AERIAL REFUELING SIMULATOR VALIDATION USING OPERATIONAL EXPERIMENTATION AND RESPONSE SURFACE METHODS WITH TIME SERIES RESPONSES

THESIS

Alexander P. Hillman, 2d Lt, USAF

AFIT-ENS-13-M-06

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-13-M-06

AERIAL REFUELING SIMULATOR VALIDATION USING OPERATIONAL EXPERIMENTATION AND RESPONSE SURFACE METHODS WITH TIME SERIES RESPONSES

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Alexander P. Hillman, BS

2d Lt, USAF

March 2013

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENS-13-M-06

AERIAL REFUELING SIMULATOR VALIDATION USING OPERATIONAL EXPERIMENTATION AND RESPONSE SURFACE METHODS WITH TIME SERIES RESPONSES

Alexander P. Hillman
2d Lt, USAF

Approved:

Dr. Ray Hill (Chairman)	date
Dr. Darryl Ahner (Member)	date

AFIT-ENS-13-M-06

Abstract

An important program in the Department of Defense is the KC-46 Supertanker. Dubbed the future of the Air Force's aerial refueling inventory, the KC-46 will replace dozens of ailing previous generation tanker aircraft. The Aerial Refueling Airplane Simulator Qualification document governs the methods by which Air Mobility Command validates its simulators, some of which will be KC-46 simulators in the near future. The methodology set forward in this thesis utilizes historical data of aircraft performance from similar air frames to gain statistical insight into the performance design space of the KC-46. Leveraging this insight, the methodology provides through a framework for validation that uses classical experimental design principles as applied to time history responses such as found in aircraft performance measures. These principles guide the generation of response surfaces from real world flight test data that can then be used to validate flight training simulators using a point by point comparison or over an entire surface of points for a variety of different aerial refueling maneuvers. This work also supports the KC-46 Tanker program by proposing statistically efficient and cost conscious experimental designs for the KC-46 flight testing. This framework is demonstrated using flight testing data from the KC-135 Aerial Refueling Simulator Upgrade testing, and is part of an Office of the Secretary of Defense initiative to add increased statistical rigor to the Department of Defense test and evaluation enterprise and specifically the acquisition community.

For my family who support me unconditionally

Acknowledgments

I would like to thank my advisor, Dr. Ray Hill, for his patience and flexibility in working with me over the past year. Without your flexibility and cooperation, things would have turned out differently for this document. Without you, this document would never have been completed.

Next, I'd like to thank Mr. Phil Jones at the Simulators and Training Division at Wright-Patterson AFB. Without a topic, it's assuredly difficult to write a thesis.

I'd like to thank Mr. Rob Sills of the Center for Operations Analysis at AFIT for his special assistance and code writing for our Matlab tool. This tool has become an invaluable part of this analysis and a key part of its success.

I'd like to thank Robert McCabe and Mark Webb for their help in getting access to the massive data requirements for this analysis. This thesis definitely falls under the "Big Data" umbrella, and without these two gentlemen this project would have never gotten off the ground.

Alexander P. Hillman

Table of Contents

Page

Abstract iv
Dedication v
Acknowledgments vi
List of Figures ix
List of Tables x
1. Introduction 1
1.1 KC-46 1
1.2 Aerial Refueling Airplane Simulator Qualification 3
1.3 Research Objectives 5
1.4 Thesis Overview 6
2. Background 7
2.1 Classical Experimental Design 7
2.2 Response Surface Methods 10
2.3 Time Series Response Data 12
2.4 Comparison of Response Surfaces 13
2.5 Statistical Rigor in the Department of Defense 13
2.6 Conceptual Modeling 15
2.7 Conceptual Model Validation 15
2.8 Levels of Validation 16
2.9 Types of Validation 17
2.10 Validation Techniques 18
2.11 Types of Simulation 20
3. Methodology 22
3.1 Bayesian Design Optimization 22
3.2 A Framework for Simulator Validation 29
4. Bayesian Design Optimization 34
4.1 Leveraging Past Data 34
4.2 Design Optimization 37

5. Simulator Validation 40
5.1 SIMCERT Process 40
5.2 Data Set 41
5.3 Validation Demonstrated 42
6. Discussion 46
6.1 Contributions 46
6.2 Recommendations 47
6.3 Suggestions for Future Research 48
Appendix A: Table 5-3, 1.a.5 50
Appendix B: Table 2-3, 1.a.3.5 57
Appendix C: Table 5-3, 1.a.8 60
Appendix D: Table 2-3, 1.a.4 65
Appendix E: Quad-Chart 68
Works Cited 69
Vita 73

List of Figures

Page

Figure 1: KC-46 Refuels C-17 (Boeing Image, 2012)..... 2

Figure 3 Sample Response Surface..... 36

Figure 2 ARASQ 2.2.5.3..... 36

Figure 4 Design Diagnostics for ARASQ Design 38

Figure 5 Diagnostics for I-Optimal Design 39

List of Tables

Page

Table 1 Levels of Validity 17

Table 2 Sample Design Matrix 34

Table 3 Coded and Uncoded Variables 35

Table 4 Sample Time Slice of Data 35

Table 5 ARASQ Design Matrix 37

Table 6 I-Optimal Design 38

Table 7 Design Comparison 39

Table 8 Matching Time Slices 42

Table 9 Sixteen Settings for Simulator Validation 43

Table 10 Validation Results 44

AERIAL REFUELING SIMULATOR VALIDATION USING OPERATIONAL EXPERIMENTATION AND RESPONSE SURFACE METHODS WITH TIME SERIES RESPONSES

1. Introduction

1.1 KC-46

The Boeing KC-46A is poised to become the future of aerial refueling for the United States Air Force. Set to replace about half of the Air Force's aging fleet of KC-135 Stratotankers, the KC-46A is a next-generation supertanker slated to become the primary aerial refueling platform for the United States Air Force. A modified Boeing 767, the KC-46A also boasts a more substantial cargo payload and a significant increase in aeromedical evacuation capability when compared to its predecessor the KC-135. According to the official website of the United States Air Force, the KC-46A will be able to refuel any fixed wing receiver capable aircraft on any mission (KC-46A Tanker 2011). The KC-46 will also provide aerial refueling capabilities for Navy, Marine Corps, and coalition aircraft.

To assist in its refueling mission, the KC-46 will be equipped with a modernized KC-10 boom. This boom is operated using a fly-by-wire control system. This boom delivers a fuel offload rate suitable for any large aircraft. The KC-46 is also outfitted with a hose and drogue system, allowing it to refuel a larger variety of aircraft. In addition, this hose and drogue system adds mission capabilities to the KC-46 that can be employed independently of the refueling boom.

The Boeing Company was awarded the contract for the KC-46 in February of 2011. Currently, the program is in the Engineering and Manufacturing Development phase. The initial test flight for the KC-46 is scheduled for late in 2014; Boeing is scheduled to deliver the first 18 KC-46 tankers to the warfighter by 2017. While the current contract requires Boeing to deliver 179 Combat-Ready KC-46A tankers in all to Air Mobility Command (AMC).

The KC-46A is an important weapon system for the United States Air Force. It has been dubbed a "supertanker," and will be the prominent tool in the future of aerial refueling for the Air Force. The program itself is also a major acquisition. The contract itself is estimated to be worth upwards of $40 billion. In today's fiscally constrained times, protecting such a program from any sort of financial overrun or engineering risk is a top priority for the United States Government.

Figure 1: KC-46 Refuels C-17 (Boeing Image 2012)

1.2 Aerial Refueling Airplane Simulator Qualification

The Aerial Refueling Airplane Simulator Qualification (ARASQ) document as published by AMC provides the acquisition community a means to evaluate the trainers and simulators used in aerial refueling (AR) and boom operator (BO) training programs. In compliance with Headquarters, Air Mobility Command Aircrew Operations and Training Division (HQ AMC/A3T), the ARASQ document is used to determine the qualification level of a specific simulator used for AR or BO training based on the fidelity and resolution of said simulator.

Prior to ARASQ, the evaluation of these simulators was performed utilizing mostly subjective criteria. Over time, it became apparent that merely subjective forms of evaluation could not be used as a reliable and dependable form of validation, especially for simulators designed to reduce flight hours spent on training and increase the quality of training spent outside the physical cockpit.

To reduce subjectivity, the ARASQ document was developed. Originally released in 1997, the ARASQ document was produced with the goal of reducing flying hours while helping augment the number of mission-capable AR qualified aircrews. To be clear, ARASQ and its five chapters apply strictly to aerial refueling simulator qualification. Simulator validation procedures and criteria for other weapons systems are contained in other documents also vetted and maintained by AMC.

ARASQ provides a list of required test events; each event has been deemed critical to the AR process. A test event is a process of interest during the AR process. As the event occurs, ARASQ explicitly details specific responses to be measured. In the past,

particularly on the KC-135 and KC-10, test data has been collected at certain instances in time. ARASQ also limits the design space for the test event according to the following four factors, also known as controls: tanker weight, receiver weight, airspeed, and altitude. ARASQ prescribes limits on the range of values each of these controls can take on as required by the individual test events. Routinely throughout the ARASQ document, factor levels are prescribed at one of three levels for the test event.

The ARASQ document also deals with different levels of certification. For example, for a simulator to be used for aircrew continuation training credit only actual flight test data may be used. ARASQ provides a checklist of test categories for Level C and Level D certification. Level C certified simulators are to provide an accurate representation of the cockpit. Level D certified simulators are required to provide a more realistic representation of the actual system in terms of audio, visual, and motion cues on top of the requirements already laid out for a Level C simulator. Simulators for the KC-46 require a Level D Certification.

There are several limitations on ARASQ and the simulator training models. First, data from flight testing has traditionally only been used to validate simulator events related to a specific flight testing event. This limits the fidelity of a simulator training model. Next, historically only a portion of the flight test data has been used in simulator validation. This same validation has also been performed on a case by case basis; for example, a single point in the design space is considered validated if the simulator's results fall within a certain tolerance at that point in the design space. This can give rise to a myriad of issues, not the smallest being the simulator's failure to capture the true nature of the physical system across all points in the design space.

A third limitation of the way the simulator training models have been built in the past is that the model building process is not reproducible. This drives a wedge in between development and implementation of a simulator as a form of training for a real world operator.

There are also limitations on the flight testing for the KC-46. Most importantly, the flight test program has prescribed that only 50 hours of flight testing will occur for each aircraft pair in the Air Force's inventory. This is a reduction in flight testing by 50% compared to the testing that was done for the KC-10 during its acquisition process. Subject matter experts (SMEs) have determined that this is hardly enough time to capture all of the ARASQ test events, let alone capture the additional data for simulator validation.

1.3 Research Objectives

The research objectives stem directly from the needs of the KC-46 program office (PO) and the Simulators Division. The goal of this research is to provide a three-pronged approach to simulator validation leveraging work done in the past for the PO and the Simulators Division. Keeping this in mind, the following three research objectives were developed along the way:

1. Evaluate current ARASQ test events employing data of like systems and propose defendable alternate designs based on efficiency, nature of the design space, cost, and subject matter expert opinion.
2. Building upon the methodology as set out by Capt Scott Storm, use flight test data to generate time series response surfaces to characterize the entire design space while optimizing a set of proposed ARASQ designs.

3. Build a framework for simulator validation using response surface methodology to augment current Simulator Certification (SIMCERT) protocol.

1.4 Thesis Overview

This document is organized according in a six chapter format. This first chapter introduces the topic of interest, the pertinent background information, and the research objectives. Chapter two details an in-depth review of the pertinent literature from the current body of knowledge. Chapter three streamlines the methodology for creating optimal test event designs to create predictive response surfaces with time history data and lay out a framework for simulator validation using response functions and surfaces. Chapter four illustrates the practicality of the proposed design optimization methodology performing a step-by-step analysis of past data. Next, chapter five focuses on the simulation validation methods through a case study. Finally, chapter six provides analysis conclusions, the contributions of this thesis, as well as the recommendations for future research.

2. Background

This chapter provides the reader an overview of foundational concepts which are relevant to this analysis through a thorough review of the body of knowledge. This chapter is divided into two sections. The first section provides a review of classical experimental design principles, response surface methods as they apply to this analysis, and the issue of time series responses in operational experimentation. It also briefly details the current push for statistical rigor in light of the initiatives from the Office of the Secretary of Defense (OSD). The second section discusses key modeling and simulation concepts applicable to this research. This includes but conceptual modeling, validation, validation techniques, terminology and related principles and techniques.

2.1 Classical Experimental Design

Experimental design (DOE) involves the use of planned experiments in which a "test or series of runs in which purposeful changes are made to the input variables of a process or system so that we may observe and identify the reasons for changes that may be observed in the output response (Montgomery 2009)." A traditional experimental design examines a set of factors and how changes in these factors affect the outcome or response for that specific process.

Designed experiments use a pre-specified plan, and often displayed as a matrix. "The factors of an experimental design are the columns or variables that have two or more fixed values or levels. The rows of a design are the treatment combinations and are sometimes called runs (Kuhfeld 2010)." Typically, analysts run experiments to make statistical insights into the effects of factor levels on outcomes or response variables.

Experimental design is widely used in industrial settings as well as in business. Originally, the groundwork for Designed Experiments was set out by Sir Ronald Fisher (Fisher, Statistical Methods for Research Workers 1958). A statistician at an agricultural firm, Fisher realized that the way some experiments were performed often biased the data in some fashion. After laying the ground work for Analysis of Variance (ANOVA), Fisher introduced three cornerstone principles of experimental design: randomization, replication, and blocking. Randomization is the idea of performing experiments in a truly random order as to minimize the systemic variation of nuisance variables that are unknown to the experimenter but still vary during the experiment uncontrollably. Replication involves repeating some treatment combinations during the experiment to glean an accurate estimate of experimental error. Blocking is a technique used to control for nuisance factors that are known but uncontrollable. It allows the experimenter to limit their effect on the experimental error estimate. Fisher's work paved the road for experimental design as we know it today (Fisher, The Design of Experiments 1966).

