3-11-2011

A Methodology Using Simulation Results for Test and Evaluation

Jonathon S. Hosket

A METHODOLOGY USING SUMULATION

RESULTS FOR TEST AND EVALUATION

THESIS

Jonathon S. Hosket, Captain, USAF

AFIT-OR-ENS-MS-11-09

DEPARTMENT OF THE AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-OR-MS-ENS-11-09

A METHODOLOGY FOR USING SIMULATION RESULTS FOR TEST AND EVALUATION

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

Jonathon S. Hosket, BS

Captain, USAF

March 2011

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-OR-MS-ENS-11-09

A METHODOLOGY FOR USING SIMULATION RESULTS FOR TEST AND EVALUATION

Jonathon S. Hosket, BS

Captain, USAF

Approved:

_____________//Signed//________________ 16 March2011___

Raymond R. Hill (Chairman) date

_____________//Signed//_________________ _16 March 2011_

Major Shay Capehart (Reader) date

AFIT/GOR/ENS/11-09

Abstract

Each year the Air Force spends billions of dollars on Test and Evaluation to ensure acquisition programs roll out the best possible products. In 1997, the National Research Council assembled to evaluate the overall procedure used in procuring various platforms with system planning, research, development and engineering (SPRDE) and program management (PM) processes. In their final report, they claimed that the full advantages of statistical practices, simulation, model-test-models, and incorporation of prior test information into current test practices have not been fully utilized. To examine one of the report's recommendations, this thesis defines and explores a methodology using simulation to augment or replace test data in lieu of operational testing. Specifically, a validated simulation model employs non-critical factor data from preliminary small sample operational testing. The simulation then generates posterior distribution data to replace the corresponding data in the final test matrix. If useful, data generated by a validated simulation model can be used in lieu of actual operational test data for selected non-critical factors. This provides T&E squadrons a means to decrease the level of live operational testing on non-critical factors. Therefore, T&E can be more efficient as less runs are needed to evaluate system factors of interest. This thesis defines methods to use test data to validate simulation results, us simulation data as evidence for subsequent operational testing, and use simulation to potentially replace test data.

AFIT/GOR/ENS/11-09

To

My Family and Friends

Acknowledgments

I would like to express my sheer appreciation to my Advisor, Dr. Raymond Hill, for all his patience and support though out this tedious process.

I also would like to thank Major Shay Capehart for his continual dedication towards my educational advancement throughout this masters program, my career, and my lifting.

Jonathon S. Hosket

Table of Contents

List of Figures

List of Tables

Page

INCORPORATION OF PRIOR TEST INFROMATION TO IMPORVE TESTING RESULTS VIA SIMULATION AND DESIGN OF EXPERIMENTS

1. Introduction

1.1 Thesis Introduction

Throughout the Air Force's history, test and evaluation (T&E) processes advance to meet the competing demands of increasing technology and the ever common reduction in the Department of Defense's fiscal budget. To counter this never ending struggle, T&E squadrons look for more inventive techniques such as design of experiments, Bayesian analysis, simulation, decision analysis, systems engineering, and advance statistical practices for innovative testing approaches. To demonstrate the important applications of Subjective Bayesian simulation principles in the test and evaluation process, this thesis applies these existing concepts to the previous research conducted in Wellbaum et al (2010). Specifically, a methodology is defined that utilizes a small sample of preliminary operational test data, a validated a simulation model, and critical test factors identified via design of experiments (DOE). The simulation is used to generate a priori evidence to support operational test results. The simulation is also used as a means to potentially screen out actual operational test events.

1.2 Problem Statement

During the system engineering process for a new platform certain test criterion must be met during the Material Solutions and Technology development phases before the program can advance to initial rate production. Since funds are generally fixed and limited, these tests can strain a program budget; going over the budget can often cancel a program. Thus, effective and less costly ways of conducting experimentations are always needed for the test and evaluation enterprise. In 1997, a National Resource Council evaluated the effectiveness of Department of Defense (DoD) testing practices and concluded that "the current practice of statistics in defense testing design and evaluation does not take full advantage of the benefits available from the use of state-of-the-art statistical methodology"(7). They further recommended that model-test-model, a technique in which simulation results augment operational testing, should be implemented more frequently in appropriate testing scenarios (7).

This thesis integrates principles from simulation, subjective Bayesian, and design of experiments to define methods for conducting test and evaluation making specific use of simulation results. If successful, such methods could be more efficient, less costly, and just as effective as results from current live test and evaluation practices.

1.3 Scope

This thesis is focused on subjective Bayesian simulation techniques applied to test data rendered from overhead watch and loiter (OWL) experiments. Specifically, the work utilizes a pre-existing simulation model validated with OWL preliminary test data, evaluates the ability of the simulation to provide a priori evidence to support test event inferences, and provides posterior data on non-critical factors, which are swapped into the final test data model. Although, this application of predictive simulation is new, predictive simulation has been applied to a variety of applications in the test and evaluation arena.

2. Literature Review

Bayesian probability, although introduced by Thomas Bayes, didn't gain popularity until the 18th century by a French mathematician Pierre-Simon Laplace (3). Since that time, there have been two major factions of Bayesian scholars; those that view probability objectively, and others that believe Bayesian probability is subjective in nature. This thesis is primarily concerned with subjective Bayesian applications; although there are traditional benefits from objective Bayes practices.

Objective Bayesian principles are founded on the belief that one can take prior information, generate posterior information with mathematics, and gain insight into the unknown. James Berger describes Bayesian analysis as, "…simply a collection of ad-hoc but useful methodologies for learning from data" (3). Berger claims that objective principles offer the following advantages : "highly complex problems can be handled, via Markov Chain Monte Carlo; very different information sources can easily be combined; multiple comparison are automatically accommodated; methodology does not require large sample sizes; and sequential analysis is much easier"(3). Objective Bayesian applications require picking the right prior distributions to generate posterior probabilities. If chosen poorly, objective Bayesian principles can lead to improper distributions which, in turn, can lead to false or less accurate statistical conclusions. These false conclusions are more prominent when modeling complex systems, or scenarios in which no subject matter expert can verify prior distribution accuracy. For these reasons, "objective Bayesian analysis is a convention we should adopt in scenarios in which a subjective Bayes analysis is not tenable" (3). This leads one to believe subjective Bayes principles, if relevant experts are available, yield a more secure estimate on posterior probability.

Subjective Bayes analysis does not significantly differ from objective Bayesian except for the premise of "verified" prior distributions. Verified in this case refers to a confidence in prior distributions when obtained through a subject matter expert (SME). However, difficulties arise in subjective practices when soliciting probability distributions from SME's. Individual biases like anchoring, familiarity with round numbers, can lead to poor prior distribution estimates. Elicitation biases can be mitigated through the use of various probability soliciting techniques such as assessing extreme probability estimates or the popular "probability wheel." In this thesis, prior distributions are derived by using a simulation model presumed to provide valid output results.

Simulation is the computer-based imitation of the operation of a real-world process or system over time (2). With simulation modeling, one can create a real-time system yielding estimates of various real world processes. The goal is to use the simulation to model real-life processes or system functions, in the hope understanding them and possibly finding ways for improving upon them in some manner. Through modeling and simulation, myriad companies have been able to analyze their business practices to improve processes, cut costs, and reduce man hours required. For example, "Knowledge modeling and resource-management techniques and tools, based on simulation and other decision analysis methodologies," yielded over 69.7 million dollars in savings (2). In this research, simulation provides an additional benefit since the model used has been validated to the real environment (via actual test results). Thus, posterior distribution data utilized in the final model are assumed to fall within the range of values one observes during actual testing using the real system. Using a valid simulation ensures that the resulting simulation-based testing yields relevant and accurate results which drive valid conclusions about the actual testing. In essence, simulation is utilized as a subject matter expert to verify and validate conclusions pertaining to the real system; a form of subjective Bayesian analysis.

