erroneous decisions to terminate a test early based on the GSPRT is an issue. Golz et al. provide a method of protecting against this early termination by what amounts to a more conservative test statistic. It is convenient, as this method only needs to be applied if the GSPRT suggests termination of the experiment (in favor of accepting or failing to accept the null hypothesis). They claim that while the simpler thresholds used with the SPRT are generally not valid for the GSPRT, their bootstrap algorithm allows use of these simpler thresholds with the GSPRT. However, they do not prove this claim. Instead, they offer limited simulation evidence in a situation where their bootstrapped GSPRT outperforms the GSPRT. This method also seems to introduce a nuisance parameter: the significance level for the bootstrap confidence interval, but this is not discussed.

Govindarajulu, Z. (1975). Sequential statistical procedures. New York: Academic Press, 1975.

This graduate level textbook covers the Sequential Probability Ratio Test (SPRT), sequential testing for simple and composite hypotheses (respectively, hypotheses that completely specify the distribution and those that do not), and sequential estimation. Additional topics include expected sample sizes, decision rules, double sampling procedures, generalized SPRTs (GSPRTs), extensions to the SPRT for grouped data, sequential decision problems, tests among three or more hypotheses, tests for two-sided alternatives, Bayes sequential procedures, non-parametric sequential test procedures, two-stage estimation procedures, sequential estimation by intervals or sets, minimax estimation, Bayes sequential estimation, sequential confidence bounds for linear regression parameters, sequential multivariate estimation, fixed width confidence intervals, non-parametric sequential estimation, and some optimal stopping problems.

Grant, A. M., Altman, D., Babiker, A., Campbell, M. K., Clemens, F., Darbyshire, J., . . . Pocock, S. (2005). Issues in data monitoring and interim analysis of trials. Health technology assessment (Winchester, England), 9(7), 1-iv.

This report is concerned with data monitoring and interim analysis of clinical trials. The report addresses the role of data monitoring committees (DMC), which may be of relevance for defense testing. In clinical trials, a DMC is used to judge the evidence and make recommendations on whether to continue or terminate a trial. Their primary purpose is to ensure that continuing a trial according to its protocol is ethical. The report lays out conclusions for DMC based on a systematic literature review, surveys, and case studies. The report provides recommendations for future research related to DMCs. The report also includes an appendix with a summary of statistical approaches to data

monitoring as it is universally accepted that monitoring of clinical trials will involve some form of statistical analysis. The report reviews four major statistical paradigms (Frequentist, Likelihood, Bayesian, and Decision-Theoretic) and how each one tackles the issue of early stopping in the face of apparent benefit of active intervention.

Hack, N., Brannath, W., & Brueckner, M. (2019). Estimation in adaptive group sequential trials (Version R package version 2.3.2): CRAN.

This document explains the use of the AGSDest R package. The R package allows the user to compute repeated confidence intervals and p-values as well as confidence intervals and p-values based on the stage-wise ordering (have an exact coverage probability) in group sequential designs (GSD) and adaptive group sequential designs (AGSD). The authors review GSD and AGSD, and provide an explanation of the use of the AGSDest package with examples.

Haldane, J. (1945). On a method of estimating frequencies. Biometrika, 33(3), 222-225.

Haldane's procedure, referred to as Inverse Binomial Sampling, is an early example of sequential estimation before the term sequential estimation was used. In particular, Inverse Binomial Sampling allows estimation of a frequency with a lower variance than is possible from traditional methods. From the paper, "If the frequency of an attribute is estimated by counting a sample until m members possessing this attribute are found, and if the total number in the sample is n, then (m-1)/(n-1) is an unbiased estimate of the frequency, and its variance is very approximately [m(n-m)]/[(n^2)(n-1)]". Hence, Haldane indicates that the estimator from this sequential estimation procedure has a lower variance compared to its non-sequential counterpart.

Hall, P. (1983). Sequential estimation saving sampling operations. Journal of the Royal Statistical Society: Series B (Methodological), 45(2), 219-223.

Hall provides a method for constructing a confidence interval with predetermined coverage and width that combines the advantages of both Stein's (1945) two-sample technique and the fully sequential procedure attributed to Anscombe (1953) and Chow and Robbins (1965). He discusses the trade-offs implicit in utilization of the two techniques and describes how his procedure can balance these trade-offs to achieve the benefits attributed to both techniques.

Hamada, M. S., Martz, H., Reese, C. S., & Wilson, A. (2008). Bayesian Reliability. New York, NY: Springer.

This book includes a chapter on reliability assurance testing. The authors explain how to plan for a reliability test using data from a previous test. The method is applicable to data following various distributions (Binomial, Poisson, exponential, and Weibull). Planning for a reliability test using the methods discussed in this chapter implies determining the number of tests needed while controlling for a specified consumer and producer risk...,

Hamada, M. S., Wilson, A. G., Weaver, B. P., Griffiths, R. W., & Martz, H. F. (2014). Bayesian Binomial Assurance Tests for System Reliability Using Component Data. Journal of Quality Technology, 46(1), 24-32.

This article presents an extension of assurance testing for binomial data. In this case, the planning for a reliability test includes the results of previous reliability testing for system's components. The authors illustrate an application of Bayesian assurance testing and show how using prior information reduces the number of trials planned for future testing.

Hansen, P. R. (2005). A test for superior predictive ability. Journal of Business & Economic Statistics, 23(4), 365-380.

This paper provides a method, which offers some improvements over White's (2000) Reality Check, to account for the effects of data snooping. In particular, Hansen makes "three contributions in this article. First, [he] provides a theoretical analysis of the testing problem and exposes some of its important aspects. [His] theoretical results reveal that the [Reality Check of White (2000)] can be manipulated by including poor and irrelevant forecasts in the set of alternative forecasts. This problem is alleviated by studentizing the test statistic and by invoking a sample-dependent null distribution. The latter is based on a novel procedure that incorporates additional sample information to identify the "relevant" alternatives. Second, [Hansen] provide[s] a detailed explanation of a bootstrap implementation of our test for [superior predictive ability]. Third, [he applies] the tests in an empirical analysis of U.S. inflation. [The] benchmark is a simple random-walk forecast that uses current inflation as the prediction of future inflation. The benchmark is compared with a large number of regression-based forecasts, and our empirical results show that the

benchmark is significantly outperformed. Interestingly, the strongest evidence is provided by regression models that have a Phillips curve structure."

Hess, J. T., & Valerdi, R. (2010). Test and evaluation of a SoS using a prescriptive and adaptive testing framework. Paper presented at the 2010 5th International Conference on System of Systems Engineering.

