Use of statistically designed experiments to inform decisions in a resource constrained environment

Laura J. Freeman, Ph.D.,[1] Karl E. Glaeser,[2] and Alethea Rucker[3]

[1]The Institute for Defense Analyses
[2]AVW Technologies, Inc.
[3]United States Air Force

Abstract: There has been recent emphasis on the increased use of statistics, including the use of statistically designed experiments, to plan and execute tests that support Department of Defense (DoD) acquisition programs. The use of statistical methods, including experimental design, has shown great benefits in industry, especially when used in an integrated fashion, for example see literature on Six Sigma. The structured approach of experimental design allows the user to determine what data needs to be collected and how it should be analyzed to achieve specific decision making objectives. This focuses decision-making processes, improves test efficiency and provides objective data for evidence-based decision making. Today the DoD Test and Evaluation (T&E) community is investigating the use of statistical methods to provide efficient and effective testing. This paper discusses the use of statistics in T&E to assist T&E practitioners and acquisition management in understanding how to improve the quantity and quality of information made available to decision makers to make risk assessments, even in a resource constrained environment.

Disclaimer: This paper is in no way representative of DoD doctrine. To avoid any confusion with any acquisition program, living or dead, the case study is purely imaginative. It is a discussion of the utility of designed experiments as a logical way to plan tests.

Keywords: Statistics, Design of Experiments, Experimental Design, Test and Evaluation

Introduction

The history of U.S military acquisition is a proud one, and the fighting force it equips is second to none. Capabilities such as the Nimitz class aircraft carriers, and Apache strike helicopters were developed and tested in a less constrained fiscal environment, and even in those resource abundant days, the DoD Test and Evaluation (T&E) community did not uncover system flaws that it might have. As budgets continue to face downward pressure over the foreseeable future, T&E must continue to provide objective information on technical and operational performance to decision makers to support sound risk-based acquisition decisions.

Some have proposed to expand the use of statistically based test planning and analysis methods to help ensure tests are planned and executed in a logical and efficient fashion, minimizing impact to cost and schedule, while maximizing the data gained from each test event and establishing test data validity. This idea is not new. Several studies have illustrated the merits of incorporating statistical methods in testing military systems. For example, the 1998 National Research Study, "Statistics in Defense, Acquisition and Testing" concluded that, "major advances can be realized by applying selected industrial principles and practices in restructuring the paradigm for operational testing…" and that "…the current practice of statistics in defense

testing design and evaluation does not take full advantage of the benefits available from the use of state-of-the-art statistical methodology."

Today the acquisition community is working to understand the value added of statistical methods, including Design of Experiments, in T&E. One misperception is that using statistical designs will impede the Program Manager's (PM) ability to expedite fielding of enhanced warfighter capability. An unintended consequence of this drive for a shorter test time can be a loss of test accuracy, and a failure to accurately capture variation within an entire production run (e.g. can we characterize the performance of all hand grenades produced through a subset of test units?). There are many documented cases where statistical design has improved test efficiency and provided higher quality data. Granted, there are times when the cost of generating statistically significant testing data in OT is prohibitively expensive, such as when testing submarine-launched ballistic missiles. In these cases it is especially important to ensure adequate testing at the subsystem and component level, and this is where statistical confidence in the results becomes even more important.

In this paper, we seek to simplify the lexicon of statistically designed experiments, and explain how developing statistically-based T&E information can assist in correctly planning T&E programs, interpreting T&E data, and identifying program risks for decision makers. By using a simple case study we demonstrate a methodology to determine the type and amount of test data to collect, and discuss ideas on how to reduce and analyze that data to support evidence-based decision making that balances risk and resource constraints. The case study is purely hypothetical but based on real-world acquisition program challenges experienced by the authors.

A Statistically Based Structured Approach for Decision makers

The DoD acquisition decision making process is a complex system that many have tried to model with limited effectiveness. Leadership must make risk-based system acquisition decisions that impact war fighting capability and in some instances the survivability of American war fighters, often with limited data. For example, at Source Selection, decision makers decide which system best balances cost and capability, without seeing the completed product. At Initial Operational Test and Evaluation (IOT&E), decision makers assess a system's war fighting effectiveness and suitability based on performance across operational scenarios. These decisions are based on a multifaceted process informed by many streams of information. Figure 1 depicts the basic decision process and shows the various types of information that may inform an acquisition decision. The goal of this paper is to focus on quantitative data, while recognizing that there are other sources of information that impact the decision making process.

