DOCUMENTED BRIEFING

Improving the Analytic Contribution of Advanced Warfighting Experiments

Thomas W. Lucas, Steven C. Bankes, Patrick Vye

Arroyo Center

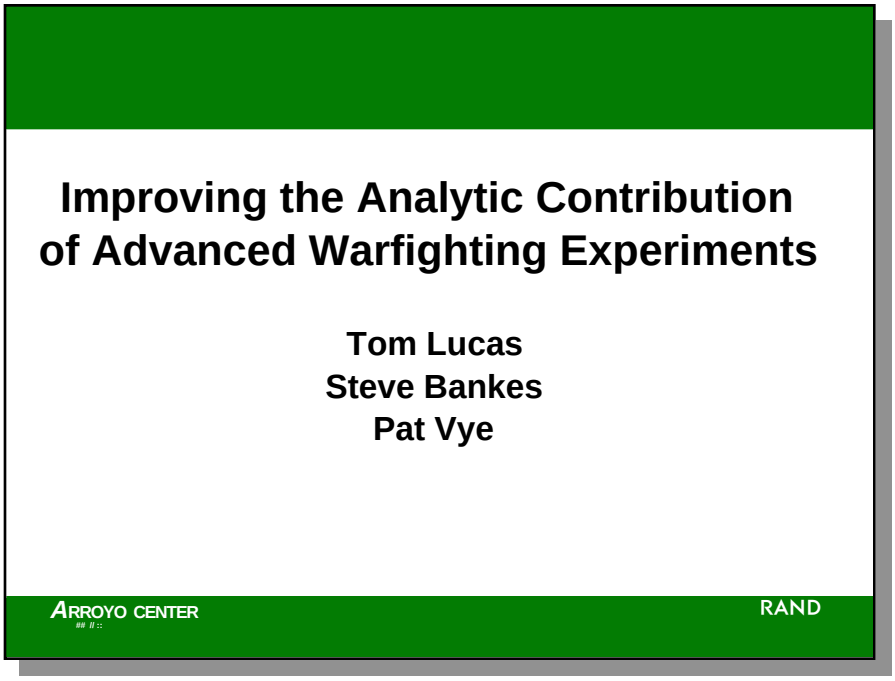
PREFACE

TRADOC Analysis Center (TRAC) asked the Arroyo Center to develop a framework for performing credible analysis using Distributed Interactive Simulation (DIS). This documented briefing reports on research that extends and demonstrates the credible analysis framework. The Focused Dispatch Advanced Warfighting Experiment (AWE), really a series of experiments, is used to conduct an exemplar analysis. The material in the report should also be relevant to those interested in combat analysis using combinations of live, virtual, and constructive simulation.

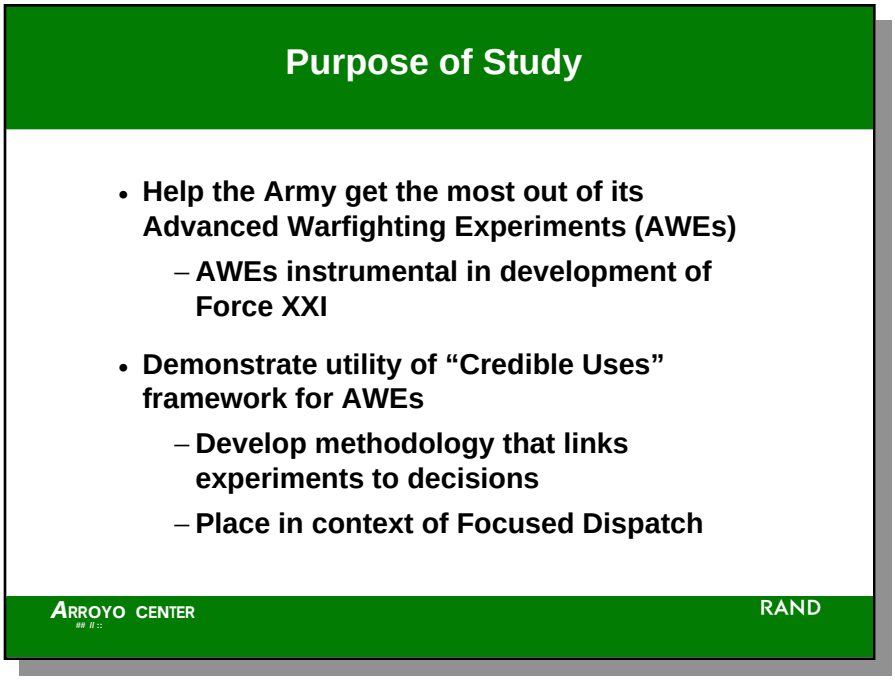
The work was carried out within the Force Development and Technology Program of RAND's Arroyo Center, a federally funded research and development center sponsored by the United States Army.

ACKNOWLEDGMENTS

The foundation of much of this work comprises numerous meetings with many Army personnel and analysts. Each contributed to the ideas within this report. At the risk of missing someone, we wish to thank the following people for their particularly important contributions. LTC Stan Ritter and Mr. Robert Bennett of TRAC-WSMR provided information on their experiences and insights with Focused Dispatch and ideas for future AWEs. Mr. Michael Bauman, our sponsor and the director of TRAC, stimulated our thinking on the general problem of few event analyses. Finally, Dr. Jim Hodges, at the University of Minnesota, and Mr. John Bondanella, of RAND, improved both the content and clarity of this briefing through reviews of an earlier draft. The responsibility for all remaining errors rests exclusively with the authors.



This briefing documents the "Credible Uses of the Distributed Interactive Simulation Environment: Phase II" research done by the Arroyo Center for the TRADOC Analysis Center. This is the final phase. Phase I findings are documented in Dewar, Bankes, Hodges, Lucas, Saunders-Newton, and Vye (1996).

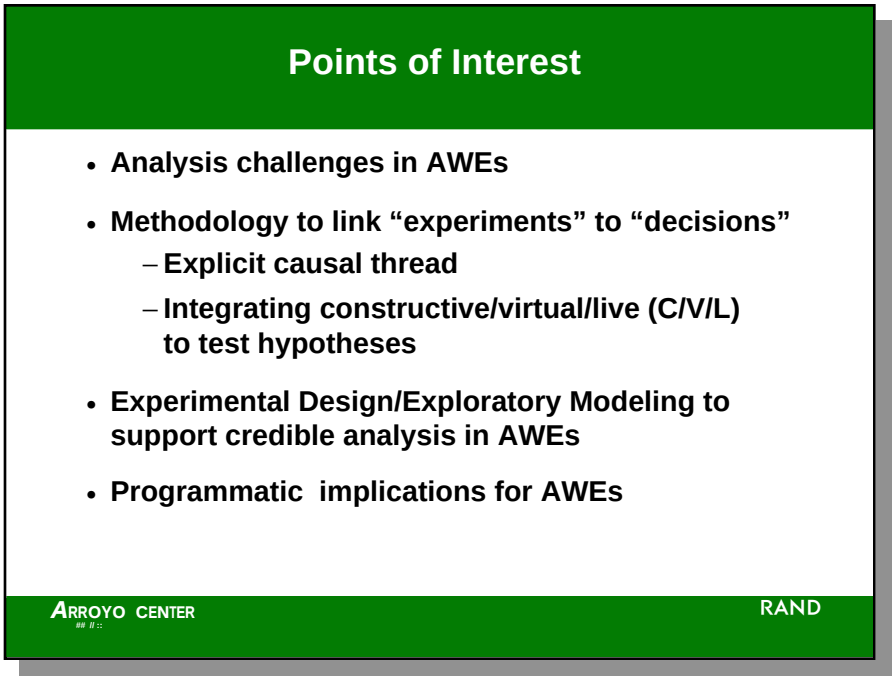


The purpose of the study is to help the Army get the maximum analytic contribution from Advanced Warfighting Experiments (AWEs)—or future similar analysis efforts involving combinations of live, virtual, and constructive simulations.

This research extends and demonstrates the framework for credible analysis using Distributed Interactive Simulation (DIS). The "Credible Uses" (CU) framework was one product of previous research documented in Dewar et al. (1996). A key component of the CU approach is that it explicitly links experiments to decisions.

In what follows we describe an example of how this framework would be applied, using for concreteness the context of an Advanced Warfighting Experiment (AWE) named Focused Dispatch (FD). The Focused Dispatch and other AWEs are really a series of experiments. Most of the experiments are constructive and virtual. The AWEs, including FD, often culminate in a live training exercise (analysis experiment). Our exemplar analysis was conducted in parallel to the actual analysis done as part of Focused Dispatch. It is notional and addresses only a small portion of the issues being studied in Focused Dispatch.

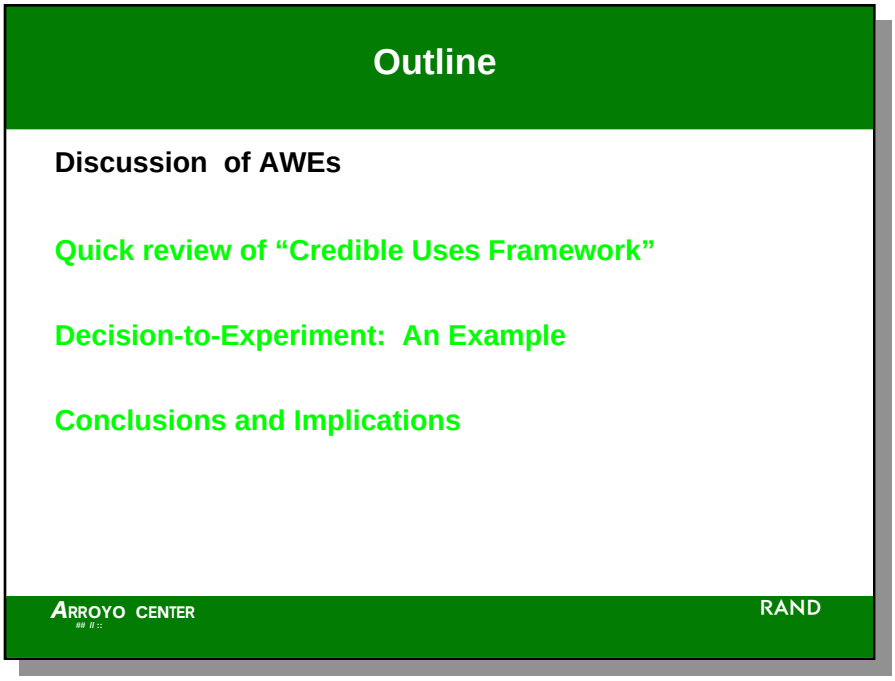
The challenge of Focused Dispatch and other AWEs is that of credible analysis using (a) models that are only weakly predictive of the results that would occur in real combat and (b) data that come from single or rare events. That is, model outcomes are not reliable enough to be considered quantitative predictions of potential real-world outcomes with known uncertainties. AWEs and similar efforts are instrumental to the Army's efforts to support the development of Force XXI. Therefore, the Army will face these challenges repeatedly in the coming years.



This briefing emphasizes four main points:

- The various challenges that must be addressed if AWEs are to become important vehicles for analysis.
- A methodology designed to meet these challenges. This methodology provides an explicit link between decisions the analysis is intended to inform and specific constructive, virtual, and/or live (C/V/L) experiments. This methodology places strong requirements on the design of the experiments that comprise an AWE.
- The experimental design requirements can be met by applying ideas drawn from the statistical literature on designing experiments and combining them with a modeling technique being developed at RAND called Exploratory Modeling (EM).
- This approach requires changes to the programmatics of conducting AWEs.

