# Better Designs for High-Dimensional Explorations of Distillations

Dr. Tom Lucas, Naval Postgraduate School
Dr. Susan M. Sanchez, Naval Postgraduate School
Major Lloyd Brown, United States Marine Corps
Major William Vinyard, United States Marine Corps

*It is the mark of an educated mind to rest satisfied with the degree of precision which the nature of the subject admits and not to seek exactness where only an approximation is possible—Aristotle*

## Introduction

Should we attack now or wait three hours for an expected reconnaissance report? If we attack now, how many people might we lose? How should we use information technologies to transform our forces? What mix of heavy and light weapons will be most survivable in 2020? When should command and control be centralized, and when should it be decentralized? These are but a few of the many critical questions that face our military leaders. The massive amount of uncertainty and the dearth of data usually associated with these questions make them difficult to answer. Consequently, decision-makers frequently turn to analysts for information to assist them in making effective choices.

Analysts, in turn, regularly rely on various types of models to examine complex military issues. Approaches include equations, massive simulations, wargames, and (more recently) *distillations*. Distillations are relatively simple simulations that attempt to capture only the salient features of the situation without trying to model all of the details that could be considered. All of these tools have substantial strengths, but also serious shortcomings. Project

Albert focuses on *Operational Synthesis*—that is, the process of combining the information gleaned from a family of diverse analytical tools to provide the most compelling analyses [1]. Much of Project Albert's efforts have involved the building of distillations, along with data farming and visualization environments in which they can be explored. Our focus is on how best to broadly search distillations—varying many more factors simultaneously than in previous approaches—in order to obtain useful information.

This paper discusses some of the challenges of gleaning information from high-dimensional models—with an emphasis on distillations of warfare. We introduce an expert-driven framework that automatically looks across a breadth of factors and adaptively focuses on the significant effects and interactions. We describe a variety of designs that may be incorporated into our overall framework, empirically characterize aspects of their performance on known response surfaces, and illustrate the potential utility via explorations on two distillations. The first exploration examines the relationship in Irreducible Semi-Autonomous Adaptive Combat (ISAAC) between squad-level intangibles and the squad's ability to reach an extraction point, while the second searches the Dewar model for non-linearities and remedies to the resultant non-monotonicities.

## Getting Useful Information From Simulations Of Warfare

*All models are wrong, but some are useful—George Box [2]*

With modern computers, it is feasible to run millions of computational experiments on distillations. Before deciding which cases to run, however, we have to think about what types of information we are trying to extract from our models. In their classic paper, "Design and Analysis of Computer Experiments," Sacks et al. [3] wrote that the three primary objectives of computer experiments are:

(i) *Predicting* the response at untried inputs,

(ii) *Optimizing* a function of the input factors, or

(iii) *Tuning [or calibrating]* the computer code to physical data.

While this may be true in many fields, how well do these three primary objectives for computer experiments apply to models of combat?

According to the dictionary [4], to predict means to "declare in advance." We can certainly run models, get numbers, and make predictions about potential future events. However, in order for our predictions to be scientifically better than those of an astrologer (who would be much faster and cheaper), we must be able to provide a credible warranty on the accuracy of our predictions. That is, we should be able to scientifically show why the decision-maker should believe our predictions over those of an astrologer. Unfortunately, this is not always easy to do with models of combat. To begin with, the data we need in order to assess the accuracy of our predictions are lacking. Furthermore, many of the factors that can be decisive in combat (e.g., morale) are difficult to know in advance—as witnessed by the consensus predictions of many thousands of allied casualties and weeks of ground conflict that were made just prior to Desert Storm.

Military analysts often strive to optimize or improve performance. For example, we might want to find the equipment, doctrine, and/or organizational structures that minimize either the expected number of casualties or the expected time to complete a mission. Using a particular model and scenario—once a set of inputs that describe the threat, time, and place of a potential battle have been specified—we may be able to find model settings that optimize some particular performance measure. Of course, since the veracity of the simulation may be uncertain, any optimum must be viewed with skepticism. Moreover, the optimum is likely based on assumptions regarding a particular opponent and scenario: future

alliances; how well the various combatants fight; what equipment they use; the reliability and effectiveness of their equipment; when and where the battle takes place; the weather; etc. This means that the so-called optimum (if it can be found) is almost certainly conditioned on many uncertain factors, while the likelihood of them all occurring is negligible. Thus, in military analyses, we are often more interested in robust solutions than in optimal ones. That is, we want equipment and tactics, for example, which work well across a broad range of plausible scenarios and conditions.

The last major objective of computer experiments identified above is calibration. Calibrating our models as data become available can be a laudable goal. However, the dearth of such data, particularly at the force-on-force level, makes quantifying intangibles such as morale, trust and leadership problematic. Moreover, calibration is a means to make our models better, not an analysis end in and of itself. At times, adding more detail to a model makes it more cumbersome and brittle, hence less likely to provide useful information in a timely manner.

Given these limitations, we cannot credibly predict, optimize, or calibrate performance for numerous military applications. How, then, should we be using these models? Major General Jasper Welch's [5] guidance is: "A model is useful if a better decision is made with the information it adds." Captain Wayne Hughes [6] states that two "primary benefits of model-based studies are to (1) help explore the issues in a structured way and organize a debate, and (2) uncover new insights and reveal surprising characteristics." Following this advice, we want our designs to help us structure and organize debates, efficiently uncover new insights, and effectively communicate our findings to decision-makers. That is, we are trying to help people think through complicated issues by illuminating the consequences of various assumptions, reinforcing or challenging intuition, and illustrating alternatives that they might otherwise not have considered. Thus, rather than optimize or predict, we seek to: (1) identify significant factors and

interactions, and (2) find regions, ranges, and thresholds where interesting things happen.

## Our Solution

*The purpose of computing is insight, not numbers—Hamming [7]*

The goals just stated proscribe the development of an efficient experimental tool kit and set of search strategies for exploring distillations. This is challenging because the number of runs required to comprehensively explore even the simplest distillation can be unmanageable. This is well illustrated by General Welch's [8] statement that "$10^{30}$ is forever." In other words, to fully evaluate all of the combinations of a model containing 100 factors, each with only two settings, $2^{100}$ [ $10^{30}$] runs of the model are necessary. Using a computer that can evaluate a model run in a nanosecond, an analyst who started making runs at the (then estimated) dawn of the universe would just be finishing his runs—hence it would have taken him forever to explore the model. Unfortunately, most of our models have more than 100 factors, many of which are continuous or can take on a large number of discrete values. Analysis is further complicated by the uncertainty corresponding to many (if not most) of the factors. Therefore, even with super computers and "simple" models, we typically cannot use brute force searches on more than about five to ten factors. Moore's Law suggests that we will be able to extend this only by about two factors (through an increase of two orders of magnitude in processing power) each decade. Thus, if we want computational experiments that look broadly across these models, we need better designs.