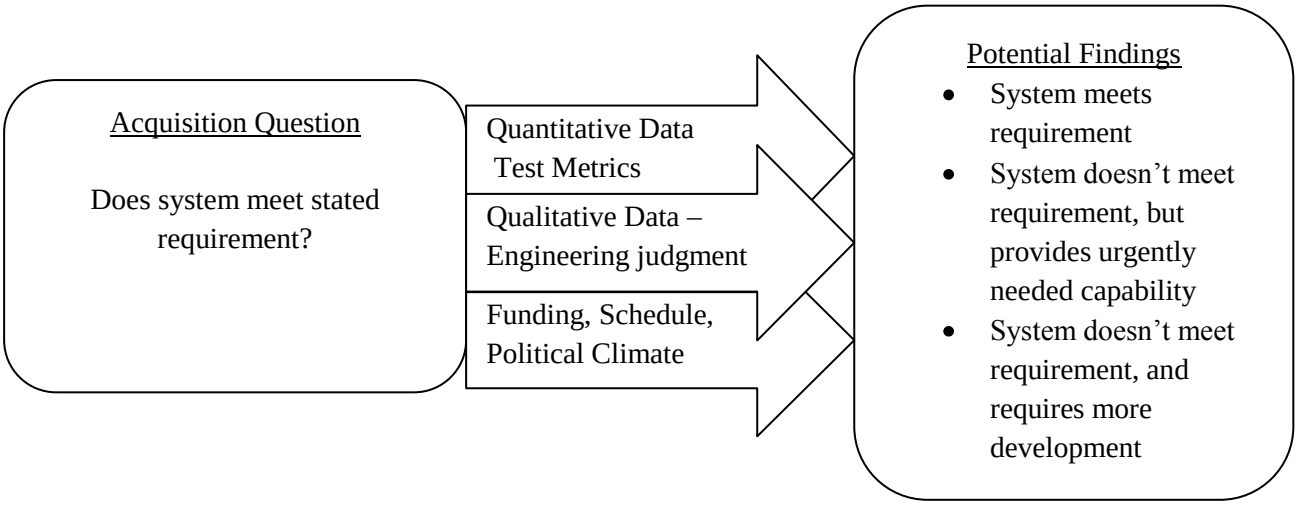

Figure 1: Elements of the T&E Decision Process

T&E teams, program managers and acquisition decision makers face many of the same questions on a daily basis. For example:

- Does the new system under test (SUT) meet system requirements?
- How does performance of the SUT vary across the intended operational envelope?
- Is the SUT reliable?
- How does one determine the answers to these questions in a logical efficient manner?

Historically, these questions have been answered through the use of a combination of subject matter expertise (SME), statistical analyses, and anecdotal case studies. However, the past approach for using statistics in making these decisions has not been well documented and often not as rigorous as the current state of the field of statistics can provide.

Back to Planning Basics

The scientific method has provided a framework for learning for hundreds of years. This simple process consists of clearly stating a question, forming a hypothesis to explain the question, planning the experiment, collecting and analyzing the data, and drawing a conclusion.

T&E is inherently related to the scientific method. In contractor and governmental developmental testing one must use a logical process to efficiently verify the technical performance of a system based on the requirements and contracts. In operational testing we must plan an efficient test program that evaluates a system's effectiveness and suitability for operational use. These decisions are based on both qualitative and quantitative data. The volume and pedigree of data required to minimize risks of drawing the wrong conclusions drives the total resources and therefore the cost of a test program. For quantitative data, the field of statistics provides an approach for qualifying those risks.

Design of Experiments is defined by Montgomery (2006) as "the process of planning the experiment so that appropriate data that can be analyzed by statistical methods will be collected, resulting in valid and objective conclusions." Montgomery (2006) and Vining (1997) outline the steps for designing an experiment that are closely related to the scientific method. We have tailored their methodologies to meet our requirements, and dubbed it the 4-D process. This is by no means a doctrinal methodology; it is simply a process framework. Others have suggested similar frameworks including the 53rd and 46th Test Wings at Eglin Air Force Base, whose framework is Plan, Design, Execute, and Analyze.