Since the usefulness of Exploratory Modeling is a central theme of this briefing, it is worth taking a moment here to define it. EM is a research methodology that uses computational experiments to analyze complex and uncertain systems. EM can be understood as search or sampling over an ensemble of models that are plausible given *a priori* knowledge or are otherwise of interest. Typically, the computational experiments are most informative when there are thousands or millions of experiments. Advances in computing technology greatly enhance our ability to effectively apply EM. See Bankes (1996, 1993) for an expanded discussion, as well as additional references.

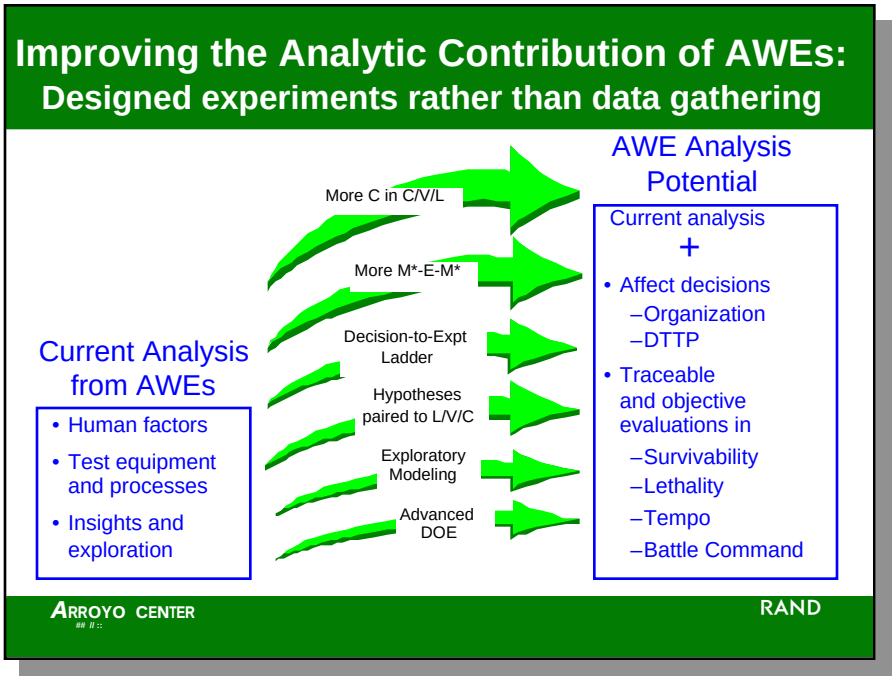


The briefing begins with a discussion of AWEs, with a focus on the challenges they present. This includes some discussion of current analysis approaches, the difficulties associated with performing analysis in training exercises, and the analytic utility of the different simulation tools—i.e., live, virtual, and constructive.

The second section briefly reviews our previous research on "Credible Uses" of combat simulations for analysis. Central to this approach is the decision-to-experiment ladder.

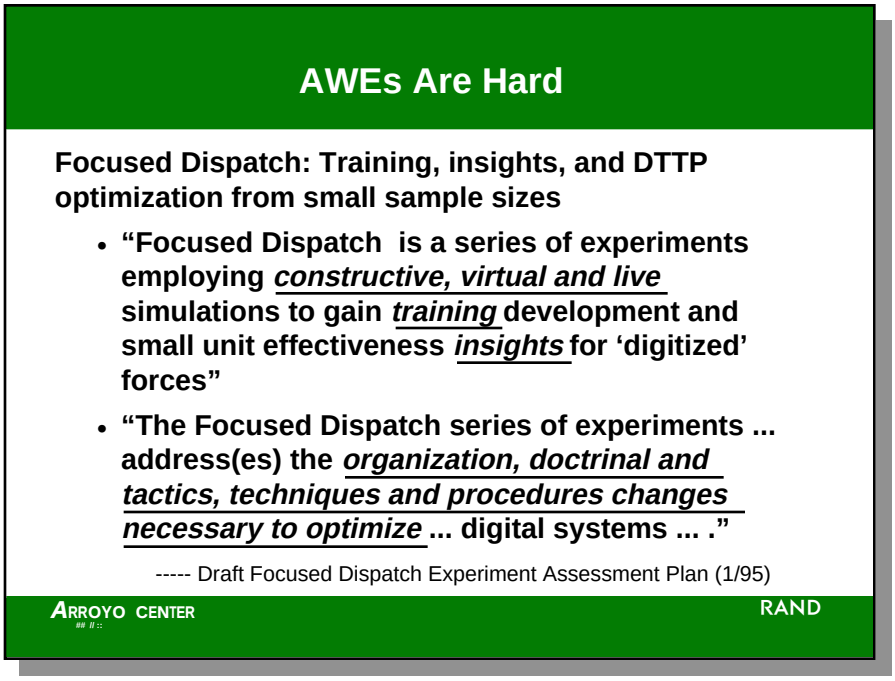
This will set the foundation for the example, which constitutes the majority of this report and is covered in the third section. Here a notional decision is made on whether all Inter Vehicle Information System (IVIS) equipped vehicles should make digital calls for fire (CFF). Included in this section is an illustration of how advanced design of experiment (DOE) techniques can efficiently explore a model space—in this case hundreds of noninteractive (batch) JANUS runs.

The final section summarizes our conclusions and suggests some programmatic changes for future AWEs and related analysis efforts.



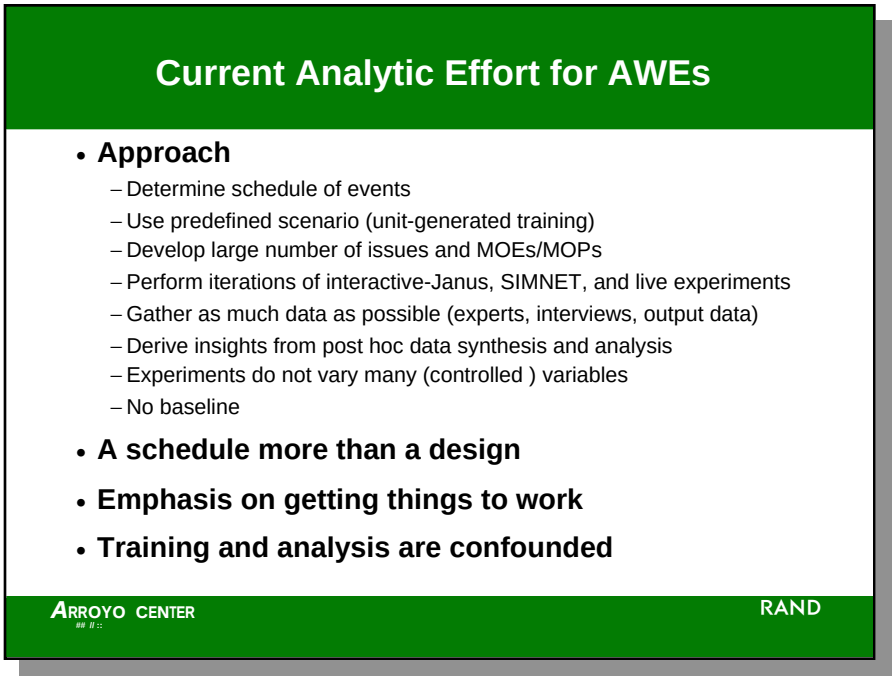
AWEs have lots of potential to support analysis, and much has been promised. Previous AWEs have clearly been successful in at least some respects. Soldiers have received training on new equipment. Analysts were able to test equipment and processes, study human factors, and obtain analytic insights into the potential effectiveness of new information-age equipment. But AWEs are very expensive, and the justification for them promises much more. In particular, AWEs are supposed to demonstrate objective and traceable "enhancements in Lethality, Survivability, Tempo, and Battle Command," as stated in the Draft Focused Dispatch Experiment Assessment Plan (1/95). Furthermore, AWEs are to assist in making and justifying decisions on force organization and doctrine, tactics, techniques, and procedures (DTTPs).

This briefing will address several ideas, methods, and techniques that move us in the direction of achieving the analytic promise of AWEs. These items are listed in the center of the chart and are defined and discussed later in the briefing. The cornerstone, we believe, is that if AWEs are to support battlefield effectiveness evaluations, there must be more emphasis on designing the experiments so that they have maximal analytic leverage to affect decisions.



AWEs are hard! AWEs rely on new and not fully tested equipment and software, which soldiers have had little experience using. Analysis of exercises where such training is being conducted is challenging, and it is further constrained by very short timelines and very small sample sizes. Moreover, employing combinations of live, virtual, and constructive simulations to measure battle effectiveness is a new analytic paradigm that the analytic community is just beginning to understand.

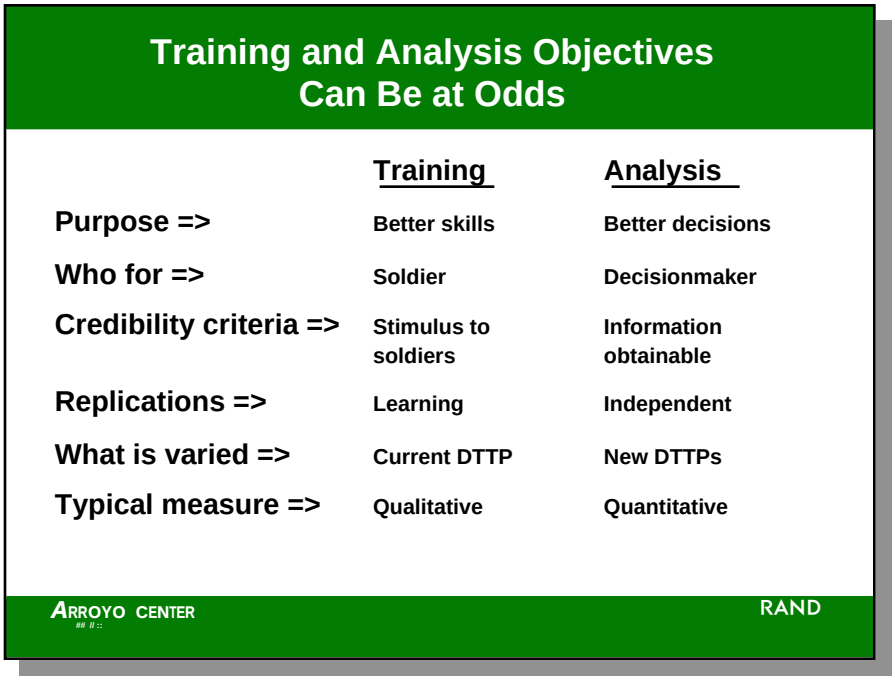
The above difficulties are compounded by the fact that AWEs have several diverse objectives. Using Focused Dispatch as an example, the goals range from training, to testing equipment, to obtaining insights, to addressing the organizational, doctrinal, tactics, techniques, and procedural changes necessary for effective use of digitized equipment. Each of these poses different, and sometimes conflicting, requirements on the training exercise (analysis experiment).