In evaluating a designed experiment, analysts use statistical criteria. Today, these criteria have come to be known colloquially as alphabetic optimality. First discovered by Smith in 1918 (Smith 1918), optimal designs allow a design to minimize prediction variance and limit bias amongst the estimators, thus yielding more applicable and useable results for the experimenter. Smith's paper, however, was 30 years before its time. It wasn't until the late 1950s when Kiefer and Wolfowitz (1959) and Kiefer (1961) proposed the idea of selecting designs based on a specific criterion. Their work initially hypothesized that designs should be selected to estimate model parameters with the most accuarcy (Kiefer and Wolfowitz, Optimum Designs in Regression Problems 1959).

Kiefer delved deeper into the development of a computer algorithm for building a D-optimal design (Kiefer, Optimum Designs in Regression Problems II 1961).

There are several alphabetic optimality criteria. For brevity's sake, only two are addressed in this analysis. The first criterion is D-optimality and the second criterion is I-optimality. In his text devoted entirely to DOE, Montgomery describes D-optimal designs as "a design that minmizes te variance of the model regresion coeeficients." I-optimal designs seek to minimize the average variance of prediction (Montgomery 2009).

Kiefer and Wolfowitz's work did not catch on at first because of the limitations in the computational power of their era. However, today, most statistical software packages include a design optimization application. These applications can use a variety of algorithms for design optimization. These algorithms can be altered for various types of problems. In the past, genetic algorithms (Todoroki and Ishikawa 2004), stochastic genetic algorithms (Jin, Chen and Sudjianto 2005), and even local search evolutionary algorithms (Dengiz, Altiparmak and Smith 1997). Today, computing power is not as much of an issue as in the 1950s, and the SAS Institute's JMP (JMP 10.0 n.d.) handles design optimization quite well.

Typically, a design is initially built to estimate the model's parameters as efficiently as possible. Designs built using this approach have been dubbed a "locally optimal" designs (Chernoff n.d.). These designs typically do not leverage knowledge of the prior distribution of the model's parameters. Such a design is a Bayesian Experimental Design (Raiffa and Schlaifer 1961). In fact, most of the alphabetical optimality criteria have a "utility-based Bayesian version (Chaloner and Verdinelli 1995)." In some experiments, the analysis is more geared towards prediction than

statistical inference or screening. For this type of problem, a predictive Bayesian approach can be used for both the analysis and the experimental design (Geisser 1993). This approach theoretically most closely relates to the approach used in the analysis section of this document.

2.2 Response Surface Methods

Building upon Fisher's seminal works, G. E. P. Box and K. B. Wilson developed a new tool for analyzing the potential relationships between explanatory control variables and one or more response variables (Box and Wilson, On the experimental attainment of optimum conditions 1951). A response surface is a "graphical perspective of the problem environment (Myers, Montgomery and Anderson-Cook 2009)." Box and Wilson developed several techniques to exploit the immediacy and sequentiality of industrial experiments; these two properties mean that experimental results can be recorded (almost) immediately during an industrial process and experiments can be run to gain informational results in a small number of treatment combinations. For this reason, DOE works well for screening experiments, or experiments in which the main goal is to determine the significance of effects.

The techniques discovered by Box and Wilson allow the experimenter to estimate the relationship between a response y and a set of controls X for a product, process, or system in form of equation (1).

$$y = f(X) + \varepsilon \tag{1}$$

where ε is a term representing the other sources of variability in the underlying process not captured in the function f (Myers, Montgomery and Anderson-Cook 2009). This function can be easily estimated using ordinary least squares.

In the popular RSM text by Myers, Montgomery and Anderson-Cook (2009), three objectives and applications of RSM are outlined:

1. Mapping a response surface over a particular region of interest.
2. Optimization of a particular response.
3. Selection of operating characteristics to achieve specific results.

"RSM is an important branch of experimental design… RSM is a critical technology in developing new processes, optimizing their performance, and improving the design and/or formulation of new products (Myers, Montgomery and Anderson-Cook 2009)."

Response surface designs usually involve two to eight continuous control variables with at least one response. Usually, *a priori*, we can assume that the model for a response surface experiment is quadratic (JMP Support 2013) at least in a sufficiently small experimental region of interest. For this reason, a second-order model is used in most response surface models.

A second-order model is used to detect curvature for a response surface (Montgomery 2009). In practice, using a second-order model to estimate the function f discussed in equation (1) is useful out of "practical experience (Myers, Montgomery and Anderson-Cook 2009)." Box and Draper go into considerable detail in their text about the flexibility of the second-order model and its ability to work well in solving a response surface problem from reality (Box and Draper, Empirical Model Building and Response Surfaces 1987). The ARASQ designs being analyzed for the KC-46 directorate will require second-order models to detect curvature within the design space.

Traditional response surface experiments use only continuous, quantitative factors. Response surfaces are used mainly for evaluating prediction, as opposed to

evaluating the efficiency of the parameter estimates. In the past, the most popular form of response surface design used the D-optimality criterion, which intuitively does not make sense. “Because I-optimal designs minimize the average variance of prediction over the region of experimentation, their focus is clearly on prediction. Therefore, the I-optimality criterion seems to be a more appropriate one than the D-optimality criterion for generating response surface designs (Jones and Goos 2012). “

2.3 Time Series Response Data

In classical experimental design, responses are usually recorded as a snapshot in time, not a continuous history of data. As noted by Box and Wilson, industrial experiments usually involved recording the outcome of a test run as a single value immediately following the experimental run (Box and Wilson, On the experimental attainment of optimum conditions 1951).

Scott Storm seems to be one of the first to apply the principles of DOE to time series experimentation. Unlike in classical DOE and RSM literature, the responses in this analysis and in Storm’s data were in the form of a function of time. This presents a “unique dilemma (Storm 2012).” Storm formulated a methodology for analyzing these time series response surfaces using discretized samples from various time steps, each corresponding to an instance in time and its own discrete matrix of inputs X_t. This allowed Storm to represent each time step as one design matrix of controls and responses.

Using historical data, Storm used these design matrices for one particular ARASQ test event to analyze the curvature of the response surface generated using pitch attitude as the response of interest. Upon visual inspection of these surfaces, if curvature existed, a three-level design for ARASQ was justified.

2.4 Comparison of Response Surfaces

A huge part of the design and testing of A/C systems occurs in wind tunnels. These test campaigns use a large number of test runs and usually take a long time to accomplish – sometimes years! Hill et al (2010) used a large legacy wind tunnel testing data set to develop a methodology for comparing two response surfaces: one from the legacy data set, and another from a smaller sample training set for validation.

Using Monte Carlo simulation to sample points from the legacy set, two response surfaces were compared using both the coefficients of their respective response surfaces and their regression equations. Using a confidence interval about the mean of the differences between each point, Hill et al showed that the differences between the surfaces are independent and *iid*. Assuming a mean of 0 for the differences, they concluded that the surfaces were roughly identical in a statistical sense (Hill, et al. 2010).

RSM models are today used occasionally for simulator validation, particularly for multi-agent social network simulations (Carley, Kamneva and Reminga 2004). However, supplementing the validation techniques currently used in social network simulations with the work of Hill et al, simulation validation and calibration can be accomplished in a non-agent based, non-social network simulation.

2.5 Statistical Rigor in the Department of Defense

The Department of Defense (DoD) Test Enterprise is responsible for test and evaluation policy as well as the planning, execution, and analysis of tests. While using the tenets of experimental design alone does not guarantee scientific adequacy or accuracy, better testing of systems results in a better allocation of resources and helps

categorize and quantify risk and uncertainty (Gilmore, Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation 2010).

While experimental design and statistical rigor are not new in the DoD (Johnson, et al. 2012), experimental design does have many practical applications within the DoD. In particular, DOE has a place within T&E for research and development (R&D), something vital to the design, development, and deployment of military systems in today's unconventional combat environment. In 2009, Hutto and Higdon concluded that "DOE can be used to great profit" in testing military systems. DOE often results in efficient and effective test programs, "even in the face of difficult and noisy test problems (Hutto and Higdon 2009)." DOE's usage throughout the DoD has become widespread, and it's no secret why.

Dr. Michael Gilmore, the director of Operational Test and Evaluation, began four test and evaluation (T&E) initiatives after taking office in 2009; one of these initiatives was the Science of Test initiative as noted in his 2013 report to Congress (2013). Its goal is simple: to support the integration of advanced statistical rigor and mathematical foundations into the test domain (Gilmore, Test and Evaluation (T&E) Initiatives 2009).

DOE and an efficient utilization of DOE can lead to a series of improvements within the DoD acquisition community and ultimately help the warfighter. DOE enables experimenters to use early results to refine future test events (Gilmore, Rigor and Objectivity in T&E: An Update 2011). Known in the academic world as a screening experiment, this idea has not always been implemented in the T&E world to its fullest extent.

2.6 Conceptual Modeling

In its simplest form, a simulation is a conceptual model of a real system. A conceptual model is the "abstraction of a model from a real or proposed system (Robinson, Conceptual Modeling for Simulation: Issues and Research Requirements 2006)." Conceptual modeling, like simulation, involves a simplification of reality. The simplification process requires input from real-world users and operators, and usually a laundry list of model assumptions developed by the analyst as well.

A model can refer to almost anything in math, statistics, or computer science. It can be any "physical, mathematical, or logical representation of a system, entity, phenomenon, or process (Zeigler, Praehofer and Kim 2000)." A model can be applied to anything, and does not refer to strictly simulation models, although today many people ultimately associate "modeling" with modeling and simulation (M&S).

A simulation is a model of an actual system, and simulation models are commonly referred to as executable forms of underlying conceptual models. Usually, simulators build models of systems to make changes to the underlying physical system and examine the results. Typically an experimenter will simulate a model where changes to the actual system are either impossible, too expensive, or impractical (Maria 1997). Using a simulation model allows the researcher to simplify the physical system and make insights into the properties of the system.

2.7 Conceptual Model Validation

The importance of simulation validation is well documented and is a common theme in the M&S literature (Robinson, Simulation: the Practice of Model Development and Use 2004). Validation ensures that the simulation model accurately portrays the

underlying system. There is widespread agreement that model validation is important; the disconnect lies in the methods by which different types of simulations are validated.

Usually, a model is validated by checking its performance under known conditions and comparing this performance to the actual system, if possible (Maria 1997). An analyst can perform statistical inference tests to judge the validity of the model. Another form of validation is to test the model's underlying statistical assumptions and then use face validation techniques (Sargent 2004).

Extensive validation not only ensures that a model is correct, but it also inspires confidence in the model's results. DoD models are not immune from model validation. DoD Instruction 5000.61 details the terminology and principles used in DoD modeling and simulation (Department of Defense 2009). The DoD definition echoes the principles discussed in this section.

2.8 Levels of Validation

Not all simulation models are created equally. Different models require different levels of verification and validation (V&V). These different levels of V&V stem from the following facets of a simulation: objectivity, repeatability, timeliness, completeness, and accuracy (Harmon and Youngblood 2005). According to Harmon and Youngblood, there are six levels of simulation validity, each with its own level, of the five facets mentioned above.

Table 1 Levels of Validity

Tier of Validity	Supporting Information	Type of Validity
0	Nothing	I have no idea.
1	Simple Statement of Validity	It works; trust me.
2	Required entities and attributes compared against the entities and attributes that the simulation represents	It represents the right entities and attributes.
3	Required entities, attributes, and dependencies compared against entities, attributes, and dependencies represented	It does the right things; its representations are complete enough.
4	Required entities, attributes, dependencies, and dependency errors compared against entities, attributes, and dependencies represented and representation errors	For what it does, its representations are accurate enough.
5	Required entities, attributes, dependencies, dependency errors, and confidences in assessment compared against represented entities, attributes, dependencies, representation errors, and assessment confidences	I'm this confident that this simulation is valid.

In an ideal world, all simulations would be in the fifth tier of validity; unfortunately, this is not always the case.

Harmon and Youngblood point out two vital assumptions for this tiered validity assessment model:

1) The quality of validation information depends upon its truthfulness and completeness, and improved truthfulness and completeness can only be achieved through improved objectivity; and
2) Reliably improving validation process objectivity requires understanding the fundamentals of that process.

These assumptions are fundamental in nature yet commonly disregarded in the modeling community.

2.9 Types of Validation

There are many uses of simulation and several different ways to validate each of them. On the surface, there are two types of validation: face validation and empirical validation (Klugl 2008).

"Face validity shows that processes and outcomes are reasonable and plausible within the frame of theoretic basis and implicit knowledge of system experts or stake-holder (Klugl 2008)." Face validity is more of a continuous loop than empirical validation. It should be started immediately once the conceptual metamodeling process has begun. It involves checking the general plausibility of the model as it relates to the underlying real-world system.

"Empirical validation uses statistical measures and tests to compare key figures produced by the model with numbers gathered from the reference system (Klugl 2008)." This involves either relating the statistical metrics from the simulation model to the real-world system, or comparing the statistical parameters from the simulation to another simulation model that has previously been validated and verified.

2.10 Validation Techniques

Different "brands" of simulations can be validated in different ways. There are several validation techniques available to the conceptual modeler.

Sargent provides a brief yet descriptive overview of the available techniques in his seminal work *Verification and Validation of Simulation Models* (Sargent 2004).

*Animation***:** this technique presents the simulation visually as it steps through time.

Comparison to other models: this technique involves either a graphical or empirical comparison to another widely-accepted model.

Degenerate Tests: this involves comparing the values of parameters as they relate to specific parameters within a model; for example, should the average number in the queue actually increase when the arrival rate is indeed larger than the service rate?

Event Validity: does the number of critical events within a model match the real-world?

Extreme Condition Tests: this method compares the model to the real-world at the extreme points of operation for the underlying system.

Face validity: this technique requires SME opinion and compares the views of the SMEs to the outputs of the model.

Historical Data Validation: if historical data exists, some of it can be withheld in the building of the model then used for empirical comparison once the model is running.