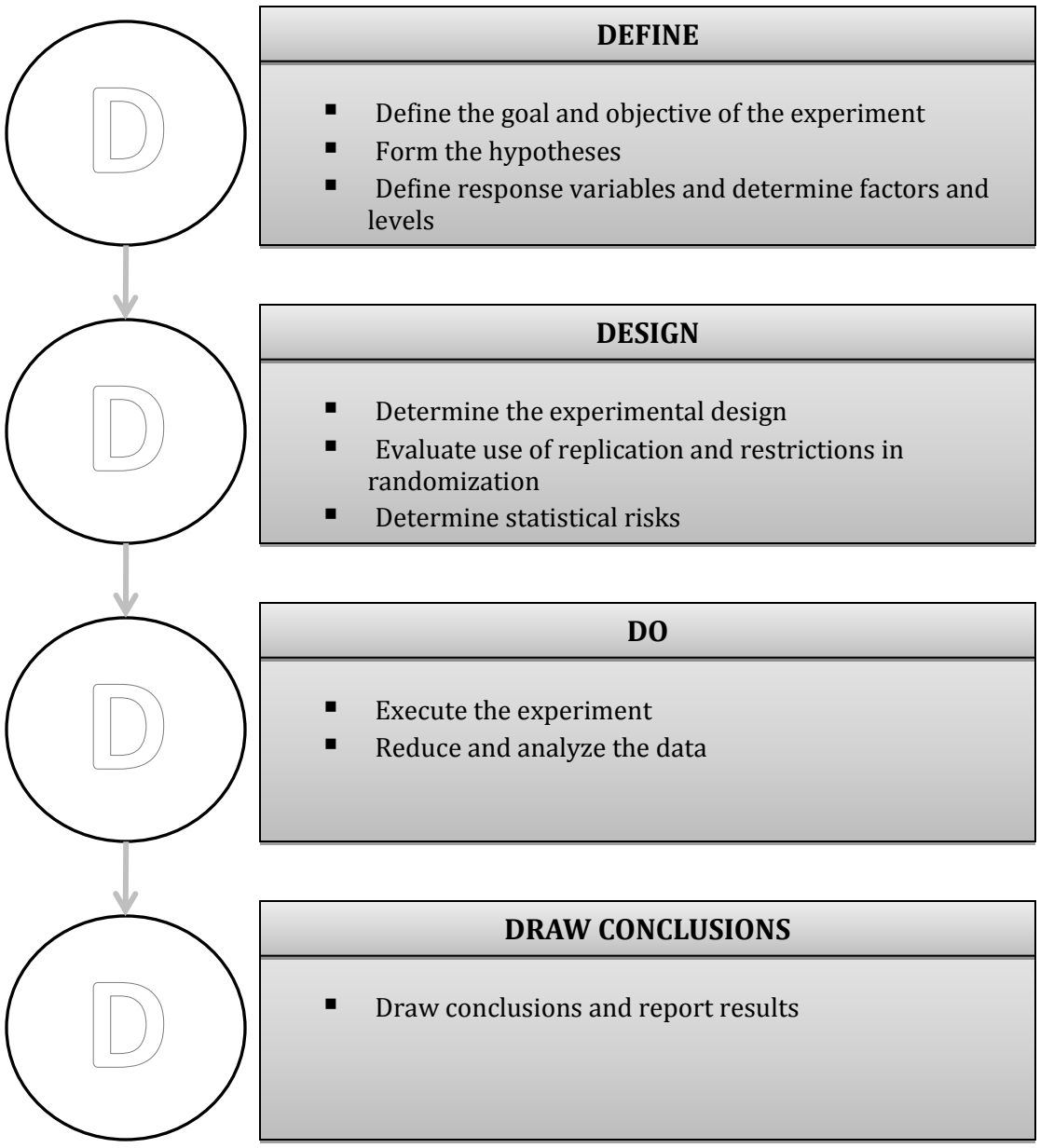

Figure 2: The 4-D process for designed experiments.

4-D is a simple and versatile process that can be applied up front and early as a foundation of an integrated test program. For maximum effectiveness, 4-D requires early collaboration among all stakeholders to identify test requirements and plan efficient test events that maximize data reuse and minimize duplication. This paper will focus on the first two phases of 4-D, as the emphasis on Define and Design truly simplifies Do and Draw conclusions.

DEFINE - A key concept in 4-D is to understand the question(s) to be answered, and then describe the question in concise quantifiable terms before planning the test. This ensures the

correct data is collected to support the analysis. The question can range from high level operational requirements (defeat enemy 5th generation fighter) or detailed technical specifications (hydrophone frequency response of 20-500 Hz).

The first step in the define stage is to determine the question that needs to be answered and the exact measures (response variables) that will be used to answer the questions. Often in T&E the response variables are specified in the requirements documents. However, this may not always be the case. In the examples above an appropriate response variable for defeating the enemy 5th generation fighter might be the exchange ratio (number of blue losses to number of red losses), whereas the response variable for the hydrophone frequency is clearly defined in the technical specifications. Determining mission oriented responses for operational requirements is a challenge that may require several iterations to get correct. When at all possible it is best to find quantitative metrics.

Once the response variable is defined, there are two possible conclusions that can be drawn from the question; the requirement is met or the requirement is not met. These two conclusions are called hypotheses. Hypothesis testing requires two hypotheses to be created, the null hypothesis and the alternative hypothesis. Formulation of the hypotheses is a critical step as the selection of the question guides the experiment and outcome.

The null hypothesis is the statement assumed to be correct and attempt to disprove. If the null hypothesis is not disproven beyond a reasonable doubt as determined by the decision maker, with the amount of data collected in the test, the null hypothesis is assumed to be correct. A common example of this concept is the legal premise of "innocent until proven guilty". The null hypothesis is the defendant is innocent and the alternative hypothesis is that the defendant is guilty. The responsibility of the prosecutor is to prove beyond a reasonable doubt the defendant is guilty. If there is reasonable doubt, then the defendant is found "not guilty". He or she may have done the crime, but there was enough ambiguity to make their guilt unclear.

Another way to think of hypotheses is as a decision maker's opinion or decision. He or she can decide that a system is good and should be accepted, or bad and returned for more work – based on the data collected in the test. Their assessment of the system performance is based on the subset of data tested. Then there is the ground truth (which is unknown, as the entire population of systems is not tested) – which is actual performance. The combination of these opinions with reality creates four potential conclusions, two of which are correct and two which are wrong, see Figure 3.

		Decision Makers Decision	
		Parachutes do not deploy properly (do not field)	Parachutes deploy properly (field to troops)
True State of Parachute Performance	Parachutes do not deploy properly	Correct Decision (Confidence Level = 1- α)	Type I Error α "User's risk"
	Parachutes deploy properly	Type II Error β	Correct Decision (Power = 1-β)

"Managers risk"	

Figure 3: Hypothesis Testing results table for a simple parachute example.

One of the values of statistical test planning is the ability to assess the likelihood of a particular test design reaching a correct decision with the resources allocated for the test. Selecting the appropriate response variable such as "time to deployment" versus "percentage of deployment" could result in a huge increase (e.g. 80-100 additional runs) in resources required! As shown in Figure 3, the metrics that provide this information are the risks associated with testing, Type I and Type II error.

The Type I error is a value (0 – 100%) typically agreed upon by the PMs based on input from T&E SMEs. It expresses leadership's risk tolerance to making a wrong decision based on limited test data. It specifically refers to the probability of rejecting the null hypothesis when the null hypothesis is true. In the parachute example above, if the null hypothesis is that the parachutes do not deploy properly then the Type I error is the largest risk that the decision maker is willing to tolerate accepting a bad parachute system. Notice the interpretation of user risk is dependent on how the null hypothesis is constructed.

The Confidence Level is related to the Type I error and is calculated by subtracting the accepted Type I error value from 100%. The resulting value is the mathematical likelihood that the test fails to reject the null hypothesis, when in fact it is true. Since we will never know the true performance of the parachute, a confidence level tells us how confident we are that the test program will capture the true performance of the system from the subset of test articles. It has nothing to do with the likelihood that the system will "meet the spec".