How do we select the best set of experiments from the vast ensemble of possibilities? The extensive body of literature on designing experiments indicates that many of the existing experimental designs have their roots in agriculture and laboratory experiments. That is, they were developed for situations with a

relatively small number of experimental units (e.g., plots of land, patients, widgets) on which experiments could be conducted. Consequently, there are not many readily available tools for high-dimensional explorations where we can take millions of runs—as is frequently the case with computational experiments. Furthermore, most of the existing designs also assume many of the following: linear effects, sparse effects, negligible higher order interactions, homogeneous Gaussian error, and a univariate response. Experience suggests that these are risky assumptions to make with models of combat.

Our approach to building search strategies is driven by the following principles:

- The design must leverage human expertise. In many cases, substantial knowledge exists, and can be harvested, about both the subject area and the model. We want to use this knowledge, but not be bound by it.
- Most searches should utilize multi-resolution designs. That is, we need different levels of information from different factors (or combinations of factors) in the model—and our designs should reflect this. Moreover, how factors affect the model's response will almost certainly be highly variable.
- The design should be sequential and adaptive. As we gain information about the model's surface, we want our search to automatically adjust the sampling scheme to focus on the regions and/or factors that seem more interesting. Furthermore, in defense analysis, we may not know *a priori* when the analysis will be due, or the original due date may change abruptly. Thus, we want designs that can provide useful information even if they cannot be run to completion.

The appropriate design depends on both the type of information needed and the nature of the model's response surfaces (or

landscapes). In general, for exploring distillations, we want designs that can look at a large number of factors, isolate interactions, identify non-linearities, and find thresholds where responses change dramatically. To accomplish this, we are devising strategies that combine some well-known designs with some newer ones. In particular, we are looking at search strategies that use adaptive mixtures of full-factorial (or grid), fractional-factorial, group screening, random perturbations, Latin hypercube (and some variants), and frequency-based designs. The following list briefly summarizes situations in which the various component designs are appropriate.

**Full-Factorial**: These gridded designs are particularly useful for looking at a modest number of factors at not too many levels. They can be good at finding higher-order non-linearities and interactions. They also provide data in a format that can be used by many visualization tools [9]. Factorial designs can have coarse grids or fine grids, depending on the numbers of levels for the factors.

**Fractional Factorial**: These designs are efficient ways of examining more factors than with full-factorial designs. The cost is that we are unable to estimate some or all higher-order interactions.

**Latin Hypercube**: These excellent space-filling designs are an efficient means to simultaneously look at many factors (20 or more) when we are concerned about possible extreme non-linearity.

**Frequency-based Designs**: These designs allow us to oscillate the levels of multiple factors across runs. They may be especially useful when we may not know the stopping time of the experiments *a priori*. The resolution for each factor depends on its unique oscillation frequency.

**Random Perturbations**: These low-resolution designs are an efficient means of determining the relative sensitivity of the surface to a large number of factors without directly assessing individual factor effects.

**Group Screening**: These low-resolution designs are an excellent means of screening a large number of factors (even a hundred or more) when we anticipate that only a few may be significant. These designs can also be imbedded within the other designs, dramatically increasing the number of factors that can be explored within one phase of experimentation.

Figure 1 illustrates the capability we are building towards. For the purposes of this paper, we consider only the distillation portion of the broader Operational Synthesis structure. Here, senior and military decision-makers work with analysts to frame questions and state hypotheses about which they would like to obtain information. Where appropriate, distillation models and scenarios are selected for investigation; potential factors and preliminary levels over which they will be varied are also identified. Expert judgment is used to partition these factors into classes based on the type of information we wish to extract from them and *a priori* beliefs about how they will affect model outcomes. This subject matter and model expertise is typically easy to elicit. The partitioning is done within the constraints imposed by processing limitations, which usually prohibit a brute-force, high-resolution design for all factors. Based on this partitioning, a series of computational experiments are run using a combination of the designs, and variants thereof.

Of course, anyone who has worked with models of warfare knows that they are constantly finding unexpected results. Thus, as we learn about the model from our initial experiments, the additional information that we want to extract from the various factors may change. In particular, some factors may be more interesting than initially anticipated. Our framework accommodates this by moving them into classes that are sampled more extensively.

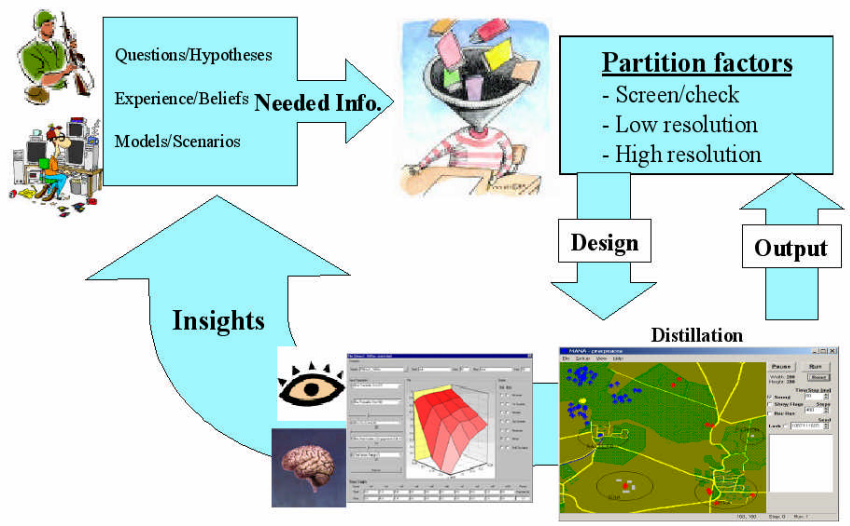


**Figure 1: Our envisioned process uses combinations of statistical designs to help provide insights into military questions.**

Conversely, other factors that we thought might be important will not have a significant effect on the response. Consequently, we will move these factors into lower resolution classes. This process continues until enough information is available or the time runs out. In the end, we can estimate effects and interactions, identify extreme points, and find thresholds. More importantly, however, is that from within the incomprehensively vast high-dimensional model space, we identify interesting sub-regions for humans to investigate further, likely with visualization tools that are particularly effective in not too many dimensions [9]. For example, from among scores of factors, we might identify the half dozen or so that have the greatest impact on responses. The analyst and decision-maker can then explore them in detail with visualization tools.