This paper implicitly defines an example of the implementation of a sequential planning procedure for System of Systems (SoS) within the autonomy literature. The authors introduce a framework that incorporates sequential elements to accommodate rapid planning and replanning. In particular, this framework enables "users to use information learned during the test process to improve the effectiveness of their own testing rather than simply follow a preset schedule. This capability is particularly attractive in the domain of Systems of Systems testing because the complexity of test planning and scheduling make frequent re-planning by hand infeasible." The authors identify several key questions, including: “How much testing is enough?”, “How do I prioritize my tests?”, and “How do I test effectively in a compressed schedule?”. They develop a Decision Support System that leverages design of experiments to “facilitate planning tests of a system by taking into account information already known”, defect modeling to determine when to stop testing, and exploratory testing to encourage “the tester to use information learned from their previous tests to select the next test case”. Defect modeling can not only be used to determine when to stop testing but can additionally be used to estimate the right amount of testing needed by considering the benefits and costs associated with finding the next failure. Design of Experiments easily accommodates adaptive techniques, such as sequential analysis, and it accommodates quick replanning while minimizing operator input. The authors summarize their framework as “tests most likely to reveal important deficiencies should have a higher priority.”

Huan, X., & Marzouk, Y. M. (2014). Gradient-based stochastic optimization methods in Bayesian experimental design. In. Ithaca: Cornell University Library, arXiv.org.

The aim of this work is to develop an approach for optimal experimental design for nonlinear systems from a Bayesian perspective, with the goal of choosing experiments that are optimal for parameter inference. The specific mathematical objective is the expected information gain in model parameters, which in general can only be estimated using Monte Carlo methods. Maximizing this objective thus becomes a stochastic optimization problem, and gradient-based stochastic optimization methods are then brought to bear.

Johari, R., Pekelis, L., & Walsh, D. J. (2019). Always valid inference: Bringing sequential analysis to A/B testing. In arXiv preprint arXiv:1512.04922.

A/B Testing is the industry term for a completely randomized controlled trial, where two versions of something (usually a control and a treatment) are compared to figure out which performs better. A/B tests are typically analyzed via frequentist p-values and confidence intervals; but these inferences become unreliable if users endogenously choose sample sizes by continuously monitoring their tests. In A/B testing practice, users are not constrained to simply analyze the output of an experiment; they can also adjust the experimental design in response to the data observed. This type of behavior can entirely undermine statistical reliability. A particularly pervasive form of this behavior is continuously monitoring the p-values and confidence intervals in order to set the sample size of an experiment dynamically. The incentive to continuously monitor experiments is strong because of the opportunity cost of longer experiments is large. The value of detecting a true effect as quickly as possible or giving up if no effect is desirable. No corrections for continuous monitoring are typically made in industrial practice. Consequently, the results are often invalid. Very high false positive probabilities are obtained. This paper introduces "always valid p-values and confidence intervals" to obtain multiple hypothesis testing control in the sequential context. Their proposed method lets users take advantage of data as soon as it becomes available, providing valid statistical inference whenever they make their decision. Their always valid p-values are derived from the sequential test: mixture sequential probability ratio test (mSPRT) (Robbins 1970), which is an extension of Wald (1947).

Johnson, N. (1961). Sequential analysis: A survey. Journal of the Royal Statistical Society: Series A, General, 124(3), 372-411.

In this review of sequential analysis, Johnson provides a simple guide to some techniques of sequential analysis that are available for practical application along with a discussion of the theory underlying such techniques. In his own words, "The development of sequential methods has been much stimulated by the sequential probability ratio test introduced by Abraham Wald, and discussion of procedures based on this test occupies the major part of this

survey (Section 3). Sequential estimation is also discussed (Section 5). Other techniques described include the stochastic approximation method of Robbins and Monro (Section 7); the "up-and- down" method of Dixon and Mood for quantal response problems (Section 7); Stein's two-sample procedure (Section 6); and inverse sampling (Section 8)."

Johnson, R. T., Hutto, G. T., Simpson, J. R., & Montgomery, D. C. (2012). Designed Experiments for the Defense Community. Quality Engineering, 24(1), 60-79.

In 2010, DOT&E published guidelines to mold test programs into a sequence of well-designed and statistically defensible experiments. This article presents the underlying tenants of design of experiments, as applied in the DoD, focusing on factorial, fractional factorial, and response surface designs and analyses. The authors emphasize the concepts of statistical modeling and sequential experimentation. The authors illustrate a designed experiment approach to building a test strategy and analyzing the data for military applications. They note that careful planning using all relevant test team representation (program management, operators, engineers, and analysts) must jointly develop the test program objectives, influential factors, responses to be measured, and the appropriate test matrices (i.e., experimental designs). They advocate for planning a sequence of test matrices to leverage knowledge gained from each test phase, feeding the findings of the previous testing into the scope of the one succeeding. As such a reasonable strategy in the development phase is to conduct a screening experiment followed by augmentation experiments to discern the true influential interactions and/or nonlinear effects. Conducting several separate, sequential experiments, each building on knowledge gained from the previous experiment is encouraged.

Johnson, T. H., Freeman, L., Hester, J., & Bell, J. L. (2014). A Comparison of Ballistic Resistance Testing Techniques in the Department of Defense. IEEE Access, 2, 1442-1455.

According to the authors, “ballistic resistance testing is conducted in the DoD to estimate the probability that a projectile will perforate the armor of a system under test. Ballistic resistance testing routinely employs sensitivity experiment techniques where sequential methods are used to estimate a particular quantile of the probability of perforation.” In this paper, the authors review and compare sequential methods, estimators, and stopping criteria used in the DoD to those found in academic literature. The response variable under consideration is binomial. The sequential methods reviewed and compared are: the Up and Down Method (Dixon and Moood, 1948), the Langlie Method (Langlie, 1962), the Delayed Robbins Monroe Method (Anbar, 1978), Wu's three phased approach (Wut and Tian, 2014), Neyer's Method (Neyer, 1994), the Robbins Monroe Joseph Method (Joseph, 2004), and K-in-a-row (Wehterill, el al., 1966). They find that Wu's three-phase optimal design (3POD) method is the most robust method of estimating multiple velocity quantiles. The 3POD method refers to three sequential parts: 1) Bound the response cure with perforations (failure) and non-perforations (success), 2) break separation (note: separation is an undesirable characteristic because it prevents a unique maximum likelihood estimate of the generalized linear model. Separation is encountered in regression models with a discrete outcome (such as logistic regression) where the covariates perfectly predict the outcome.), 3) refine estimate of quantile interest. All of the methods described are for a single factor design. The authors conclude that future work should investigate multi-factor sequential methods available for Ballistic resistance testing so that an additional factor (e.g., obliquity angle) could be added to the test and evaluation because a single factor sequential design executed numerous times within a multi-factor factorial is not a defensible design.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Transactions of the ASME--Journal of Basic Engineering, 82(Series D), 35-45.

The Kalman filter, first described in this paper, is a recursive estimation procedure that is used to produce and sequentially update a linear projection for a dynamic system as new data is observed and is used in many applications for navigation, control, signal processing and many other topics of study.

Kim, J. H., Seo, K.-M., Lee, T.-E., & Choi, B.-W. (2018). Achieving new insights into combat engagement analysis via simulation-based sequential experimentation. Military Operations Research, 23(4), 51-80.