Simulation-based subjective Bayesian applications "…have been around for some time, but have been increasingly applied and developed in recent years" (3). This is due to the advantages simulation offers to improve prior distribution certainty. Notably, there can never be absolute certainty about prior distributions; they are subjective. However, validated models offer additional confidence in prior distribution selection. This increased confidence from simulation platforms has impacted recent distribution projections in fields such as healthcare, logistics, transportation, distribution, and military applications. In some cases, real data distributions are used as the preliminary foundation upon which the simulation subsequently runs. The next case utilizes simulation maps GPS routes in cars.

Palagummi (9) applied simulation and Bayesian techniques to assess the viability of GPS devices to predict driving routes along avenues of low congestion. In his study, the entire map of an area of interest to a driver is divided into grids. The next grid that a person drives into is generated and mapped via the GPS, and the simulation uses the current status and history of the prospective grid as prior information. With this information, the simulation generates posterior prediction information used by the GPS to plan routes for the driver. The information required includes static and dynamic data such as topology, signal control, and vehicle flow rates. At the beginning of each simulation run, the avenues are divided into overlapping “simulation windows”. “Each ‘road link’, defined by starting and stopping coordinates between two intersections, is defined as a “the essential resolution within a simulation window” (9). Each simulation window stores the information o road links within that window. Palagummi (8) defines an active region as, “the set of simulation windows that are currently simulated by the vehicle.” Furthermore, each road link in the active region is dubbed an “active link”, and continuous data for these links is obtained for the simulations. All this continuous information will influence the different outcomes of the simulator.

The simulator, first, updates information on all active links and windows, then discards any old active windows. Prior information needed for the simulation is then downloaded. The simulation then generates all posterior information for the region of interest based on the prior information obtained earlier. This process continues until the predefined simulation stop time is reached when all results are recorded and the simulation ends. These results, based upon using different initialization techniques, are then compared in the final evaluation.

Palagummi (9) defines three different initialization techniques called “empty grid” initialization, “simulation with flow rates”, and “simulation with flow rates and queue lengths”. Empty grid initialization entails starting the simulation with unpopulated windows that populate as vehicles enter and exit the windows. Simulation with flow rates incorporate flow rates based on mean vehicular headway where vehicles are distributed uniformly across a road link by the mean vehicular gap (9). The third initialization technique (simulation with flow rates and queue lengths) incorporates flow rates and queue lengths of slowly moving traffic, based on continuous mean queue length data, on the way to traffic lights. Results from these three initialization techniques are compared to ground truth, the actual transversal time of an active link, as well among one another. Palagummi found that empty grid initialization underestimated the ground truth. The other two initialization methods yielded vehicle travel times more relevant to the actual situations.

Pengfei Li (8) uses simulation, with prior distribution information, to keep drivers out of what he termed the “Dilemma Zone” (DZ). The DZ “…is an area at high-speed signalized intersections, where drivers are indecisive of stopping or crossing when presented with yellow indicator” (8). Li utilizes a simulation-based, Markov process as a way to predict the number of drivers in the DZ. This posterior prediction data, in turn, indicates the best time to transition the light to yellow to decrease collisions amongst vehicles traveling though the intersection. The equation used to predict the hourly number of vehicles in the DZ is where, at step time t, is the predicted number of vehicles caught in the DZ, is the current green light duration, is the calculated number of vehicles caught in the DZ over an hour, is the time loss between green lights, and is the average green light durations on conflict phases (8). If the number of vehicles in the DZ is less than predicted, then the green light period ends. But if the number of vehicles in the DZ is “minimally equal” to the predicted value, then the green light period is extended one time step. To keep the predicted value accurate, Li uses a “rolling horizon” technique which “collects state transitions during the (head) time of each stage, updates the matrix according to new data, and then applies the new matrix during the (tail) time” (8). This algorithm was deployed in VISSIM which fed real time data to into the algorithm and then evaluated when to change the light depending on what output data it received. To model current traffic volume patters, data were collected every fifteen minutes, over a 9 hour span, from Peppers Ferry Road and fed into VISSIM. The measurement parameters of interest were: “probabilities of max outs in an hour” (lights that change green because they reached their allotted time), and “the average number of vehicles caught in the dilemma zone” (8). The results of the simulation were compared to a “green extension system,” using advance detectors, to extend the green light, to circumvent a collision caused by a car in the DZ. Li concluded that the green extension system failed to minimize max-out ratios, whereas the prediction model kept more vehicles out of the DZ in heavy traffic and max outs below 8% (8). Clearly predictive simulation offers great advantages when applied to traffic patterns; but studies have shown that the public health department can also benefits from predictive simulation when modeling population trends.

Bohk (5) created the “probabilistic population projection model (PPPM)” to predict the future demographic of an area based off past trends, from 1990 to the jump off year of 2006, to make projections from 2007 to 2048 (5). The algorithm required a large number of input parameters to effectively predict future populations: current birth rate, mortality rates, fertility rates, sexual birth proportion of males and females, as well as immigration trends. The model also required a set of rules, or “assumption paths,” that contain estimated future values of a certain input parameter (5). Assertion paths represent possible evolutions during the projection horizon which were determined by a subject matter expert involved in the modeling. After all constraints and inputs were defined, the model was simulated via Monte Carlo. The first “limited type” simulation differed from the second (open type), in that the yielded projections were not influenced by improper pairing of assumptions due to the addition of “set types”. For each set type, which was essentially population propagation rules, the modeler would define consistent assumptions so that each input parameter was included into a corresponding set type. An example would be a set type labeled “fertility rates”, which restricts the introduction of births to individuals over the age of eighteen. Results showed that the limit type simulation predicted a population between 65.51 and 69.3 million people, while the open type yielded a 65.58 to 69.1 million estimates. Significant emphasis was put on the fact that the limited type showed a 7% smaller variance. Bohk claims that the matching of improper inputs to assumptions paths caused an averaging effect in the data from the “open type” simulation which could explain the greater variance.

An important issue in the medical field is the evaluation of drug effectiveness in patients. Bayesian simulation is used to predict the correct level of medication to prescribe a patient. Historically, patients must visit a doctor for multiple follow up appointments in order to determine if the prescription drug is working at desired levels. This procedure is costly, time intensive, and uncomfortable for the patient since blood work is usually required while over prescribing medication can cause discomfort. Blau (4) created a subjective Bayesian model-based methodology, using simulation, to determine the optimal drug dose for an individual while minimizing the required invasive procedures.

Blau’s model required existing Pharmaco-Kinetic/PharmacoDynamic (PK/PD) population data, available during the drug development phase, as prior distribution information. Then, using traditional Bayesian principles and Markov Chain Monte Carlo sampling techniques, posterior probability distributions for individuals were created to determine the drug levels after each dose. The effectiveness of this technique relies on the concept of a “therapeutic window”, which is the desired “drug plasma concentration, which is less than an acceptable risk of a toxic side effect and greater than an acceptable level of efficacy” (4). By working within the therapeutic window, Blau demonstrates the effectiveness of his prediction model.