For purposes of discussion, we can characterize the general research plan used by Focused Dispatch and other AWEs as follows. Driven by scheduling constraints, a fixed schedule of events is generated, along with a "canonical" scenario. Early stages of planning attempt to preserve flexibility by generating maximal lists of issues and desired measurables, which are subsequently pruned, driven by pragmatic constraints. The various events are conducted: In the case of Focused Dispatch, these were interleaved JANUS and SIMNET experiments leading to a final live training exercise (analysis experiment). As much data as possible are gathered, and analysis using these data is done post hoc, deriving insight where possible.

This approach is more a plan for coordinating the elements of an AWE than a design for an analysis. The emphasis in most AWEs has been on bringing the live exercise off successfully and on getting equipment to work. Further, training needs drive the exercise design, so training and analysis are confounded—greatly restricting the kinds of analysis that can be obtained.

AWEs are typically arduous endeavors, with most of the effort expended in bringing them off successfully. Getting soldiers and equipment to the field, making the new systems work reliably, and bringing the soldiers up to a level of proficiency on the new equipment consumes the majority of the effort, with only marginal resources left for the analysis itself. This is due to the logistical and managerial difficulties presented by the live experiments and the challenges of the SIMNET-based virtual experiments, which often are breaking new ground in the evolution of DIS capabilities.



In some AWEs, such as Focused Dispatch, the goals include obtaining both training and analysis from the exercises / experiments. Training and analysis can sometimes be at odds. Fundamentally, training is for the soldier, while analysis is for a decisionmaker. The credibility criterion for simulations used in training is, Do they provide the right stimulus for the soldier to learn better skills? If they do, they meet their objective, even if the resultant outcomes are unrealistic. The credibility criteria for analysis, by contrast, relate directly to the capacity of the simulations to provide information about potential real events.

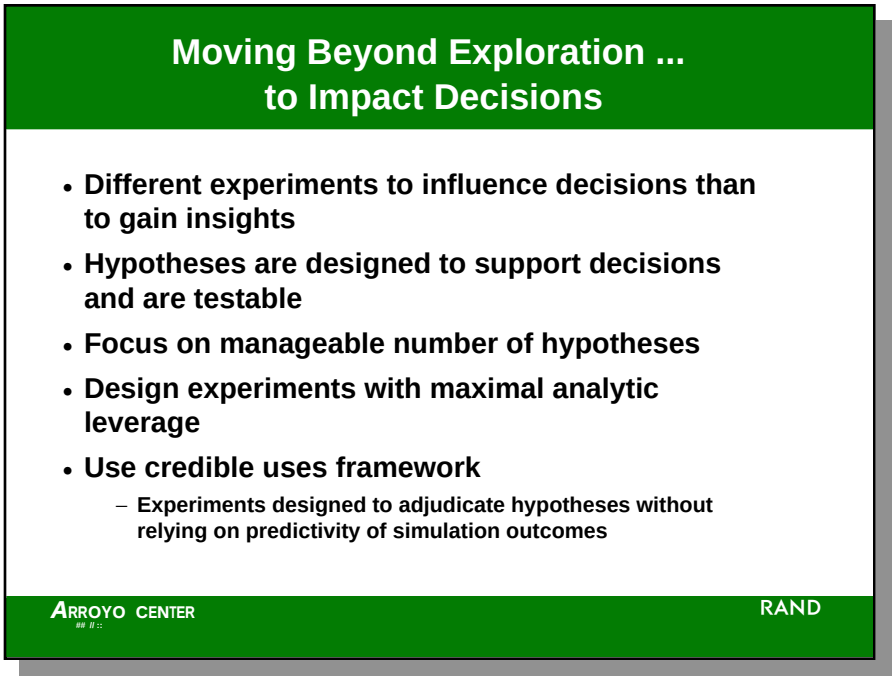
Effective training can require providing feedback for errors, for example, having an unrealistically strong opposing force (OPFOR) that will exploit any mistake the soldiers make. This has the consequence of biasing some analytic measures. For training exercises one should always be aware of the adage “Do not take tactical lessons from the training.”

There are other fundamental differences between training exercises and analysis experiments. In training, the objective is to improve skills, so learning should occur between replications. This learning can confound analytic comparisons among replications. Additionally, training often involves repetition on approved doctrine. Conversely, optimizing force organization and DTTPs requires examining multiple variations.

This chart illustrates that training and analysis necessarily conflict in the live portions of AWEs; therefore, we have to do more analysis in the constructive part of AWEs. Moreover, the scenarios should be designed for maximal analytic leverage given the training constraints.

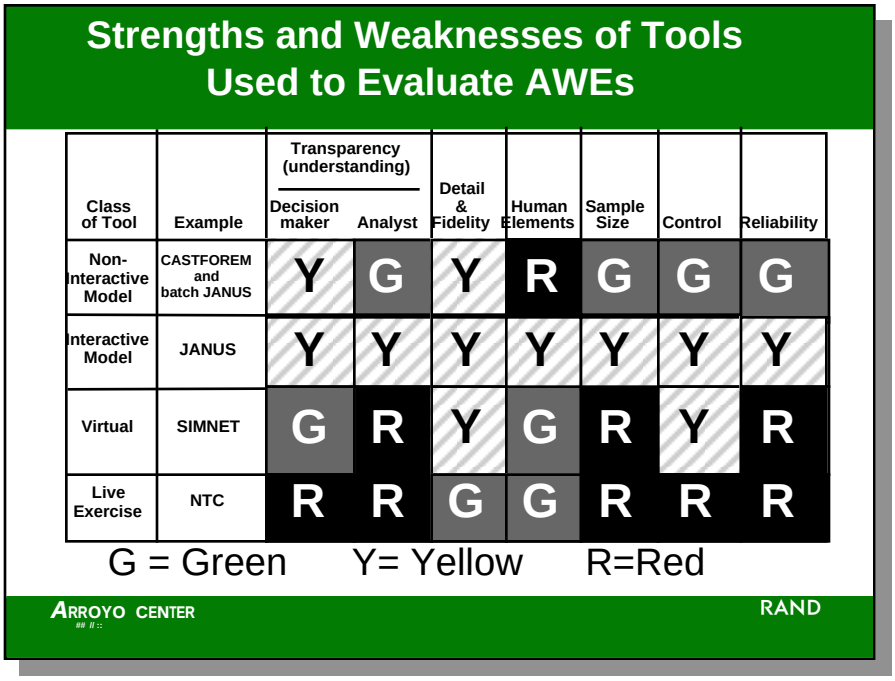
This is not to say that it is impossible to provide good analysis with AWEs. Indeed, even under present conditions, good analysis is being done. However, the constraints imposed by the current design of AWEs limits the effectiveness of analysis, often to micro issues, that is, smaller issues (such as human factors) that contribute to the bottom-line macro issues (such as lethality and survivability). Classes of analysis that have been credible in previous AWEs include studies of human factors, measurements of process-oriented information (such as delays in transferring information), lessons learned from implementation problems (such as equipment or software failure), and qualitative insights obtained from the pattern of outcomes.

Classes of analysis that are difficult to do simultaneously with training include comparisons of the battlefield effectiveness of various systems, organizations and DTTPs, measurements of lethality, survivability, and tempo, statistically valid quantitative outcomes, comprehensive sensitivity analyses, or exploratory modeling. Unfortunately, some of these classes of analyses are expected from the AWEs. We will address how supplementing the current experiments with more closed constructive simulations can help compensate for these difficulties.



The focus of this briefing is not the production of qualitative analytic insights, which we will refer to as hypothesis generation. Instead, our intent is to demonstrate how to design analyses to adjudicate hypotheses, which is the step after their discovery. Although some hypothesis adjudication might be possible with the experiments within Focused Dispatch, in general, hypothesis adjudication will require that different experiments be performed than for hypothesis generation. Hypothesis adjudication will require focusing on a manageable number of specific hypotheses (in contrast to the very long list of issues and hypotheses used in Focused Dispatch) and the experiments must be specifically designed to provide maximal analytic leverage to address those hypotheses.

Hypothesis adjudication is a challenging task because no combat model can be presumed to accurately predict the outcomes of real combat. Indeed, due to safety restrictions and lack of knowledge of future threats, even live simulations must be regarded as at best weakly predictive. Last year we produced a framework for understanding how to use weakly predictive models for analysis. Here, we will apply and demonstrate that framework through a mock analysis.

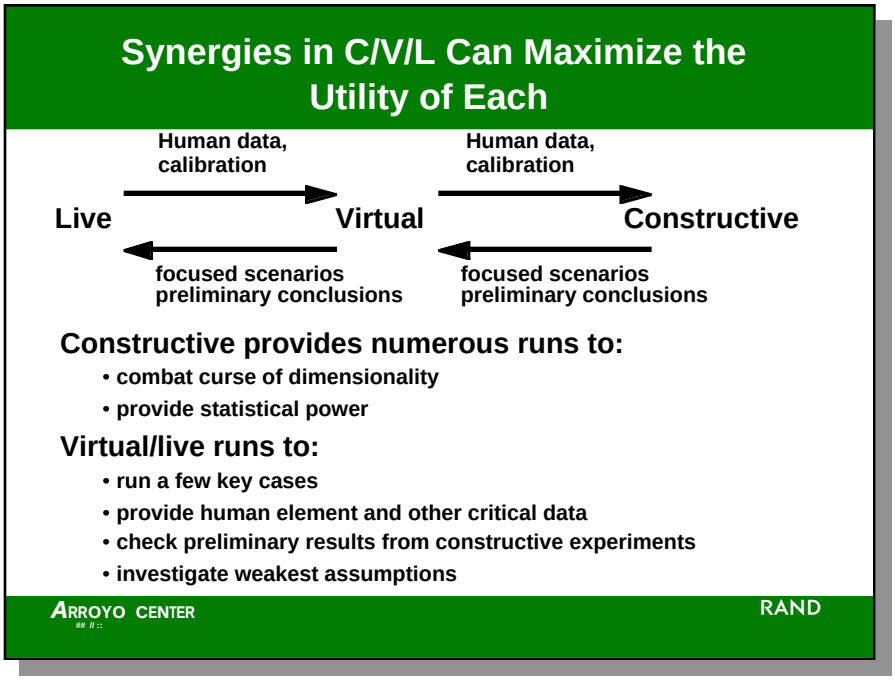


While no component of AWEs is strongly predictive, by combining multiple analysis tools we can use the strengths of one to compensate for the weaknesses of another. This matrix displays the strengths and weaknesses of potential AWE analysis tools in terms of several attributes important for good analysis. Green cells indicate that the class of simulation (row) has good capability in the analytic attribute (column). Yellow and red cells represent moderate and poor capabilities. While the specific colors assigned to different cells of the matrix can be debated, clearly the variety of available tools have differing strengths and weaknesses. Basically, as we move along the columns in the table, the simulation types allow for an increased capability to take large samples of controlled experiments. The simulation types toward the bottom of the table provide more detail and realism, particularly with respect to human elements. Furthermore, no single tool can provide all the attributes that may be required in an analysis. So, credible analysis may require combining the strengths of different tools.