This paper discusses practical simulation analysis via sequential experimentation for a military study, the basic philosophy for which is that an engagement scenario should be sequentially evolved with changes in operational concepts and tactics during simulation. The authors propose a process for simulation-based sequential experimentation, classified into 1) general phases - i.e., designing, modeling, simulation, and analysis; and 2) cycling to maximize the flexibility of experimental factors as well as the scenario itself. The focus of the approach is to use

sequential simulation to identify the factors that are most important in representing a target system and, thus, determining which model is the most suitable among several models to be adapted with the target system. The authors use the proposed simulation experimentation to obtain new insights into a naval warfare combat engagement covering various types of combat entities - e.g., submarines, warships, torpedoes, and decoys. In this simulation, the key question is how to evolve the scenarios sequentially by deploying false targets, or decoys, efficiently from a defender side and distinguishing false targets accurately to hit a real target from an attacker side. With extensive experiments, the paper shows various empirical investigation and lessons learned, which provides a guide to those who desire to analyze combat effectiveness and make decisions about future warfare.

Kirk, J. L., & Fay, M. P. (2014). An Introduction to Practical Sequential Inferences via Single-Arm Binary Response Studies Using the binseqtest R Package. The American Statistician, 68(4), 230-242.

This paper provides the main ideas of sequential testing through the simple case of binary response, and show examples using the authors' R package binseqtest. The authors provide a helpful summary on the differences between a fixed sample size design and a sequential design. They provide an overview of some sequential testing methods that use stopping rules (for example, O'Brien and Fleming (1979)), and show how to apply the methods by using their R package binseqtest.

Lai, T. L. (2001). Sequential Analysis - Some Classical Problems and New Challenges with rejoinder. Statistica Sinica, 11(2), 303-408.

Since a large amount of theoretical work regarding sequential testing was addressed prior to 2001, this review still provides a thorough review of sequential analysis and points to many useful references. In particular, Lai provides an analytic and historical overview of sequential analysis and several subtopics including sequential estimation, sequential testing, and sequential design of experiments. Further, an attached comment by Herman Chernoff (beginning on page 49 of the review) expands upon the discussion of sequential experimental design, and other comments in the attached rejoinder expand upon other topics. Additionally, Lai discusses advances, challenges, and opportunities for sequential analysis as the field moves forward.

Lo, A. W., & MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. The Review of Financial Studies, 3(3), 431-467.

Lo and Mackinlay discuss a particular type of data snooping and its associated consequences that commonly appear in analyses in financial economics. The authors adopt the definition of data snooping described by the situation in which you know the null distribution of a test statistic T(A) for some fixed value of A, but you instead use some A that is chosen based on the information contained in the data. They show that this can lead to rejection of the null hypothesis with high probability even when the null hypothesis is true. In particular, tests of financial asset pricing models often first involve grouping assets into portfolios and then performing testing on the returns of the portfolios. The assignment of assets to portfolios is usually not random; instead assignment is usually based on some empirical characteristics of the assets. This often results in a form of data snooping that creates biased test statistics, and the authors show that “using prior information only marginally correlated with statistics of interest can distort inferences dramatically” when construction of the test statistics is influenced by this information. The authors extend this idea by noting “when scientific discovery is statistical in nature, we must weigh the significance of newly discovered relations in view of past inferences,” but correcting for the effects of specification searches is difficult in practice, “since such searches often consist of sequences of empirical studies undertaken by many individuals over many years” leading to data-instigated pretest biases as “future research is often motivated by the successes and failures of past investigations.”

Lovell, M. C. (1983). Data Mining. The Review of Economics and Statistics, Vol. 65, No. 1 (Feb., 1983), pp. 1-12.

This paper investigates some of the consequences of data mining, a procedure defined by the reuse of a data set for purposes such as inference or model selection. Data mining is another term for data snooping. Lovell notes that such data reuse can mislead the analyst into believing that statistically significant results exist when, in fact, they do not. In particular, Lovell notes "It is ironic that the data mining procedure that is most likely to produce regression results that appear impressive in terms of the customary criteria is also likely to be the most misleading in terms of what it asserts about the underlying process generating the data under study." A Bonferroni type bound is used to establish a simple rule of thumb for calculating the actual size of a test with some nominal significance level when data mining is present.

Lu, C. Y., Penfold, R. B., Toh, S., Sturtevant, J. L., Madden, J. M., Simon, G., . . . Kulldorff, M. (2018). Near Real-time Surveillance for Consequences of Health Policies Using Sequential Analysis. Medical Care, 56(5), 365.

This case study uses the maximized sequential probability ratio test (maxSPRT) to perform sequential analyses on a longitudinal administrative healthcare data. Although this was a proof of concept, the authors suggested that the effects of health policy can be assessed with sequential analysis of observational data in a timely manner. They confirm the results of their sequential analysis with the results from a previous study that used the entire dataset for the analysis.

Mahalnobis, P. C. (1940). A sample survey of the acreage under jute in Bengal, with discussion on planning of experiments. Proc. 2nd Ind. Stat. Conf., Calcutta, Statistical Publishing Soc.

Malek, A., Katariya, S., Chow, Y., & Ghavamzadeh, M. (2017). Sequential multiple hypothesis testing with type I error control. Paper presented at the International Conference on Artificial Intelligence and Statistics.

The authors of this article propose extending multiple hypothesis testing to a situation where data is collected sequentially. The idea is to sequentially apply multiple hypothesis testing to obtain sequential p-values, while controlling for Type I error, false discovery rate, and family-wise error rate. The authors review hypothesis testing, present the problem and proposed their method, including the theory, examples, and the algorithm.

MIL-HDBK-781A, Reliability Test Methods; Plans, and Environments for Engineering Development, Qualification, and Production, Department of Defense, Washington, DC, 1996.

MIL-HDBK-781A provides guidance on test methods, test plans, and test environmental profiles which can be used in reliability testing during the development, qualification, and production of systems and equipment. Chapter 5 provides information and guidance for selecting methods and plans for reliability testing. Reliability test methods and plan selections discussed include: sequential probability ratio test (SPRT) plans, short-run high-risk SPRT plans, and fixed-duration test plans. The plans presented assume that the underlying distribution of times-between failures is exponential. MIL-HDBK-781A recommends a fixed-duration test plan selection when it is necessary to obtain an estimate of the true MTBF demonstrated by the test, as well as an accept-reject decision, or when total test time must be known in advance. MIL-HDBK-781A recommends a sequential test plan selection when it is desired to accept or reject predetermined MTBF values with predetermined risks of type I and type II error (α,β), and when uncertainty in total test time is relatively unimportant. An SPRT plan will save test time as compared to fixed-duration test plans having similar risks when either the demonstrated MTBF is high or very low. A short-run high-risk SPRT describes a test scenario in which the test time is limited and both the producer and the consumer are willing to accept relatively high decision risks. For an SPRT plan, total test time may vary significantly; therefore, because of the practical requirements of real-world test programs (e.g., cost and schedule) test length must be planned to truncation. MIL-HDBK-781A includes a thorough sequential test plan example.

Montgomery, D. C. (2017). Design and analysis of experiments (Ninth ed.): Hoboken, NJ: John Wiley & Sons, Inc.