First, data collection on an individual must be taken to estimate his PK/PD parameters. With this information one can predict the individual’s therapeutic window, determine the proper doses available, and “…candidate dose intervals convenient to the individual to find a regimen that maximizes the therapeutic window” (4). However, instead of collecting real data, Blau generated all required information on 8 subjects using simulation and design of experiments. Data derived using a full, two-level factorial design over “reasonable” parameters was entered into ModQuest to predict posterior distributions for the PK/PD parameters. The results were compared to “the posterior probability distribution obtained where the means of the individual posterior parameter distribution for the eight subjects were averaged and standard deviation obtained” (4). Blau’s method used was able to determine the correct posterior PK/PD distribution for the eight subjects. He states, “the personalized pharmacokinetic parameters are in good agreement with the values used to generate them”, and rarely was more than one test for data needed.

Steffens (10) designed a tactical prediction system based on data mining and simulation. The posterior results strive to reduce the cognitive work load placed on a commander, by predicting future tactical scenarios. In his methodology, a user can classify various similar states into cluster sets which are then checked for ambiguity using the k-means-algorithm (MacQueen 1967) (10). After aerial reconnaissance and communication data are acquired, the system stores a state relative to the field conditions. Using a function, “c (A)” (defined by Steffens), a state can be mapped into a cluster if the similarity between the cluster and the state does not fall below a predetermined threshold. Then “using a Markov graph, the system presents the probabilities of future situations and graphically depicts the fitness values of these situations” based on the fitting of clusters to states (10). The advantage of this process is that little actual online computing is done. Most of the scenarios grouped into clusters are defined off line by subject matter experts leaving only aerial reconnaissance and matching completed online. This saves time and effort by not bogging down the military online community which tends to see a lot of action during tactical scenarios, but also incorporates data to future mapping predictions.

Celik and Son (10) used a Monte Carlo-based, dynamic-data-driven-adaptive, multi-scale simulation (DDDAMS) to control the fidelity states of overloaded systems in supply chains. Fidelity is defined as how closely the simulation model imitates the true environment. Therefore, the higher the reported fidelity, the closer the DDDAMS system showed, predicts the actual states of the supply chain. Celiks and Sons methods “…1) handle the dynamicity issue of the system by selectively incorporating up-to-date information into the simulation-based real-time controller, and 2) introduce adaptive simulations capable of adjusting their level of detail according to the altering conditions of a supply chain in the most economic way. (6)” Sensors on the shop floor report fidelity states to the DDDAMS system which analyzes the data using four imbedded algorithms.

The first algorithm detects noise and any abnormal status of the system via the reported sensor data. The second algorithm selects the correct fidelity of the system using a Bayesian Belief Network. The third algorithm examines the available resources of the system and then chooses the available fidelity for each component. Finally, the fourth algorithm predicts the future performance of the system and selects the optimal control tasks to complete based on the identified fidelity of the system.

In addition to the sensory data used above, DDDAMS also used performance data which "…shows the cumulative effect of the successive changes in a system state or sensory data." This data, unlike sensory data, were collected at all times regardless of the fidelity state of a system. Following the culmination of all the information the DDDAMS system, an optimal fidelity state was achieved.

Celik and Son tested this system on a manufacturing supply chain where the goal was to find "the best preventative maintenance scheduling and part routing" (6). Using historical data for prior information, DDDAMS was applied to the supply chain to form the initial fidelity measurements. Celik and Son conclude, that "Monte-Carlo based fidelity selection would lead to highly accurate results while saving computational resources and time" (6).

The previous literature review highlights advantages and areas of application in which subjective Bayesian simulation techniques have been used for system prediction. The main difference in the proposed research from that of the past, shown above, is the influence to design of experiments. In addition to using the simulation to generate (predict) distributions as evidence for a real test, simulation can augment (replace) actual test data provided the simulation is valid and it is accredited for such use. The subsequent methodology focuses on augmenting test results leaving the accreditation challenge to future research.

3. Background

This effort focuses on the advantages of implementing simulation techniques to reduce the amount of time, runs, and data to be collected in actual experiments. Part of the research extends the work of Wellbaum et al (11). Therefore, a brief discussion of the overhead watches and loiter system (OWL), operation center, data collection, testing issues, and the simulation model is warranted. The limited OWL test data is used in Chapter 5 to demonstrate (in a limited manner) the methodology of Chapter 4.

3.1 OWL Platform

The platform all the data was collected on is called the overhead watch and loiter system (OWL). This is a modified configuration of the type A RAVENS used in the Area of Responsibility (AOR). Following the implementation of the RAVEN version B, A versions were disengaged and returned to the U.S. Once state side, AFRL over-purchased a large amount of the platforms after removal of the classified systems. From this surplus, the Air Force Institute of Technology acquired four RAVENS and made additional avionics modifications to tailor the platform to future research efforts.

3.1.1 Modified Avionics System

The Procerus Kestrel avionics system (OWL shown in figure 1) serves as the autopilot once the OWL has been hand launched. It combines air data sensors, accelerometers, and gyroscopes to navigate missions streaming from the operations base. In return, the system provides continuous updates on airspeed, altitude, orientation, and body measurement back to the user.

3.1.2 OWL Specifications and Operations

The OWL platform has roughly a four foot wingspan and a body length of three feet. As seen in Figure 1, the OWL lacks landing equipment and thus requires a soft terrain to land in order to prevent damage to the body. The propulsion system is located behind the body to push the platform during flight. Once airborne, OWL receives and relays information via the sensor in the nose cone. This information is then relayed to the avionics system located behind the orange plate on the side of the platform next to dual 2100 mili-amp-hours batteries. The avionic system then controls the speed, elevation, and direction of the OWL for the duration of the flight via the propeller and the flap located on the tail of the platform. Each avionics system can relay information via different communication channels to prevent confusion of systems during multiple OWL flights. Following mission completion, the OWL is disassembled and placed into a 2’x6”x1’ travel box stored in the operations base trailer.

Figure 1. OWL

3.2 Operations Base

The operations base is a converted mobile trailer roughly forty feet in length, twenty feet in width, and six and a half feet high. The rear half of the trailer was converted into a work shop to repair the platforms and recharge the OWL batteries. In contrast, the front of the trailer contained all the computer hardware, software, and monitors used to control and document the OWLs flight.

3.2.1 Computer Software

Virtual Cockpit is the main program for controlling the OWL. In this system, the user plots the course of the mission, and then uploads it into the database. Before the OWL is launched, the flight controls are given over to the computer system which relays the series of mission coordinates for each OWL to fly. Simultaneously diagnostics from the OWLs are returned to the computer system and recorded in a database.

3.2.2 Video Surveillance Monitors

The video feedback from the OWLs is relayed to base operations and then displayed on a standard 30” Samsung flat screen monitor. Each signal is displayed on a quarter of the total surface area of the screen in order to capture up to four video relays at one. Figure 2 shows the flow of information and relay of signals between the monitors in the operations base to the OWLs.