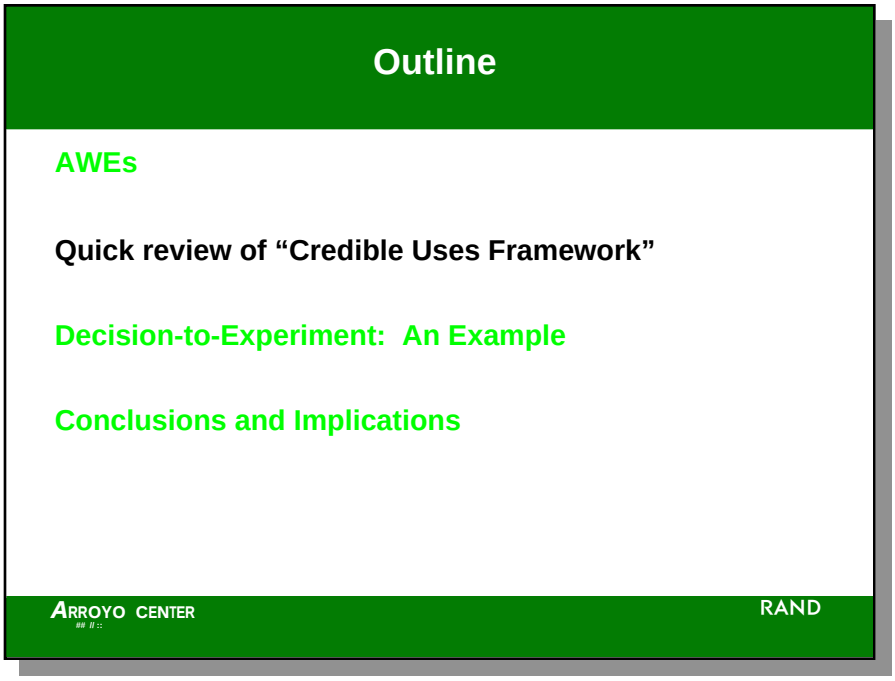
Most AWEs have not formally included noninteractive constructive models, such as batch JANUS. Batch JANUS is a noninteractive version of the Army standard interactive JANUS. Excluding noninteractive constructive models severely restricts analysts' ability to perform many experiments, examine the effects of many different variables and scenarios, control for nuisance factors, and replicate results. Moreover, as constructive simulations are less dependent on the training proficiency of the soldiers as well as unreliable (at least, not fully tested) hardware and software, they are well suited to compensate for some of the problems introduced by training. We believe that explicitly using this relatively

inexpensive tool (as compared to virtual and live simulations) to supplement the others will greatly strengthen the analysis contributions of AWEs.

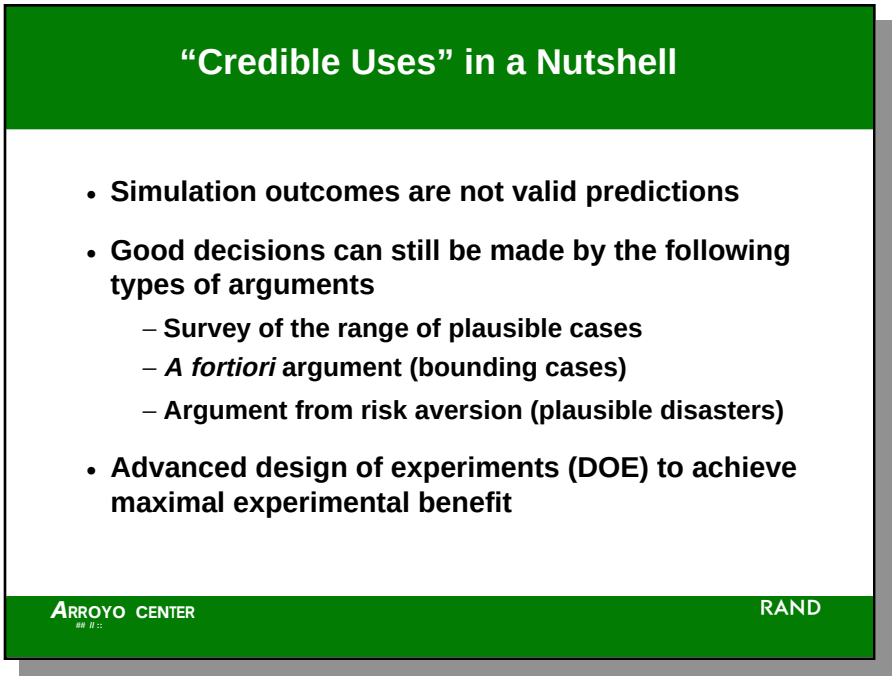
Constructive simulations, of course, must be used carefully. Many constructive models will not directly and explicitly simulate events and factors that the analyst is interested in. Therefore, there must be careful thought given to how they are used to characterize new pieces of equipment, human elements, and new doctrine.



In devising an experimental design, each source of information can potentially be used to enhance, focus, or compensate for the limitations of other resources. The constructive simulations can support many more runs than virtual or live. These runs can be used to (a) explore many scenarios, (b) identify key cases to be further examined in the limited number of live or virtual experiments, and (c) generate sample sizes sufficient to produce statistically defensible conclusions. Conversely, where virtual or live experiments can provide superior realism, they can be used to (a) check preliminary conclusions based upon constructive simulation, (b) provide data for the key cases nominated through constructive exploration, and (c) inform constructive simulations by measuring key human element (and other) data. Thus, the demands of the analysis being conducted could dictate doing experiments of different types in a variety of sequences.



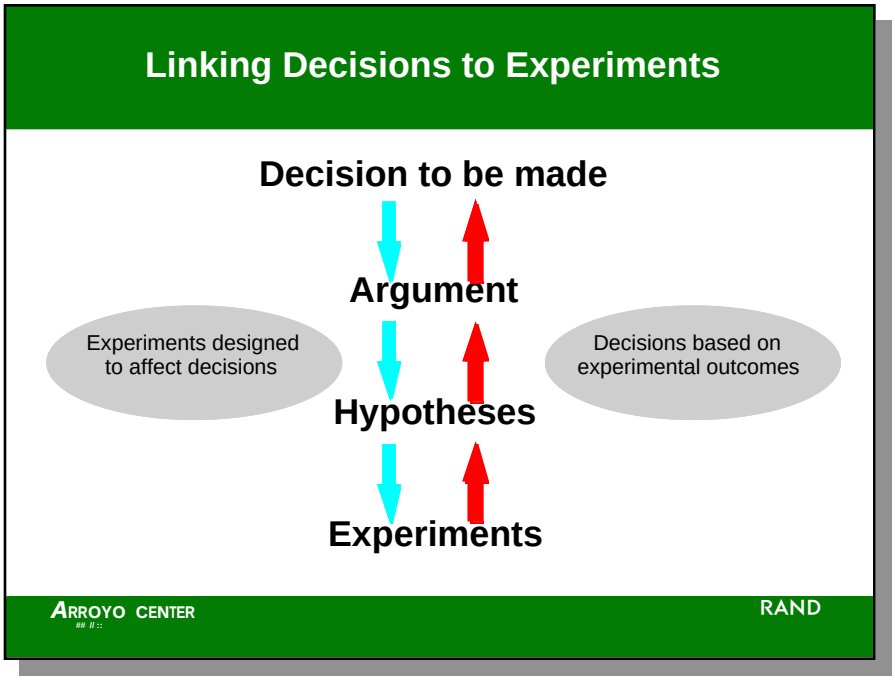
Next, we briefly review the "Credible Uses" framework.



The underlying philosophy of our approach is that most combat simulations cannot be regarded as accurate predictions of real-world outcomes, and thus we should not reason as if they were. Models are called weakly predictive in cases where there is sufficient knowledge to—when represented in a model—result in model behavior that is interesting and informative, but where too little is known for model outputs to be credible predictions of real-world behavior. Thus, weakly predictive uses are those requiring only that outcomes be consistent with all information that is available and seen as salient to the analysis at hand. This requirement is associated with various other terms that are in use, including "realism" and "face validity."

Even though most combat models cannot be viewed as generators of reliable predictions of real-world outcomes, they can be valuable analytic tools. This does, however, require us to use different research strategies than we would with predictive models. Credible research strategies for using weakly predictive models are typically driven by patterns of argument based on reasoning under uncertainty. Thus, to pick one example, a reasoning strategy could be based on discovering plausible disaster scenarios and using them to drive subsequent analysis. Weak predictivity suffices to demonstrate the plausibility of the discovered scenarios, and the reasoning strategy provides a context for model use that compensates for the lack of predictive accuracy in the models. Research strategies with weakly predictive models typically require many runs. Doing so efficiently benefits from advanced design of experiments.

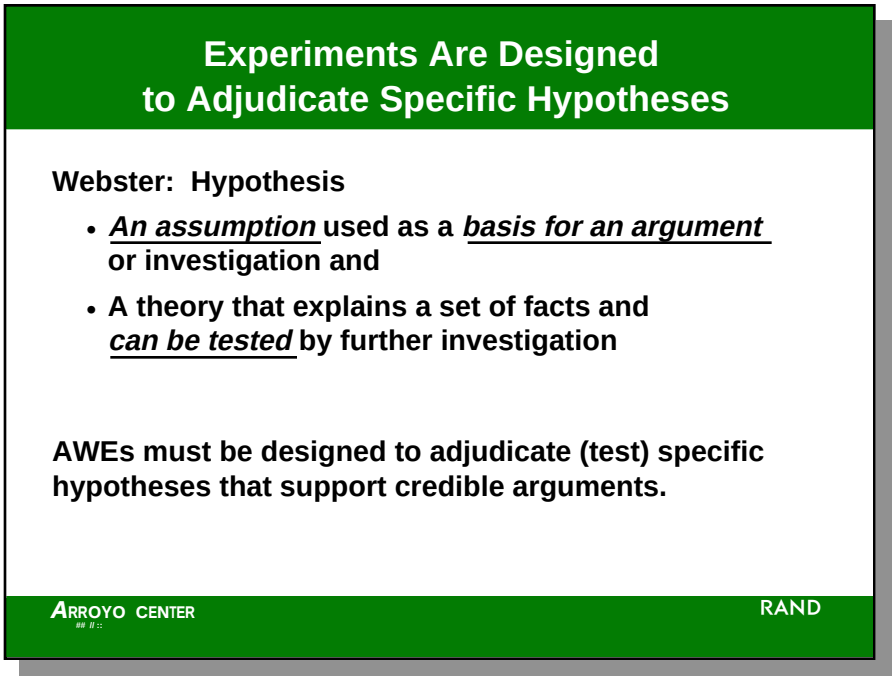
This approach contrasts strongly with the naive use of weakly predictive models to make "pseudo predictions" that could result in very misleading conclusions. For an expanded discussion of this see Dewar et al. (1996) and Hodges and Dewar (1992).



The question is, How can we construct experiments with nonpredictive tools? The outline of an analysis using weakly predictive tools looks something like this figure. A decision requiring analytic support leads to a proposed argument for making the decision. Some assertions in that argument will, inevitably, be of unclear truthfulness and thus constitute hypotheses, some of which may be subject to adjudication through experiment. The experiments are designed, including case selection and the data to be obtained, to maximize the experiment's ability to adjudicate the hypotheses. Going the other direction, if the experiments successfully resolve the hypotheses of interest, this will buttress the provisional argument, providing a traceable basis for making the decision.