Montgomery's book is well-regarded as a go to source for instruction on the Design and Analysis of Experiments. Throughout the book, Montgomery makes an argument for applying a sequential approach to the planning an execution of an experiment. In fact, he makes the statement that experiments are usually iterative and that it is unwise to design too comprehensive of an experiment at the start of a study. Montgomery specifically states, “A single comprehensive experiment requires the experimenters to know the answers to a lot of questions, and if they are wrong, the results will be disappointing. This leads to wasting time, materials, and other resources and may result in never answering the original research questions satisfactorily." For this reason, Montgomery recommends no more than 25 percent of the available test resources be invested in the first experiment. Regarding ways to implement a sequential experimental design strategy, Montgomery suggests starting with a screening design where many factors can be tested in order to assess their importance. Using the results from the screening design, testers can augment the design by adding additional experimental runs. An augmented design can help determine if higher order terms are needed in the statistical model. There are many ways to plan for sequential data collection. For example, testers might also consider the use of a fractional factorial designs, where a subset of the data points are collected and analyzed. Once analyzed, testers can determine if they need to collect the remaining data points to complete the factorial design. Montgomery also points to the use of RSM methods.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). Response surface methodology: process and product optimization using designed experiments (Fourth ed.). New York: Hoboken, New Jersey: Wiley, 2016.

Identifying and fitting an appropriate response surface model from experimental data requires some knowledge of statistical experimental design fundamentals, regression modeling techniques, and elementary optimization methods. Response Surface Methodology is the integration of these three topics. That is, RSM is a collection of statistical and mathematical techniques useful for optimizing processes and product designs. The RSM approach is based on a philosophy of sequential experimentation, with the objective of approximating the response with lower-order polynomial in a relatively small region of interest that contains the optimum solution. This book covers topics that allow for a phased approach to sequential experimentation. Where, for example, phase zero involves applying a screening experiment to identify the important independent variables, phase one looks to further determine if the current settings or process result in an optimum, and phase two begins with the process near the optimum and, at this point, the experimenter looks to find a statistical model that will accurately approximate the true response function.

National Research Council. (1998). Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements: National Academies Press.

This report, which was sponsored by DOT&E, and compiled by the National Research Council's Panel on Statistical Methods for Testing Evaluating Defense Systems, examines the DoD's entire approach to testing and evaluating defense systems. It brings the application of statistical principles and practices, such as sequential testing, to the topic of defense acquisition. The report notes that sequential tests have the property that given a confidence level and power, the expected sample size will be smaller than the comparable fixed sample size test. In the application of operational testing, wide use of sequential design and testing could result in substantial savings of test dollars and a decrease in test time. The panel notes though, that such practical considerations as the need to obtain expedited analysis of test results and the scheduling of soldiers and test facilities make sequential testing and sequential experimentation, that is the use of a sequential design of experiments procedure, difficult to apply. Acknowledging that there are practical constraints, the panel believes that these methods should be examined for more wide spread use. To ignore sequential testing would be unfortunate given their advantages in reducing test costs.

Pauwels, E., Lajaunie, C., & Vert, J.-P. (2014). A Bayesian active learning strategy for sequential experimental design in systems biology. BMC systems biology, 8(1), 102.

The authors propose an automated Bayesian method to choose the next design point when working with kinetic parameter estimation in dynamic models, and continuous data. This method uses a criterion that averages over the parameter space. The design points are weighted in terms of costs and minimization of error functions. The authors offered ways of implementing their method, including an R package.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. Biometrika, 64(2), 191-199.

This is one of the earliest, if not the first, manuscripts on group sequential designs. In this paper, Pocock proposes the group sequential design. He explains that his method can be applied by calculating the number of individuals needed in a clinical trial and then dividing the individuals into groups of equal numbers. The accumulated data is analyzed and a decision is made: stop or continue the trial. Pocock's method is based on repeated significance tests and works for groups with two treatments, in which the outcome follows a normal distribution with known variance. The paper also presents how the method can be modified and applied to other types of response variables.

Porter, D., McAnally, M., Bieber, C., Wojton, H., & Medlin, R. (2020). Trustworthy Autonomy: A Roadmap to Assurance. Part I: System Effectiveness. Institute for Defense Analyses.

This document serves as a framework that the Test and Evaluation (T&E) community can follow in order to provide evidence that artificial intelligence (AI)-enabled and autonomous systems function as intended. At times the authors echo broad policy recommendations made by others as they will also enable T&E activities. In other places, the author makes more specific recommendations relating to test planning and analysis. The authors discuss the challenges and possible solutions to assessing system effectiveness. Of note, in relation to sequential analyses, the authors note, because obtain, verify, validate, and accredit (OVVA) testing will involve a large amount of exploratory testing, testers will need tools that help them adaptively plan test points over time, rather than monolithically designing large test events years in advance, as is often the case now. Here, sequential design of

experiments is a tool that can help testers adaptively, efficiently, and effectively execute the OVVA process. However, though sequential DOE has been used by industry for decades, there remain methodological challenges to adapting it to DoD AI&A T&E (Ahner & Parson, 2016; Hess & Valerdi, 2010), such as how to meaningfully perform factor screening on categorical variables.

Pugachev, V. S. (1984). in Probability Theory and Mathematical Statistics for Engineers, pp 242-272, Chapter 7 - Estimator Theory, Section 7.3.1 Recursive estimation of an expectation

Ray, D. M., & Roediger, P. A. (2018). Adaptive testing of DoD systems with binary response. CHANCE, 31(2), 30-37.

In this paper, the authors provide an updated review of sensitivity testing literature as related to the DoD. They note, “Sensitivity testing is a time-honored subject in military and hardware testing.” Adaptive sensitivity testing seeks to generate a set of data that supports a predictive regression model for the probability or quantile associated with various stimulus levels. It adapts to the results as each data point is generated and analyzed, thereby significantly reducing the risk of generating useless or unbalanced data sets. The U.S. Army statisticians have developed a customized sensitivity testing implementation in R called Gongo, equipping experimenters to perform four adaptive protocols in real time: Neyer test, Wan, Tian, and Wu's newly revised 3pod2.0. Application of adaptive sensitivity testing to areas beyond DoD use include psychoacoustics (e.g., hearing tests) and dose-ranging in pharmaceutical research. One drawback, noted by the authors, of the adaptive sensitivity methods is that currently, procedures are only available for experiments with single factor.

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58(5), 527-535.

This paper is often cited for introducing the two-armed bandit problem as a formulation of experimental design and publicizing the Robbins-Monro (1951) stochastic approximation method. This paper provides much historical context in sequential analysis and sequential design of experiments. In particular, Robbins references Dodge and Romig (1929, 1941) and Mahalanobis (1940) as providing early work on sequential design, but credits Wald with the first significant contribution to the theory of sequential design. Robbins notes that Wald's (1947) book *Sequential Analysis* "states the problem in full generality and gives the outline of a general inductive method of solution." Additionally, Robbins provides an exposition for the concept that it is important to account for prior knowledge in sequential decision problems. "As Professor Lai points out, the two-armed bandit problem raises the issue 'Is it ultimately economical to sacrifice, by pulling the apparently poorer arm, in the hope of thereby getting more useful information?' ... The Robbins paper illuminated the more general issue of sequential experimentation which is at the very heart of inductive inference and scientific progress. How should one use past experience to help select the next experiment to perform, or to decide to stop experimentation? Previously, sequential analysis problems had been formulated in terms of when to stop repeating a given experiment, and no serious thought was given to deciding among alternative experiments" (Chernoff in a comment to Lai, 2001).