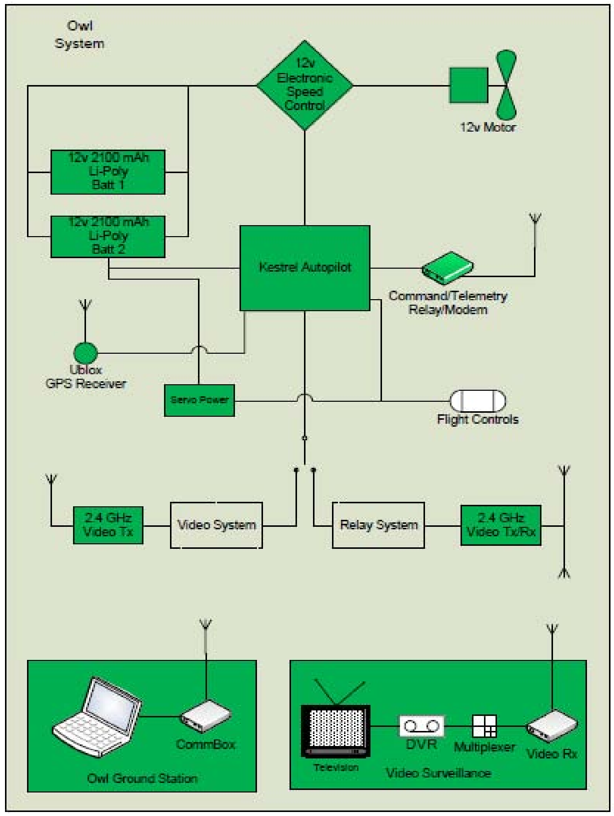

Figure 2: System Dynamics

3.3 Testing

Testing presented a multitude of problems since the entire procedure was created from scratch and had to abide by both the OWL flight regulations and Camp Atterburry safety standards. Therefore, test members, determined the correct UAV launch protocol, testing location, interruption mitigation techniques, and metrics to measure OWL performance prior to any tests.

3.3.1 Preflight Set Up and Diagnostics

Before testing could commence, a preflight checklist and test flight was conducted to ensure safety during the mission. The preflight checklist verified that each OWL was oriented and responding appropriately to the computer software in the operations base. Following completion of the checklist, a manual flight was launched to assess if the platform was responding appropriately to the remote stimulus. After successful completion, the preflight is not conducted again unless any malfunctions or significant breaks occurred during testing.

3.3.2 Testing Scenarios

The testing scenarios are designed in order to observe the added benefit of multiple UAVs operated solely by one person. Therefore, each testing scenario consisted of deploying one, two, or three UAVs to observe a forward location for some duration of time; and measuring the resulting time over target and the value added time for each scenario. The more time over target and total value added time observed indicated there was additional added benefit, to the user, or deploying the corresponding number of OWLs.

3.3.2.1 Time over Target

Time over Target (TOT) is defined as the time an OWL reached the designated marked area until it is instructed to return to the operations base. Transit time is not counted in this metric as the quality and availability of the video feed varied due to weather.

3.3.2.2 Total Value Added Time

During the course of the mission, the operator watches the relayed video feed on the monitor. This is exactly what “Value added time” pertains to; the time the operator spends visually assessing the target. Thus, by stopwatch, the amount of time the operator spent in the control center is recorded during deployment scenario as Total Value Added Time (TVAT) for each test.

3.3.4 Testing Location

Several local locations near Wright Patterson Air Force Base were proposed to test the OWLs for data collection. However, due to DoD regulations, the nearest airstrip cleared for testing was located at Camp Atterburry in Indiana (longitude:086-02’18”, Latitude:39-17’15”) . Located 709 feet above elevation, the airstrip offered ample room for multiple flights up to 739 feet in elevation. Additionally, few flights occupied the airspace which left data collection primarily uninterrupted. The main disadvantage, however, is the 3 hour distance from the camp Atterbury to the nearest parts store in Cincinnati, Ohio. Therefore, careful planning must account for all replacement parts of the OWLs and operation centers.

3.3.5 Testing Issues

Generally the OWLs were allowed to complete all mission without interruption. Occasionally, though, mission essential and commuter aircraft reserved the right to land in the airstrip. To mitigate these interruptions, the operators changed the flight path of the OWLs in order to conserve the current mission without conflicting with the additional aircrafts. Since they were able to preserve the current elevations and total distance the platform flew to the target, no abnormal battery usage occurred. Therefore the validity of the data was preserved and used for the sequential validation and simulation efforts.

3.4 OWL Simulation

Wellbaum (11) created an ARENA simulation used to model time over target and added value time of the operator and the OWLs during various scenarios. The user entered the number of OWLs on the mission and the successive time between launches. The simulation returned the resulting time over target, value added time, repair time, and battery life for the specified duration. The only issue discovered with the simulation was it based all results on an unrealistic battery life distribution (Cottle 2011).

3.4.1 Changes in Battery Life Distribution

Simulation battery distributions differed from operational testing results as they were derived by running the OWLs indoors, mounted on a platform, until the batteries were completely drained. This created problems with comparing the simulation output with the operational output for two main reasons.

First, in the operational environment, there existed extraneous factors, like wind, that caused a non-constant drain on the battery power required to sustain flight. The simulation did not account for these factors which, in turn, rendered inconsistent results compared to observed values.

Second, the mission life was determined based on a distribution that modeled the battery life until failure. This does not consider the amount of power used for transit time to and from the target. Additionally, the batteries drained at a non-constant rate after 10.6

amp-hours remained. Therefore, for the safety of the OWLs, the operator instructed aircraft to base when the battery life dropped below 10.8 amp-hours.

4. Methodology

This thesis defines methods to implement Bayesian statistics to exploit the advantages of simulation data in lieu of operational test data. To accomplish this task, the simulation data must be validated against observed operational test data; otherwise all sequential efforts will be in vain. Following successful validation, the information will be utilized to gain further insight into probability outcomes based on prior information obtained during testing. Finally assessment, analysis of results, and comparison of the results to the operational DOE design is completed to determine the validity of using simulation data in lieu of prior operational test data.

4.1 Simulation Validation

The preliminary step in implementing simulation data in lieu of operational test data is the determining the validity of the simulation output. To accomplish this task, the simulation is replicated and the response output is fit to a distribution. Then, the response expected value is determined along with a ninety percent confidence interval about that mean. Finally, observed test data is compared against the constructed confidence interval to assess compliance of the simulation to operational test data. If enough operational data is collected to determine the result distribution, e.g., mean, and standard deviation of the operational data, then the simulation data can be updated to more precisely model the observed testing data. However, if small data sets interfere with distribution estimation, the simulation can only be "checked" by assessing whether the value of the observational metric falls in a ninety percent confidence interval of its' simulation output counterpart. This latter approach is used in the Chapter 5 example.

4.2 Posterior Predictions

If significant discrepancies occur between the simulation output and the operational data collected, it is highly suspect to deem the simulation validated and assume that the observational data is drawn from the simulation output distributions. However, if the operational data falls within a ninety percent confidence interval of the generated simulation output, the observed data is assumed adequately modeled by the corresponding simulation output distribution. This prior information is used to update predictions on future events using Bayesian probability. Specifically, future outcomes are further scrutinized using previous data observations to enhance the knowledge of obtaining certain events based on the equation

$$\frac{\qquad}{\qquad}\,. \tag{1}$$

In this equation, *X* is the random variable from the simulation output; T is the proposed time threshold of the simulation distribution; *Y* is the observed random variable assumed from the same distribution as X; and *t* is the observational recorded time. This posterior knowledge should not only increase confidence in obtaining various TOT and TVAT thresholds, but add additional information to design of experiments matrices. The Chapter 5 example demonstrates the use of prior information, such as from a simulation, updated and using real test data. Interpretation of the posterior information is provided.

4.3 DOE Analysis

The validated simulation data is also used to determine changes in critical factors. Again, this procedure should only be used for a validated simulation since invalid simulation output cannot be modeled correctly to account for operational data. This fact can also be complicate by the sparse data collected which limits the approximation of determining a distribution to fit the operational data. For the valid simulation data, the mean TVAT and TOT times are substituted into the real test response matrix, initially one metric at a time. Then combinations of mean TVAT and TOT values are swapped into the DOE matrix and analyzed until the matrix is composed strictly of validated simulation data respectively. Analysis of the results indicates the impact of utilizing data from a validated simulation in lieu of operational test data.