Type II error is likelihood a test fails to reject the null hypothesis, when the null hypothesis is false. Power is related to the type II error and is calculated by subtracting the Type II error form 100%. Power is the likelihood a test rejects the null hypothesis, when the null hypothesis is false. Power is calculated in test planning, and takes into account the effect size desired to be detected, standard deviation (σ), and the sample size (N). In the parachute example above a Type II error would be if the parachutes were determined to not meet requirements, when in fact they did, this could be interpreted as manager's risk.

It is important for the decision maker to understand both types of risks when approving testing resources because both types of test risks have programmatic implications. In the parachute example, one risk could result in fielding a bad system. The other could result in increased development time or canceling of a good program, in the most extreme cases.

Clearly, performance for all systems is not uniform across all operating conditions. Factors are the independent variables that impact the outcome of the test. Determining the factors that may impact performance is not a trivial task, and should be done by a team of SMEs, PM's and T&E professionals. General suggestions that normally result in a comprehensive list of independent variables (factors) are:

- Use a cause-and-effect (fishbone) diagram to facilitate a discussion on all potential active factors (see ASQ Quality Tools for example)
- Refine factors into a subset of testable factors by some sort of factor management process

- Design sequential experiments that allow for the testing of the appropriate factors in the right testing.
- Allow for early tests to have large numbers of factors and be less powerful, later tests can refine the active factors after less important factors are screened out

The specified values of the factors are called levels and are determined using engineering judgment based on the experiment. In general we recommend a small number of levels, when possible (2-3).

DESIGN **-** Now that we have defined the question to be answered, the factors and levels that impact the process and the desired confidence and power, the next step is to actually design a test that will provide data that answers the question with a minimum acceptable number of test events. This is the step where tradeoffs between risk and cost must be evaluated. The design must provide knowledge of how well the test has captured the variability of the system, so the decision maker can understand the risks involved in collecting a small amount of data.

Additional tradeoffs that decision makers can discuss in the context of statistical risk are:

- How much better (or worse) than the requirement does a system need to be before a decision maker will reject the system
- How much risk (Type I and Type II) is the decision maker, tester, PM, etc. willing to accept when testing the system. Some have mentioned that the default value should be 80% confidence level and 80% power. For some systems this may not be appropriate – for example parachutes or body armor! Many researchers use 95% confidence and power.

When designing a test, consider the use of replication, randomization and blocking in the design phase to improve test data precision. Replication of test points allows for estimation of system variability and test procedure error. Randomization reduces the likelihood of introducing bias to the experiment by randomizing effect of uncontrolled variables, such as unplanned weather effects. Blocking provides another way to address variability and improve our estimates. Examples of factors which one might which to block over are: time of day(night/day), time of year, and climate.

There is one additional use for statistical test designs – the ability to determine resources required to efficiently test to particular level at an early stage of the program lifecycle – often before program initiation. If leadership determines the level of knowledge desired within a test program, the T&E lead can calculate with some accuracy the costs to achieve that fidelity. At that point leadership can conduct an educated discussion on cost versus risk.

The bottom line is – plan to collect data based on the anticipated analysis, ensure the data collected is adequate with a margin of error, and ensure test program risk is characterized.

DO **-** After all of the planning that occurs in the Define and Design stages the execution of the test should be well defined. However, rarely do tests execute fully as planned. For example, the weather may be extremely windy on a test day, preventing the collection of the fully planned test data. It is important whenever a change in testing occurs that we reiterate through the DESIGN stage, even if just briefly to account for any changes in the data collection, to ensure data

collected can still adequately answer the key questions outlined in the DEFINE stage. A robust experimental design will allow for missing data.

Once the data is collected we need to analyze the data. In a designed experiment we use a statistical model to assist in analyzing the data and determining which factors impact the process, and to what level they impact. Depending on the data increasing complex statistical models can be used to better account for variability in the dataset. The analysis should result in statistically defensible results because the data collection was designed with the analysis methodology in mind.

DRAW CONCLUSIONS **-** Finally, we must draw conclusions based on the data, update our hypotheses for any future testing, and determine how well our test likely captured the true performance of the system. As discussed previously, only two possible conclusions can be drawn from a hypothesis test; reject the null hypothesis or do not reject the null hypothesis. Note that we do not use the term "accept" the null hypothesis as only definitive conclusions can be drawn when rejected, much like the jury logic. The likely accuracy of the performance data documented in the test program will also be reported in the form of a confidence interval.

PARACHUTE DEPLOYMENT CASE STUDY

With the 4-D process established, let us walk through a simple T&E testing example - the FGR-1 parachute. This example will be used throughout the remainder of the article and has been simplified for academic purposes. Since we will never know the true state of FGR-1 performance without testing every parachute, we must plan a test that selects a sample of parachutes sufficient to characterize the performance, with enough repetitions to account for the variability of the production process.

DEFINE - Assume the stated requirement is for a notional FGR-1 parachute to open within three seconds of actuation 90% of the time when actuated via primary method, from 1000-25,000 ft Mean Sea Level (MSL) and 50-250 Knots Indicated Airspeed (KIAS). In this case the question becomes: Does the FGR-1 parachute deploy within three seconds 90% of the time when operated as directed within the operating environment? This question defines the goal of the experiment.

If we were to use a "binary" (yes/no) response variable: "did the parachute deploy within three seconds 90% of the time?" then hundreds of tests would be necessary to obtain reasonable statistical power. A smart statistician on the test team showed the test lead that for 90% of the parachutes to deploy in three seconds, with a standard deviation of 0.5 second, the mean deployment time must be no greater than 2.4 seconds. This "derived" requirement can be used to dramatically reduce the sample size to answer the stated requirement.