Robbins, H., & Siegmund, D. O. (1974). Sequential tests involving two populations. Journal of the American Statistical Association, 69(345), 132-139.

The authors provide a class of sequential probability ratio tests (SPRTs) for testing which of two normally distributed treatments with a common variance has a larger mean response. They show that "the expected number of observations on the “inferior” treatment must be at least one-half that required by pairwise sampling" and provide a class of sampling rules which almost attains this theoretical bound.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. Econometrica, 73(4), 1237-1282.

Conducting multiple hypothesis tests on the same data can result in the incorrect rejection of one or more hypotheses with high likelihood; Romano and Wolf propose a stepwise procedure to account for this multiple testing issue. From the authors: "In econometric applications, often several hypothesis tests are carried out at once. The problem then becomes how to decide which hypotheses to reject, accounting for the multitude of tests. This paper suggests a stepwise multiple testing procedure that asymptotically controls the familywise error rate. Compared to related single-step methods, the procedure is more powerful and often will reject more false hypotheses. In addition, we advocate the use of studentization when feasible. Unlike some stepwise methods, the method implicitly captures the joint dependence structure of the test statistics, which results in increased ability to detect false hypotheses. The

methodology is presented in the context of comparing several strategies to a common benchmark. However, our ideas can easily be extended to other contexts where multiple tests occur. Some simulation studies show the improvements of our methods over previous proposals. We also provide an application to a set of real data."

Ryan, E. G., Bruce, J., Metcalfe, A. J., Stallard, N., Lamb, S. E., Viele, K., . . . Gates, S. (2019). Using Bayesian adaptive designs to improve phase III trials: a respiratory care example. BMC Medical Research Methodology, 19.

The authors show the use of Bayesian adaptive designs with stopping rules for futility and efficacy. They implement the method using a phase III clinical trial data set and show that a Bayesian adaptive design yields a smaller sample size. The use of the Bayesian posterior predictive probabilities as the stopping criterion at each interim analysis is useful since Bayesian methods allow for the incorporation of prior information. The interpretations are also straight forward.

Schuler, S., Kieser, M., & Rauch, G. (2017). Choice of futility boundaries for group sequential designs with two endpoints. BMC Medical Research Methodology, 17.

The authors state that stopping a trial for efficacy has received more attention than stopping the trial early for futility, and that it is important to tailor the assessment of the futility boundaries based on a specific criterion. In general, a trial is stopped for efficacy when there is enough information to reject the null hypothesis because there is a treatment effect, and it is stopped for futility when there is enough information suggesting that it is unlikely the null hypothesis will be rejected (fail to reject) because a treatment effect will not be found. The authors propose a method for stopping a trial based on an optimal futility criterion for a two-stage group sequential designs with two endpoints. Their method maximizes the probability of correctly stopping a trial for futility, and minimizes the power loss and the probability of wrongly stopping the trial.

Sieck, V. R. C., & Christensen, F. G. W. (In Press). A framework for improving the efficiency of operational testing through Bayesian adaptive design. Quality and Reliability Engineering International.

This paper presents a method for applying Bayesian adaptive designs in conjunction with the experimental design methods currently used in defense testing. Sieck and Christensen propose moving from a frequentist method to a Bayesian method when designing experiments for an operational test. They propose choosing priors based on the operational environment of the system under test. Their method allows for analysts to look at the data during predetermined interim analyses, and to stop collecting data due to efficacy or futility. They use predictive probabilities to make the decision of stopping the test early.

Silva, I. R., & Kulldorff, M. (2019). Exact Sequential Analysis for Poisson and Binomial Data (Version R package version 3.1): CRAN.

This document introduces a new R package to perform exact sequential analyses for Poisson and binomial data. It allows for the calculation of exact critical values, statistical power, expected time to signal, and required sample sizes, for group sequential analyses and other types or methods with rejection boundaries. The package uses alpha spending functions (a function that specifies the distribution of the Type I error across interim analyses; Lan and DeMets (1983)) based on Wald-type upper boundary, the Maximized Sequential Probability Ratio Test (a variant of the Wald Sequential Probability Ratio Test (1947)), or user defined alpha spending function. The user can specify the characteristics for the sequential test to be conducted (for example, number of interim analysis planned) or analyze the data at each interim analysis.

Simpson, J. (2018). Testing via Sequential Experiments Best Practice and Tutorial. Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

The paper is intended for the defense acquisition community. The paper defines sequential experimentation as a series of tests, where the information gained at each stage of experimentation should be considered in how to continue the investigation. The paper encourages the use of a staged or phased process of testing that allows for pauses, ideally with time for analysts to make any needed corrections in secondary phases. If learning takes place during test with data analysis confirmation, there may be cause to alter the plan. The paper points out the risk to a one-stage test then analyze strategy is that all the evidence gathering risk resides in one place. However, even if there is no alternative but to run all the tests in a single test session with no planned immediate follow-up, that is the execution of a one-stage test event, there are principles of sequential experimentation that one can apply, such as prioritizing the arrangement of the test runs within the full set of test events.

Simpson, J. (2020). Sequential Testing and Simulation Validation for Autonomous Systems. Paper presented at the DATAWorks, Virtual.

This presentation was delivered at the DATAWorks webinar series in 2020. The author notes that autonomous systems expect to play a significant role in the next generation of DoD acquisition programs. New methods need to be developed and vetted for two groups we know well that will be facing the complexities of autonomy: a) test and evaluation, and b) modeling and simulation. For test and evaluation, statistical methods that are routinely and successfully applied throughout DoD need to be adapted to be most effective in autonomy. One method is sequential testing and analysis. In the presentation, Simpson covers the topics of sequential experimentation, stating, it is often better to conduct a series of experiments to be most efficient and effective in test. His definition of sequential experimentation involves combining knowledge of design of experiments augmentation strategies with knowledge gained regarding factors during test. Real-time sequential test design may benefit T&E for autonomous systems, but it will require new processes and policy to be put in place. To implement sequential experimentation, we must be willing to adapt and respond quickly. The author then provides a review of methods for simulation validation, which includes covering the operational space and experimental designs, and the relationship between the truth and simulated data. Unfortunately, the talk does not provide an example of how to apply sequential experimentation. It only lays out the basic construct, which can be found in many Design Books including: Montgomery (2017) and Myers et al. (2016).

Simpson, J. R., Listak, C. M., & Hutto, G. T. (2013). Guidelines for Planning and Evidence for Assessing a Well-Designed Experiment. Quality Engineering, 25(4), 333-355.