5. Results and Analysis

The previous chapters highlight the methodology and reasoning behind the findings in this chapter. This chapter presents a preliminary case study using the very limited OWL data available. The first step in evaluating the methodology proposed above is validating the simulation output since both the integrity of both posterior predictions and DOE analysis depend on the results. Then, given correct application of the validation technique, Bayesian statistics is applied to gain more information on posterior predictions. In turn, this should increase user confidence in obtaining TOT and TVAT objectives which can be utilized via DOE to gain more insightful information about OWL characteristics. Finally, validated simulation data is substituted into a simple 3^1 DOE model to demonstrate the effectiveness of valid simulation data in lieu of operational data. The results should show no significant difference between simulated data and operational data, or change in critical factors between the original DOE matrix and the augmented matrix.

5.1 Data Validation

The simulation was validated in two increments, (Wellbaum et al. 2010) and (Cottle 2011), and the results showed the simulation data to be representative of operational data observed from preliminary OWL testing. Therefore, in this instance, one should not expect any significant difference between the operational data and the simulation data that would indicate the simulation was an invalid representation of the OWL tests. However, one cannot simply assume the OWL simulation is valid since the sequential effort's results depend on the accuracy of the simulations output to the operational data. Therefore, the OWL simulation is validated for compliance with new operational test findings below.

The simulation ran for one hundred iterations for delay between launch settings of 5, 20, and 30 minutes using two OWLs. The total time over target, TOT, and total value added time, TVAT, output was analyzed in jmp version 8 to determine the output distributional characteristics. In each test case, there was insufficient evidence to reject the null hypothesis and conclude that the data was not drawn from a Weibull distribution (shown in figures below). This was based on a large value of .25 which exceeded the alpha critical value of .05. Therefore, ninety percent confidence intervals and expected value estimates were calculated for both metrics, TVAT and TOT, on each test. Based on the results below, the TOT and TVAT from test one, and TVAT from test three did not fit into the corresponding confidence intervals (highlighted in red). In fact, the observational data points, for test one, fell so unrealistically far outside the confidence intervals that there is no reason to accept that the simulation data is a valid representation of its operational counterpart. However, the test three TVAT metric is substantially close to the lower bound of the ninety percent confidence interval. Since ten percent of the data is expected fall outside the interval, there is insufficient evidence to reject that this metric does not come from the proposed Weibull distribution. Therefore, although a discrepancy exists, the TVAT value from the operational test three was included for further analysis unlike the test one values which showed an enormous conflict with the simulated data distributions.

These conflicts may have occurred for several reasons. First, the simulation is assumed validated against the operational activities. If any part of the simulation does not capture the true nature of the OWL, and its tasks, then the simulation will produce data inconsistent with the operational outcome. Second, although test one went very smoothly, the simulation may not account for the problems that can occur during testing like dangerous wind velocities, or interruptions during testing. Lastly, fitting a distribution to a single data point is impossible. If the simulation is correct, and that single data point was recorded in error or occurred from an unlikely series of events, the simulation data will still be considered invalid.