## A Tool To Sow The Seeds For Successful Data Farming

*"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay."—Sherlock Holmes*

A version of this search tool readily usable by Project Albert and other model explorers is still under development. While we have a general framework, considerable work must be done on what search designs work best, and when, with the various distillations. Those findings will then need to be implemented in a form readily accessible to analysts and decision-makers. Towards that end, we are empirically studying the effectiveness of the various designs along two axes. First, we are assessing the designs' ability to detect effects on known response surfaces. When we generate the surface, we know the underlying truth; thus, we can determine not only what our explorations find, but also what they miss. We are doing this on surfaces with features that we expect to find in distillations—that is, high-dimensional surfaces with many factors, nonlinear responses, rich or sparse effects, significant high-order interactions, and complex error structures. Second, we are studying the effectiveness of portions of the designs on prototype distillations.

In this section, we illustrate the type of experiments we are running on known surfaces and summarize our preliminary findings. We defer our findings on the explorations of distillations until the following section.

One approach that plays an important role in our developing framework is the class of Latin Hypercube (LH) designs. McKay et al. [10] developed these constrained, random designs specifically for computational experiments. The basic LH design works as follows. Suppose that we have $k$ continuous input factors $X_i$ that we want to search with $n$ samples (or input combinations). For each of the $k$ factors, a probability density function $f_i$ is chosen, and the range of the factor is divided into $n$ equal probability segments. Within each of the probability segments, an input point

is selected—we usually select the median. For each of the *n* input combinations, for every factor, $X_i$, independent of the other factors, an input point is selected by random sampling without replacement. This ensures that the entire range of the factor is explored proportional to the weight we assign to it. For our explorations, we usually use Uniform $f_i$. A Uniform distribution has the greatest entropy of all distributions whose range is restricted to a finite interval. A Uniform distribution also minimizes the largest gap between each factor's input values. Thus, it is good at catching sharp spikes in a surface.

How can we best utilize LH designs in our searches? For example, how many samples are needed to detect various effects? To answer these questions, we are running LH experiments and cataloguing the results on a variety of known surfaces. That is, the underlying truth is known, though random error is added to our simulated responses. Thus, we can evaluate how often the design correctly (or incorrectly) identifies the significant effects and interactions. Figure 2 displays an illustrative example of one such experiment. Here, the underlying surface is $E[Y_i] = 20 + 2x_1 + 3x_1^2 + x_2 + .5x_3 - 3x_4^2 + x_8x_9 - x_1x_2$, with the inputs' domains all being in the interval [–1.0, 1.0]. However, the responses are corrupted by random Gaussian error (with mean zero and standard deviation one). We want to find the probability that an LH design will detect each of seven different effects and interactions in a nine-dimensional space as a function of the sample size (*n*).

In these experiments, quadratic regression is used, and factors with p-values of less than .05 are counted as detections. We can see that, in this particular situation, we need about 60 samples to have any chance at all of detecting any of the effects.

From this, we can infer that for nonlinear surfaces with interactions, using quadratic regression to identify effects, LH designs should have more samples than the combined number of possible linear, quadratic, and interaction effects. After exceeding

this threshold, strong effects are easily identified with a few more samples. For weaker effects, the probability of detection increases steadily with the sample size. With a large number of samples, even small effects can be found with high probability.

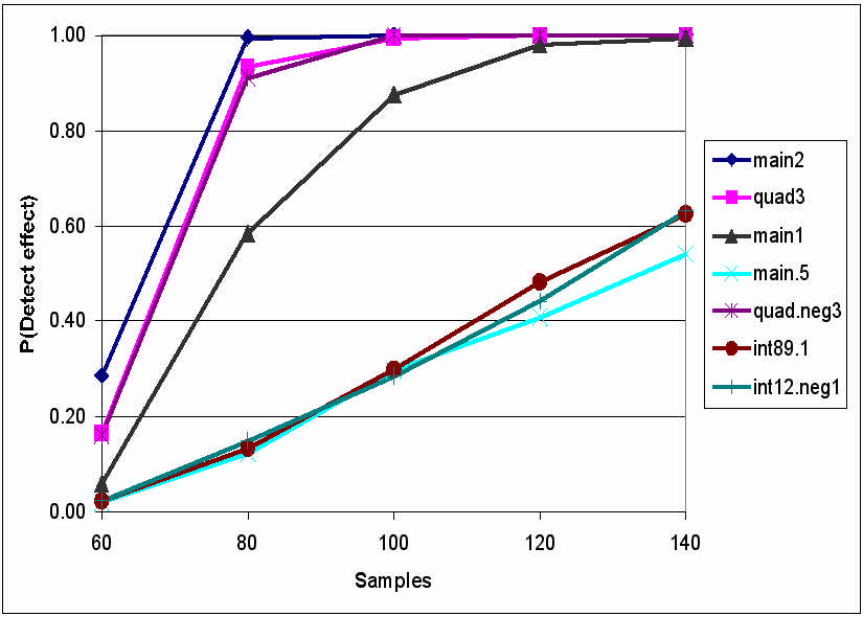


**Figure 2: The probability that linear, quadratic, and interaction effects are detected as a function of the number of Latin Hypercube samples. The factors in the legend are ordered (top to bottom) as they appear in the equation.**

While the goodness of a design must be judged by what it tells us about the model's surface, we can also assess the general quality of a design without regard to a particular surface. That is, we expect designs that are orthogonal and have good space-filling properties to work well on a variety of diverse surfaces. For LH designs, Figure 3 displays the average maximum correlation between any two input factors as a function of the dimension of the space (number of factors) and the sample size. Smaller correlations enhance our ability to fit meta-models to the data. Given a limit on the average maximum correlation that we are willing to accept, we

can determine how many samples are required for a given number of factors—or, conversely, how many factors we can explore for a given number of samples. Note that when we can take thousands of LH samples, the correlations between input factors using LH designs are very small even in high-dimensional spaces.