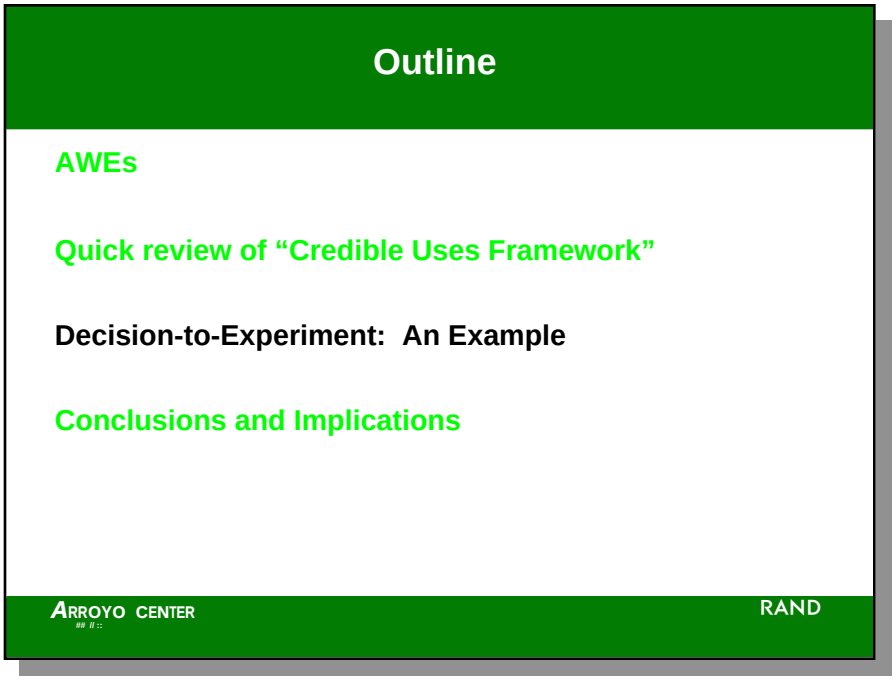
We believe that analysis plans for AWEs should formally make this decision-to-experiment ladder. There is nothing new here. This is the scientific method—portions of which analysts do informally all the time. Formalizing this process ensures internal consistency, external traceability, and efficiency of resources. Moreover, potential decisions are explicitly written down. Analysts, and others, know exactly the strength of the argument needed to make the decision. This process also explicitly uses the decisionmaker's prior beliefs. For situations where strong prior beliefs exist, it may be that very little experimental evidence is needed to make the decision—perhaps some assurance that nothing unexpected will happen when decisions are implemented in the field. If substantial evidence is needed, this process may reveal that it is not possible to adjudicate the hypotheses within the cost, time, and analytic constraints of a given AWE. All of this is much preferred to the hazardous approach of seeing what can be inferred after the experiments have been conducted.

It has been noted that this process may result in showing that there are not sufficient resources to resolve all the decisions of interest. Knowing this in advance is a good thing. By prioritizing the decisions to be made (or issues to be addressed) one can efficiently use the experiments. Issues that cannot be evaluated by designed experiments can still be addressed much as they are currently—with postexperiment investigation.

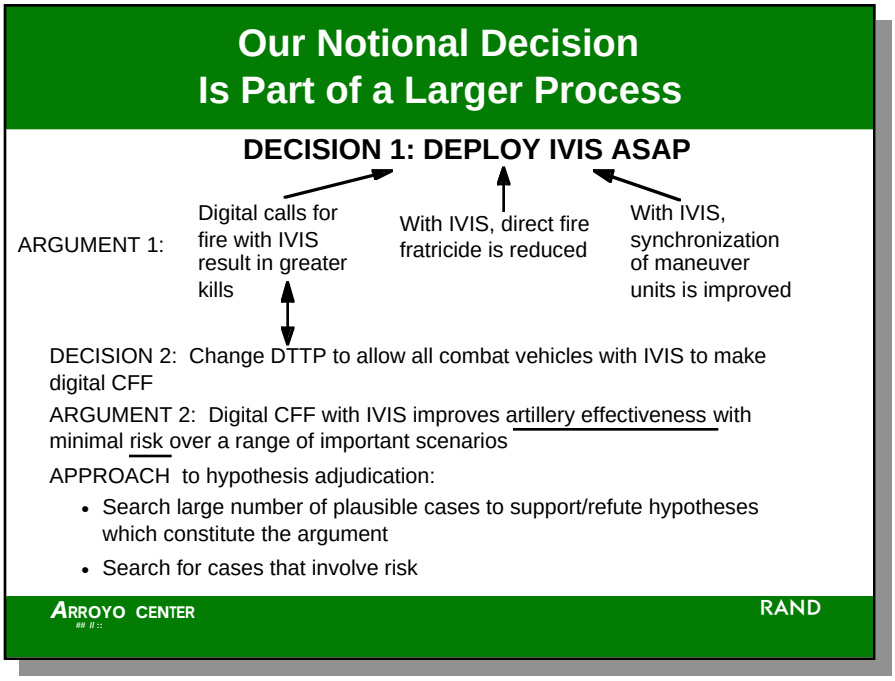


One tangential point merits mention here. The meaning of the word "hypothesis" appears to have been drifting somewhat in recent usage. Among hypotheses appearing in some lists are issues of interest that are either obviously true ("improved lethality leads to better battle outcomes") or are not specific enough to be subject to experimental adjudication ("If procedural, functional, and organizational changes in fires, intelligence, logistics, and battle command are implemented as a result of digital connectivity, then even greater enhancements in lethality, survivability, and tempo will result.").

If we refer to the dictionary definition (*Webster's II New Riverside Dictionary*, 1984), the role of a hypothesis to bridge between argument and data is clear. Our use of the word will be limited to situations that satisfy both of these criteria.



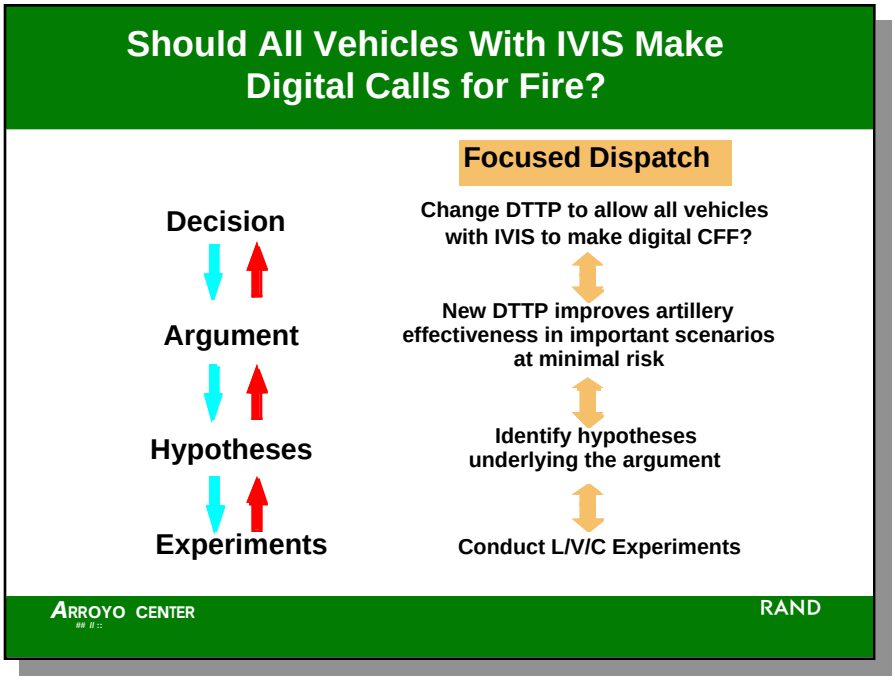
We now will demonstrate this process on an issue related to procedures using information technologies. The example is for pedagogic purposes only.



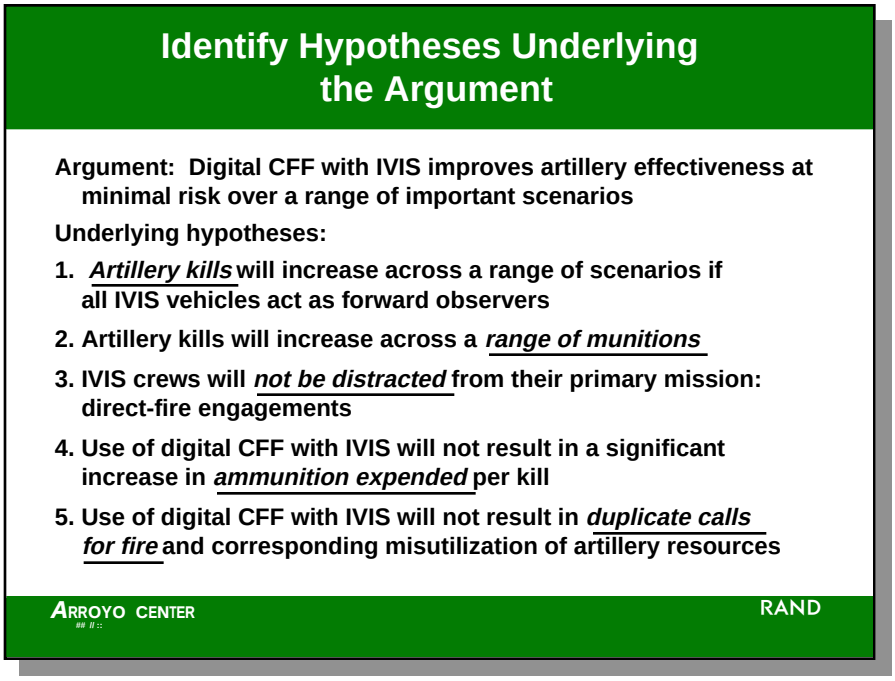
Our example relates to a potential decision that could be addressed by a "Focused Dispatch"–type AWE. The notional decision will be whether to change the DTTP to allow all vehicles with IVIS to make a digital call for fire (CFF). Of course, this decision does not exist in a vacuum; it is part of a larger process. The diagram above shows how this AWE-specific DTTP decision might fit in the context of multiple AWEs. The high-level decision, decision 1, is whether the Inter-Vehicle Information System (IVIS) should be bought and deployed as soon as possible. The decision will eventually made by senior military and civilian leaders.

The notional high-level argument (argument 1) says IVIS should be bought and deployed if it results in more indirect fire kills, reduced direct fire fratricide, and improved synchronization of maneuver. While innumerable additional items can be added to the argument, this diagram displays what is notionally determined as the minimal support for making the decision. It is better to have a few good hypotheses and a lot of cases for each hypothesis than the other way around. An excessive number of hypotheses makes their evaluation infeasible. Ideally, the list of hypotheses will be necessary and sufficient.

The lower-level argument (argument 2) to make the DTTP decision (decision 2, to allow all combat vehicles with IVIS to make digital CFF) is that it will improve artillery effectiveness with minimal risk over a range of important scenarios. Specific hypotheses, shown on subsequent slides, are contained in the elaboration of this argument.