Historical Methods: this category involves three techniques – rationalism, empiricism, and positive economics. Rationalism deduces logical conclusions to judge validity of the model from the model's underlying assumptions. Empiricism requires that all underlying assumptions be not only rationally justified but empirically proven. Positive economics requires that the model is able to predict its outputs correctly.

Internal Validity: this involves replications of the model to determine the internal variability of the model.

Multistage Validation: this technique rolls all three historical methods into a multistage process.

Operational Graphics: operational parameters and their levels are shown graphically as the model progresses in time. This can easily be prepared by time slice to the real-world system.

Parameter Variability: this sensitivity analysis checks the variance of the outputs based on the model's inputs.

Turing Tests: This method requires SMEs to discriminate between model and real-world outputs.

2.11 Types of Simulation

While many types of simulation exist, certain validation techniques are not always appropriate for different simulation brands. Some popular types of simulation are live, virtual, or constructed, or some combination of these three categories. The DoD traditionally uses a simulations with varying resolution from any of these three categories.

For example, manufacturing simulation models typically involve discrete-event modeling techniques. This is often referred to as a "job-oriented" world view (Fowler and Rose 2004). This type of simulation is typically stochastic in nature, and validation of the internal variability will be a particularly important feature of the validation.

Another type of simulation is deterministic simulation. In a deterministic simulation, the model produces results without variance. This helps limit the complexity of the simulation as well as the computational intensity of the model. However, occasionally deterministic simulations require extensive computing power. The popular RSM technique "Kriging" is used to "detrend" the data using linear regression (Beers and Kleijnen 2002). This is an example of extensive empirical validation using degenerate testing.

Another type of simulation is agent-based simulation. Often these simulations contain feedback loops between particular entities, or agents, and their environment. This

makes validation difficult because of the non-linear effects on parameter estimates within these models (Klugl 2008). Another difficulty of validating agent-based simulations is their reliance upon historical data. If this data is not readily available, validation, the most central feature of a "good" simulation, is nearly impossible. As they tend to study the hierarchy of interactions within a modeling environment, agent-based simulations require historical data validation testing the model's underlying statistical assumptions as well as the model's outputs. In his survey paper, Heath et al provides a detailed discussion of the validation of agent-based models from 1998-2008 (2009).

Another type of simulation is the Man-in-the-Loop style of simulation. These simulations are typically virtual and constructed, like a flight simulator trainer (Knepell and Arangno 1993). These simulators require validation with SME opinion as well as empirical validation. Another commonly used technique with these models is the validation using DOE (Schatzoff 1975). Using replicated designs in both the real-world system and the simulator, the modelers can produce statistically comparable results for the same parameters. Coincidentally, both AR and BO training simulators fall into this category of simulations.

3. Methodology

A unique feature of the ARASQ test events is the predicament of response variables as a function of time (Storm 2012). The ARASQ document requires a specified time interval of flight testing data for each maneuver being tested. As previously noted by Storm, traditional response surface methodology (RSM) literature has not yet truly addressed how to deal with time series response variables. The training simulators being validated with ARASQ testing data are dynamic systems that obviously must change over time to portray a cohesive representation of their underlying physical systems. For this reason, using time history test events appears to be a logical and inherent approach to modeling this type of dynamic system as well as validating such a model. This chapter lays out a streamlined approach for optimizing a design and generating response surfaces for simulator validation. This chapter is presented in two parts. The first section details how to leverage past data for design optimization and provide a criterion for design optimization. The second section steps through the methods for simulator validation using response surfaces generated with data from the optimized design.

3.1 Bayesian Design Optimization

The first section of this chapter walks through how to optimize the designed experiment for an ARASQ test event. This is done in two steps:

1. Leverage historical data from similar air frames to analyze via analogy the experimental design space for the KC-46.

2. Using statistical insights into the operating envelope of the KC-46, optimize the designed experiment for a particular test event according to the optimality criterion of interest.

3.1.1 Leveraging Historical Data to Glean Insights about the Design Space of a Similar Air Frame. While not identical, the KC-135 and KC-10 come from the same class of aircraft (A/C) as the KC-46. this class of A/C is used for aerial refueling, and their primary mission is to provide American and coalition air forces aerial refueling capability. According to subject matter experts, the flight control surfaces of these A/C will exhibit similar performance when doing similar maneuvers. With this logic, a comparison by analogy can be drawn between the previous generation and the future generation of AR A/C.

Given a particular test event of interest for the KC-46, there exists a similar test event performed in the past using either the KC-10 and/or the KC-135. These test events are organized as designed experiments, similar to the designed experiments planned for the KC-46 based on ARASQ revision C. These past test events involve time series response data using controlled factors for experimentation. These test events utilize a data structure that allows us to examine each time slice as its own discrete design matrix. Using control variable response data, we build a design to estimate the following relationship

$$Y^t = f(X_t) \tag{1}$$

where Y^t is a n x l matrix of response. The index l corresponds to the response while n corresponds to test run or observation number. The superscript t attached to Y indicates

the time slice. The function f implies some transformation of the design matrix X_t. This transformation is just an exploitation of the relationship between Y and X. X_t is organized in the following fashion:

$$X_t = \begin{pmatrix} x_{1,1,t} & x_{1,2,t} & \cdots & x_{1,p,t} \\ x_{2,1,t} & x_{2,2,t} & \cdots & x_{2,p,t} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1,t} & x_{n,2,t} & \cdots & x_{n,p,t} \end{pmatrix}.$$

The index p corresponds to the independent variables in that column of the X matrix.

According to subject matter experts, the biggest source of engineering risk in the implementation of an Air Force-wide flight training simulator lies in the ability to capture the nuances of the design space. In statistical terms, this means the ability of a test to detect curvature of a response surface within the design space is of huge importance.

To detect this curvature based on historical data we generate response surfaces with the data. For each of the responses of interest, at each time slice $y^{l,t}$, where l represents the response variable and t represents the time slice for which the equation is generated, a second-order response function can be estimated using ordinary least squares (OLS). Using OLS, the X matrix takes the form:

$$X_t = \begin{pmatrix} 1 & x_{1,1,t} & x_{1,2,t} & \cdots & x_{1,p,t} \\ 1 & x_{2,1,t} & x_{2,2,t} & \cdots & x_{2,p,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1,t} & x_{n,2,t} & \cdots & x_{n,p,t} \end{pmatrix}.$$

The response function estimated is presumed to be nonlinear and estimated using equation (2).

$$f^{l,t}(X) = \beta_0 + \sum_{j=1}^{J} \beta_j X_{j,t} + \sum_{j=1}^{J} \beta'_j X_{j,t}^2 \tag{2}$$

In previous ARASQ experiments, the controls were Tanker Weight, Receiver Weight, Airspeed, and Altitude. The X_j terms that correspond to the final third of the right side of equation (2) are the terms used to estimate the quadratic effects. They are calculated using this simple equation:

$$(x_{j,t} - \mu_j)^2 \tag{3}$$

Here the index j corresponds to the control of interest, while μ_j is the mean for that control across the entire maneuver. This operation controls for multicollinearity between each $x_{j,t}$ and $x^2_{j,t}$ columns within X_t which can inflate experimental error estimates and cause incorrect estimation of model effects. All historical designs employed a three-level design, making estimation of quadratic effects possible.

Using Microsoft Visual Basic for Applications (Visual Basic for Applications 2007), the time series control and response data were imported into Microsoft Excel for pre-processing (from original .ASCII file format). Visual Basic provides an automated and reproducible avenue for extracting Tanker Weight, Receiver Weight, Airspeed, and Altitude as well as the values for twelve other responses from these files.

After grooming the data and preprocessing it, Microsoft VBA was used to loop across the time slices, sending each matrix of controls and responses to MATLAB for regression analysis (MATLAB 2012). Using MATLAB's *regstats* command, the regression coefficients for the second order nonlinear model were estimated along with their p-values.

At this point in the analysis, the p-values of the quadratic effects are analyzed with lenient scrutiny to determine the statistical significance of their coefficient estimates. For each quadratic effect for each response at each time slice, we perform the following hypothesis test:

H_0: The quadratic effect is equal to 0

H_A: The quadratic effect is not equal to 0

A p-value lower than a pre-specified significance level α allows us to reject the null hypothesis and conclude that there potentially is curvature from that quadratic effect for that time slice.

MATLAB provides the ability to generate these response surfaces graphically. A special MATLAB tool[1] was built for the analysis of these response surfaces. Pictorially, any sort of hill or valley in the graph of the response function is evidence of curvature in the design space for that time slice that can be used to help reinforce evidence shown by the model's p-value for that time slice. The MATLAB tool automates the plotting of response surfaces to visually detect curvature.

The bottom line is that if response surface within the design space for any time *t* exhibits any evidence of curvature, a three-level design is reasonable. If not, a two-level design can be employed in a new proposed ARASQ test event with a corresponding reduction in the size of the ARASQ test design.

3.1.2 Optimize the ARASQ Test Event of Interest Using a Specific Optimality Criterion This resulting design from this section is termed "Bayesian" Optimal because it leverages prior knowledge based on historical data from a similar ARASQ test

[1] This tool was developed by the Center for Operations Analysis at AFIT.

event on a similar airframe to gain insights into the design space for a developmental airframe. For KC-46 ARASQ, we can use the response surface analyses on previously collected test data to infer information about the KC-46 and its response surfaces. Using this information, a primary goal is to build response surfaces to validate the network of training simulators ultimately used by KC-46 air crews.

If an analysis relies upon response surfaces for the bulk of the statistical inferences, a Box-Behnken design or a Central Composite Design is preferred because of their ability to accurately estimate the quadratic effects within a specified design space. However, these designs are "expensive" – they require more test runs than often possible in today's fiscally constrained acquisition environment. The current ARASQ designs are fractionated three-level designs but have poor variance properties in general; the fractions employed do not appear to have been developed using statistical rigor or analytical objectivity. For this reason, computer generated optimal designs are used in this investigation to improve the variance-based efficiency of the ARASQ designs.

A computer generated optimal design uses an algorithm to search among and compare the different mixes of test points from a pool of candidate test points to find some best design matrix within the design space with respect to a particular optimality criterion. The optimality criterion of interest in this analysis is Integrated Optimality (I-optimality). Another popular design criterion is Determinant Optimal (D-optimal).

An I-optimal design attempts to minimize the average prediction variance over the design space; to compare to another common optimality criterion, D-optimal designs try to maximize the determinant of the information matrix. While a D-optimal design is handy in screening parameter estimates, an I-optimal design minimizes the variance of

prediction for the response function discussed in equation (2). This allows the analyst to more accurately fit the response surface to the actual data, which at the same time helps decrease the size of the prediction interval associated with a new observation.

A design is I-optimal when

$$Max(I) = Trace[(X'X)^{-1}M] \tag{4}$$

where M is defined as

$$M = \int_R x^{(m)} x^{(m)}{}' dx \tag{5}$$

and $x^{(m)}{}'$ is the row of interest from the design matrix. The trace of a matrix is simply the sum of that matrix's diagonal elements. The objective value for an I-optimal design comes from the prediction variance of a particular design. As these I-optimal designs minimize the average scaled prediction variance across the design space, they tend to develop better predictors in the center of the design space. Unfortunately, this can increase the prediction variance at the extreme points of a feasible space.

In this analysis, an I-optimal design is preferred to an I-optimal design because an I-optimal design typically places fewer test runs near the extreme points of the feasible region for the particular design and the design focuses on the variance of the surface estimates. I-optimal designs provide the best characterization of a response surface, especially one that is used for prediction.

In this investigation, SME directed ARASQ test events are evaluated statistically using JMP in terms of D-efficiency and average scaled variance of prediction. D-efficiency is given by the following equation:

$$D_e = 100 \times \frac{|X'X|^{1/p}}{N_D} \tag{6}$$

with p as the number of parameters being estimated and N_D as the number of test runs in the designed experiment. The scaled prediction variance for any location in the design space x_o is given by the following equation:

$$v(x_0) = \frac{N \operatorname{var}(f(x_0))}{\sigma^2} = N x_0'(X'X)^{-1} x_0 \tag{7}$$

with N as the total sample size and $f(x_0)$ is the predicted response at the point x_o in the design space. The value of this variance operator is averaged across the design space.

Using JMP 10.0 and the functionality of the DOE tab (JMP 10.0), JMP can be utilized to generate an I-optimal design. Using the custom design in JMP to build a design with the desired number of factors with a pre-specified number of levels (as investigated in the first step of this methodology chapter), we can build a "better design" with the same number or fewer runs as specified in the ARASQ document. The word better is in quotes because this is a relative term: JMP will definitely build a design that minimizes the average variance of prediction across the entire design space, but this is usually a tradeoff between D-efficiency and prediction variance.

3.2 A Framework for Simulator Validation

Using the data from an ARASQ test event as performed on the KC-46, using the same style of analysis as in the preliminary step of this methodology, response functions can be estimated according to the following response function:

$$f^{l,t}(X) = \beta_{0,t} + \sum_{j=1}^{J} \beta_{j,t} X_{j,t} + \sum_{k=j+1}^{J} \sum_{j=1}^{J-1} \beta_{j,k,t} X_{j,t} X_{k,t} + \sum_{j=1}^{J} \beta'_{j,t} X_{j,t}^2 \tag{8}$$

This response function is generated using OLS with one subtle exception. This response function contains first order interaction effects. This is better for a response surface used for simulator validation; it characterizes not only curvature in the feasible region but also rotation and twisting in that same space. However, this is an ideal case. Typically, equations following the form of equation (8) require more runs than estimate parameters. A small sample size design will not have enough degrees of freedom (runs) to estimate these effects in an unbiased fashion.

The analyst must determine the exact model form to fit to the ARASQ data. This determination considers each of the required model terms and the available ARASQ response data for a test event.

The ARASQ design, or some alternatively proposed design, can be used to collect responses from the simulator. This data are then used to fit models according to equation (8) for the simulator data using the same methods used to build the response surface models built from ARASQ flight test data.