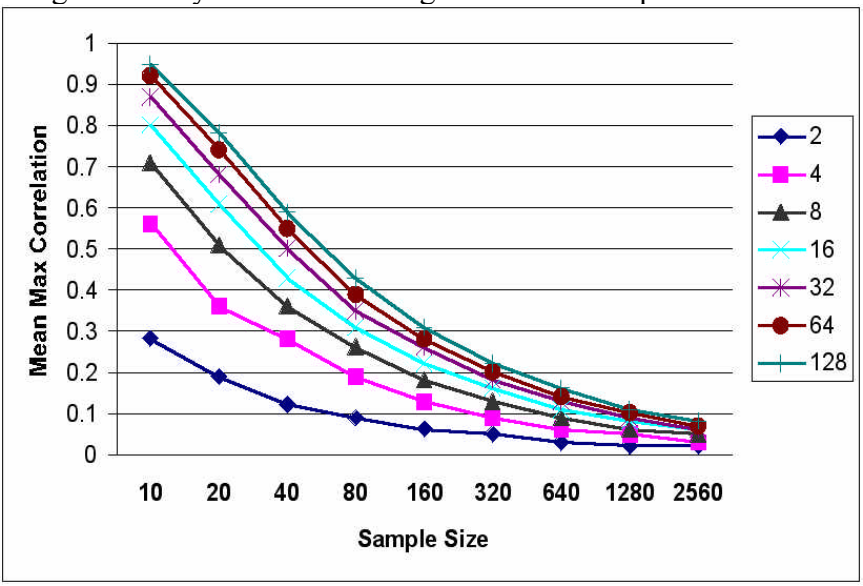


**Figure 3: The mean of the maximum pairwise correlation between any two factors for Latin Hypercube designs, as a function of the number of factors and samples. Each point on the graph is the result of 200 Monte Carlo replications.**

We are also examining designs that set factor levels and evaluate the responses in the frequency domain. As before, suppose that we have $k$ continuous factors with associated (finite) ranges. If we view the runs as an indexed set ($t=1,2,\ldots$), then we can assign a unique driving frequency $\omega_i$ to each factor $X_i$. The value of $X_i$ oscillates between its low and high levels according to $X_{i,t}$ = midpoint$_i$ + (half-range$_i$)(cos($2\pi t\omega_i$)). Figure 4 depicts this graphically for an experiment involving three factors, where all have been scaled so that the low and high levels correspond to –1 and +1, respectively.

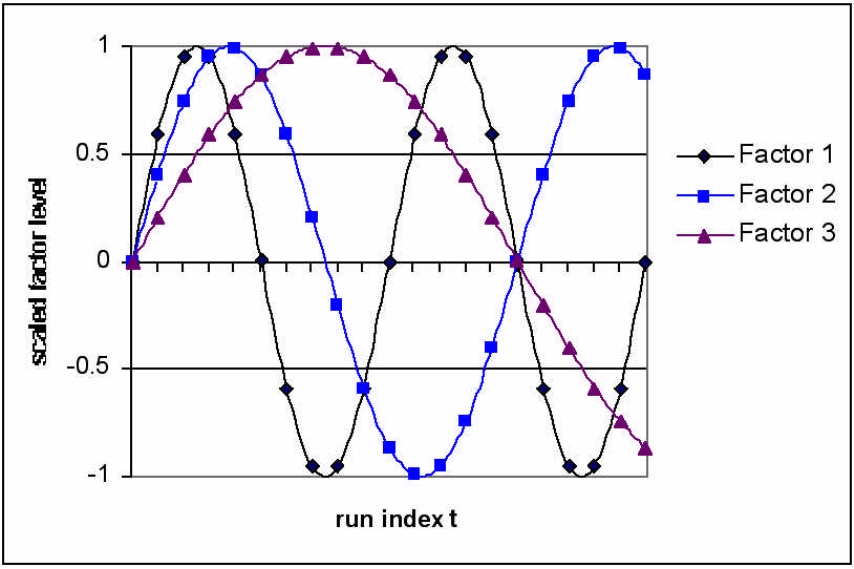


**Figure 4: Choice of levels for a frequency-based experiment involving three factors.**

Once the sequence of experiments is complete, the variability in the performance measure is decomposed into its spectral components. Important main effects appear as spikes in the spectrum at the corresponding driving frequencies, while important quadratic, interaction, and cubic effects will result in spikes at predictable indicator frequencies. For fixed *k*, the driving frequencies can be chosen to allow complete estimation of all second-order or third-order terms and interactions [11].

We now describe one set of frequency-based experiments involving a third-order response surface. There are a total of 285 potential effects (including main, quadratic, cubic, and interactions). We fix the total sample size at two complete cycles of the lowest frequency—i.e., 8192 observations. We then examine the probability of detecting an effect (p-value < .05) for the following model: $E[Y] = 2x_1 + x_2 + .5x_3 + x_1^2 + x_4^2 + x_1x_2 + x_2x_6 + x_1^3 + x_5^3 + x_1^2x_2 + x_6^2x_7 + x_1x_2x_3 + x_8x_9x_{10}$. The response is

then contaminated with additive noise with standard deviation ranging from 1.0 to 8.0. Results are provided in Figure 5. It is not surprising that we can detect all effects when the error standard deviation is comparable in magnitude to the factor effects since the number of data points is large. However, the method remains powerful even as the system becomes noisy. The cubic terms are the most difficult to detect under high noise and are likely to be identified as main effects rather than cubic effects. The dotted line for "$x_5$ (main)" illustrates this behavior.