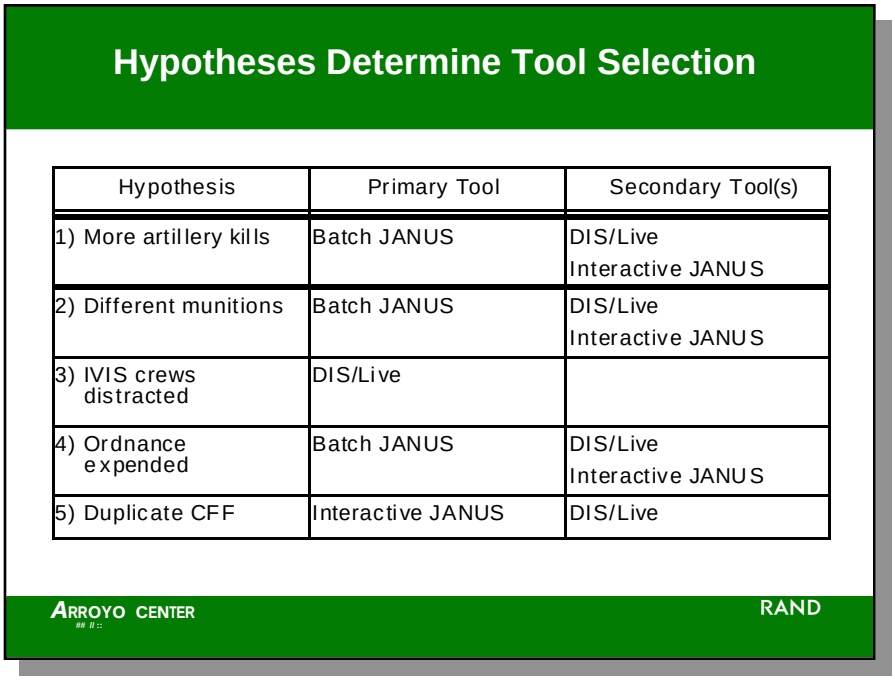


On the right is a notional example of this process drawn from the domain of Focused Dispatch. Here the decision is whether all vehicles with IVIS should make digital CFFs. This leads an analyst to posit that such deployment is desirable because these systems could plausibly improve battle outcomes for important scenarios, and because there is little risk. Such an argument contains hypotheses, some of which can be adjudicated by experiments using various combinations of live, virtual, and constructive simulation-based experiments. To adjudicate these hypotheses with weakly predictive models, our research strategy will see if the hypotheses hold over a range of plausible cases, while looking for plausible disasters that could result from the general implementation of the DTTP. The design of experiments includes both what things will be varied and what data (measures) will be extracted.



For our notional example we assume that the argument is decomposed into the above hypotheses. Of course, these hypotheses do not include all those of potential interest. An extensive list, however, would be prohibitive within the time, budget, and other constraints of most studies. One role of the analyst is to help make the argument as efficient as possible. *A tradeoff must be made between stronger arguments for decisions and the number of decisions for which the experiments can provide information.* For our purposes we will take these five hypotheses as sufficient to make the decision.

It should be emphasized that hypothesis 4 is stated in terms of ammunition expended per kill. This is done so as at roughly the same rate of ordnance per kill to not penalize the new DTTP if it results in significantly more kills and, hence, more ordnance used.

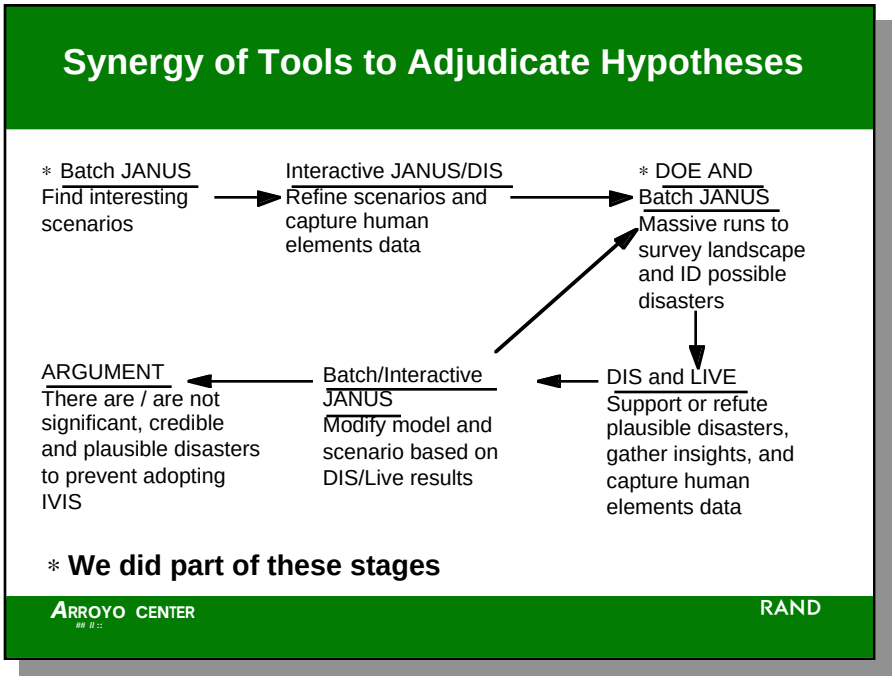


The hypotheses determine which analytical tool, or combination of tools, will be used to adjudicate them. Some hypotheses, such as number 1, require a large number of plausible scenarios to be examined. Thus they *must* be *primarily* tested using noninteractive simulations—such as batch JANUS. Of course, all closed models contain many limitations and assumptions, especially regarding human behavior and decisionmaking. The argument can be strengthened by showing that the results remain consistent over a range of scenarios and assumptions and are consistent with and informed by the results of live and virtual experiments.

Since there are very few experiments available with humans in the loop (HIL), each experiment must provide information on many hypotheses. Since there may be more hypotheses than HIL experiments available, the ability to effectively inform on many hypotheses may be feasible only if process information is being examined, e.g., the delay in direct fire reaction times (as relates to hypothesis 3). Each HIL experiment will generate multiple occurrences of direct fire reaction time—and similar data. Additional output from the same experiment should, where possible, be used to study other hypotheses. The key is to use the HIL experiments to provide as much information as possible about elements that they address well—for example, human effects.

There are important aspects here which are different from the usual model-experiment-model (M-E-M) paradigm. Briefly, the usual M-E-M approach uses the first M to focus the experiments and the second M to extend the results and provide sensitivity analysis. Both of these features are very good and are incorporated in our approach. In practice, a large portion of the traditional M-E-M effort goes into calibrating the M to the E and relating measures of

effectiveness (MOEs) to general issues. Our approach focuses on determining what tools or combination of tools can best adjudicate specific hypotheses. This may or may not involve calibrating the models and any particular simulation sequence. Differences between models can be not only tolerated, but used for better decisionmaking! Furthermore, many of the hypotheses are evaluated primarily with a single model rather than a combination of models. This allows the Credible Uses (CU) approach to address a broader range of issues. Moreover, and more important, with CU the experiments are designed explicitly to inform decisions. These advantages are illustrated on the next few slides.

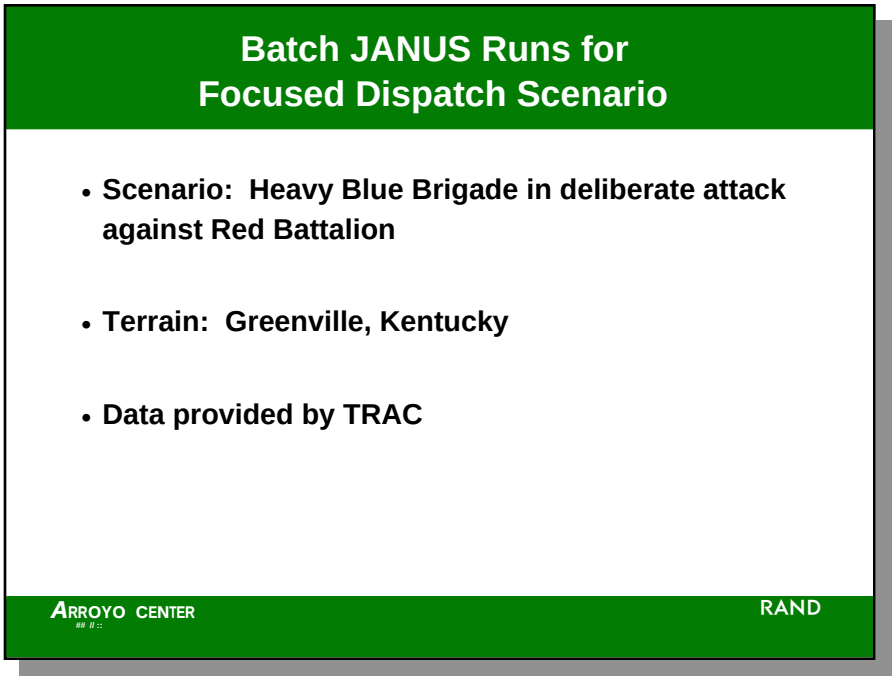


There is typically a natural temporal ordering of the experimental tools; one such ordering for our example is illustrated here. The assets that are plentiful (the batch constructive simulations) can be used to examine multiple scenarios and variables. Finding the scenarios that provide the most analytic leverage to adjudicate the hypotheses is one of the most important aspects of the analysis. *If scenario selection is not done carefully, one can easily become trapped in a scenario dominated by factors of little interest, thus masking the effects important for the analysis.* On the other hand, the scenario should not be one that *a priori* the experimenters know will turn out as desired. Such an experiment would (1) be uninformative, thus not helping the decisionmaking process, and (2) cause the objectivity and thus credibility of the analysis to be called into question.

Once informative scenarios have been determined, and perhaps modified as a result of preliminary virtual simulations (interactive JANUS and perhaps DIS), a massive number of batch runs can be efficiently taken to explore the outcome measures over a wide range of plausible scenarios and to find plausible disasters. Preliminary conclusions on some of the hypotheses may be obtained. Additionally, these batch runs can be used to focus the few live and virtual experiments on cases where they can be the most informative. Data and insights from these runs can be used to tune the constructive models. The tuning might consist simply of data calibration and scenario adjustment, or possibly even to change aspects of the model itself (e.g., create code to simulate a critical observed factor in more detail). The modified constructive models are then rerun to provide further information about the hypotheses. This process repeats as time and budget allow. Of course, to be able to implement such a process requires

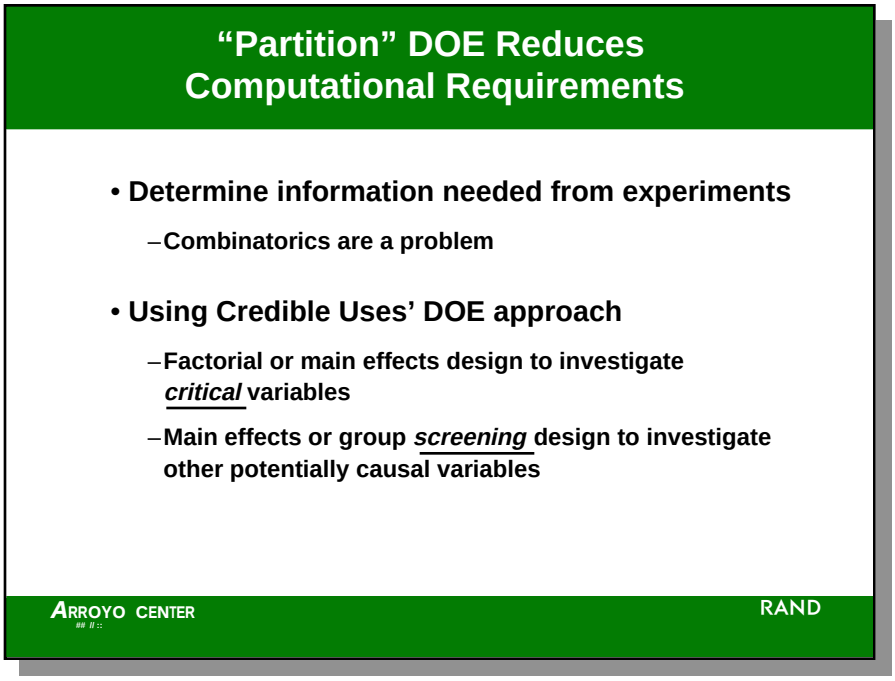
resources that are beyond most analysis studies. Thus, implementation of such a process is intended for large analysis efforts, e.g., AWEs, such as Focused Dispatch, which last over a year and have budgets of several millions of dollars.