Validation now proceeds by comparing the ARASQ and simulator response functions (and their corresponding surfaces) for general agreement. We can assess response function agreement using two different methods.

3.2.1 Validation Using Point by Point Comparisons The first validation method is a comparison of a test point couple sampled anywhere in the design space. This point couple is produced using its control levels first simulator. Next, the flight test response value is produced using the response function built from the ARASQ test event corresponding to the control levels in the simulator. Because the value of the response for flight testing comes from the previously generated response function, no additional flight

test time is needed. This method is robust enough to compare any point from the simulator run in the operating envelope of the KC-46 to the "Ground Truth" of the ARASQ flight test event. This technique can be used for a single point-by-point comparison for proof-of-fit.

Using a *(1-α)%* prediction interval, we can account for the variability of the distribution $Y^{l,t}|(X = x_t)$. This conditional distribution is built around $E(Y^{l,t}|X = x_t)$, but it also considers the uncertainty in the fitted values of *Y*. Given a sample point run within the simulator, this prediction interval can be used to validate individual samples or sample data not sufficient to generate a comparable response surface for validation.

Using the fitted values of the response function $f^{l,t}(\widehat{x_t})$, we can build this prediction interval around any fitted value. The *(1-α)%* prediction interval is built according to the following equation:

$$\hat{y}^{l,t} \pm t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2(1 + \widehat{x_t}'(X'_t X_t)^{-1}\widehat{x_t}} \tag{9}$$

where $\widehat{x_t}$ is the column vector of inputs corresponding to the sample point and $\hat{y}^{l,t}$ is the fitted response being analyzed at time *t* matching the time slice of the input values. This prediction interval is calculated using critical value from the student's t-distribution with probability *α/2* and *n-p* degrees of freedom where *n* is the number of test runs in the "ground truth" response surface and *p* is the number of parameters estimated in that response function.

Using this prediction interval as our guide, if a response value as modeled in the simulator fits within this *(1-α)%* prediction interval, we can conclude that the simulator captures the actual response.

3.2.1 Validation Using Surface Comparisons The second form of validation proposes the use of a mesh grid as discussed by Hill et al (2010). This method calls for the comparison of the response surface from ARASQ flight testing for the KC-46, denoted as the ground truth, or *GT*, and the surface generated in the simulator from the optimized designs for minimized prediction variance (I-optimal). Let this surface be called the optimized surface or *O*.

The difference between responses can be denoted as $D_{l,m,t} = Y_{GT}^{l,m,t} - Y_{O}^{l,m,t}$ for *m=1,2,…,M*. This *M* is the total number of points that are being compared for validation. As the pairs of responses are *iid*, the $D_{l,m,t}$ are also *iid* with

$$\mu_{D,l,t} = E\left[D_{l,m,t}\right] = E\left[Y_{GT}^{l,m,t}\right] - E[Y_{O}^{l,m,t}] \tag{10}$$

and

$$\begin{aligned}\sigma_{D,l,t}^2 = V\left(D_{l,m,t}\right) &= V(Y_{GT}^{l,t}) + V(Y_{O}^{l,t}) - 2COV(Y_{GT}^{l,t}, Y_{O}^{l,t}) \\ &= \sigma_{GT,l,t}^2 + \sigma_{O,l,t}^2 - 2\rho\sigma_{GT,l,t}^2\sigma_{O,l,t}^2 \end{aligned} \tag{11}$$

The unbiased estimator of the mean of the differences is $\overline{D}_{l,t}$ as its expected value is equal to μ_D and the unbiased estimator of the variance is $\frac{1}{M}\sigma_{GT,l,t}^2 + \sigma_{O,l,t}^2 - 2\rho\sigma_{GT,l,t}^2\sigma_{O,l,t}^2$. As discussed by Hill et al, $Y_{GT}^{l,m,t}$ and $Y_{O}^{l,m,t}$ are roughly normally distributed for any given *m*, the difference $D_{l,m,t}$ also must be normally distributed. Thus, the test statistic follows the student's t-distribution with *M-1* degrees of freedom as shown in equation (12).

$$t = \frac{\overline{D}_{l,t} - \mu_{D,l,t}}{\frac{1}{M}\sigma_{GT,l,t}^2 + \sigma_{O,l,t}^2 - 2\rho\sigma_{GT,l,t}^2\sigma_{O,l,t}^2} \tag{12}$$

Using this test statistic, the null hypothesis would imply that the average difference between responses for a given response at a given time is equal to 0, or $\mu_{D,l,t} = 0$. A rejection of this null hypothesis signifies that the surfaces are statistically identical at some strict alpha level, resulting in a validated set of data from the training simulator. Confidence intervals can be built around $\overline{D}_{l,t}$, and it is true in practice that if the confidence intervals contain 0 or a value sufficiently close to 0 that the simulator data is statistically identical to the ground truth data from ARASQ flight testing.

4. Bayesian Design Optimization

This chapter illustrates the design optimization methods detailed in chapter three. The sample data used for this analysis comes from Table 5-3 1.a.2 from Flight Testing, using just the KC-135 in flight for the test. The title of the test is Boom Operator Control Characteristics in Free Air. It corresponds to section 2.2.5.3 (Boom Operator Control: Elevation – Free Air) of ARASQ revision C.

4.1 Leveraging Past Data

The test matrix for the historical data is shown in Table 2. Unlike most other ARASQ test events, this event uses only a tanker A/C. For this reason, there are only three controls in the sample design matrix.

Table 2 Sample Design Matrix

Airspeed	Altitutde	Weight
289	24865	260749
289	25068	261126
288	24984	260157
290	24974	259701
291	24914	259885
290	25008	259487
289	24979	259042
290	25078	253532
291	25013	253324
290	24951	259238
291	25065	253780
290	25052	253632
289	25000	258810
323	24913	256064
322	24889	256244
322	24952	255851

Using Microsoft Visual Basic for Applications, the time series control and response data were imported into Microsoft Excel for pre-processing. Table 3 lists the responses of interest in this maneuver:

Table 3 Coded and Uncoded Variables

Coded Response	Uncoded Response
AOA	Angle of Attack
AX	Corrected Longitudinal Acceleration
AY	Corrected Latitudinal Acceleration
BMAZM	Boom Azimuth Deflection
BMELV	Boom Elevation Deflection
BMFAX	Boom Longitudinal Force
PitchAttitude	Pitch Attitude
PitchRate	Pitch Rate
RollAttitude	Roll Attitude
RollRate	Roll Rate
YawRate	Yaw Rate

The data was imported into Excel and organized as shown in Table 4.

Table 4 Sample Time Slice of Data

Time	Airspeed	Weight	Altitude	Airspeed^2	Weight^2	Altitude^2	PitchAttitu	BMAZM	BMELV	BMFAX	AX	AY	PitchRate	RollRate	YawRate	RollAttitu	AOA
0	293.113	260751	24843.4	85915.23077	67991084001	617194523.6	2.80557	0.284828	32.8105	2571.89	0.049883	0.009756	0.063188	-0.07539	0.008949	-0.84027	2.80662
0	293.355	261128	25083.6	86057.15603	68187832384	629186989	2.59923	-0.23262	32.719	2358.47	0.039966	0.006657	-0.0125	-0.1205	0.002539	-1.17071	2.46099
0	292.931	260159	24986.5	85808.57076	67682705281	624325182.3	2.49716	-0.35022	32.8105	2211.63	0.046142	0.009051	0.107813	-0.07635	0.018446	-0.38313	2.58407
0	294.489	259704	24990.6	86723.77112	67446167616	624530088.4	2.19218	-8.8646	33.2074	-4978.86	0.044114	0.001936	-0.02156	0.032389	0.056104	-0.89431	2.46577
0	295.228	259887	24898.9	87159.57198	67541252769	619955221.2	3.07291	-9.89949	33.3905	-5831.44	0.04928	0.003983	0.116312	-0.13333	-0.0167	-0.9323	2.73611
0	294.365	259490	24988.3	86650.75323	67335060100	624415136.9	2.50958	-9.38204	31.9558	-5214.45	0.04373	0.002113	0.017462	-0.16836	0.00438	-0.90725	2.51126
0	293.333	259044	24964.1	86044.24889	67103793936	623206288.8	2.99025	9.38721	30.3684	9368.31	0.049038	0.01104	-0.00545	-0.25801	0.054715	-0.44209	2.59507
0	294.196	253533	25089.4	86551.28642	64278982089	629477992.4	2.57055	11.2464	31.3122	11410.3	0.044155	0.010632	-0.31407	-0.16096	-0.16795	-1.09347	2.43049
0	295.19	253325	24987.2	87137.1361	64173555625	624360163.8	3.04893	10.776	31.0069	11120.9	0.050979	0.011967	-0.02636	-0.67569	-0.10826	-0.35008	2.59412
0	294.804	259240	24966	86909.39842	67205377600	623301156	2.04587	9.99874	31.8947	10258	0.042758	0.010672	0.183324	-0.10377	0.131199	0.760402	2.49904
0	295.275	253781	25032.9	87187.32563	64404795961	626646082.4	2.74861	10.682	30.5795	10691.8	0.045344	0.015964	0.091708	0.150444	-0.10809	-1.32362	2.50355
0	294.408	253633	25056	86676.07046	64329698689	627803136	2.32809	11.1053	31.2816	11310.9	0.051025	0.010484	0.285268	-0.17456	-0.08005	-0.15281	2.61072
0	293.075	258812	25006.1	85892.95563	66983651344	625305037.2	2.5006	10.0928	30.0631	9613.36	0.046608	0.013488	0.051153	0.177366	-0.08781	-1.88642	2.54153
0	328.099	256066	24904.9	107648.9538	65569796356	620254044	1.25289	1.01396	32.8105	2418.91	0.027489	0.004618	0.040742	0.538578	-0.00983	-1.97703	1.54163
0	327.878	256246	24910.4	107503.9829	65662012516	620528028.2	1.93973	-0.53839	33.1769	2582.99	0.026968	0.008436	-0.11595	-0.03216	0.034396	-0.72565	1.49961
0	327.829	255853	24963.4	107471.8532	65460757609	623171339.6	1.6574	-1.17344	32.0779	1977.07	0.032131	0.002033	0.038814	-0.2819	0.128999	1.03576	1.66595

As shown in Table 4, the time slice is followed by the design matrix which is then followed by the matrix *Y*, or the matrix of responses.

For each of the responses of interest at each time, a second-order response function is estimated using ordinary least squares. This function corresponds to equation (1).

After grooming the data and preprocessing it, Microsoft VBA loops across each time slice and sends the data to MATLAB to be evaluated with the *regstats* command. Using a rather lenient standard of significance, with an alpha level of .2, we see that a large proportion of these equations display significance in the second order effects.

Using these equations, response surfaces were generated graphically for visual inspection. The figure below represents one of the thousands of surfaces generated for this maneuver. As you can see in the figure, there is severe curvature in the design space. Using script files in JMP or Matlab, these surfaces are generated for every response at each time slice for inspection.

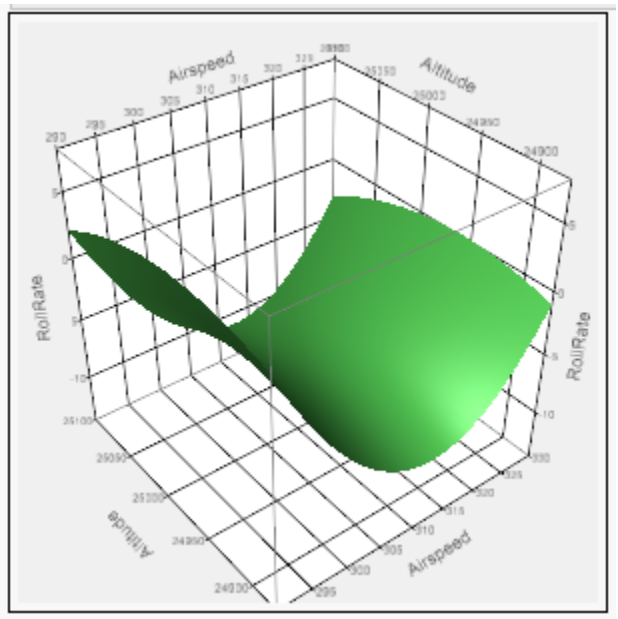


Figure 3 Sample Response Surface

After this inspection and the statistical significance of the second-order models, it was clear that for this maneuver there is curvature in the design space. This justifies a 3-level experimental design for ARASQ 2.2.5.3 that will be run with the KC-46.

4.2 Design Optimization

The design matrix for ARASQ 2.2.5.3 can be seen in Table 5. Note that the design contains 19 runs.

Table 5 ARASQ Design Matrix

Airspeed	Boom Rate	Boom Azimuth	Boom Extension
-1	-1	0	0
-1	0	0	0
-1	1	0	0
0	-1	0	0
0	0	0	0
0	1	0	0
1	-1	0	0
1	0	0	0
1	1	0	0
0	-1	-1	0
0	0	-1	0
0	1	-1	0
0	-1	1	0
0	0	1	0
0	1	1	0
0	-1	0	-1
0	0	0	1
0	-1	0	-1
0	0	0	1

Using JMP to evaluate this design, Figure 4 shows the diagnostics for the ARASQ design. As shown, this is a D-optimal design with 19 runs. This analysis assumes that we are estimating both main effects and quadratic effects.

Design Diagnostics	
D Optimal Design	
D Efficiency	26.49139
G Efficiency	44.32506
A Efficiency	14.95845
Average Variance of Prediction	0.522222
Design Creation Time (seconds)	0

Figure 4 Design Diagnostics for ARASQ Design

As previously discussed a recommended optimality criterion for this design is I-optimal as we will use results to generate response surfaces for simulator validation.