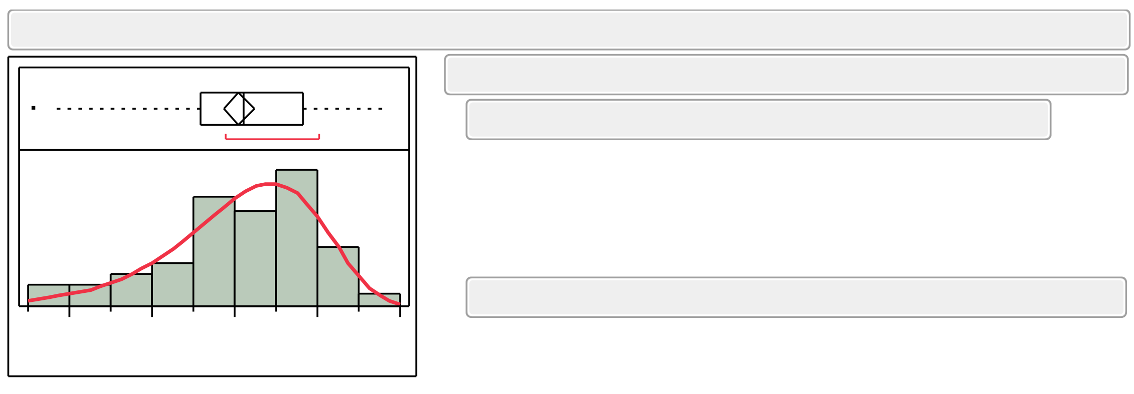

Figure 3: 5 Minute Delay TVAT Distribution Estimate

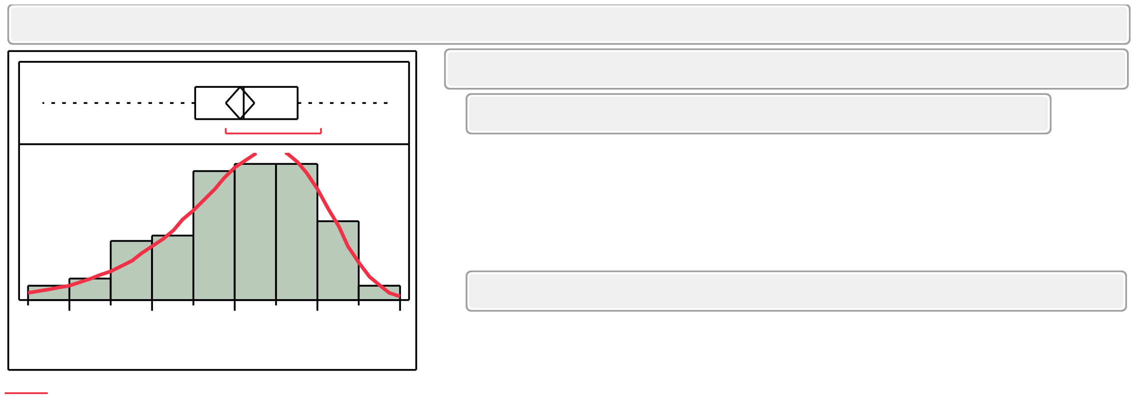

Figure 4: 20 Minute Delay TVAT Distribution Estimate

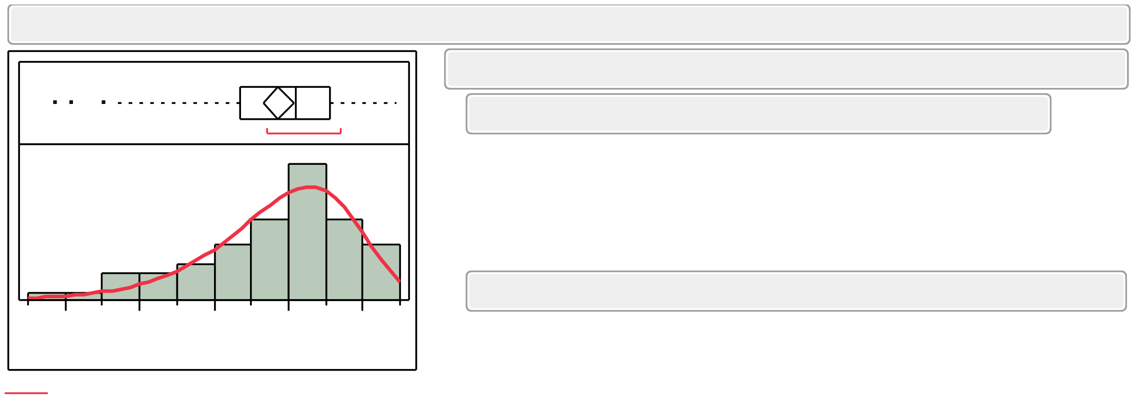

Figure 5: 30 Minute Delay TVAT Distribution Estimate

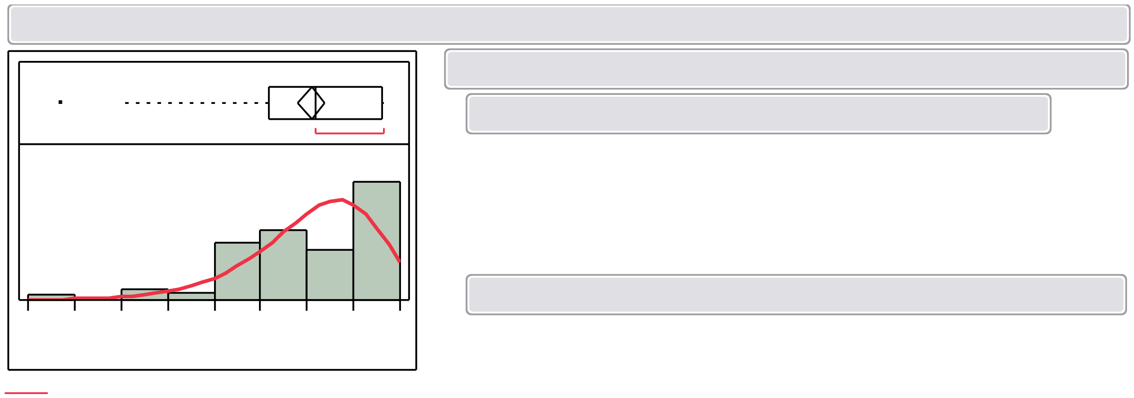

Figure 6: 5 Minute Delay TOT Distribution Estimate

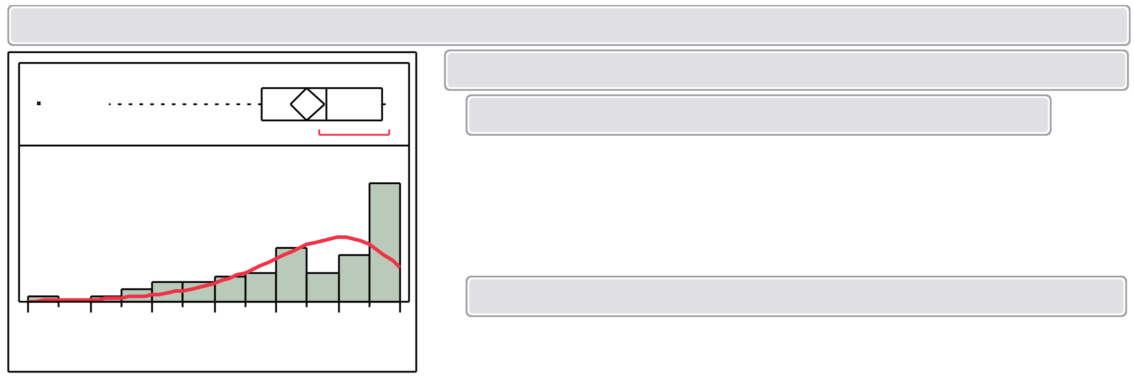

Figure 7: 20 Minute Delay TOT Distribution Estimate

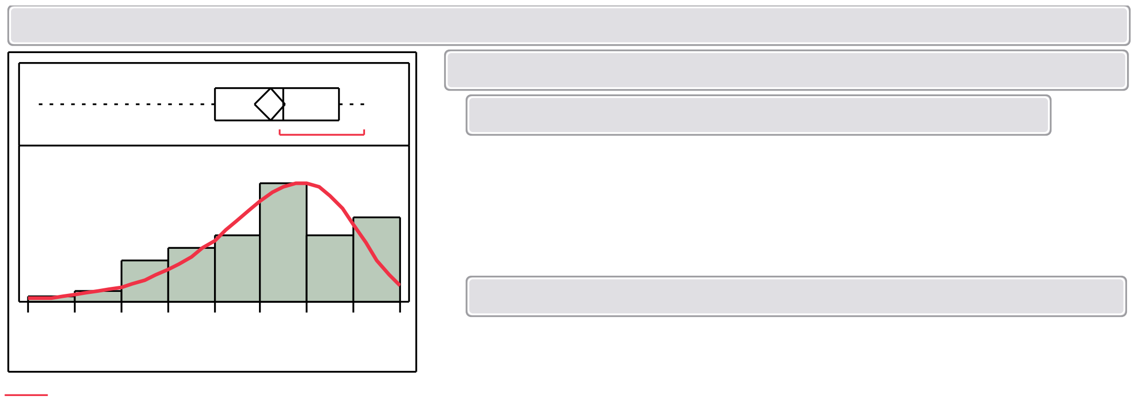

Figure 8: 30 Minute Delay TOT Distribution Estimate

Table 1: TOT & TVAT Comparison of Operational and Simulation Data

Test Number	Delay Time	Metric	Lower Bound	Upper Bound	Mean	Observed Value
1	5 Minute Delay	TVAT	86.521	101.68	95.26574	69.35
1	5 Minute Delay	TOT	103.006	124.96	115.6092	84.24
2	20 Minute Delay	TVAT	96.43	111.7	105.245	109.5
2	20 Minute Delay	TOT	116.59	139.51	127.7268	128.39
3	30 Minute Delay	TVAT	105.21	121.04	114.3589	104.58
3	30 Minute Delay	TOT	121.25	147.16	136.1308	129.49

5.2 Posterior Prediction Estimates

Since four of the six metrics in the previous section are assumed to come from their corresponding identified distributions, additional insight can be gained with respect to probability outcomes. One expects the chances of obtaining certain TVAT and TOT thresholds to increase or decrease depending on the location of the observed value with respect to the mean of the corresponding distribution. In any case, the updated probability outcomes should be more informative for each threshold identified below when compared to the prior probabilities. Therefore, one expects to observe a change in the posterior probabilities when compared to the prior probabilities which would indicate a benefit from prior knowledge with respect to probability outcomes.

With a validated simulation observational data may used to predict posterior TOT and TVAT probability outcomes. Subsequent posterior TVAT and TOT probabilities are compared to prior probabilities of TOT and TVAT exceeding a certain time using the Bayesian equation listed above. This result showed the probability of the OWLs yielding a TVAT and TOT of a certain number of minutes listed in the chart below. The results, highlighted in green, show an increased probability in obtaining a certain threshold given an operational time was observed, in every case except the TVAT metric in test three. Note that even intervals were not used across each test measure in order to show the impact of additional information across each differently defined simulation distribution. Furthermore, although included to indicate the significance of prior information, test one metrics cannot be considered valid.