The following slides contain the results of batch JANUS runs that were made as part of the constructive portion of our mock analysis.



The first phase of Focused Dispatch (the first of three interactive JANUS phases—JANUS 1) was concerned with artillery issues. An issue arose as to whether digital calls for fire should be routed directly from an observer to the guns as opposed to going through a Fire Support Officer (FSO) and a Fire Direction Center (FDC). It became clear during the JANUS 1 phase that there was potential to overwhelm the guns with indirect fire requests. A subsequent question was whether all vehicles equipped with IVIS should be allowed to make digital calls for fire or only the traditional forward observers. This issue was confirmed by the Mounted Battlespace Battle Lab, TRAC-WSMR, and TEXCOM as being important.

To demonstrate the methodology, we decided to use the same scenario used in the JANUS 1 phase to evaluate our notional hypotheses. The "Greenville Kentucky" scenario, used throughout Focused Dispatch, comprised three scenarios: a defense, a movement to contact, and a deliberate attack. We used the deliberate attack scenario. The scenario and data that served as the starting point of our analysis was provided by TRAC-WSMR and was what they used at Fort Knox for the JANUS 1 runs. The batch JANUS runs were conducted and evaluated in the Military Operations Simulation Facility (MOSF) at RAND.



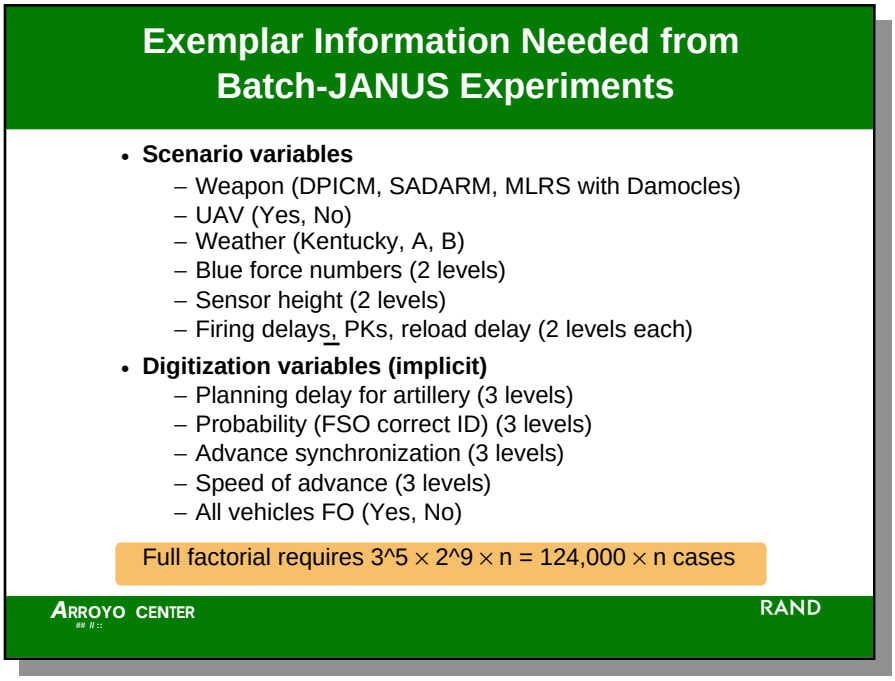
Even with constructive models, the information needed from the experiments requires large amounts of computation. This is especially true when one is making arguments with weakly predictive tools—such as most combat models. Inevitably, the combinatorics of the number of variables we wish to vary are enormous. It is typically not feasible to run all the cases we would like to, particularly with respect to uncertainties in the model.

To use limited computational resources effectively, guided by expert knowledge we partitioned the ensemble of model variables into two classes:

(1) Variables whose effects it is *critical* we evaluate.

(2) Variables we wish to *screen* for effects.

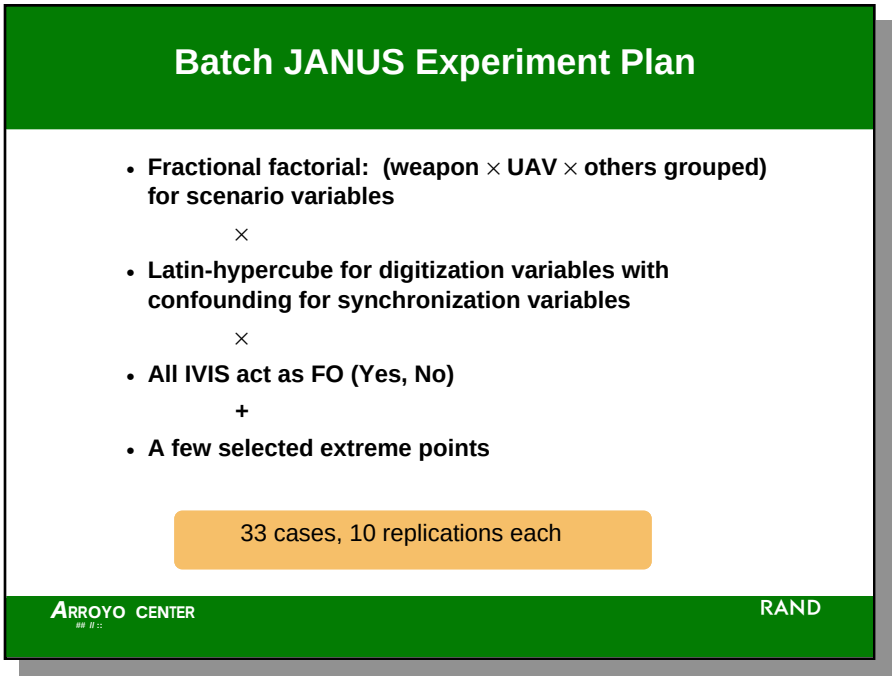
This is an approach that allocates the model runs to provide varying amounts of information on the variables depending on expert judgment of their potential importance. For additional design classes (not considered here), see Dewar et al. (1996).



For our example we needed expert information on the above variables. These variables and their values were gleaned from discussions with TRAC analysts on their experiences in the JANUS portion of FD, a variety of active duty personnel, and RAND's JANUS team. Recall that our research strategy is to see whether digital IVIS CFFs result in improved MOEs over a range of plausible scenarios. Since we could not explicitly change the scenario (i.e., move to a different terrain, force mix, etc.), we varied the following scenario variables: weapon type, UAV used, weather conditions, blue force level, sensor heights (to check for terrain effects), and multiple firing delays and probabilities of kill. Weapon type and UAV were believed *a priori* to be potentially causal and were thus classified as critical.

JANUS does not explicitly model IVIS and other digital equipment. Therefore, we had to characterize some of the effects of digitization in terms of variables that JANUS does represent. Variables that may be sensitive to the improved situational awareness that digitization is supposed to provide include planning delays, the probability that the fire support officer (FSO) correctly correlates incoming CFFs with previous CFFs from other platforms, and the timing and speed of advance of different units. When we made these runs it was not clear whether the FSO would actually be better off with all this information, or how accurate that information would turn out to be in the field. Ideally, information on planning delays and probability of correct correlation would be gathered from the live and virtual simulations. The variable of primary interest is whether or not all IVIS vehicles act as a forward observer (FO), i.e., whether all IVIS units make digital CFFs.

To run a full-factorial experiment on just these variables requires 124,000 × n runs, where n is the number of runs per setting. Not all two-value factors are shown explicitly in the slide, for there are multiple firing delays. It was not feasible to make 124,000 × n runs, so an alternative design had to be used.



As is typical with most analytic efforts, time and processing constraints caused us to run fewer cases than would be ideal—by at least an order of magnitude or two, perhaps even three! This is what we believe it typically requires to sufficiently explore combat models.

The key to testing our hypotheses is to efficiently vary as many variables as possible while ensuring that we can measure first-order effects for the critical variables and identify large effects for the other variables. Furthermore, we are looking for plausible disasters, so we look at a few extreme points.

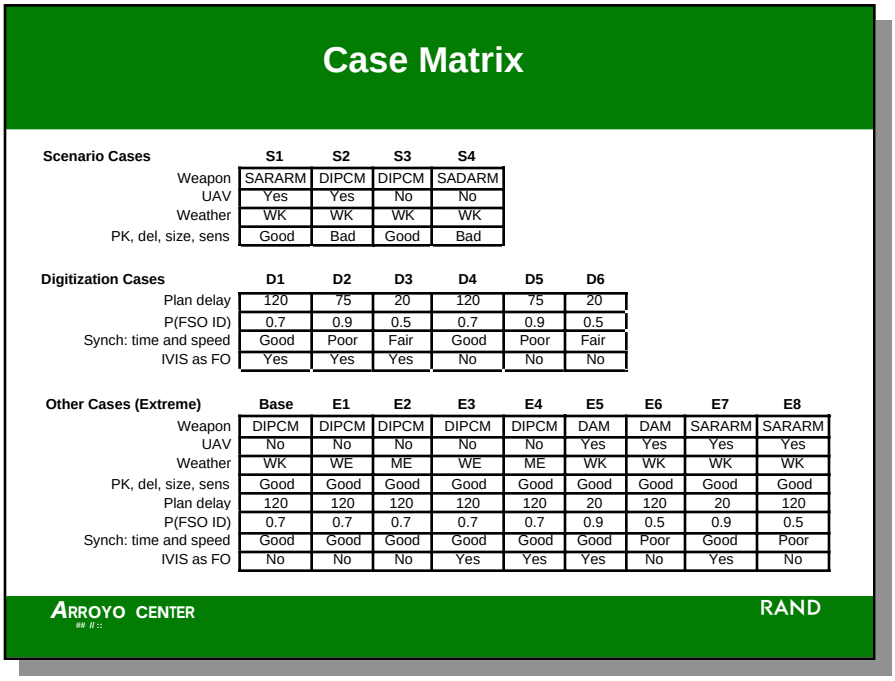
Our design used a fractional factorial on the scenario variables, with blue force level, sensor height, firing delays (two levels on several delays), PKs, and reload delays all grouped and varied together. That is, the two settings for each of these variables were varied together, thus confounding any resultant effects. Additionally, the extreme weather conditions and Damocles weapon were run only in the extreme cases. This design uses four different scenario input combinations and is used to measure main effects. Higher-level effects and interactions are confounded with the main effects and require more runs to evaluate.

The four scenario input settings were crossed with the digitization input combinations. The three digitization combinations form a (simple) Latin-hypercube on the digitization variables, with the time-advance and speed-of-advance variables grouped. This too allows us to estimate the first-order effects of these variables.