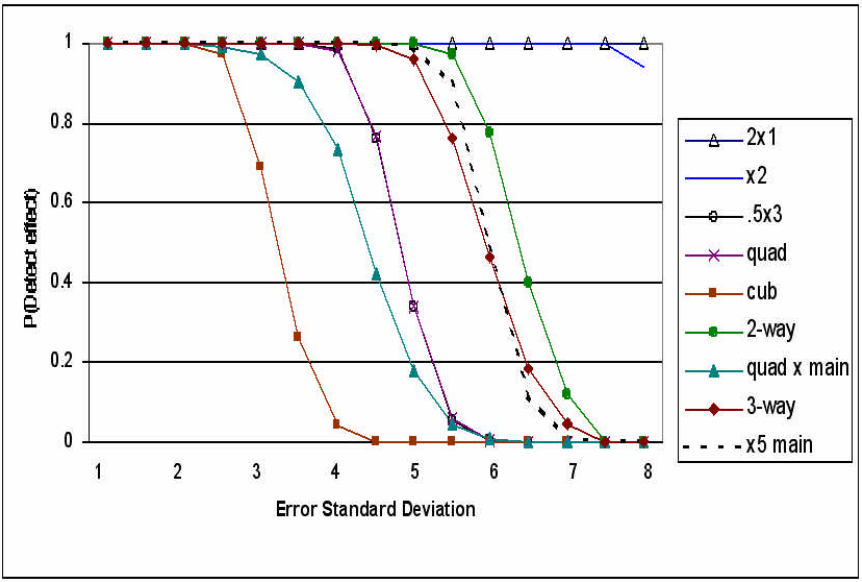


**Figure 5: The probability that linear, quadratic, cubic and interaction effects are detected as a function of the error standard deviation.**

These frequency-based designs have some added advantages for exploring distillations. The runs are easily parallelizable (segmenting on the run index *t*), and the designs can be stored concisely in terms of the driving frequencies. The output spectrum provides lack-of-fit information if, for example, we have attempted

to fit a third-order model, but strong fourth-order effects are also present. Finally, should we switch frequency assignments as part of an adaptive strategy, results can still be tested using regression rather than the (computationally efficient) Fast Fourier Transform. We continue to investigate the impact of running partial cycles of these designs, as well as their performance under alternative error structures.

These are but a few of the many experiments we have been running on known surfaces. Some general findings include the following:

- The appropriateness of the design depends critically on the shape of the model's surface and the feasible number of samples. There is no one-fits-all design.
- For relatively smooth surfaces, fractional factorial designs are an efficient means of looking at a dozen or so factors.
- For high-dimensional surfaces with sparse effects, group screening designs work well.
- When large samples are feasible (hundreds of thousands or millions), regular Latin Hypercubes work very well, particularly on highly non-linear surfaces.
- For high-dimensional searches of highly nonlinear surfaces when only a few hundred or a few thousand runs can be taken, special near-orthogonal LH designs work well.
- Frequency-based designs also work well on highly non-linear surfaces when moderate or large samples are feasible, even in the presence of substantial error.
- The adaptive sequential framework is richer and more powerful than any single one-stage design. One-stage designs would correspond to categorizing all factors into two classes: those evaluated (typically at a common level of resolution) and those ignored.

## Exploring Distillations

This section provides an overview of explorations on distillations using some of the designs that play important roles in our framework. The first example uses and contrasts what can be found with full-factorial and fractional factorial designs in an exploration of intangibles in ISAAC. In the second example, literally billions of computational experiments—utilizing Latin Hypercube and fractional factorial designs—search for extreme non-linearity and ways to mitigate the resultant non-monotonicity in the Dewar model.

### *Intangibles in ISAAC*

*The moral is to the physical as three is to one—Napoleon*

The essence of war is a clash between human wills [12]. Marine Corps warfighting doctrine encompasses the notion that uncertainty and intangibles will always be present on the battlefield. Unfortunately, quantifying how bravery and other human dimensions affect combat outcomes has proven difficult. However, since uncontrollable human dimensions will always be present, we must be able to function effectively with them [13]. The Marine Corps, through Project Albert, is studying the human dimension of land warfare with distillations, such as the prototype agent-based combat simulation ISAAC [14]. ISAAC is designed to allow the user to explore the evolving patterns of unit behavior that result from the collective interactions of individual agents.

In ISAAC, there are commander and subordinate agents, each of which makes decisions on where to move and who to engage. Commander agents also issue orders to their subordinates. Agents are given personalities that affect their propensities to hear orders, listen to orders, and move towards and away from enemy agents, friendly agents, and a goal. There are also variables that define the agents' capabilities, such as their ability to see the battlefield,

move, and attrite enemy agents. In total, there are scores of factors that can be varied in ISAAC.

To explore how changing the personalities of leaders and subordinates in ISAAC affects the Blue agents' ability to reach a goal, an urban (see Figure 6) and a desert scenario were developed. The desert scenario is similar in all aspects, except that the terrain has been removed to simulate a terrain-less environment.

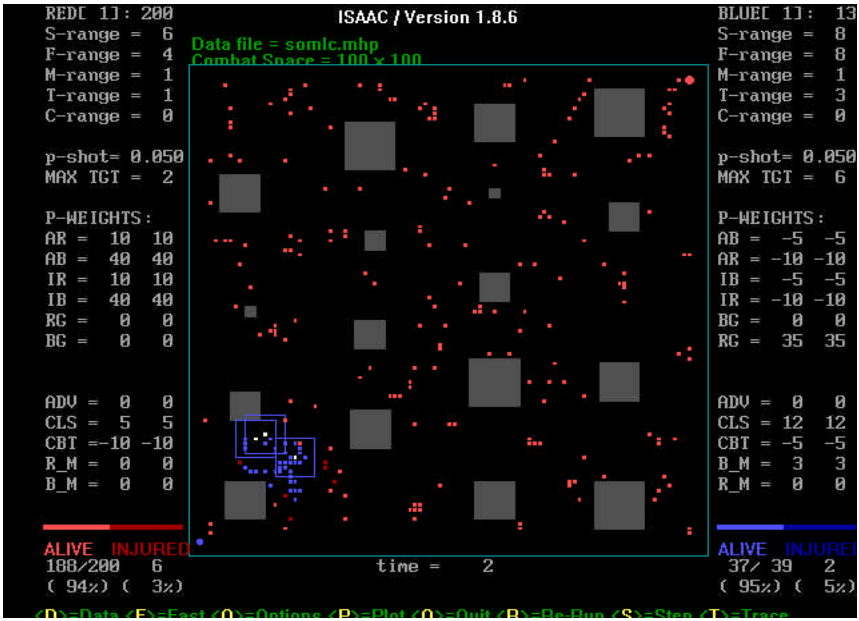


**Figure 6: Urban scenario in ISAAC. Three squads of 13 Blue agents, each with a local commander, are up against 200 loosely organized Red forces. The Blue forces are maneuvering through the urban environment to reach their goal (upper right-hand corner).**

In these scenarios, the Red forces are greater in number, less technologically capable, and have a loosely organized command and control structure. The Red forces use aggressive personalities that are held constant throughout all of the runs. The Blue forces are smaller in number, technologically more advanced, and have a

structured command and control system. The Blue forces are divided into three squads, each with a local commander (LC).