Table 2: TVAT & TOT Prior & Posterior Probability Comparison

Test	Delay Time	Measurement	T	Prior Probability	Posterior Probability	Change
1	5 Minute Launch Delay	Total Value Added Time	70.0000	0.99975	0.99995	0.00019
1	5 Minute Launch Delay	Total Value Added Time	80.0000	0.99290	0.99310	0.00019
1	5 Minute Launch Delay	Total Value Added Time	90.0000	0.87074	0.87091	0.00017
1	5 Minute Launch Delay	Total Value Added Time	100.0000	0.13987	0.13989	0.00003
2	20 Minute Launch Delay	Total Value Added Time	100.0000	0.86923	1.00000	0.13077
2	20 Minute Launch Delay	Total Value Added Time	110.0000	0.14107	0.79298	0.65191
2	20 Minute Launch Delay	Total Value Added Time	115.0000	0.00123	0.00692	0.00569
2	20 Minute Launch Delay	Total Value Added Time	120.0000	0.00000	0.00000	0.00000
3	30 Minute Launch Delay	Total Value Added Time	90.0000	0.99945	1.00000	0.00055
3	30 Minute Launch Delay	Total Value Added Time	105.0000	0.95272	0.99470	0.04198
3	30 Minute Launch Delay	Total Value Added Time	120.0000	0.09721	0.10149	0.00428
3	30 Minute Launch Delay	Total Value Added Time	135.0000	0.00000	0.00000	0.00000
1	5 Minute Launch Delay	Total Time over Target	90.0000	0.99701	1.00000	0.00299
1	5 Minute Launch Delay	Total Time over Target	105.0000	0.92601	0.92670	0.00069
1	5 Minute Launch Delay	Total Time over Target	120.0000	0.27861	0.27882	0.00021
1	5 Minute Launch Delay	Total Time over Target	135.0000	0.00000	0.00000	0.00000
2	20 Minute Launch Delay	Total Time over Target	115.0000	0.96498	1.00000	0.03502
2	20 Minute Launch Delay	Total Time over Target	129.0000	0.47190	0.91508	0.44318
2	20 Minute Launch Delay	Total Time over Target	135.0000	0.08138	0.15780	0.07643
2	20 Minute Launch Delay	Total Time over Target	140.0000	0.00138	0.00268	0.00130
3	30 Minute Launch Delay	Total Time over Target	120.0000	0.95960	1.00000	0.04040
3	30 Minute Launch Delay	Total Time over Target	130.0000	0.80127	0.98259	0.18133
3	30 Minute Launch Delay	Total Time over Target	140.0000	0.34965	0.42878	0.07913
3	30 Minute Launch Delay	Total Time over Target	150.0000	0.01138	0.01395	0.00257

(T is in minutes)

5.3 Implementation of Design of Experiments

Since four of the six metrics were determined as representative of the operational test data, they can be utilized in future DOE-based analysis. Stated simply, comparing the test matrix composed solely of operational data to the matrices augmented with simulation data shows the impact of simulation data in DOE. Additionally, since the data is validated, there is no reason to suspect a change in identified critical factors. This indicates that simulation data can be used in lieu of operational data, for non critical factors, in DOE.

The mean of each simulation output, described in section 5.1, was substituted into a simple 3^1 DOE model consisting of single and all combinations of valid simulation means for the corresponding operational response variables. The TVAT and TOT simulation data from test one were excluded from this analysis primarily because they are sure to change the characteristics of the factors in a design of experiments model. The results displayed below, for both TVAT and TOT models, show overlapping of confidence intervals between the original TVAT and TOT models and their simulation data counter parts. Further analysis shows there is quite a vast overlapping consistency across TOT and TVAT models. Therefore, there is insufficient evidence to conclude that swapping means of valid simulation data, into a DOE model, will change the outcome of the factors for a DOE model. Hence, there is evidence that valid simulation data can be used in lieu of operational data without jeopardizing the quality of the DOE analysis outcomes.

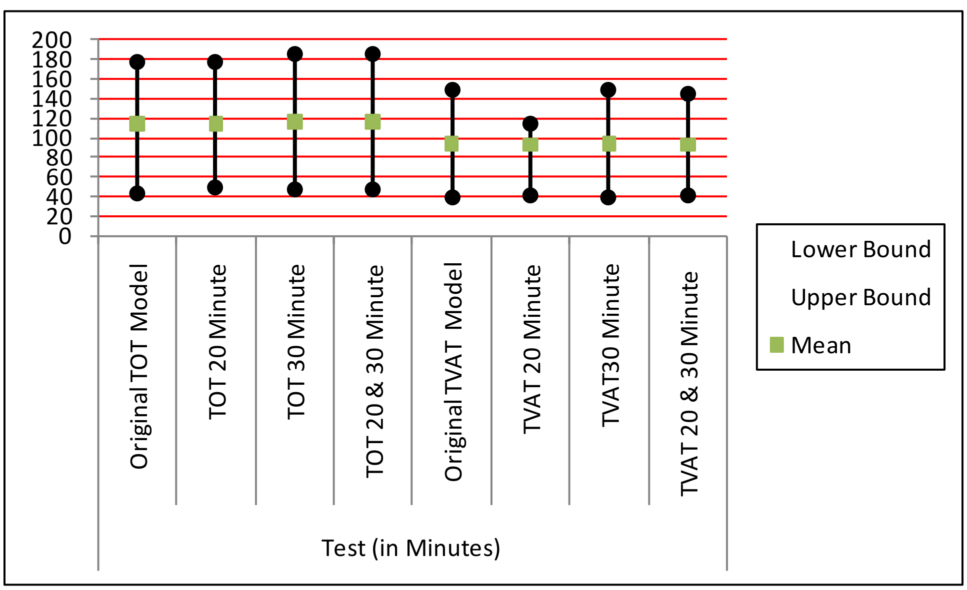

Figure 9: 95% DOE Confidence Interval Comparison

6. Future Recommendations

Based on the results above, there exists evidence supporting the use of valid simulation output and prior operational output to predict posterior probabilities and aide in DOE analysis. However, simulation is not the only operations research specialty area that can be applied to UAV testing. Future efforts should be geared toward all focus areas of operation research. Specifically, future efforts should incorporate decision analysis, optimization via linear programming, optimization via simulation, and design of experiments focused on enhancement of OWL performance and functions. Only through the combination of all these concentrations simultaneously can the full operational potential of the OWL be determined.

6.1 Decision Analysis

The systems engineering department of the Air Force Institute of Technology was interested solely in maximizing value added time and total time over target. However, there was very little research performed to answer the age-old dilemma of “ability” versus “need”. Just because you can obtain a certain degree of a metric does not mean there is any added benefit past a certain point. Therefore, a decision analysis study should be performed to determine if maximizing those metrics yields the most benefit to the operator or if there are additional metrics of interest. One may find that the operator is actually interested in other important metrics that were overlooked in the early stages of test planning. Future efforts can utilize value focused thinking, or even expected utility, to establish, quantify, and measure the current needs of UAV operators in the AOR.

Forming this preliminary foundation will yield a new set of ranked preferences, goals, and cost analysis that will guide future OWL research.

6.2 Linear Programming Optimization

Following establishment of user goals, additional optimization techniques should be performed to analyze the various numbers of users, OWLs, and OWL components to achieve desired thresholds for a various number of targets while considering budget and resource constraints. One way to accomplish this task is through linear programming (LP). Following the identification of system measurements, goal programming along with other LP techniques can be utilized to optimize the OWLs performance in accordance with strategic goals. This would lead to not only a leaner system, but possibly several optimal scenarios that would increase flexibility in the protocol for OWL deployments.

6.3 Simulation Optimization

After preliminary goals and metrics have been established, simulation can be employed in a different context than in this work. Specifically, simulation should be applied to predict how future changes in OWL deployment scenarios would affect the accomplishment of the mission. Manipulating the number of OWLs, number of users, the flying altitudes, battery types, launch times, and the camera types should yield different optimal outcomes of interest to the mission. However, the current simulation must be incrementally validated for future research, giving a simulation thesis more of a twofold purpose.