This article is written for the defense community and proposes guidelines and evidence that span all phases of the experiment cycle, which can inform assessment of experiment planning soundness. The authors use the experiment cycle of plan, design, execute, and analyze to structure their discussion. The paper is geared toward an audience familiar with the design of experiments method. Checklists are provided for each experiment cycle phase coupled with descriptions of what would constitute evidence of successful implementation. As part of the "planning" discussion, the authors encourage the use of a sequential design of experiments strategy. They note that a sequential strategy avoids waste by leveraging new knowledge to choose the points needed to adequately model the response surface. The authors note that a series of test conducted according to well-understood principles is the best way to limit risk, manage chaos, and maximize the likelihood of correct conclusions. They note that the factor space for many of our weapon systems undergoing T&E is vast and that testers seldom know which of the variables will matter the most in driving performance. Therefore, an experimental best practice is to structure the overall test program in such a way as to test in stages with appropriate objectives and experimental designs for each stage, thereby providing periods for analysis, understanding, and redesign. The information gained at each stage is invaluable in considering how to continue the investigation. Many times, at the outset of the test, there is limited knowledge of which factors are important, the appropriate factor level ranges, the degree of repeatability or noise in the process, etc. Sequential experimentation helps build the knowledge in stages to that experiments are increasingly beneficial and much more effective than one-stage test. A one-stage test being a test where all test points / resources are exhausted in a single test event and no immediate planned follow-up for investigation. The article provides several principles of sequential assembly to follow. A limitation of the article is that there is no case study example provided for testers to follow.

Smith, C., Kelly, D., & Dezfuli, H. (2010). Probability-informed testing for reliability assurance through Bayesian hypothesis methods. Reliability Engineering & System Safety, 95(4), 361-368.

The authors present a Bayesian method based on hypothesis testing that is used to calculate the number of trials needed to demonstrate that a system met its reliability requirement. This method considers the complexity and cost of overall system testing, and incorporates prior information in the planning for a new test.

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. The Annals of Mathematical Statistics, 20(4), 502-522.

Sobel and Wald investigate a sequential test for choosing one of the three mutually exclusive and exhaustive hypotheses. They refer to this class of problems as multi-decision problems. They note that while the sequential decision procedure that they provide is not an optimum procedure, it is simple to carry out, and the results indicated that the sequential procedure requires substantially fewer observations on average to reach a final decision compared to its non-sequential counterpart.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. The Annals of Mathematical Statistics, 16(3), 243-258.

Stein presents a two-sample test with the property that the power is independent of the variance and the size of the second sample depends upon the result of the first. Such a test is not possible for samples of fixed size. The same procedure, referred to as Stein's Double Sampling procedure is used to determine a confidence interval of preassigned length and confidence coefficient for the mean of a normal distribution with unknown variance. Stein notes "In order to make the power of a test or the length of a confidence interval exactly independent of the variance, it appears necessary to waste a small part of the information. Thus, in practical applications, one will not use a test with this property, but rather a test which is uniformly more powerful, or an interval of the same length, whose confidence coefficient is a function of [the variance], but always greater than the desired value, the difference usually being slight, at the same time reducing the expected number of observations by a small amount. Any two-sample procedure, such as that discussed in this paper, can be considered a special case of sequential analysis developed by Wald (1945)."

Tal, O., McCollin, C., & Bendell, T. (2001). Reliability demonstration for safety-critical systems. IEEE Transactions on Reliability, 50(2), 194-203.

Among the two generally accepted methods for reliability demonstration testing - fixed-duration testing and sequential testing - sequential testing, which the authors equate with Wald's Sequential Probability Ratio Testing (SPRT), is more efficient in terms of the mean number of tests. Both methods, however, are based on the concept of dual, comparable risk, to a consumer and producer, for instance, as well as two corresponding unreliability values, a specified failure rate and a minimum acceptable failure rate. The authors claim that this framework is less appropriate for safety-critical systems than for non-safety-critical systems, because - among other reasons - success or failure of the system is often more relevant than its time to failure or corresponding failure rate. More specifically, failure in a safety-critical system can result in extensive non-reversible damage in terms of human life and/or environmental effects, meaning that the risk on one side may dominate the other. Single Risk Sequential Testing, or SRST, effectively eliminates one of the risk factors, not from the problem but from the model formulation. The authors describe the model in relatively simple yet rigorous terms, using modified Ball-Urn models as a framework. They claim - and show by empirical results on avionics software data - that within the context of safety-critical systems, SRST provides a higher chance of successful testing than PRST, takes less time and fewer iterations to complete, and still satisfies all the requirements of PRST.

Tartakovsky, A. G. (2014). Nearly optimal sequential tests of composite hypotheses revisited. Proceedings of the Steklov Institute of Mathematics, 287(1), 268-288.

Tartakovsky notes that the quality of a sequential test can be judged on the basis of its error probabilities and the moments of the sample size, and he is interested in tests that minimize asymptotically the moments of the stopping time distribution (up to some order). He studies two sequential tests: the multihypothesis generalized sequential likelihood ratio test (MGSLRT) and the multihypothesis adaptive sequential likelihood ratio test with one-stage delayed estimators (MASLRT). While the latter loses information compared to the former, it has an advantage in designing thresholds to guarantee given upper bounds for the probabilities of errors, which is practically impossible for the GLR (generalized likelihood ratio) type tests. He shows that both tests have asymptotic optimality properties minimizing moments of the stopping time as probabilities of errors vanish.

Telford, J. K. (1992). The number of tests needed to detect an increase in the proportion of defective devices. Johns Hopkins APL Technical Digest, 13(2), 326-331.

It is a frequent problem in clinical, industrial, and defense testing programs to determine whether a change has occurred in a population proportion relative to a known or accepted initial value. Simple binomial sampling to estimate, for instance, a proportion of defective devices from a population entails taking a random sample and then classifying the devices as either defective or nondefective. The number of tests one must conduct to detect an increase or decrease in the relevant proportion, the necessary sample size, the desired confidence level in the conclusion, the various risks, and the associated testing costs are all important variables to be considered, and these along with the specific question to be answered by the testing can lead to different statistical methods. This article surveys five widely used methods for detecting changes in population proportions, motivated primarily by missile defense testing scenarios but more broadly applicable across a variety of testing programs. The five surveyed methods include: 1. Fisher's Exact Test, used for testing whether two samples have the same proportion of

defectives/non-defectives; 2. One-Sample Neyman-Pearson Test, which tests one sample against two specific proportions of defectives, the initial defective proportion and an increase by a given amount; 3. Two-Sample Neyman-Pearson Test, which tests two samples for an increase in the defective proportion by a given amount between the first and second samples; 4. Double Sampling Plan Test, a version of a Neyman-Pearson Test where a sample is tested against two specific proportions of defectives, carried out in groups, with a statistical test made after the first group to determine whether the second one is needed (this is a simplified version of sequential testing); 5. Sequential Testing, where a sample is tested against multiple specific proportions of defectives in sequential order, with a statistical test applied after each observation to determine if the present value is accepted or if another sample test is needed.

Terejanu, G., Upadhyay, R. R., & Miki, K. (2012). Bayesian experimental design for the active nitridation of graphite by atomic nitrogen. Experimental Thermal and Fluid Science, 36, 178-193.