After learning about the derived requirement the test lead decides the null hypothesis should be that the mean deployment time is less than or equal to 2.4 seconds across the operating environment, and the alternative hypothesis is that the mean deployment time is greater than 2.4 seconds across the operating environment.

After going through the factor selection process described previously, the FGR-1 SMEs determined three factors that are likely to impact parachute deployment time: altitude, airspeed and body weight. Each factor is continuous, and was assigned three levels. Selecting at least

three levels for each factor ensures the test will detect non-linear performance. To summarize the results of the Define phase:

- Goal: Answer question "Does the parachute deploy within three seconds ninety percent of the time when operated as directed within the operating environment?"
- Null Hypothesis: parachute deploys properly (mean deployment time is $\leq$ 2.4 seconds)
- Alternative Hypothesis: parachute deploys improperly (mean deployment time > 2.4 seconds)
- Response Variable: Time from actuation until complete FGR-1 deployment.
- Type I error represents manager's risk because it is the probability the decision maker concludes that the parachute is bad at one or more of the operating conditions, when the true performance is good.
- Type II error is the user/warfighter risk – because it is the probability that a decision maker concludes that the parachute is good when the true performance is bad at one or more of the operating conditions.
- Factors/Levels:
  - Altitude : 1000 ft, 12,500 ft, 25,000 ft MSL
  - Airspeed: 50, 150, 250 KIAS
  - Body Weight: 100 lbs, 150 lbs, 200 lbs

Notice that in order for us to evaluate the null and alternative hypotheses at each combination of the factors and levels would require a good deal of testing because it would require multiple replicates at each of the operating conditions. However, through the use of designed experiments we can combine data across multiple operating conditions, through the use of a statistical model. This allows us to conduct smaller tests than if each condition were replicated independently. In fact, in one of the designs proposed below there is no replication of the operating conditions.

The null hypothesis is: none of the factors impact parachute performance. The alternative hypothesis is that at least one of the factors significantly impacts performance. If we fail to collect enough data in this context we risk failing to identify a potentially problematic region of the operational envelop.

DESIGN - Now that we have defined the question to be answered, the next step is to actually design the test. There are several designs that can provide robust data but a full discussion of the benefits of each model is beyond the scope of this discussion. One simple design would be to use a 3^3 full factorial (27 runs). A full factorial design produces very high quality data, where all possible factor combinations are tested at least once. Full factorial designs are typically used in DT and OT when the total number of factors and factor combinations is not too large. A full factorial design allows for the estimation of all main effect and interaction terms in the model. However, in a test with several factors and levels, the number of test runs required can quickly become untenable. A second alternative would be a 22 run D-optimal design. A D-Optimal design can estimate all main effects and two-way interactions (all other interactions are assumed to be insignificant). A third option would be to use the initial 3^3 full factorial, with replicated corner points to improve power, for 39 total runs. This design provides the highest power of all three designs due to the replicated points. Replication of any portion of a test will provide a quick improvement in power by providing an assessment of process/product variability.

However, replication of points also raises the number of total runs. All three designs allow testers to observe non-linear performance (if present).

In a resource constrained environment there are several options for reducing the design;

1. Use a fractional factorial design vice the full factorial. The fractional factorial uses substantially fewer resources, and has less power to detect active factors. Additionally it can create partial confounding of factors, depending on the degree of fractionation. Any unacceptable confounding can be resolved by conducting an additional fraction of the full factorial. Regardless, the resources required to resolve confounding and generate confidence will be less than using a full factorial. Even if the test does not execute as planned, analysis techniques exist to develop useful risk information. Ideally, fractional designs will retain the ability to estimate all main effects and two-way interactions (all other interactions are assumed to be insignificant).
2. Reduce the number of replicated points. If leadership is willing to accept higher risk of detecting a difference when there truly is one, we can reduce replications.
3. Use a smaller design type. Optimal designs and small Central Composite Designs provide targeted knowledge, but once again there is a trade-off between risk and cost.
4. As a last resort, eliminate factors. Ideally if a factor is eliminated it is because previous testing has shown it to be inactive. This is the ultimate risk because we will obtain no information on the factors eliminated from the design.

Figure 4 provides the power curves comparing the three possible designs power. The figure uses a response surface analysis for the deployment time of the parachute. Previous data collected have shown the standard deviation we can expect is approximately 0.5 seconds. In the power chart below, the x-axis is the detectable effect size and shows the ability to detect deployment time differences ranging from zero to three seconds. Notice for small time differences all three designs do poorly but for large differences all three designs perform well. Simply put, this means that the ability of a given test to detect small changes in performance (0.2 sec deployment time difference) is low, but the ability to detect large changes in performance time (1.0 sec) is much higher. If the desire is to detect small performance changes, then the program manager must be prepared to conduct a large number of tests, often in the 100's. This is why there must be coordination between the requirements process and the test organization from program inception to determine what is testable, and at what cost. Let's say that for the parachute example the test team has decided that a detectable difference of 0.5 second is adequate. In that case, at the 95 percent confidence level, the full factorial design has 63% power, the D-optimal design has 41% power, and the replicated full factorial design has 87% power for detecting an active factor.