6.4 Small Data Set DOE

This thesis sought to utilize DOE and simulation to predict the impact simulation can have on testing and evaluation. However, several interruptions, uncooperative weather, and contracting issues handicapped the size of the operational data set collected. Therefore, design of experiments should be applied to the testing of the OWL with a goal to minimize testing while maximizing the use of quality data. Through smaller yet more informative tests, critical factors can be identified and further explored where bigger test have failed due to lack of data. This application will yield a plethora of information on which test avenues should be explored to utilize the simulation procedure listed in the methodology. Furthermore, future DOE testing should incorporate more than just two variables. Before any testing commences, the test committee should consult systems engineering documents to determine which components are tied to functions that may cause changes in OWL performance. Identifying these function influencing components should lay the ground work for a complete DOE map of factors to explore. In turn, the test design will be geared toward minimal data collection with the intent of maximizing benefit from data, which will be beneficial considering how volatile OWL data collection has been.

6.5 Summary of Future Work

In the past several years, a lot of work has been accomplished on various aspects of the OWL platform. However, as mentioned above, the accomplishment of the OWL mission can be scrutinized through various operations research techniques which have not been applied to date. Through the application of simulation, decision analysis, liner programming, and design of experiments the full potential of the platform can be achieved. This, in turn, should influence improvements and processes on the OWL platforms currently in the AOR to increase mission effectiveness.

Appendix A: Blue Dart

Test and evaluation (T&E) is costly to the DOD and the United States Air Force. New, innovative uses of simulation technology have emerged as a partial solution to the challenges facing T&E. This research develops and discusses a methodology to utilize minimal data sets augmented with simulation, Bayesian analysis, and design of experiments, to reduce the level of live testing required. A small fairly notional data set is used to discuss the methodology.

Validated simulations are crucial if simulation hopes to augment T&E. This research discusses some simulation practices and how T&E data can be exploited to validate simulation models.

While Design of Experiments (DOE) has been underutilized in the past for T&E, recent policy changes require its use. This work takes a preliminary look at how simulation can affect a test design both in terms of providing prior evidence of system performance and in replacing components of the actual test.

T&E practices need to evolve to meet current DOD fiscal budget restraints. Simulation, coupled with statistical techniques, offer a viable solution method to help achieve DOD T&E goals.

Bibliography

1. Andradottir, S. a. (2000). Applying Bayesian Idea in Simulation. *Simulation Theory and Practice 8* , *8* (3-4), 253-280.

2. Banks, J. J. (2010). *Discrete-Event System Simulation.* Upper Saddle River, New Jersey: Prentice Hall.

3. Berger, J. (2006). The Case for Objective Bayesian Analysis. *Royal Statistics Society* , 385-402.

4. Blau, G. a. (2009). A Beyesian Pharmacometric Approach For Personalized Medicine-A Proof of Concept Study With Simulated Data. *2009 Winter Simulation Conference*, (pp. 1969-1976). Austin, TX.

5. Bohk, C. a. (2009). Probabilistic Populatin Projection With James II. *2009 Winter Simulation Conference*, (pp. 2008-2019). Austin, TX.

6. Celik, N. a.-J. (2009). Sequential Monte Carlo-Based Fidelity Selection In Dynamic-Data-Driven Adaptive Multi-Scale Simulations (DDAMS). *Winter Simulation Conference*, (pp. 2281-2293). Austin, TX.

7. Cottle, A. (2011). *Operational Validation Of A UAV Mission Optimizatoin Model.* Wright Patterson Air Force Base: AFIT.

8. Li, P. a. (2009). A Markov Process Based Dilemma Zone Protection Algoritym. *2009 Winter Simulation Conference*, (pp. 2436-2445). Austin, TX.

9. Palagummi, V. a. (2009). Technuques For Rapid Initialization In In-Vehicle Traffic Simulators. *2009 Winter Simulation Conference*, (pp. 2446-2456). Austin, TX.

10. Steffens, T. a. (2007). Real-Time Prediction In A Stochastic Domain Via Similarity-Based Data-Mining. *Winter Simulation Conference,* (pp. 1430-1435). Washington, D.C.

11. Wellbaum, C. M. (2010). *System Analysis and Prototyping For Single Operator Management of Multiple Unmanned Aerial Vehicles Operating Beyone Line of Sight.* Wright Patterson Air Force Base, OH: AFIT.

Vita

Capt Jonathon S. Hosket graduated in 2001 from the International School of the Americas in San Antonio Texas. Although nominated to the United States Air Force Academy, he enrolled at Baylor University on a full ride scholarship. After achieving an undergraduate degree in applied mathematics, he was commissioned through Detachment 810 with a Regular Commission in 2006.

His first assignment was at Wright Patterson AFB as a Scientific Research Analyst for the Warfighter Interface Visualization branch of the Human Effectiveness directorate. There, he improved on night visualization capabilities for the ANVIS-14s and the multispectral monocular thermal visual system used by the ARMY. Shortly following the two years spent in this branch he was promoted to a division level management position.

During his second assignment as Chief of Division Operations, Capt Hosket aided in the evaluation of division programs and annual division expenditures. He was also responsible for the establishment of hazardous waste and material management protocol and contributed to advancements of cognitive warfighter improvements in the cockpit. In August of 2010, he entered the Graduate School of Engineering and Management, Air Force Institute of Technology. There he specialized in statistics and decision analysis applications and is projected to graduate in May 24, 2010. Upon achievement of his masters' degree in Operations Research, he will be assigned to the Studies Analysis Squadron (AETC) at Randolph Air Force Base.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
04-09-2011	**Graduate Research Project**	Sept 2009 – Mar 2011

4. TITLE AND SUBTITLE

A METHODOLOGY USING SIMULATION RESULTS FOR TEST AND EVALUATION

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Hosket, Jonathon, S., Captain, 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, Building 642
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-OR-MS-11-09

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

AFIT/ENS
Attn: Dr. Raymond Hill
2905 Hobson Street, Bld 642 DSN: 255-6565
WPAFB OH 45433-7765 e-mail: Raymond.hill@afit.edu

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Each year the Air Force spends billions of dollars on Test and Evaluation to ensure acquisition programs roll out the best possible products. In 1997, the National Research Council assembled to evaluate the overall procedure used in procuring various platforms with system planning, research, development and engineering (SPRDE) and program management (PM) processes. In their final report, they claimed that the full advantages of statistical practices, simulation, model-test-models, and incorporation of prior test information into current test practices have not been fully utilized. To examine one of the report's recommendations, this thesis defines and explores a methodology using simulation to augment or replace test data in lieu of operational testing. Specifically, a validated simulation model employs non-critical factor data from preliminary small sample operational testing. The simulation then generates posterior distribution data to replace the corresponding data in the final test matrix. If useful, data generated by a validated simulation model can be used in lieu of actual operational test data for selected non-critical factors. This provides T&E squadrons a means to decrease the level of live operational testing on non-critical factors. Therefore, T&E can be more efficient as less runs are needed to evaluate system factors of interest. This thesis defines methods to use test data to validate simulation results, us simulation data as evidence for subsequent operational testing, and use simulation to potentially replace test data.

15. SUBJECT TERMS

Design of Experiments, Bayesian Analysis, Simulation

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 Hill, PhD. (ENS)
					19b. TELEPHONE NUMBER *(Include area code)*
U	**U**	**U**	**UU**	56	(937) 255-6565, ext 4314; e-mail: Stephen.Chambal@afit.edu