Using the DOE tab in JMP for design optimization, we produced the design found in Table 6 I-Optimal Design

Table 6 I-Optimal Design

Airspeed	Boom Rate	Boom Azimuth	Boom Extension
0	0	0	1
1	1	0	0
-1	-1	0	0
-1	-1	0	-1
-1	0	-1	1
-1	0	1	0
1	-1	1	0
1	0	1	-1
1	-1	-1	1
0	1	0	1
0	1	-1	0
0	0	-1	0
1	0	0	-1
0	-1	1	-1
-1	1	1	1
-1	1	-1	-1

Using JMP to judge the goodness of this design, we see that it is indeed I-optimal, and it actually offers an improvement in relative D-efficiency while allowing us to lower

our average scaled prediction variance (SPV) using just16 runs vice the 19 shown in the current ARASQ design as shown in Table 5.

Design Diagnostics	
I Optimal Design	
D Efficiency	41.51379
G Efficiency	86.60254
A Efficiency	26.8323
Average Variance of Prediction	0.414931
Design Creation Time (seconds)	0.016667

Figure 5 Diagnostics for I-Optimal Design

Comparatively, the I-optimal design is a better design for its purpose, as shown in Table 7.

Table 7 Design Comparison

Design	D-efficiency	SPV	Relative Efficiency Improvement	Decrease in Prediction Variance
ARASQ	26.49	0.522222	56.72%	20.55%
I-optimal	41.514	0.4149	-	-

This new design represents a proposed improvement to the current ARASQ test matrix. It offers improvement in efficiency and prediction with a decrease in the number of test runs. Not only is this a more efficient design, this new design will more accurately capture the curvature of the design space. Properly modeling this curvature is a priority from both the analyst's and the program management's perspective, and using this I-Optimal design allows for this.

This new design is also leaner than the previous ARASQ design. It will be cheaper to test and easier to implement. However, this is merely an example analysis. Other forms of this deep-dive analysis are included in the appendices of this document.

5. Simulator Validation

Chapter five of this thesis will demonstrate the first validation technique. Using this prediction interval comparison technique is a more stringent method for validating individual simulator runs. The data used in this validation comes from Volume III of the Kohlman Systems Research Inc. report on Proof of Match for the KC-135 simulator (Kohlman Systems Research 1998).

5.1 SIMCERT Process

In the United States Air Force, each Major Command (MAJCOM) has its own responsibility to validate its own simulators and training devices. Some MAJCOMs do a better job than others in terms of validation. Within AMC, the job of simulator validation is tasked to specific SIMCERT teams. AMC employs equipment specialists that deal with most of the aircraft systems testing and assist with the objective testing review. Historically, these personnel were “Blue-suit” Air Force personnel with a specialization in simulator maintenance. However, this specialty designator was eliminated. Today, SIMCERT teams are typically composed of one or two experienced pilots, one or two senior enlisted aircrew members. The pilots on SIMCERT teams are usually not flight test qualified pilots, and the senior enlisted testers are usually former load masters or boom operators (as appropriate for the weapon system). Each SIMCERT team is specific to a single type of aircraft or training system. There are teams for pilot training, aircrew training, and maintenance training; respectively, each of the certification processes are handled by a different SIMCERT team.

SIMCERT validation today involves subjective proof of match testing between simulator and flight test data. As Kohlman Systems Research originally showed proof of match in the late 1990s, today data is compared on a time slice by time slice basis. If the data from the simulator falls within a specified tolerance of the flight test data, the simulator is said to be in compliance with the requirements of ARASQ for that test event.

However, these methods fail to capture the probability distribution associated with a response for a given test event or the conditional distribution that is associated with each response in a sufficiently small region of the design space. The current subjective methods for validation apply minimal statistical rigor providing a somewhat shallow framework for validation using two sets of data.

5.2 Data Set

The data used in this portion of the analysis was originally generated for a proof of match report on the KC-135 simulator in 1998. This report is Volume III of the KC-135 vs. KC-10, and the data for this comparison is housed in its appendices. The specific portion used in this validation chapter corresponds with Boom Operator Control Characteristics in Free Air. This is the same ARASQ test event as in Chapter four. The actual flight test data comes from Table 5-3 1.a.2; this test event is described in 2.2.5.3 in ARASQ Revision C.

Within the data set, Kohlman Systems Research replicated four of the test runs associated with the requirement outlined in Revision C of ARASQ. Unfortunately a data set containing only 4 test runs does not meet the data requirements for building a response function to compare to the one generated using Table 5-3 1.a.2. However, the

response function from Table 5-3 1.a.2 is used to build a prediction interval for the response of interest given a set of controls.

In this case, the most interesting response is Corrected Latitudinal Acceleration based on statistical significance. In this analysis, the response values are compared from four test runs at ten matching time slices. The result is 40 individual comparisons with a pass/fail judgment for each observation.

Kohlman Research Systems concluded that all four validation scenarios passed validation. This fact is important, and it will be touched on again in the next section.

5.3 Validation Demonstrated

The ten random matching time slices of comparison for this test event are displayed in Table 8.Table 2

Table 8 Matching Time Slices

Time Slices
0
0.288
2.88
5.088
8.928
13.44
20.448
21.12
27.168
36.384

As this maneuver consisted of 16 runs of flight testing, each time slice contains 16 individual data points.

Shown in **Table 9**, you can see an example from the ten time slices of data from flight testing.

At each time slice in Table 8 a response function is generated using the historical flight test data using equation (1).

Using equation (9) to generate a 95% prediction interval, it is shown in Table 10 that not all of the test runs passed validation according to the (*1-a%)* prediction interval criterion. Table 10 shows the data from the simulator along with the corresponding prediction interval for each observation.

Table 9 Sixteen Settings

Time	AY	Airspeed	Weight	Altitude	Airspeed^2	Weight^2	Altitude^2
36.38	0.0094027	293.576	260638	24886	48.782978	10000655	9948.2731
36.38	0.010622	293.377	261015	25000	51.602403	12527220	214.8854
36.38	0.0031306	294.682	260046	24910	34.556542	6606860	5751.8618
36.38	0.0020734	294.497	259591	24996	36.765805	4474838	105.24646
36.38	0.0017198	293.305	259775	25034	52.642008	5287154.3	2319.2864
36.38	0.0018942	294.042	259372	25004	42.490598	3596262	359.44254
36.38	0.009563	293.954	258927	24984	43.645594	2106507.7	2.3747736
36.38	0.013426	294.749	253476	25056	33.773314	15996950	5035.1754
36.38	0.0150882	295.14	253268	25083	29.381617	17704055	9439.8654
36.38	0.0117525	294.305	259123	24938	39.131046	2713865.1	2222.2767
36.38	0.0125181	294.637	253725	25073	35.08763	14067141	7666.5732
36.38	0.0112803	295.017	253577	25056	30.730184	15199228	4922.2811
36.38	0.0113503	293.32	258695	25011	52.424569	1486890.7	668.68633
36.38	0.0070489	330.457	255955	24825	893.80183	2312281.2	25837.679
36.38	0.0081873	330.489	256135	24858	895.71624	1797258.5	16139.423
36.38	0.0057347	330.577	255742	24876	900.9914	3005433.8	11911.764

Table 10 Validation Results

Time Slice	AY	Airspeed	Weight	Altitude	Lower 95%	Upper 95%	Pass/Fail	Validation Scenario
0	0.00741	288.256	260156	24984.6	-0.001370483	0.050136994	Pass	1
0.288	0.00744	288.269	260156	24984.5	-0.008029931	0.030789403	Pass	1
2.88	0.00932	288.398	260156	24983.4	0.000621941	0.04107653	Pass	1
5.088	0.0068	288.473	260156	24981.8	0.004361576	0.033456188	Pass	1
8.928	0.00867	288.635	260156	24979.5	-0.005587225	0.026498094	Pass	1
13.44	0.00575	288.839	260156	24977	-0.001503785	0.039923524	Pass	1
20.448	0.00877	289.482	260156	24966.2	0.002868258	0.042847398	Pass	1
21.12	0.00799	289.563	260156	24963.4	-0.002831572	0.031003467	Pass	1
27.168	0.00822	290.665	260156	24925	-0.010839915	0.034419992	Pass	1
36.384	0.00664	289.929	260156	24948.3	0.004124071	0.060017154	Pass	1
0	0.01023	288.447	260156	25001.2	-0.001138164	0.047998874	Pass	2
0.288	0.00987	288.461	260156	25000.9	-0.00747429	0.02948742	Pass	2
2.88	0.01249	288.515	260156	24999.4	0.000647544	0.03987508	Pass	2
5.088	0.01008	288.492	260156	25001.1	0.004578489	0.033254322	Pass	2
8.928	0.01146	288.227	260156	25011.6	-0.005058994	0.029059047	Pass	2
13.44	0.01152	287.72	260156	25034.6	-0.00135407	0.048467461	Pass	2
20.448	0.01437	287.568	260156	25055	0.00531668	0.052475752	Pass	2
21.12	0.01394	287.625	260156	25053.1	-0.001478108	0.037438672	Pass	2
27.168	0.01049	288.532	260156	25022.5	-0.017326684	0.044990716	Pass	2
36.384	0.01271	288.55	260156	25019.7	0.001665439	0.071794123	Pass	2
0	0.00853	322.496	260156	24883	-0.031742513	0.009705157	Pass	3
0.288	0.00837	322.457	260156	24884.7	-0.017294796	0.014774948	Pass	3
2.88	0.00984	322.202	260156	24895.7	-0.028688687	0.007445479	Fail	3
5.088	0.0078	322.181	260156	24899.4	-0.019432274	0.006557089	Fail	3
8.928	0.00835	322.514	260156	24893.8	-0.014954519	0.015026299	Pass	3
13.44	0.00901	323.137	260156	24874.4	-0.025883978	0.01190987	Pass	3
20.448	0.00887	324.259	260156	24843.9	-0.036481218	0.00818224	Fail	3
21.12	0.00875	324.356	260156	24840.3	-0.025609564	0.012266044	Pass	3
27.168	0.00493	325.103	260156	24811.6	-0.041045499	0.029939691	Pass	3
36.384	0.00882	324.525	260156	24837.6	-0.062384697	0.008931037	Pass	3
0	0.00803	322.465	260156	24951.3	-0.033326874	0.008622746	Pass	4
0.288	0.0082	322.463	260156	24951.6	-0.020930039	0.011645497	Pass	4
2.88	0.00927	322.408	260156	24955.8	-0.027977369	0.008696088	Fail	4
5.088	0.01153	322.357	260156	24960.5	-0.020961365	0.005525702	Fail	4
8.928	0.01011	322.395	260156	24964.9	-0.016619114	0.013888562	Pass	4
13.44	0.00862	322.419	260156	24966.8	-0.028396148	0.010710724	Pass	4
20.448	0.01	322.72	260156	24960.8	-0.032093176	0.007423661	Fail	4
21.12	0.00994	322.8	260156	24959	-0.023993775	0.009629407	Fail	4
27.168	0.00908	323.59	260156	24933.8	-0.036696389	0.023643141	Pass	4
36.384	0.00804	324.82	260156	24887.3	-0.060367851	0.008529733	Pass	4

Unlike the results as displayed by Kohlman Research Systems in Appendix A of Volume III, these four simulator tests do not match the reality of the *GT*. As shown in

Appendix A, KSR made blanket statements that these four scenarios all passed validation when using a time series comparison of response parameters. However, when using a more detailed form of analysis, it is shown that in the 4th test scenario only 60% of the test points can be validated using a prediction interval and in the 3rd test scenario only 70% of the test points match the ground truth at 95% confidence.

6. Discussion

In this thesis, we have proposed a new methodology for flight simulator validation using time series responses and their response surfaces to evaluate the fit of the simulator model to the ground truth as recorded in flight testing while streamlining the approach for optimizing ARASQ test events as originally proposed by Storm using a simple-to-use software package. This chapter summarizes the contribution of this analysis to the body of knowledge, recommendations gleaned from the analysis, and suggested areas for future research.

6.1 Contributions

The primary contribution of this work is the practical method with which it addresses simulator validation for the KC-46. It expands upon a previous method for evaluating ARASQ test events and provides a step-by-step process from start to finish. In this case, the "start" is analyzing the ARASQ test event. It is then followed by that test event, and the "finish" is the use of the data from that test event to validate the KC-46 flight training simulator.

The KC-46 Directorate can use these insights to potentially improve each of the several dozen ARASQ test designs. As shown in this work, it is possible to build a design matrix that has more desirable variance properties, sometimes in fewer runs, translating to a design that costs less and allows for a more accelerated schedule for the KC-46 flight test program. In today's cost conscious culture, this is a major contribution and will be valuable if utilized properly.

These insights can also be used to ensure that the KC-46's simulator is as accurate as possible. In today's Air Force, the goal of a training simulator for an air crew is to ensure as many revenue bearing flights as possible. Using the proposed methodology for simulator validation ensures accuracy of the simulator across the entire design space and the operating envelope. Using a computer-generated I-optimal design, prediction variance is minimized, and the data from that flight test's design matrix will produce a more accurate representation of the "ground truth" when compared to today's proposed, highly fractionated designed experiments. The biggest takeaway from the proposed methods for validation is the ability to compare simulator data at any point in the design space for roughly any time slice. This allows a more objective and accurate form of validation.

This work can be applied to eliminate a portion of the engineering risk associated with flight testing while entertaining an approximately equal (if not more desirable) cost and schedule to what is currently proposed. This work is a prime example of the use of designed experiments in the defense community and applies rigor and objectivity in a statistical sense in test and evaluation for a major United States Air Force acquisition.

That being said, the fruits of this labor are applicable to programs outside of the KC-46 directorate. The United States Air Force employs simulators for many air frames in its inventory. Changing a handful of assumptions, these methods can be applied to not only current assets in this inventory but also to future A/C acquisitions.

6.2 Recommendations

We recommend using this template for analysis for each and every ARASQ test event. This would increase the probability of meeting the proposed schedule and budget for flight testing for the KC-46.

Within the analysis moving forward, a computer-generated I-optimal design should be used for these flight test events. This ensures that the data captured during flight testing will more accurately portray the underlying truth when displayed as a response surface.

We also recommend using a surface-to-surface comparison for proof-of-match simulator validation if possible. Intuitively, we think that it is a more robust method than the point-by-point comparison even though the point-by-point comparison uses statistical rigor to make objective inferences.