Bayesian experimental design allows us to evaluate different sequences of design points and choose the one that best fits the experimental purpose. The method applied in this paper uses decision theory to help analysts choose the optimal design with the goal of estimating the model parameters, or deciding when to stop collecting data. In particular, the authors use k-nearest neighbor to estimate the dependence between the model parameters and predictions, quantify this dependence by mutual information, and measures like Kullback-Leibler information to decide when to stop the data collection.

U.S. Food and Drug Administration (2019). Adaptive designs for clinical trials of drugs and biologics guidance for industry. US Department of Health and Human Services, Federal Registrar.

The Food and Drug Administration (FDA) prepared this document to provide guidance on the use of adaptive designs for clinical trials of drugs and biologics. The guidance defines adaptive designs as "prospectively planned modifications to one or more aspects of the design based on accumulating data," and later explains different types of study designs that allow for adaptation. The guidance also explains the motivation behind the use of adaptive designs; for example, one could stop a trial early if a new treatment is shown to be effective or futile. The guidance also includes suggestions regarding the information the FDA needs to evaluate clinical trials.

Wald, A. (1945). Sequential tests of statistical hypotheses. Annals of Mathematical Statistics, 16, 117-186.

This paper introduces the sequential probability ratio test (SPRT). Wald also includes the following origin story of the SPRT: "In March 1943, the problem of sequential analysis arose in the Statistical Research Group, Columbia University, in connection with a specific question posed by Captain G. L. Schuyler of the Bureau of Ordnance, Navy Department. It was pointed out by Milton Friedman and W. Allen Wallis that the mere notion of sequential analysis could slightly improve the efficiency of some current most powerful test." Wald used this idea to develop the Sequential Probability Ratio Test (SPRT). While some sequential procedures were considered prior, this introduction of the Sequential Probability Ratio Test (SPRT) is often considered the seminal paper formalizing sequential analysis. Wald defines a sequential test of a hypothesis as "any statistical test procedure which gives a specific rule, at any stage of the experiment (at the n-th trial for each integral value of n), for making one of the following three decisions: (1) to accept the hypothesis being tested (null hypothesis), (2) to reject the null hypothesis, (3) to continue the experiment by making an additional observation. Thus, such a test procedure is carried out sequentially. An essential feature of the sequential test, as distinguished from the current test procedure, is that the number of observations required by the sequential test is not predetermined, but is a random variable due to the fact that at any stage of the experiment the decision of terminating the process depends on the results of the observations previously made." In particular, the SPRT frequently results in considerable savings in the number of observations as compared with the current most powerful test.

Wald, A. (1947a). Foundations of a general theory of sequential decision functions. Econometrica, 15(4), 279-313.

Traditional decision theory studies the problem of constructing a statistical decision function which associates each sample point x with a decision d(x) so that the decision d(x) is made when the sample point x is observed. This paper extended the theory of statistical decision functions "to the case where the number of observations required for a decision is not determined in advance, but is made dependent on the outcome of the observations. A decision function for which the number of observations needed to reach a decision depends on the outcome of the observations is called a sequential decision function."

Wald, A. (1947b). Sequential analysis: New York, J. Wiley & Sons, Inc.; London, Chapman & Hall.

This book is highly cited, being the first comprehensive textbook treatment of sequential analysis. Topics covered include the Sequential Probability Ratio Test (SPRT), expected sample sizes, decision rules, sequential testing for simple and composite hypotheses (respectively, hypotheses that completely specify the distribution and those that do not), application to acceptance inspection of a lot where each unit is classified into one of two categories, testing the difference between the means of two binomial distributions, testing that the mean of a normal distribution with known standard deviation is less than or equal to a given value, testing that the standard deviation of a normal distribution does not exceed a given value, selecting a sequential sampling plan, multi-valued decisions (choosing a hypothesis from a collection of mutually exclusive hypotheses), and sequential estimation by intervals or sets. New editions were published in 1966, 1973, and 2004. From the publisher: "In 1943, while in charge of Columbia University's Statistical Research Group, Abraham Wald devised Sequential Design, an innovative statistical inference system. Because the decision to terminate an experiment is not predetermined, sequential analysis can arrive at a decision much sooner and with substantially fewer observations than equally reliable test procedures based on a predetermined number of observations. The system's immense value was immediately recognized, and its use was restricted to wartime research and procedures. In 1945, it was released to the public and has since revolutionized many aspects of statistical practice. This book is Professor Wald's own description of the system. Part I contains a discussion of the general theory of the sequential probability ratio test, with comparisons to traditional statistical inference systems. Part II discusses applications that illustrate the general theory and raise points of theoretical interest specific to these applications. Part III outlines a possible approach to the problem of sequential multi-valued decisions and estimation. Sequential Analysis offers statistical researchers a time and money saving approach, introduces students to one of the major systems in contemporary use, and presents those already acquainted with the system with valuable background information."

Wald, A., & Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. The Annals of Mathematical Statistics, 19(3), 326-339.

While Wald (1945) introduces the Sequential Probability Ratio Test (SPRT), this paper proves its optimality for tests deciding between two simple alternatives. In particular, Wald and Wolfowitz (1948) show that among all tests with the same power for choosing between two simple alternatives, the SPRT requires the fewest number of observations on average. This paper is one of the first optimality results for SPRT related tests.

Wallis, W. A. (1980). The statistical research group, 1942–1945. Journal of the American Statistical Association, 75(370), 320-330.

The paper provides a detailed account of the origins of sequential analysis in 1943 during the Second World War. In the paper, sequential analysis is described as being one of the most powerful and seminal statistical ideas of of the 20th century. The problem of sequential analysis arose in connection with a specific question posed by Captain G. L. Schuyler of the Bureau of Ordnance, Navy Department. In response, Wald devised the sequential probability ratio test. Because of the substantial savings in the expected number of observations effected by the sequential probability ratio test, and because of the simplicity of the test procedure in practical applications, the National Defense Research Committee considered these developments sufficiently useful for the war effort to make desirable to keep the results out of the reach of the enemy, at least for a certain period of time.

Wason, J. M. S. (2015). OptGS: An R package for finding near-optimal group-sequential designs. Journal of Statistical Software, 66(2), 1-13.

This paper presents an R package for group sequential designs in clinical trials. The package uses a two-parameter stopping boundary function proposed by the author, which allows the futility and efficacy boundaries to be different. In general, a trial is stopped for efficacy when there is enough information to reject he null hypothesis because there is a treatment effect, and it is stopped for futility when there is enough information suggesting that it is unlikely the null hypothesis will be rejected (fail to reject) because a treatment effect will not be found. The stopping boundaries are the guideline for stopping the trial. The paper includes an introduction to the methods for obtaining the near-optimal and balanced designs, as well as examples of the use of the R package. Some of the limitations of the method are that it can be sensitive to the starting values for the algorithm, the sample size at each interim analysis should be equally spaced, and it works for normally distributed outcomes.

Weitzman, M. L. (1979). Optimal search for the best alternative. Econometrica, 47(3), 641-654.