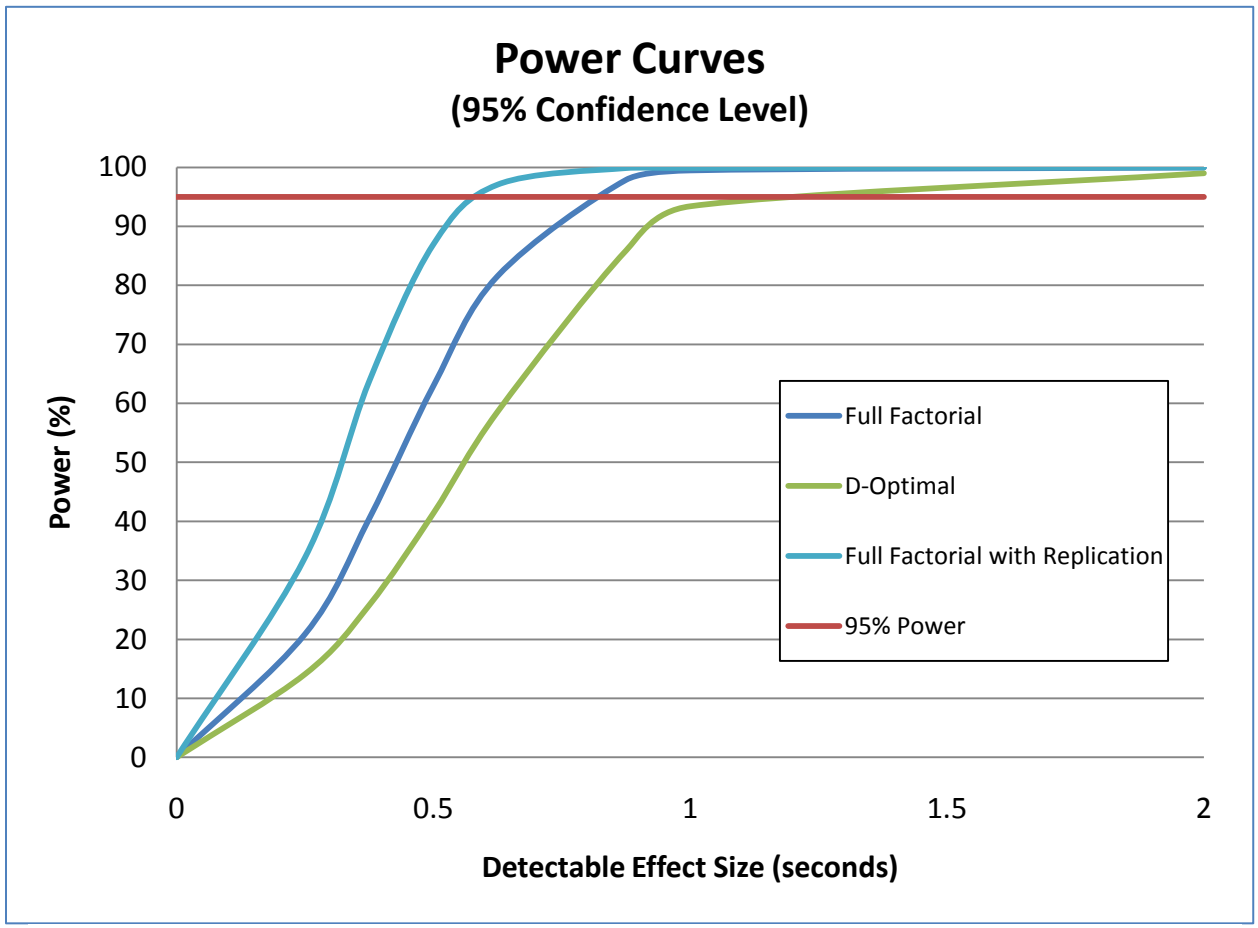

Figure 4: Power curves for detecting the effect of an active factor on Parachute Performance

Figure 4 illustrates the risk decision makers are assuming for allowing smaller experiments is that we may miss an active factor in determining the performance of the parachute. If this occurs, by definition, the factor had a non-constant effect on performance. Therefore, at certain levels of the factor the impact on performance is negative. For example, consider body size, if small body types have the effect of reducing the likeliness that the parachute will deploy within 2.4 seconds and we do not test enough to pick up that difference then we could falsely conclude that the parachute meets the requirement across the operating environment, when in fact fails to meet the requirement for small body types.

So after considering the options and balancing the critical nature of parachute performance with available funding, the PM established a desired confidence level of 95%. The FGR-1 Test Lead determined they would conduct the full factorial test without replicates as an efficient way to develop that confidence. This provides 63% power for main effects to detect a 0.5 second difference in the response time, and allows observation of all main effects, 2 factor interactions, and quadratic effects.

DO The Test team proceeded to the range with the 27 parachutes specified in the test plan and 3 spares. In spite of superb planning and perfect jump conditions the aircraft broke down with 5 jumps remaining. The maintenance crew completed all repairs overnight, and following a post-

maintenance check flight aircraft serviceability issue, was ready to fly the next morning. The weather was 10 deg F cooler, so the temperature was recorded as a nuisance factor in the design, in case the change in temperature caused a difference. Fortunately, there was no impact on deployment time observed that could be attributed to the temperature. The mean deployment time observed was 1.9 seconds, and 85.2% (23/27) of the parachutes deployed within three seconds. This summary information, while traditionally used does not fully characterize performance of the system.

Figure 5 shows the average time to deploy the parachute as a function of altitude and body weight for the low airspeed (50 KIAS). The lines are contours along which the average response (deployment time) is constant. Figure 6 shows the average time to deploy as a function of altitude and body weight for the high airspeed (250 KIAS). Notice for the high airspeed all of mean deployment times are less than two seconds, while for the low airspeed the average deployment time increase to above four seconds for the higher body weights.