The final recommendation is to use the current MATLAB tool for a side-by-side visual comparison of both the simulator response surfaces and the ARASQ flight test response surfaces at every stage of the validation process. Even though the methods proposed for validation are statistically sound, it is likely that there will arise an instance where the experienced opinion of a subject matter expert should trump the statistical insight of an analyst who has no background in flight test or computational fluid dynamics.

6.3 Suggestions for Future Research

The first suggestion for future research would be directly applying the methods for validation to other programs around the Air Force. This could be beneficial and provide a better goodness of fit for the simulator models used in the Air Force.

A second and more realistic suggestion is to apply the discussed computer-generated design optimization techniques to the ARASQ test events in order to obtain a phased approach to flight testing for the KC-46. However, if a sequentially planned scheme is the end goal for flight testing, Bayesian D-optimal designs should be used for

the first phase. This will help capture the behavior of the responses at the boundaries of the design space. This will hopefully help ensure that the extreme points within our design space are captured initially. In the second phase of flight testing, these D-optimal designs should be augmented with runs concentrated in the center of the design space. This will help capture curvature, twisting, and interaction effects of the design space. This data will then be supplemented with the data previously collected to build response surfaces for simulator validation according the methods set forth in this document.

Appendix A: Table 5-3, 1.a.5

This appendix details the deep-dive analysis of Table 5-3, 1.a.5 using flight test data with the KC-135 as a Tanker A/C and the KC-10 as the receiver A/C. This part of the flight testing corresponds to 2.2.5.6 of revision C of the ARASQ document. The title of this test is Boom Operator Control – Elevation. This test was undertaken in disturbed air with the A/C in their respective pre-contact positions.

The design matrix from flight testing is shown in Table 11.

Table 11 Flight Test Design Matrix

Airspeed:		Altitude:		Weight:	
KC-135R	**KC-10A**	**KC-135R**	**KC-10A**	**KC-135R**	**KC-10A**
290	289	25050	24920	260954	414154
290	288	25056	24915	234135	476735
291	290	25024	24912	261238	414644
290	289	25031	24898	240668	391897
290	288	25044	24915	260647	413628
290	289	25051	24922	233588	475539
290	289	25046	24915	260243	412932
290	289	25065	24916	233432	475199
289	288	25054	24935	260470	413322
290	289	25037	24917	259917	412372
289	287	25053	24919	233125	474528
290	289	25026	24892	259517	411756
289	288	25055	24915	232542	473367
291	290	25044	24911	259739	412094
289	289	25048	24919	232771	473794
290	289	25036	24909	258964	410913
287	286	25067	24920	232193	472718
317	317	24907	24836	242930	386659
326	326	24953	24873	222304	453711
318	318	24906	24834	243128	386938
326	326	24968	24876	222907	454599
317	317	24914	24838	241810	384690
325	325	24973	24867	222008	453276

The response of interest for this maneuver are displayed in Table 12.

Table 12 Variable Descriptions

Coded Response	Uncoded Response
AOA	Reciver Angle of Attack
AX	Corrected Longitudinal Acceleration
AY	Corrected Latitudinal Acceleration
BMAZM	Boom Azimuth Deflection
BMELV	Boom Elevation Deflection
BMFAX	Boom Longitudinal Force
PitchAttitude	Pitch Attitude
PitchRate	Pitch Rate
RollAttitude	Roll Attitude
RollRate	Roll Rate
T_AOA	Tanker Angle of Attack
YawRate	Yaw Rate

This maneuver contained 359 time slices of useable data for analysis.

Table 13 provides an example time slice of the data for time *t=0*. Each time slice contains the same type of data. Using this data, we generate response functions according to equation (2).

A design that initially proposes 3 factor levels for any of the controls assumes nonlinearity in the design space. To make an informed recommendation on the future uses of the ARASQ test matrix 2.2.5.6, the curvature of similar design spaces can be leveraged to apply statistical knowledge of a flight test matrix to a future test event. This technique is typically used in developing Bayesian Optimal Designs.

Using a very lenient standard for statistical significance with an alpha level α=.20, it is clear across the entire design space that there is curvature in the surfaces using the second order response surface model. This provides justification for the use of a 3-level experimental design.

Table 13 Example Time Slice

e (sec)	PitchAttitude	BMAZM	BMELV	BMFAX	AX	AY	PitchRate	RollRate	YawRate	RollAttitu	AOA	T_AOA	Tanker Weight	Receiver Weight	Airspeed	Altitude	TW^2	RW^2	Airspeed^2	Altitude^2
0	3.19638	-0.15829	31.1687	1405.16	0.047822	-0.01609	0.07089	-0.77326	-0.10715	-0.27816	2.26191	2.58051	260955	414158	289.183	24918.4	293196253.9	336533977.2	66.1011615	332.1827408
0	3.72506	0.808631	31.2684	1972.4	0.063466	-0.02609	-0.42508	-0.4759	-0.1097	0.125673	2.71914	2.1682	234137	476739	288.405	24918.7	93993520.49	1956835868	79.3571333	343.2082694
0	3.19638	0.406202	29.2074	1095.97	0.065923	-0.01351	0.301223	-0.42011	0.041557	0.337332	2.36261	2.58181	261240	414647	289.415	24909.9	303037574.3	318831822.8	62.38254389	94.5927626
0	2.74262	-0.3133	32.4762	2561.84	0.04745	-0.00439	-0.19365	0.540752	0.006331	0.016046	2.00962	2.1966	240670	391901	289.259	24902.2	9998405.605	1648511232	64.87113766	4.104194112
0	3.46013	0.876607	28.7189	1588.28	0.067346	-0.011	-0.11375	-0.3781	-0.03417	1.21661	2.43853	2.62549	260649	413631	288.144	24914.8	282810629.5	356147191.2	84.0753671	213.9163971
0	3.81299	0.314704	31.6348	1694.12	0.068755	-0.01505	0.299128	-0.73744	-0.07529	0.125673	3.07981	2.11515	233590	475544	289.131	24918.8	104899087.5	1852539524	66.94941276	346.9234457
0	3.2843	-9.78547	30.3769	-5700.2	0.041496	-0.00253	-0.04495	-0.3979	-0.00087	-0.45402	2.08872	2.56231	260245	412935	288.753	24916.5	269385730.2	382901239.7	73.27808689	266.5343927
0	4.16465	-9.84609	33.7106	-6401.95	0.062148	-0.00812	-0.34152	0.523169	0.198136	3.20315	3.03707	2.10251	233434	475203	289.072	24917.8	108118935.4	1823301749	67.91840068	310.6716835
0	3.2843	-10.0133	33.5728	-6506.03	0.062291	0.00139	-0.32251	0.253724	-0.09245	-0.80573	2.33179	2.5902	260472	413326	287.909	24935	276888749.6	367752052.3	88.44014528	1212.841992
0	3.54804	-9.53703	31.7106	-5674.78	0.063859	-0.0022	0.015213	-0.43834	0.122312	0.689047	2.4427	2.6013	259919	412376	288.822	24917.1	258790746.8	405090590.9	72.10153172	286.48545
0	4.25256	-9.46389	32.6421	-5927.27	0.067499	-0.00948	-0.20241	0.518896	0.039553	0.653238	3.26076	2.14309	233127	474532	287.474	24919.9	114597572.1	1766448406	96.81107829	389.110384
0	3.2843	10.6846	28.7495	9334.46	0.046777	-0.01213	-0.32833	0.418571	0.038002	-0.19023	2.14977	2.61841	259519	411759	289.23	24891.1	246081167.3	430307828.1	65.33912586	82.33963434
0	4.16465	10.6975	31.5432	10392.7	0.069296	-0.00799	-0.14315	0.040625	0.090071	1.44459	3.12354	2.12992	232543	473371	287.621	24916.7	127442098	1670204669	93.93994424	273.1047452
0	3.10847	11.0844	32.0159	11397.3	0.049581	0.001505	0.023354	0.101319	0.099011	-0.54195	2.25571	2.55827	259741	412097	289.627	24908.1	253095467.9	416399221.1	59.07862477	62.81959075
0	3.9009	10.8048	32.3674	10960.6	0.064062	-0.00722	-0.23049	-0.55099	-0.09468	0.47738	3.00513	2.13467	232773	473798	288.617	24920.1	122302046.2	1705288388	75.62497418	397.0407364
0	3.2843	11.0109	31.7449	11259	0.062144	-0.00645	-0.17477	0.541504	-0.04619	-0.36609	2.27019	2.59358	258966	410916	288.271	24907.2	229037182.5	465992629.1	81.76250354	49.36300482
0	4.07673	10.7651	32.5906	10746.4	0.070542	-0.0284	-0.28337	-0.02207	0.049109	1.26873	3.10139	2.13166	232194	472721	286.398	24920	135443638.8	1617498590	119.1429463	393.0655602
0	1.78974	-0.32293	30.6116	1190.26	0.032092	0.013949	-0.31864	-0.1083	0.073521	-0.19023	1.15082	1.59747	242932	386662	316.607	24835.3	810045.998	2101384667	372.2483229	4208.651307
0	2.58217	0.463176	31.5222	2262.98	0.042152	-0.00949	-0.43381	-0.52767	-0.07363	0.829095	1.73442	1.13506	222306	453714	325.761	24875.9	463369776.2	449912357.4	809.2737927	589.2328498
0	1.96556	0.304769	30.3884	1275.73	0.027762	0.001339	0.065639	0.172017	0.025515	-0.54195	1.35291	1.58058	243130	386941	318.255	24836.9	492839.8786	2075883307	438.5563871	4003.614126
0	1.96676	0.65722	31.5126	2277.1	0.036968	-0.01076	0.138913	-0.56124	-0.06129	-1.63289	1.42039	1.11101	222909	454602	325.815	24878.9	437772998.3	488371881.8	812.3490644	452.5881362
0	2.05348	-0.32293	31.8327	1681.43	0.026315	0.007829	-0.13049	0.816419	-0.02465	0.337332	1.38426	1.52634	241812	384694	316.822	24845.8	4080503.239	2285687326	380.5908552	2956.544809
0	2.84591	1.00121	31.0299	2436.17	0.044378	0.006005	-0.39685	0.067382	-0.04243	-0.84154	1.91046	1.09866	222011	453279	325.012	24864.1	476157156.2	431647892.8	767.2200823	1301.342057

Table 14 shows an illustration from a sample set a time slices of the curvature of the response surface models for Pitch Rate as analyzed in the initial flight testing. Note that the starred p-values are significant at the 20% alpha level.

Table 14 P-Values for Quadratic Effects

	P-Values			
Time Slice	TW^2	RW^2	Airspeed^	Altitude^2
24.288	.151*	0.202	0.526	0.439
24.384	.122*	.156*	0.482	0.274
24.48	.110*	.131*	0.506	.137*
24.576	.087*	.082*	0.348	.037*
24.672	.085*	.052*	.176*	.006*
24.768	.097*	.039*	.106*	.001*
24.864	.196*	.055*	.080*	.001*
24.96	0.393	.108*	.076*	.001*
25.056	0.618	0.246	.155*	.005*
25.152	0.661	0.314	0.285	.013*

As this type of distribution of significance is seen across the entire design space for each response, there is significant justification for a three level design for this maneuver. This technique was performed for each of the 12 responses at each of the 359 time slices.

The significance of these quadratic regression coefficients using historical data from similar flight control surfaces and tests provides the analyst insight for related flight testing on similar airframes in the future. This flight testing was done with the KC-135 as a tanker A/C and the KC-10 as the receiver. According to subject matter expertise, the curvature in the design space for this maneuver is assumed to exist in the design space of the next generation tanker of the USAF the KC-46.

The KC-46, like its older counterparts, will undergo flight testing in order to capture ARASQ test events for simulator validation for different A/C pairings. The

ARASQ document specifically prescribes the test events. According to ARASQ 2.2.5.6, the coded design matrix is shown in Table 15.

Table 15 ARASQ Design 2.2.5.6

Receiver	Boom	Azimuth Limit
-1	-1	0
-1	0	0
-1	1	0
0	-1	0
0	0	0
0	1	0
1	-1	0
1	0	0
1	1	0
0	-1	-1
0	0	-1
0	1	-1
0	-1	1
0	0	1
0	1	1

As shown in the table, this design contains 15 runs and three levels for each factor.

This ARASQ 2.2.5.6 design was evaluated using JMP version 10.0 (JMP, 2012).

Design Diagnostics	
D Optimal Design	
D Efficiency	35.83281
G Efficiency	65.81076
A Efficiency	21.21212
Average Variance of Prediction	0.367407
Design Creation Time (seconds)	0

Figure 6 ARASQ 2.2.5.6 Design Evaluation

As shown in Figure 6, the design itself is D-optimal with an average variance of prediction of .367. This D-optimal design minimizes the variance of the parameter estimates in the response surface generation.

However, as the data collected from these tests is used for simulator validation, according to this methodology, the design in question should use I-optimality as response surfaces are being generated for simulator validation. Note that these efficiencies come from a design that is forced to estimate both main effects and quadratic effects.

Using JMP to optimize this design with I-Optimality as the design generation criterion yields the design as shown in Table 16 Proposed I-Optimal Design for ARASQ 2.2.5.6

Table 16 Proposed I-Optimal Design for ARASQ 2.2.5.6

Receiver	Boom	Azimuth Limit
0	0	1
1	0	0
-1	0	0
0	0	-1
0	1	-1
1	0	-1
1	1	1
0	-1	1
-1	0	1
-1	1	0
0	-1	0
-1	1	1
-1	-1	-1
0	1	0
1	-1	1

Using JMP to evaluate this design, we see that it is I-optimal with an average variance of prediction of .339 according to Figure 7.