A series of full-factorial experiments were run at the Maui High Performance Computing Center (MHPPC). Due to input constraints that existed at that time, only full factorial designs of up to five factors could be run. For our runs, we looked at five factors, each at five levels. With a hundred replications, this requires $5^5*100 = 312{,}500$ runs. The parameter sets we looked at included: the Blue LC's personality weights; the Blue subordinate's personality weights; and a mixed parameter set that consisted of a combination of interesting personality weights and sensor range parameters.

Our analysis focused on determining which ISAAC parameters significantly influence the number of Blue agents killed and the time it takes for Blue to complete the mission. Of the factors that were examined in ISAAC, the ones with the biggest effects on the outcome, in both scenarios, are the LC's propensities to move toward alive Blues, away from alive Reds, and toward the Red goal. Losses are reduced for an LC that: (1) has a strong propensity to mass his forces while maneuvering away from the enemy, and (2) assigns a relative degree of importance to the mission of reaching the objective without letting this objective dominate his actions. This type of movement propensity relates directly to the concept of maneuver warfare.

An interesting interaction was uncovered in the exploration of ISAAC. Friction—that intangible element that is always present in stressful combat environments—influences the battlefield in both scenarios. Keeping in mind that the battle involves abstractions of reality (e.g., the LCs are just dots), ISAAC models friction by inhibiting the subordinates' ability to listen to their LC. In our ISAAC scenarios, higher friction levels are positively correlated with Blue losses. However, friction's effect on losses is mitigated by (i.e., interacts with) a strong bond between an LC and his

subordinates. Bond is the degree of importance a subordinate places on staying close to his LC. Thus, even if the subordinate agents cannot hear, comprehend, or otherwise act on the commander's orders, their losses are reduced if they stay with him.

All of these potential insights must be qualified as being internal to ISAAC. Are they real? We cannot tell without additional data involving real people, perhaps under real combat conditions. Nonetheless, there are some interesting insights gleaned regarding the effectiveness of potential designs. A close look at the Analysis of Variance (ANOVA) tables in [15] reveals that most of the variation in the cases examined result from linear and first-order interactions. By using a three-level one-third fractional factorial design, we can identify all of the same effects and interactions with only 8100 ($3^{5-1}$*100) runs. Moreover, about 90 percent of the variation was in the linear terms; thus, most of the same conclusions could have been reached with a two-level fractional factorial design, needing only 1600 ($2^{5-1}$*100) runs. Another way to look at this is that when these designs are available at MHPCC, we will be able to simultaneously vary many more factors. For example, a one-eighth fractional factorial design on 14 variables, each taking two levels, with a hundred replications, requires 204,800 runs. More detail on these ISAAC explorations can be found in [15].

### *Non-linearity in the Dewar Model*

*For want of a nail . . . the battle was lost*

Extreme non-linearity—even chaos—may be a characteristic of combat. Indeed, it has been shown to be a characteristic of combat models (see [16] and [17]). A combat model that exhibits chaotic behavior in an appropriate way seems, on an intuitive level, to be more realistic than a model that does not. Dewar et al. [18] studied the behavior of a relatively simple deterministic combat model that

can be thought of as a distillation of some of the aggregate-level combat models currently in use. They found that chaos in the model generated non-monotonic behavior. Non-monotonicity is defined as capability added to one side resulting in worse outcomes for that side. Of course, this can be troubling for modelers because non-monotonic behavior might make a decision-maker question a model's validity. Furthermore, it is an inescapable fact that, no matter how careful our measurements, the data used in our analyses are subject to errors. In chaotic systems, even if the magnitude of these errors is extremely small, the uncertainty associated with them creates uncertainty about our knowledge of the system in the future.

We now briefly summarize the findings in [19] of high-dimensional explorations of the relatively simple Dewar model. In particular, we address two questions: (1) how widespread is non-monotonicity in the model? and (2) can the response surface in non-monotonic regions be made more amenable to interpretation by decision-makers without destroying the chaos that may be inherent to both real combat and the model?

**Searching for Non-monotonicity**

The original Dewar model is a deterministic time-step simulation of a homogeneous Lanchester square law battle. In addition to the attrition rate coefficients, the model has parameters for initial force sizes, reserves, reinforcement levels, reinforcement delays, and decision thresholds. At each time step, depending on the engaged force ratio and force levels, each side makes decisions on whether to withdraw or call in reinforcements. There is a natural symmetry in the parameters: for each Blue parameter there is a corresponding Red parameter. The previous research focused on the initial forces subspace (i.e., the two-dimensional space defined by initial Blue force level and initial Red force level) and the binary outcome measure 'who wins.' Figure 7 shows the non-monotonic output of this subspace in the Dewar model. In this two-dimensional graph,

initial Red force levels vary from 10 to 3500 in increments of 10. Initial Blue force levels vary from 10 to 2000, also in increments of 10. Thus, the model was run 69,451 times to generate this surface. The black region represents those initial force levels that result in a Red win. Consider the following scenario. Fix the initial Blue force level at 450, and vary Red from 700 to 1800. The response trend goes from Blue wins to Red wins, and back and forth many times. This non-monotonic trend seems to make it impossible for a decision-maker to decide whether or not adding more Red forces is a good idea.