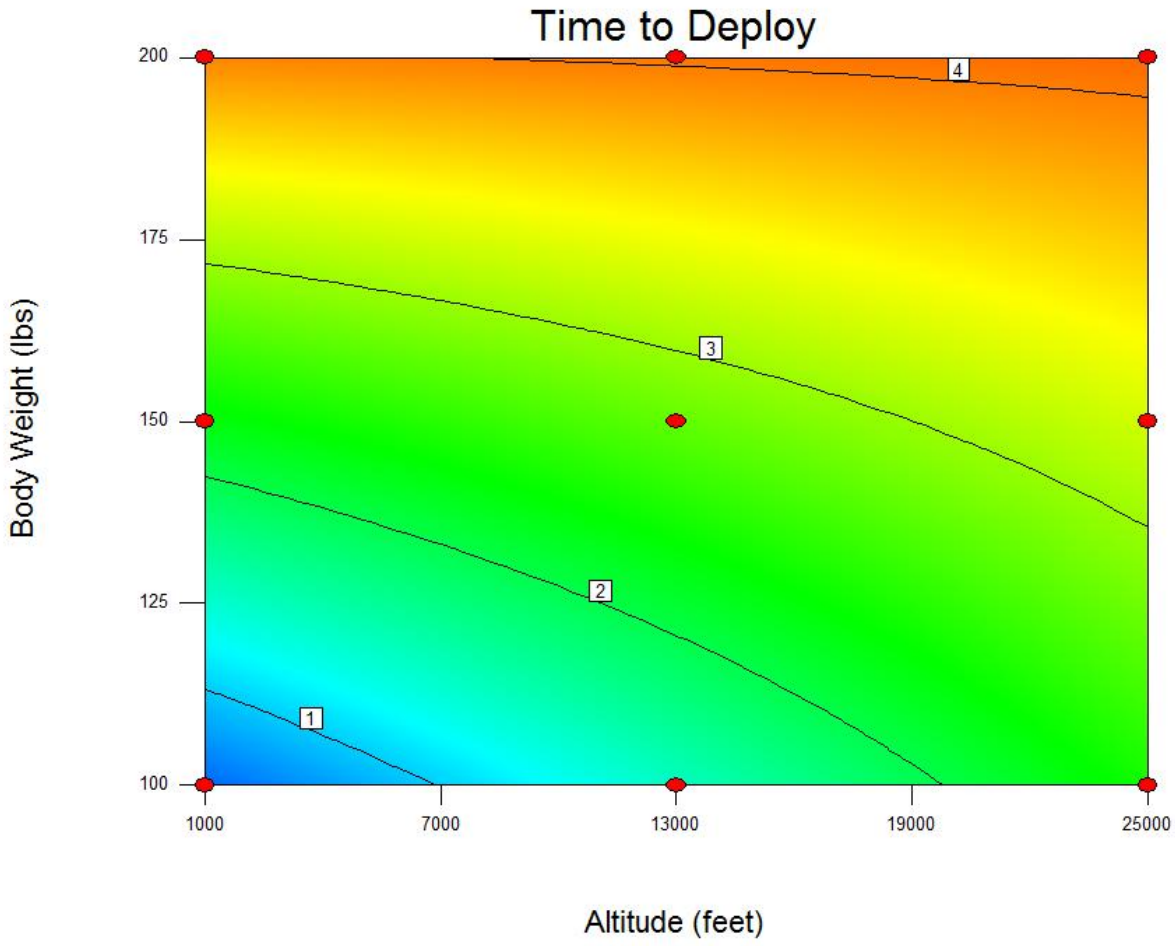

Figure 5: Average Time to Deploy for Low Airspeed (50 KIAS)