As you can see, the optimized design provides a relative improvement in prediction variance as well as D efficiency while providing an I-optimal design. This improvement is quantified in

Design Diagnostics	
I Optimal Design	
D Efficiency	42.7021
G Efficiency	83.666
A Efficiency	27.92244
Average Variance of Prediction	0.338889
Design Creation Time (seconds)	0.033333

Figure 7 Design Diagnostics for I-Optimal Design

Table 17 Design Comparison

Design	Criterion	D Efficiency	Prediction Variance	Relative Improvement of D Efficiency	Relative Improvement of Prediction Variance
Original	D-Optimal	35.833	0.367407	-	-
Optimized	I-Optimal	42.702	0.338889	19.17%	7.76%

This new design can be sent forward as a proposed improvement to the current ARASQ test matrix. It offers improvement in efficiency and prediction without an increase in the number of test runs. Not only is this a more efficient design, this new design will more accurately capture the curvature of the design space. This curvature, according SME opinion, is the biggest source of engineering risk in all of flight testing. As previously discussed, properly modeling this curvature is a priority from both the analyst's and the program management's perspective.

Appendix B: Table 2-3, 1.a.3.5

This appendix details the design optimization analysis of Table 2-3, 1.a.3.5 using flight test data with the KC-135 as a Tanker A/C and the KC-10 as the receiver A/C. This part of the flight testing corresponds to 2.2.1.1 of revision C of the ARASQ document. The title of this test is Test Initialization Point. It is part of the Aerial Refueling portion of ARASQ. The design matrix for 2.2.1.1 from ARASQ is shown in Table 18. This design contains 36 runs.

Table 18 ARASQ Design

Control State	Tanker	Receiver
-1	1	-1
-1	1	-0.75
-1	1	-0.5
-1	1	-0.25
-1	1	0
-1	1	0.25
-1	1	0.75
-1	1	1
-1	-1	-1
-1	-1	-0.75
-1	-1	-0.5
-1	-1	-0.25
-1	-1	0
-1	-1	0.25
-1	-1	0.75
-1	-1	1
1	1	-1
1	1	-0.75
1	1	-0.5
1	1	-0.25
1	1	0
1	1	0.25
1	1	0.75
1	1	1
1	-1	-1
1	-1	-0.75
1	-1	-0.5
1	-1	-0.25
1	-1	0
1	-1	0.25
1	-1	0.75
1	-1	1

In the typical fashion, using JMP to evaluate this design we see the design evaluation shown in Figure 8.

Design Diagnostics	
D Optimal Design	
D Efficiency	81.14668
G Efficiency	84.48825
A Efficiency	75.25424
Average Variance of Prediction	0.076389
Design Creation Time (seconds)	0

Figure 8 Design Diagnostics for ARASQ 2.2.1.1

Using I-Optimality criterion for design generation, JMP was used to generate the design in Table 19 Proposed I-Optimal Design

Table 19 Proposed I-Optimal Design

Control State	Tanker	Receiver
-1	-1	-1
-1	1	-1
-1	1	1
1	1	1
1	1	-1
1	-1	1
1	-1	-1
1	-1	-1
-1	-1	1
1	-1	-1
1	1	-1
-1	-1	-1
-1	-1	-1
-1	1	-1
-1	-1	1
-1	-1	-1
1	1	1
-1	1	1
1	-1	1
1	-1	1
1	1	-1
-1	1	1
1	-1	1
-1	-1	1
-1	1	1
1	1	-1
-1	1	-1
1	1	1

Using JMP to evaluate this new design, we see an improvement in prediction variance as well as D-efficiency in fewer runs. This is shown in Figure 9.

Design Diagnostics

I Optimal Design	
D Efficiency	100
G Efficiency	100
A Efficiency	100
Average Variance of Prediction	0.071429
Design Creation Time (seconds)	0

Figure 9 Design Diagnostics for Proposed I-Optimal Design

This design is an all around better choice for this portion of ARASQ. With over 20% fewer runs and better variance properties, it is an improvement over the current design.

As shown in Table 20 Relative Design Improvement, this design clearly is an improvement over the current ARASQ design.

Table 20 Relative Design Improvement

Design	Criterion	D Efficiency	Prediction Variance	Relative Improvement of D Efficiency	Relative Improvement of Prediction Variance
Original	D-Optimal	81.14668	0.076389	-	-
Optimized	I-Optimal	100	0.071429	23.23%	6.49%

Appendix C: Table 5-3, 1.a.8

This appendix will detail the deep-dive analysis of Table 5-3, 1.a.8 using flight data with the KC-135 as a Tanker A/C and the KC-10 as the receiver A/C. This part of the flight testing corresponds to 2.2.5.9 of revision C of the ARASQ document. The title of this test is Boom Operator Control – Elevation (Contact). This test was undertaken in disturbed air with the A/C in their respective contact positions.

Below in Table 21 is the design matrix for flight testing.

Table 21 Flight Testing Design

Airspeed:		Altitude:		Weight:	
KC-135R	**KC-10A**	**KC-135R**	**KC-10A**	**KC-135R**	**KC-10A**
290	289	25042	24928	254833	403005
293	292	25010	24905	255045	403438
289	288	25041	24925	239361	389615
290	289	25046	24919	254585	402607
291	289	25027	24904	254155	402029
290	289	25064	24958	254355	402297
287	286	25045	24928	253883	401662
291	290	25042	24935	253483	401125
289	288	25068	24938	230044	467744
287	286	25036	24930	253687	401399
291	290	25043	24919	230208	468324
291	290	25047	24947	253185	400723
290	289	25047	24920	229616	466216
290	289	25044	24952	267592	354007
320	320	24876	24819	239064	380578
325	325	24972	24914	225052	457349
319	318	24931	24870	239810	381065
325	325	24982	24910	225768	458241
325	325	24955	24876	225275	457627
319	319	24886	24824	237928	379654

Using Microsoft Visual Basic for Applications, the time series control and response data was imported into Microsoft Excel for pre-processing. Using VBA, the data was organized in the standard manner. With our prescribed methodology, regression models were generated according to equation (2). These regression models are second order response functions, capturing the curvature within the design space. Using these equations, response surfaces were generated graphically for visual inspection. This figure looks similar to hundreds of other surfaces for this maneuver.

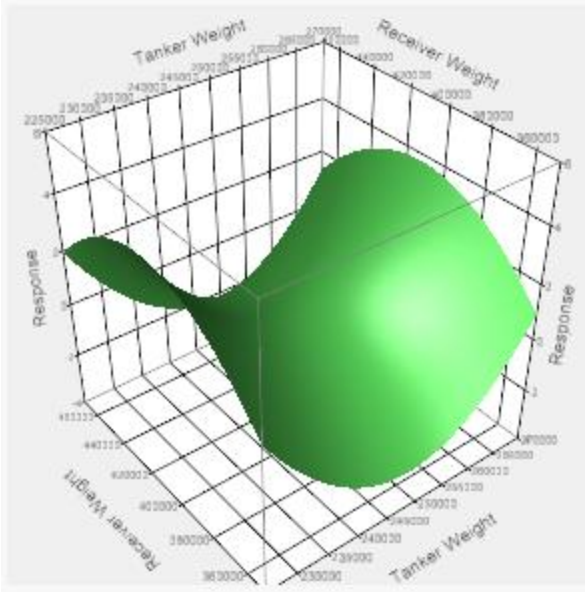


Figure 10 Example Response Surface

A design that initially proposes 3 factor levels for any of the controls assumes nonlinearity in the design space. To make an informed recommendation on the future uses of the ARASQ test matrix 2.2.5.9, the curvature of similar design spaces can be leveraged to apply statistical knowledge of a flight test matrix to a future test event.

Using a very lenient standard for statistical significance with an alpha level α=.20, it is clear across the entire design space that there is curvature in the surfaces using the second order response surface model.

To provide an example, the shows an illustration from a sample set a time slices of the curvature of the response surface models for Pitch Rate as analyzed in the initial flight testing. Note that the starred p-values are significant at the 20% alpha level.

Table 22 Example P-Values

Time *t*	TW^2	RW^2	AS^2	Alt^2
2.16	0.019*	0.035*	0.112*	0.026*
2.208	0.014*	0.026*	0.087*	0.029*
2.256	0.008*	0.016*	0.068*	0.029*
2.304	0.009*	0.018*	0.101*	0.029*
2.352	0.009*	0.018*	0.083*	0.021*
2.4	0.003*	0.006*	0.043*	0.010*
2.448	0.003*	0.007*	0.050*	0.017*
2.496	0.002*	0.005*	0.039*	0.010*
2.544	0.002*	0.004*	0.028*	0.012*
2.592	0.001*	0.004*	0.033*	0.011*

After graphically inspecting some of the response surfaces and these P-values, it is clear that a three-level design is reasonable, perhaps even justified. This distribution of significance is seen across the design space for each of the various responses.

The KC-46, like its older counterparts, will undergo flight testing in order to capture ARASQ test events for simulator validation for different A/C pairings. The ARASQ document specifically prescribes the test events. According to ARASQ 2.2.5.9, the coded design matrix from ARASQ will be a highly fractionated 3-level design in ten runs. This design is shown in Table 23.

This ARASQ 2.2.5.9 design was evaluated using JMP version 10.0 (JMP, 2012). The results are shown in Figure 11.

Design Diagnostics	
D Optimal Design	
D Efficiency	71.13118
G Efficiency	74.31637
A Efficiency	63.36634
Average Variance of Prediction	0.29375
Design Creation Time (seconds)	0

Figure 11 Design Evaluation for ARASQ 2.2.5.9

Table 23 ARASQ 2.2.5.9

Receiver	Boom Operator	Azimuth Limit
-1	-1	0
-1	0	0
0	-1	0
0	0	0
1	-1	0
1	0	0
-1	-1	-1
-1	0	-1
-1	-1	1
-1	0	1

As shown in Figure 11, the design itself is D-optimal with an average variance of prediction of .29375. This D-optimal design minimizes the variance of the parameter estimates in the response surface generation. However, as the data collected from these tests is used for simulator validation, according to this methodology, the design in question should require I-optimality as response surfaces are being generated for simulator validation. Note that these efficiencies come from a design that is forced to estimate both main effects and quadratic effects.

In this ARASQ test event, a decrease in prediction variance is not seen with the use of a computer-generated I-optimal design. These results are shown in Figure 12. For this reason, the current ARASQ test design is sufficient for generating response surfaces

in the future which will work nicely with the simulator validation methodology prescribed in this work.

Design Diagnostics	
I Optimal Design	
D Efficiency	.
G Efficiency	.
A Efficiency	.
Average Variance of Prediction	0.459737
Design Creation Time (seconds)	0.016667

Figure 12 Design Diagnostics for Computer-generated I-Optimal Design

Obviously when comparing the previous Figure 12 to Figure 11 it is clear that the average prediction variance for the ARASQ design is less than the I-Optimal design. The impact of these results is actually still significant. These results show that the current ARASQ design is statistically strong and should remain as a part of the ARASQ testing regiment.

Appendix D: Table 2-3, 1.a.4

This appendix details the design optimization analysis of Table 2-3, 1.a.4 using flight test data with the KC-135 as a Tanker A/C and the KC-10 as the receiver A/C. This part of the flight testing corresponds to 2.2.1.6 of revision C of the ARASQ document. The title of this test is Acceleration/Deceleration Effects (Pre-Contact). It is part of the Aerial Refueling portion of ARASQ. The design matrix for 2.2.1.6 from ARASQ is shown in Table 24.

Table 24 ARASQ 2.2.1.6

Closure Rate	Tanker	Receiver
-1	-1	-1
-1	-1	-0.66
-1	-1	-0.33
-1	-1	0
-1	-1	0.33
-1	-1	0.77
-1	-1	1
-1	1	-1
-1	1	-0.66
-1	1	-0.33
-1	1	0
-1	1	0.33
-1	1	0.77
-1	1	1
1	-1	-1
1	-1	-0.66
1	-1	-0.33
1	-1	0
1	-1	0.33
1	-1	0.77
1	-1	1
1	1	-1
1	1	-0.66
1	1	-0.33
1	1	0
1	1	0.33
1	1	0.77
1	1	1

Using JMP in the typical fashion, the diagnostics for this design are shown in Figure 13 Design Diagnostics for ARASQ 2.2.1.6

Design Diagnostics	
D Optimal Design	
D Efficiency	82.51157
G Efficiency	87.48913
A Efficiency	77.5497
Average Variance of Prediction	0.085227
Design Creation Time (seconds)	0

Figure 13 Design Diagnostics for ARASQ 2.2.1.6

This design originally contains 28 runs. However, using the I-optimality criterion we cut the number of runs from 28 to 24 and produce a design with preferable variance properties. This design can be seen in Table 26 Optimized ARASQ 2.2.1.6 and the diagnostics for this design are shown in Figure 14 Design Diagnostics for Optimized 2.2.1.6

Design Diagnostics	
I Optimal Design	
D Efficiency	100
G Efficiency	100
A Efficiency	100
Average Variance of Prediction	0.083333
Design Creation Time (seconds)	0

Figure 14 Design Diagnostics for Optimized 2.2.1.6

This new design is a statistically preferred design. As shown in a comparison between Figures 13 and 14, the new design provides more D efficiency and improved prediction variance in roughly 15% fewer runs. This comparison can be seen on page 67 in Table 25.

Table 25 Design Comparison

Relative Improvement of D Efficiency	Reduction in Runs
21.20%	14.29%

Table 26 Optimized ARASQ 2.2.1.6

Tanker	Closure Rate	Receiver
1	1	1
1	1	-1
-1	1	1
1	-1	-1
-1	1	1
1	1	1
-1	-1	-1
-1	1	-1
-1	1	-1
1	-1	1
1	-1	1
-1	-1	1
1	1	-1
-1	1	-1
-1	-1	-1
1	1	-1
-1	1	1
1	-1	-1
1	-1	-1
-1	-1	1
1	1	1
-1	-1	-1
1	-1	1
-1	-1	1

As shown with the comparisons, this design as shown in Table 26 is to be pushed forward as an alternate ARASQ design to be used in the testing of the KC-46.