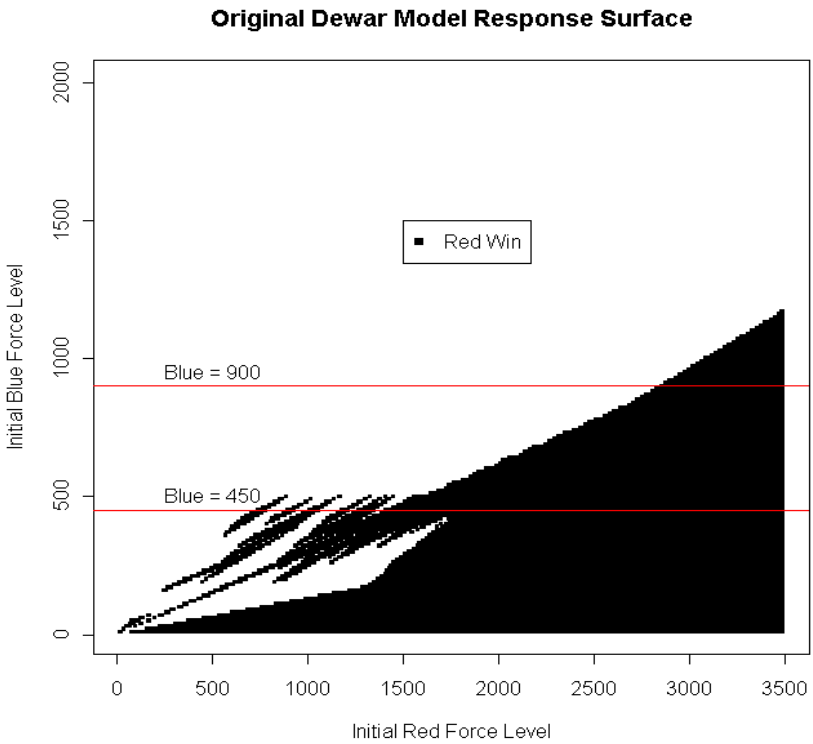


**Figure 7: The response surface of the original Dewar model contains a region of extreme non-monotonicity.**

Of the many investigations of chaos and non-monotonicity in this model, only one or two of the model's 18 dimensions had been examined prior to [19]. Thus, the question must be asked: are these findings rare anomalies that occur only in the small portions of the model that have been examined, or do they generalize to other measures and dimensions?

The combinatorial possibilities of main effects and interactions among the 18 dimensions are too great to examine en masse. The Dewar model contains 153 pairs of variables. We choose to search the subspaces associated with the nine natural pairs of variables, with a natural pairing consisting of the same parameter for Red and Blue. For each of these pairs, we want to see if the surface is monotonic over a range of settings for the other 16 parameters. To do this, we use Latin Hypercube designs on the remaining 16 parameters. The LH designs each use 16 input combinations.

For each of the nine natural two-dimensional subspaces, 16 surfaces are generated and assessed for non-monotonicity. In total, 9*16*69,451 = 10,000,944 battles are simulated to generate 144 surfaces. In designing the sample, we adhere to the original Dewar model's basic structure of a smaller, more efficient force opposing a larger, less effective force; or, if you prefer, a smaller Blue defensive force opposing a larger Red attacking force. To preserve the original model's tension between opposing forces, we restrict the domain of the remaining variables to fairly thin hyperplanes, centered at the nominal values that generated Figure 7.

So, are the previously found results rare anomalies, or do they generalize? It turns out that non-monotonicity, with respect to the measure 'who wins,' is prevalent in the model, with non-monotonicity for 'who wins' being found in seven of the nine two-dimensional subspaces explored. In fact, five of the nine subspaces exhibit pervasive non-monotonicity, with it showing up in more than 80 percent of the surfaces checked. In total, 54 percent of the surfaces generated contain non-monotonic regions.

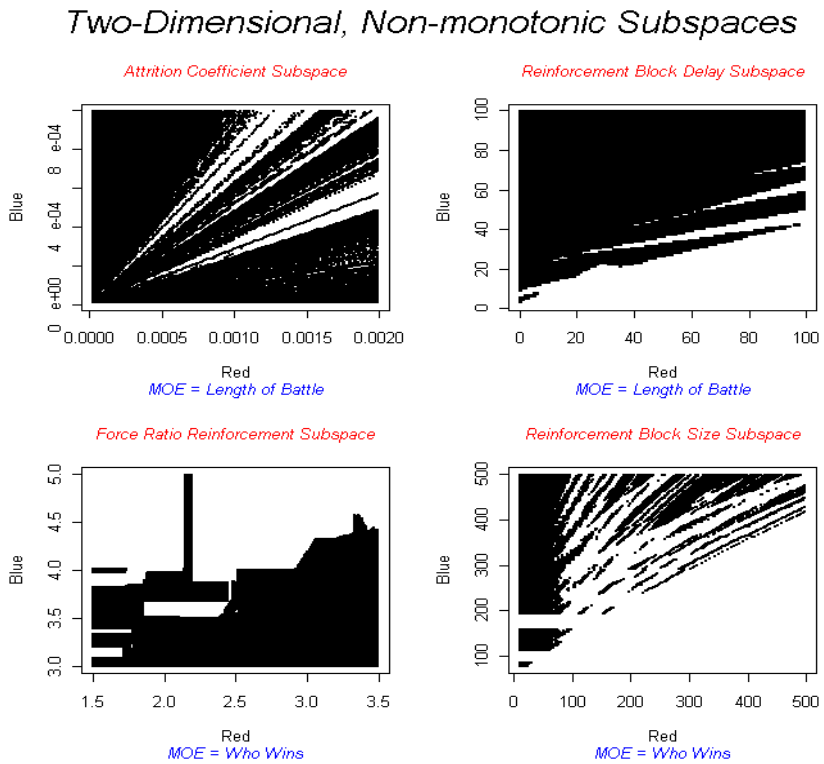


**Figure 8: Four of the many two-dimensional subspaces exhibiting non-monotonic behavior.**

Figure 8 shows some striking examples of the newly discovered widespread non-monotonicity in the model. The two graphs in the top row show examples of non-monotonicity with respect to the MOE 'length of battle.' The two bottom graphs in Figure 8 exhibit non-monotonicity with respect to the MOE 'who wins' in the force ratio reinforcement and reinforcement block size subspaces.

**Mitigating Non-monotonicity**

The fact that so many subspaces contain non-monotonic behavior is cause for concern. The Dewar model contains some of the same basic processes that many of the larger models use, such as decision thresholds and attrition processes. If the interaction of

these processes in the Dewar model generates such widespread non-monotonic behavior, then the larger, more complex models may also be affected by similar non-monotonicities. What can be done to generate interpretable responses? It has been suggested that stochastic modeling can be a useful way to deal with non-monotonic behavior in both the simple Dewar model and other, more complex models (e.g., see [16]).

To examine how making parameters stochastic affects non-monotonicity in the Dewar model, we ran a fractional factorial experiment to determine the effect of stochastic modeling on the trends of the response surface. The experiment varies nine factors consisting of the nine types of parameters in the model. Each factor has two levels, deterministic and stochastic. To efficiently search all nine factors simultaneously, a $2^{9-3}$, resolution V, fractional factorial design is used. This requires 64 different input settings. Each surface consists of 69,451 points; at each point, 1000 replications are used to obtain a precise estimate of the probability that Red wins. In total, this exploration simulates almost 4.5 billion battles.