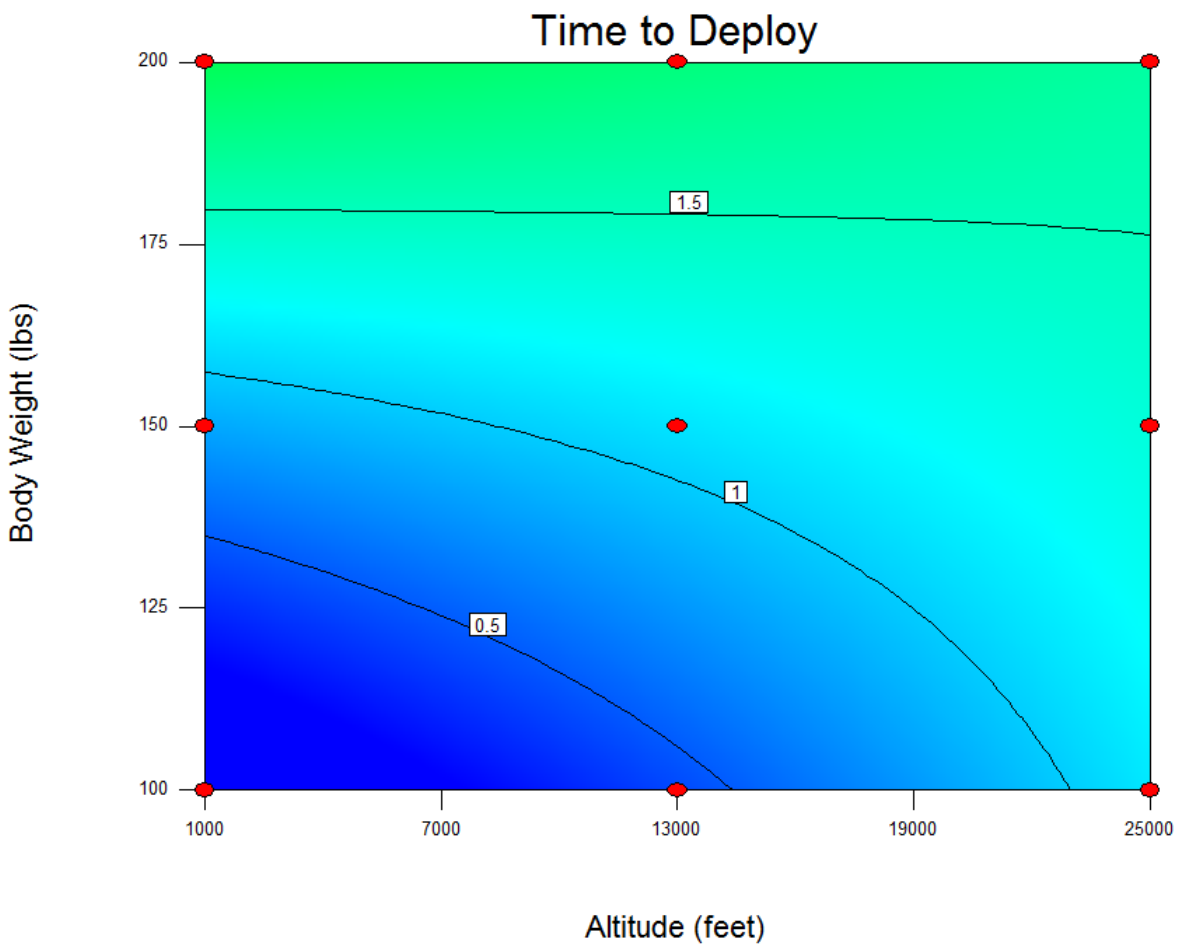

Figure 6: Average Time to Deploy at High Airspeed (250 KIAS)

DRAW CONCLUSIONS

Given the test data and the robust nature of the test plan, the T&E lead had no problem convincing his PM the FGR-1 had achieved the desired deployment time across the majority of the operational envelope. However, the four failures to deploy properly were closely analyzed, and due to the full factorial design, the data showed an interaction between body weight and airspeed that dramatically impacted deployment time. This allowed the PM to state an operational limitation on the use of the FGR-1 for jumpers in the "large" category of 150 lbs greater, in a low speed environment. (This is all notional and done purely to illustrate the kind of information that can be generated).

Conclusion

In this paper we provided a "common man's" interpretation of statistical terminology that can be useful in T&E. We then proposed a structured "4-D" process for answering key T&E questions through a statistical approach. We applied the 4-D process in a simple case study to demonstrate the thought process, discussed the benefits and risks associated with several test designs, and

mentioned how these designs can help develop an effective and efficient test program budget that allows decision makers to pay only for the level of test fidelity they require.

The processes inherent in statistical test design can help facilitate better coordination and cooperation among the DoD requirements, acquisition and T&E communities, and help ensure alignment of acquisition strategies, budgets, resources and test plans. It provides a quantifiable method to understand the risks of making incorrect decisions based on poor planning and insufficient information. Additionally, the 4-D process ensures the data collected will be adequate to answer the questions posed by the decision maker and requirements process. While we freely admit not all decisions can be made solely based on quantitative risk assessments; but when combined with expert engineering judgment and experience, the quantitative assessment of risk should play a key role in the DoD acquisition decision making process.

Biographies

Dr. Laura Freeman is a Research Staff Member at the Institute for Defense Analyses. She currently works in support of the Director, Operational Test and Evaluation on the use of statistics, specifically designed experiments in operational test and evaluation. She has a B.S. in Aerospace Engineering, a M.S. in Statistics and a Ph.D. in Statistics, all from Virginia Tech. Her Ph.D. research was on design and analysis of experiments for reliability data.

Karl Glaeser is a retired Navy Commander with 25 years of operational, training and testing experience, and 2500 flight hours as P-3 Orion Mission Commander. His last tour was as Aviation Branch Head, Office of the Naval T&E Executive (OPNAV N091). His branch oversaw the T&E of all USN/USMC aviation acquisition programs, including weapons and UAS. He is now employed by AVW Technologies Inc., Chesapeake, VA, and provides aviation T&E support to N091 at the Pentagon. He holds a B.S. from the University of MD, College Park, and a M.A. from the U.S. Naval War College, Newport , R.I.

Alethea Rucker is a Staff Engineer for the Policy and Programs Division, Headquarters United States Air Force Test and Evaluation Directorate, Pentagon, Washington D.C. She has held previous flight test and flight dynamics positions at Air Force Flight Test Center, Edwards AFB, CA, Northrop Grumman Corporation, Palmdale, CA and Parker Hannifin (Aerospace), Irvine, CA. She holds a B.S. and M.S in Aeronautical and Astronautical Engineering from Purdue University and is a proud military spouse and mother.

Disclaimer: The authors' affiliations with the Director, Operational Test and Evaluation, The Institute for Defense Analyses, United States Air Force, United States Navy, and AVW Technologies, Inc. are provided for identification purposes only and are not intended to convey or imply the above organizations' concurrence with or support for the positions, opinions, or viewpoints expressed by the authors.

Acknowledgements
The authors would like to thank our many reviewers for their constructive comments and quick turnaround times on review. We would especially like to thank Dr. Ray Hill, Dr. James Simpson and Mr. Greg Hutto for their detailed reviews.

References

American Society for Quality (ASQ) Quality Tools Website: Fishbone Diagram:

http://asq.org/learn-about-quality/cause-analysis-tools/overview/fishbone.html

Montgomery, D. C. (2006) Design and Analysis of Experiments (6th ed). Hoboken, NJ: John Wiley & Sons, Inc.

National Research Council. "Statistics, Testing, and Defense Acquisition:New Approaches and Methodological Improvements" © 1998. Editors: Michael L. Cohen, John E. Rolph, and Duane L. Steffey.

Vining, G.G. (1997) *Statistical Methods for Engineers* (1st ed). Pacific Grove, CA: Duxbury Press