Appendix E: Quad-Chart

AERIAL REFUELING SIMULATOR VALIDATION USING OPERATIONAL EXPERIMENTATION AND RESPONSE SURFACE METHODS WITH TIME SERIES RESPONSES

2LT Alexander Hillman
Advisor: Dr. Raymond Hill
Reader: Dr. Darryl Ahner
Department of Operational Sciences (ENS)
Air Force Institute of Technology

Problem Discussion

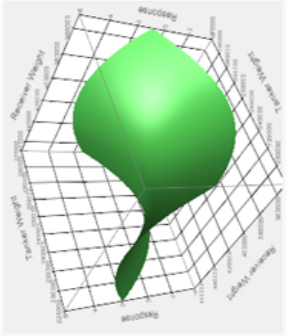
- Current Aerial Refueling Simulator Validation process too subjective – mostly done on a point by point matching basis.
- Some Aerial Refueling Airplane Simulator Qualification (ARASQ) flight testing inefficient – most tests highly fractionated designed experiments non-statistically generated.
- Potential insights into design space for KC-46 lie within the design space of similar aerial refueling air frames, namely the KC-10 and KC-135. Historical data can be leveraged to judge curvature and justify 3-level DOE.
- Research falls under Science of Test initiative for the Operational Test and Evaluation community.

Significance

- New methodology for simulator validation is statistically sound and can automate validation process if necessary.
- Evaluating past designs to gain statistical insights helps reduce program engineering risk.
- From a DOE perspective, ARASQ designs can be evaluated to produce better results with preferable variance properties in fewer runs.

Bayesian Design Optimization

- Using historical data response surfaces can be generated for each time slice.
- Curvature can be visually seen or examined using P-values on quadratic coefficient estimates.



- Curvature justifies use of I- optimal 3-level design.

Research Results

- Optimized several ARASQ designs for use in testing the KC-46 to record data for simulator validation.
- Two new, efficient, statistically objective and defendable simulation validation methodologies provided to KC-46 for new simulator.
- Expands upon past work to include analysis of ARASQ test events at thousands of time slices for each response. Can be applied to any test event that meets data requirements.

Design Evaluation

- Computer-generated designs can be compared along statistical measures such as D-efficiency and Average Scaled Prediction Variance.

Design Diagnostics	
D Optimal Design	
D Efficiency	26.49139
G Efficiency	44.32506
A Efficiency	14.95845
Average Variance of Prediction	0.522222
Design Creation Time (seconds)	0

Response Surface Comparison

- Sampling from *Ground Truth* and Simulator data, surfaces can be compared statistically using the test statistic:

$$t = \frac{\bar{D}_{i,t} - \mu_{D,l,t}}{\frac{1}{M}\sigma^2_{GT,l,t} + \sigma^2_{O,l,t} - 2\rho\sigma^2_{GT,l,t}\sigma^2_{O,l,t}}$$

- where D is the difference between the surfaces at the control settings l at time t and GT denotes the *Ground Truth* from flight testing and O represents the simulator data.
- Using this test stat, we can build a confidence interval around D. If it contains 0, the surfaces are statistically similar.

I-Optimal Designs

- Designs that minimize average scaled prediction variance.

$$v(x_0) = \frac{N\,\mathrm{var}(f(x_0))}{\sigma^2} = N x_0'(X'X)^{-1}x_0$$

Simulator Validation

- Using prediction intervals, simulator data can be validated:

$$\hat{y}^{l,t} \pm t^{\alpha}_{\frac{\alpha}{2}, n-p}\sqrt{\hat{\sigma}^2(1+\hat{x}_t'(X_t'X_t)^{-1}\hat{x}_t}$$

- Observations outside of interval are not "valid."

D-Optimal Designs

- Designs that minimize the variance of the regression coefficients are said to be D-optimal.
- D Efficiency is given by

$$D_e = 100 \times \frac{|X'X|^{1/p}}{N_D}$$

Works Cited

Beers, C.M. van, and Jack P.C. Kleijnen. *Kriging for Interpolation in Random Simulation.* Tilburg, The Netherlands: Tilburg University, 2002.

Boeing Image. "Boeing KC-46 Tanker Program Successfully Completes Preliminary Design Review." *The Boeing Company.* May 8, 2012. http://boeing.mediaroom.com/index.php?s=13&item=1930.

Box, G. E. P., and K. B. Wilson. "On the experimental attainment of optimum conditions." *Jounral of the Royal Statistical Society, Series B*, 1951: 13:1-45.

Box, G. E. P., and N. R. Draper. *Empirical Model Building and Response Surfaces.* New York: John Wiley and Sons, Inc., 1987.

Carley, Kathleen M., Natalia Y. Kamneva, and Jeff Reminga. *Response Surface Methodology.* Research Paper, Pittsburgh: CASOS Technical Report, 2004.

Chaloner, Kathryn, and Isabella Verdinelli. "Bayesian Experimental Design: A Review." *Statistical Science*, 1995: 273-304.

Chernoff, H. "Locally optimal designs for estimating parameters." *Annals of Mathematical Statistics* 586-602.

Dengiz, B, F Altiparmak, and A. E. Smith. "Local search genetic algorithm for optimal design of reliable networks." *IEEE Transactions on Evolutionary Computation*, 1997: 179-188.

Department of Defense. "Department of Defense Instruction 5000.61: Department of Defense (DoD) Modeling and Simulation (M&S) Verification, Validation, and Verification." Department of Defense, 2009.

Fisher, R. A. *Statistical Methods for Research Workers.* Edinburgh, Scotland: Oliver and Boyd, 1958.

The Design of Experiments. New York : Hafner, 1966.

Fowler, John W., and Oliver Rose. "Grand Challenges in Modeling and Simulation of Complex Manufacturing Systems." *Simulation*, no. 80 (2004): 469.

Geisser, Seymour. *Predictive Inference: An Introduction.* London: Chapman and Hall, 1993.

Gilmore, J. Michael. "Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation." *Memorandum for Record.* Washington, D.C.: Office of the Secretary of Defense, October 19, 2010.

Gilmore, J. Michael. "Rigor and Objectivity in T&E: An Update." *ITEA Journal*, 2011: 237-240.

Gilmore, J. Michael. "Test and Evaluation (T&E) Initiatives." *Memorandum for DOT&E Staff.* Washnignton, D.C.: Office of the Secretary of Defense, November 24, 2009.

Harmon, S.Y., and Simone Youngblood. "A Proposed Model for Simulation Validation." *JDMS* 2, no. 4 (2005): 179-190.

Heath, Brian, Ray Hill, and Frank Ciarallo. "A Survey of Agent-Based Modeling Practices (January 1998 to July 2008)." *Journal of Artificial Societies and Social Simulation* , 2009.

Hill, Raymond R, Derek A Leggio, Shay R Capehart, and August G Roesener. "Examining improved experimental designs for wind tunnel testing using Monte Carlo sampling methods." *Quality and Reliability Engineering International.* John Wiley and Sons Ltd., Dec 29, 2010.

Hutto, Gregory T., and James M. Higdon. *Survey of Design of Experiments (DOE) Projects In Developmental Test CY07-08.* Research Paper, Eglin AFB, Florida: American Institute of Aeronautics and Astronautics, 2009.

Jin, Ruichen, Wei Chen, and Agus Sudjianto. "An efficient algorithm for constructing optimal design of computer experiments." *Journal of Statistical Planning and Inference*, 2005: 268-287.

"JMP 10.0." The SAS Institute Inc.

JMP Support. "Creating Response Surface Experiments." *JMP: Statistical Discover from SAS.* 2013.

Johnson, Rachel T., Gregory T. Hutto, James R. Simpson, and Douglas C. Montgomery. "Designed Experiments for the Defense Community." *Quality Engineering*, 2012: 60-79.

Jones, Brad, and Peter Goos. *I-optimal versus D-optimal split-plot response surface designs.* Research Paper 2012-002, Antwerp: University of Antwerp, 2012.

KC-46A Tanker. United States Air Force. May 18, 2011. http://www.af.mil/information/factsheets/factsheet.asp?id=18206.

Kiefer, J. "Optimum Designs in Regression Problems II." *Annals of Mathematical Statistics*, 1961: 298-325.

Kiefer, J., and J. Wolfowitz. "Optimum Designs in Regression Problems." *Annals of Mathematical Statistics*, 1959: 271-294.

Klugl, Franziska. "A Validation Methodology for Agent-Based Simulations." Fortaleza, Cear´a, Brazil: SAC, 2008.

Knepell, Peter L., and Deborah C. Arangno. *Simulation Validation: A Confidence Assessment Methodology.* Los Alamitos, CA: IEEE Computer Society Press, 1993.

Kohlman Systems Research. *Proof of Match Report.* Lawrence, KS: ASC, 1998.

Kuhfeld, Warren F. "Experimental Design: Efficiency, Coding, and Choice Designs." In *Marketing Research Methods in SAS*, by Warren F. Kuhfeld, MR-2010C. SAS, 2010.

Maria, Anu. "Introduction to Modeling and Simulation." Edited by S. Andradottir, K. J. Healy, D. H. Withers and B. L. Nelson. *Winter Simulation.* Winter Simulation, 1997.

"MATLAB." *R2012A.* Mathworks, 2012.

Montgomery, Douglas C. *Design and Analysis of Experiments.* Hoboken, NJ: John Wiley and Sons, Inc., 2009.

Myers, Raymond H., Douglas C. Montgomery, and Christine M. Anderson-Cook. *Response Surface Methodology: Process and Design Optimization Using Designed Experiments.* Hoboken, NJ: John Wiley and Sons, Inc., 2009.

Raiffa, H., and R. Schlaifer. *Applied Statistical Decision THeory.* Boston, MA: Division of Research, Harvard Business School, 1961.

Robinson, Stewart. "Conceptual Modeling for Simulation: Issues and Research Requirements." Edited by L. F. Perrone, F. P. Wieland, J. Liu, B. J. Lawson, M. Nicol and R. M. Fujimoto. *Winter Simulation.* IEEE, 2006.

—. *Simulation: the Practice of Model Development and Use.* Chichester, UK: John Wiley and Sons, Inc., 2004.

Sargent, R. G. "Validation and Verification of Simulation Models." Edited by R. G. Ingalls, M. D. Rosetti, J. S. Smith and B. A. Peters. *Winter Simulation.* Piscataway, NJ: IEEE, 2004. 17-28.

Schatzoff, M. "Design of Experiments in Simulator Validation." *IBM Journal of Research and Development* 19, no. 3 (1975).

Smith, Kirstine. "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance They Give Towards a Proper Choice of the Distribution of Observations." *Biometrika*, 1918.

Storm, Scott. *Evaluating Aerial Refueling Simulator Test Designs by Extending Response Surface Methodology to Analyze Time History Responses.* Dayton, OH: Air Force Institute of Technology, 2012.

Todoroki, Akira, and Tetsuya Ishikawa. "Design of experiments for stacking sequence optimizations with genetic algorithm using response surface approximation." *Composite Structures*, 2004: 349-357.

"Visual Basic for Applications." Microsoft, 2007.

Zeigler, Bernard, Herbert Praehofer, and Tag Gon Kim. *Theory of Modeling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems* . San Diego: Academic Press, 2000.

Vita

2d Lt Alexander Hillman graduated from De La Salle Collegiate High School in Warren, Michigan. He entered undergraduate studies at the United States Air Force Academy in Colorado Springs, Colorado where he graduated with a Bachelor of Science degree in Economics with a Russian Language Minor in May 2011. Here he was commissioned an officer in the United States Air Force. Alexander graduated from the Air Force Academy with academic distinction, signifying a grade point average in the top 10% of his class.

In the summer of 2011, Alexander studied as a United States State Department Critical Language Scholar in Kazan, Russia. In October of 2011, Alexander entered the Graduate School of Engineering and Management, Air Force Institute of Technology. Upon graduation, he will be assigned to Detachment 1 of the 53rd Test Management Group at Barksdale Air Force Base.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
21-03-2012	Master's Thesis	OCT 2011-MAR 2013

4. TITLE AND SUBTITLE
Aerial Refueling Simulator Validation Using Operational Experimentation and Response Surface Methods with Time Series Responses

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Hillman, Alexander, P., 2d Lt, USAF

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENS-13-M-06

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Secretary of Defense
Attn: Dr. Catherine Warner
1700 Defense Pentagon — commercial #: (703) 697-7247
Washington D.C. 20301 — e-mail: catherine.warner@osd.mil

10. SPONSOR/MONITOR'S ACRONYM(S)
OSD

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

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. UNLIMITED DISTRIBUTION.

13. SUPPLEMENTARY NOTES
DESTRUCTION NOTICE – For unclassified, limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
An important program in the Department of Defense is the KC-46 Supertanker. Dubbed the future of the Air Force's aerial refueling inventory, the KC-46 will replace dozens of ailing previous generation tanker aircraft. The Aerial Refueling Airplane Simulator Qualification document governs the methods by which Air Mobility Command validates its simulators, some of which will be KC-46 simulators in the near future. The methodology set forward in this thesis utilizes historical data of aircraft performance from similar air frames to gain statistical insight into the performance design space of the KC-46. Leveraging this insight, the methodology provides through a framework for validation that uses classical experimental design principles as applied to time history responses such as found in aircraft performance measures. These principles guide the generation of response surfaces from real world flight test data that can then be used to validate flight training simulators using a point by point comparison or over an entire surface of points for a variety of different aerial refueling maneuvers. This work also supports the KC-46 Tanker program by proposing statistically efficient and cost conscious experimental designs for the KC-46 flight testing. This framework is demonstrated using flight testing data from the KC-135 Aerial Refueling Simulator Upgrade testing, and is part of an Office of the Secretary of Defense initiative to add increased statistical rigor to the Department of Defense test and evaluation enterprise and specifically the acquisition community.

15. SUBJECT TERMS
Aerial Refueling, Validation Testing, Response Surface Methodology, Optimal Designs, Statistical Rigor, Time-Series Response Data, Simulator Validation, Science of Test

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Raymond R. Hill, Ph.D. (ENS)
U	U	U	UU	85	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 7469; e-mail: Raymond.Hill@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18