Can stochastic modeling provide interpretable results without destroying the underlying chaos? Yes, if done carefully. Stochastic perturbation usually dramatically reduces the non-monotonic behavior of the response surface, but can, by some measures, exacerbate it. The attrition coefficients are the model parameters, over the values investigated, that have the greatest effect on the reduction in non-monotonic behavior. Figure 9 shows the same surface as Figure 7, with all of the parameters stochastic. Unlike the previous graph, Figure 9 appeals to our intuition¾the outcome remains uncertain until one side or the other quits the field of battle. See [19] for more details on these explorations.

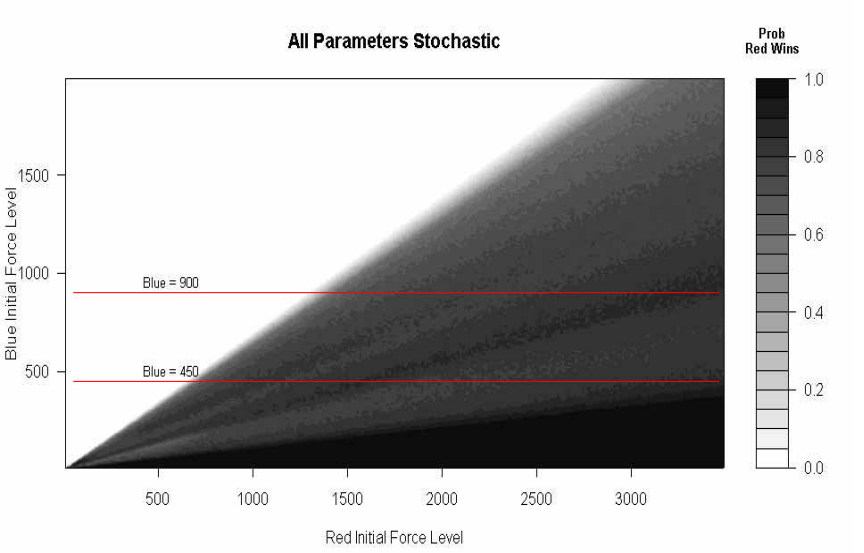


**Figure 9: When all parameters are stochastic, the response surface represents the probability of a Red win. In this graph, the regions where Red or Blue wins are clearly delineated, and the surface appeals to our intuition.**

## Conclusions

*Operations Research is a scientific method of providing executive departments with a quantitative basis for [decision-making]—Morse and Kimball [20]*

Military decision-makers frequently must make decisions about complex issues that involve billions of dollars and put many lives at risk. Our job, as analysts, is to assist them in making the best possible decision despite a plethora of uncertainties. Towards that end, we are building a framework that allows analysts to explore models in ways that have previously not been feasible. In particular, we will be able to see the effects of simultaneously exploring a large number of factors. Our initial results look promising, and the long-run success of this effort will be easy to

judge. Our success will be measured by whether or not analysts and decision-makers find this framework a useful tool for exploring distillations—and underpinning decisions.

## References

[1] Brandstein, Alfred, "Operational Synthesis: Applying Science to Military Science," *PHALANX*, Vol. 32, No. 4, December 1999.

[2] Box, George, "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics*, Editors Launer and Wilkinson, Academic, 1979.

[3] Sacks, Jerome, William Welch, Toby Mitchell, and Henry Wynn, "Design and Analysis of Computer Experiments," in *Statistical Science*, Vol. 4, No. 4, November 1989.

[4] *Webster's Third New International Dictionary*, Unabridged, Merriam-Webster, 1993.

[5] Welch, Jasper, quoted in "Overview," *Military Modeling for Decision Making*, 3rd Edition, Edited by Wayne Hughes, Military Operations Research Society, 1997.

[6] Hughes, Wayne "Overview," in *Military Modeling for Decision Making*, Edited by Wayne Hughes, 3rd Edition, Military Operations Research Society, 1997.

[7] Hamming, Richard, *Numerical Methods for Scientists and Engineers*, McGraw-Hill, 1962.

[8] Welch, Jasper, quoted in Hoeber, F., *Military Applications of Modeling: Selected Case Studies*, Military Operations Research, 1981.

[9] Meyer, Theodore and Sarah Johnson, "Visualization for Data Farming: A Survey of Methods," in Maneuver Warfare Science 2001, Edited by Dr. Gary Horne and Ms. Mary Leonardi, Marine Corps Combat Development Command, Quantico, Virginia, 2001.

[10] McKay, M.D., R.J. Beckman, and W.J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, Vol. 21, No. 2, May 1979.

[11] Jacobson, Sheldon, Arnold Buss, and Lee Schruben, "Driving Frequency Selection for Frequency Domain Simulation Experiments," *Operations Research*, **39**(6), 1991.

[12] Marine Corps Doctrinal Publication 1 (MCDP), Warfighting, Quantico, Virginia, Marine Corps Combat Development Command, 1997.

[13] Marine Corps Doctrinal Publication 6 (MCDP), Command and Control, Quantico, Virginia, Marine Corps Combat Development Command, 1996.

[14] Ilachinski, Andrew "Irreducible Semi-Autonomous Adaptive Combat (ISAAC): An Artificial-Life Approach to Land Combat, *Journal of the Military Operations Research*, Vol. 5, Number 3, 2000.

[15] Brown, Lloyd, "Agent Based Simulation as an Exploratory Tool in the Study of the Human Dimension of Combat," Master's Thesis, Naval Postgraduate School, Monterey, California, 2000.

[16] Saeger, Kevin and James Hinch, "Understanding Instability in a Complex Deterministic Combat Simulation," *Journal of the Military Operations Research*, Vol. 6, Number 4, 2001.

[17] Sandmeyer, Richard, "Simtech Project: Application of Supercomputers to Division/Corps Level Combat Simulation," *AMSAA, Interim Note No. C-159*, October 1990.

[18] Dewar, James, James Gillogly, and Mario Juncosa, "Non-Monotonicity, Chaos, and Combat Models," *Journal of the Military Operations Research*, Vol. 2, Number 2, 1996.

[19] Vinyard, William, "Reducing Non-monotonicity in Combat Models," Master's Thesis, Naval Postgraduate School, Monterey, California, 2001.

[20] Morse, Phillip and George Kimball, *Methods of Operations Research*, MIT Press, 1951.