This paper introduces an optimal solution to a large class of sequential search problems in the context of Economics, and the reservation price discussed here is analogous to the Dynamic Allocation Index of Gittins (1979). In fact, the two papers were published the same year independently. In particular, "this paper completely characterizes the solution to the problem of searching for the best outcome from alternative sources with different properties. The optimal strategy is an elementary reservation price rule, where the reservation prices are easy to calculate and have an intuitive economic interpretation. ... A broad class of economic search problems can be cast in the following form. There are a number of different opportunities or sources, each yielding an unknown reward. The uncertainty about the reward from a source can be eliminated, at a fee, by searching or sampling. Each source has its own independent probability distribution for the reward, search cost, and search time. Sources are sampled sequentially, in whatever order is desired. When it has been decided to stop searching, only one opportunity is accepted, the maximum sampled reward. Under this formulation, what sequential search strategy maximizes expected present discounted value? A powerful solution concept applies to the above model. Each source is assigned a reservation price - an invariant critical number analogous to an internal rate of return. The reservation price of a source is easily computed, depends only on the features of that source, and has an intuitive economic interpretation. The selection rule is to search next that unsampled source with highest reservation price. The stopping rule is to terminate search whenever the maximum sampled reward is above the reservation price of every unsampled source. This simple characterization of an optimal policy is the basic result of the present paper. Fundamental properties are derived and interpreted."

Wetterslev, J., Jakobsen, J. C., & Gluud, C. (2017). Trial Sequential Analysis in systematic reviews with meta-analysis. BMC Medical Research Methodology, 17.

The authors propose the use of Trial Sequential Analysis (TSA) when working with meta-analysis and the required sample size has not been reached. TSA treats meta-analysis as interim analysis where each time a new study is included, an interim analysis is performed. At each interim analysis, researchers decide if they have enough data or if more studies should be included in the meta-analysis. The paper includes a good overview of TSA, which, according to the authors, provides better control of Type I and Type II errors than the traditional methods for meta-analysis.

White, H. (2000). A reality check for data snooping. Econometrica, 68(5), 1097-1126.

White discusses data snooping and provides a procedure for model selection that accounts for the effects of data snooping. According to White, "Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection. When such data reuse occurs, there is always the possibility that any satisfactory results obtained may simply be due to chance rather than to any merit inherent in the method yielding the results." For example, good performance observed for a forecasting model which was obtained by an extensive specification search may simply be due to luck rather than any inherent forecasting ability for the same reason that flipping a large number of coins will result in a coin that always comes up heads with high likelihood. White notes, "This problem is practically unavoidable in [some situations]. It is widely acknowledged by empirical researchers that data snooping is a dangerous practice to be avoided, but in fact it is endemic. ... [The method presented here], the Reality Check, provides simple and straightforward procedures for testing the null that the best model encountered in a specification search has no predictive superiority over a given benchmark model, permitting account to be taken of the effects of data snooping."

Whitehead, J., & Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. Biometrics, 39(1), 227-236.

The theory underlying the use of sequential probability ratio tests (SPRT) and a subsequent closed modification of SPRT with a triangular continuation region in the design and analysis of sequential clinical trials is based on a large-sample analogy with the situation of independent normal observations. It ignores any overshoot of the stopping boundary after each test in the sequence - that is, the difference between the position of the sample path and the boundary at termination. Provided that test statistics are calculated and checked against the stopping boundaries frequently, any such overshoot will indeed be small and this framework will be adequate. However, if inspection of the data occurs only at widely spaced intervals - the situation that is often the case in group sequential trials, for instance - overshoot is likely to be large and this framework unreliable. In this paper, the authors introduce and demonstrate the accuracy of a class of overshoot corrections, comparing with simulation results for further validation. The correction method is presented within a context of comparing two new treatments or a new treatment with an

accepted existing one, though the mathematical formulation is easily adapted to clinical trials beyond this scope. A case study in evaluating an anesthetic procedure using this procedure is also briefly described.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. Journal of applied probability, 25(A Celebration of Applied Probability), 287-298.

This paper extends the work of Gittins (1979) with discussion and examples for the case where multiple projects can be operated at once and the projects can evolve even if they are not being operated. These are termed "Restless Bandits," and a choice of project can be thought of as a choice of experiment. In particular this allows not only information to be gained by the projects under operation but also allows information regarding the projects not in operation to be lost. The goal is to choose the projects in operation at each point in time so as to maximize the expected reward under a constraint on the number of projects that can be operated. Similar to Gittins (1979) and Weitzman (1979), an index is dynamically assigned to each project, and the optimal choice of projects is given by the projects with the largest current index. When the projects not being operated are static, the index reduces to the results of Gittins (1979). The author provides the following illustrative example: "suppose m aircraft are trying to track the positions of n enemy submarines, where m < n, so that aircraft must change task from time to time if all submarines are to be monitored. We regard this as a case of the operation of exactly m projects out of n, in that exactly m submarines out of the n are under surveillance at a given time. The problem is to allocate this surveillance. For this problem, the bandits are restless in the most literal sense. While a submarine is under observation, information on its position, etc., is being gained. While it is not, information is usually being lost, because the submarine will certainly be taking unpredictable evasive action."

Xun, H., & Marzouk, Y. M. (2013). Simulation-based optimal Bayesian experimental design for nonlinear systems. Journal of Computational Physics, 232(1), 288-317.

The authors present a method to find optimal experimental designs using Bayesian methods and optimization strategies, while taking into account the statistical model when designing the experiments. Their method allows them to work with noisy and incomplete data, and to incorporate different sources of information. They use utility functions to plan for the next design point or set of design points. They use approximations and simulation techniques to evaluate these utility functions.

Yuan, Y. (2016). Group Sequential Analysis Using the New SEQDESIGN and SEQTEST Procedures. Rockville, MD: SAS Institute.

This document explains the use of two SAS procedures for designing (SEQDESIGN procedure) and analyzing (SEQTEST procedure) group sequential clinical trials. The document includes an introduction to clinical trials and to group sequential clinical trials, followed by a review of the methods to design group sequential designs (for example, Armitage, McPherson, and Rowe (1969), and O'Brien and Fleming (1979)).

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
10-12-2020	Final	June - December 2020

4. TITLE AND SUBTITLE

A Review of Sequential Analysis

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Rebecca Medlin; John Dennis; Keyla Pagan-Rivera; Leonard Wilkins

5d. PROJECT NUMBER

BD-9-2299

5e. TASK NUMBER

2299(90)

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

NS D-20487

H 2020-000515

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Heather Wojton

14. ABSTRACT

The majority of statistical analyses involves 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. Unlike these traditional methods, sequential analysis is concerned with 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. Expanding the use of sequential analysis in DoD testing has the potential to save substantial test dollars and decrease test time. In this paper, we review the sequential analysis literature as a first step in providing a recommendation for its use in DoD test and evaluation.

15. SUBJECT TERMS

Design of Experiments; Operational Testing; Statistics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Rebecca Medlin
Unclassified	Unclassified	Unclassified	Unlimited	50	19b. TELEPHONE NUMBER *(Include area code)* 703-845-6731