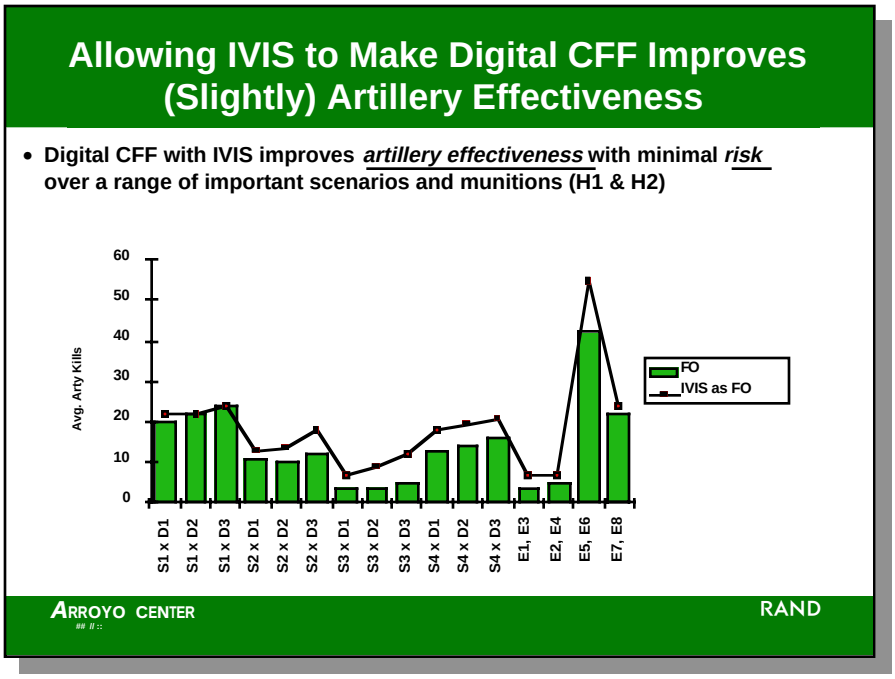
The twelve combinations of scenario and digitization input variables were run with all IVIS vehicles acting as forward observers (FO) and only the scout vehicles acting as FO. Four additional extreme points were taken, each one with and without all IVIS vehicles acting as FO. The total resulted in 16 contrasts (plausible cases) on whether or not all IVIS act as FO.

Ten replications were taken in each design cell. Since JANUS has stochastic elements (e.g., acquisition and attrition), replication is necessary to determine whether differences in model outcomes are real (with respect to the model—not necessarily the real world) or the result of random variation. Ten was selected based on processing constraints, the expected signal-to-noise ratio, and JANUS setup constraints. The JANUS setup constraints made it impossible to automatically generate new cases; thus we could examine many fewer distinct cases than we would have liked. In a perfect world, given 330 runs were available, we would have reduced the number of replications and examined additional cases.

More details on the design are contained on the next slide.

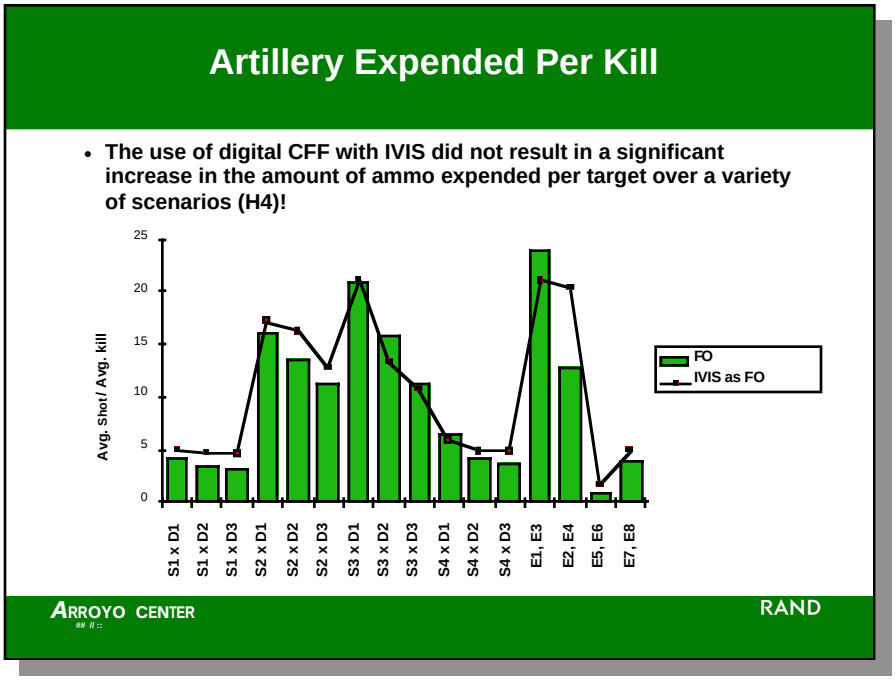


This is the case matrix for the batch JANUS runs. It consists of four scenario levels, six digitization levels, and eight extreme cases. The four scenario levels were run against each of the six digitization levels for a total of 32 cases (counting the eight extreme cases). This allowed us to evaluate the decision on whether or not to allow digital CFF over 16 contrasts of plausible scenarios. Each of the 32 settings (plus a baseline) was run ten times. A typical run required about an hour on a SPARC 2. Furthermore, it took several weeks to modify the data sent from TRAC-WSMR to accommodate RAND's version of batch JANUS.



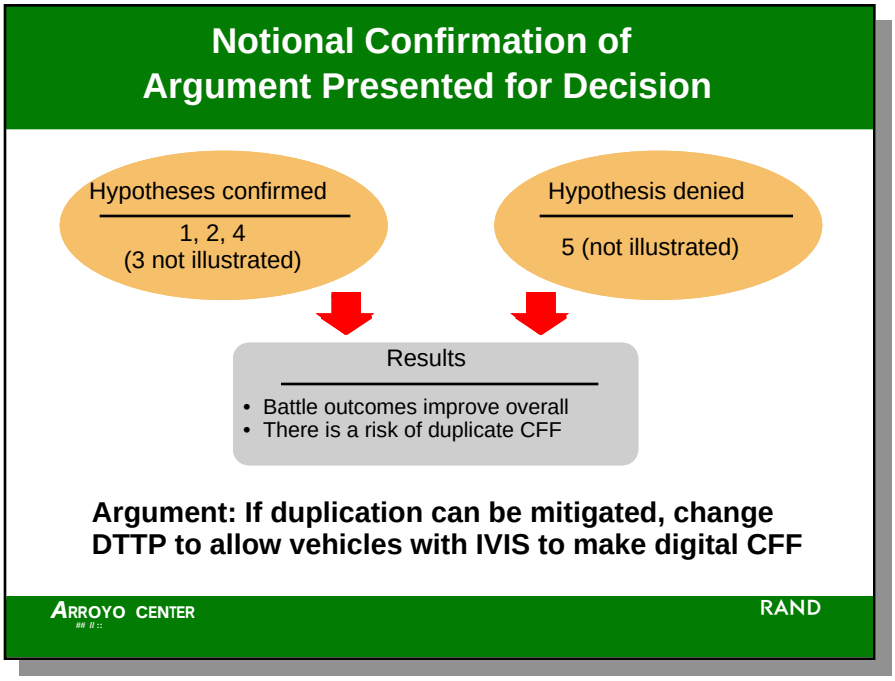
We now return to our argument on digital CFF with IVIS. Here, we plot the average of ten runs for 16 scenarios when all vehicles with IVIS make digital CFF and when only the FOs do. There is nontrivial variation around these averages, as discussed later. This is a total of 16 plausible contrasts. The bars represent the current situation, i.e., only FIST teams make CFFs. The line represents the situation where every IVIS vehicle makes digital CFFs.

For this notional example, over the 16 contrasts of various plausible combinations of scenario and digitization variables, digital CFF with IVIS improves artillery effectiveness. That is, there are consistently more indirect fire kills with the new DTTP. Furthermore, no plausible scenario was uncovered where the new DTTP might cause an unacceptable risk. Thus, we have evidence that supports the affirmation of hypotheses 1 and 2 without assuming that the model outcomes are reliable predictions of what might happen in actual combat. Of course, increased confidence would be obtained by investigating many more plausible scenarios with batch JANUS and information from the virtual and live simulations. In particular, the virtual and live simulations would be invaluable in establishing the credibility of JANUS and what different constructive scenarios should be examined.

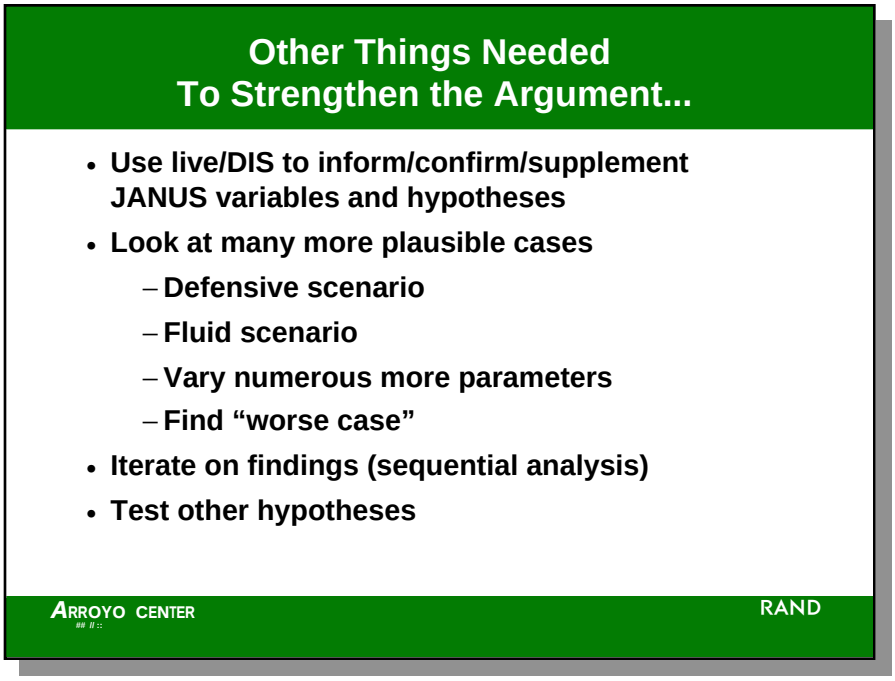


This chart shows how allowing all IVIS vehicles to make CFF (the line) compares to the current situation (the bars) with respect to artillery ordnance expended per indirect fire kill. Although the cases where all vehicles with IVIS make CFF typically expended more artillery, the average per kill is not significantly higher (in our judgment) over the range of cases. Thus, we have evidence that supports hypothesis 4 in our argument. Here too, live and virtual runs would be invaluable in determining whether or not JANUS is systematically biased. A bias could result in erroneous conclusions.

It may at first seem counterintuitive as to why more ammunition is expended per kill when all IVIS vehicles make CFF. The reason turned out to be a large number of additional CFFs. Since the FSO correlates the various tracks with error, there are inevitably unintentional redundant engagements. As a result, with all IVIS vehicles making CFF there were more CFF, more opportunities for unintentional redundant engagements, more kills, more ammunition expended, and, on average, slightly more artillery shots per kill. A notable exception is the contrast between E4 and E2. This is the baseline case in a Middle Eastern environment. For this situation (good weather with flat terrain) there is little difficulty in detecting threats. Thus, the extra reports do not dramatically affect what and when targets are detected; however, since many Blue vehicles can detect a given Red threat, there were often many redundant CFFs—and many more shots per indirect fire kill. For situations with excellent intervisibility, if the FSO correlates with error, one may want to limit the number of people who can make a CFF.



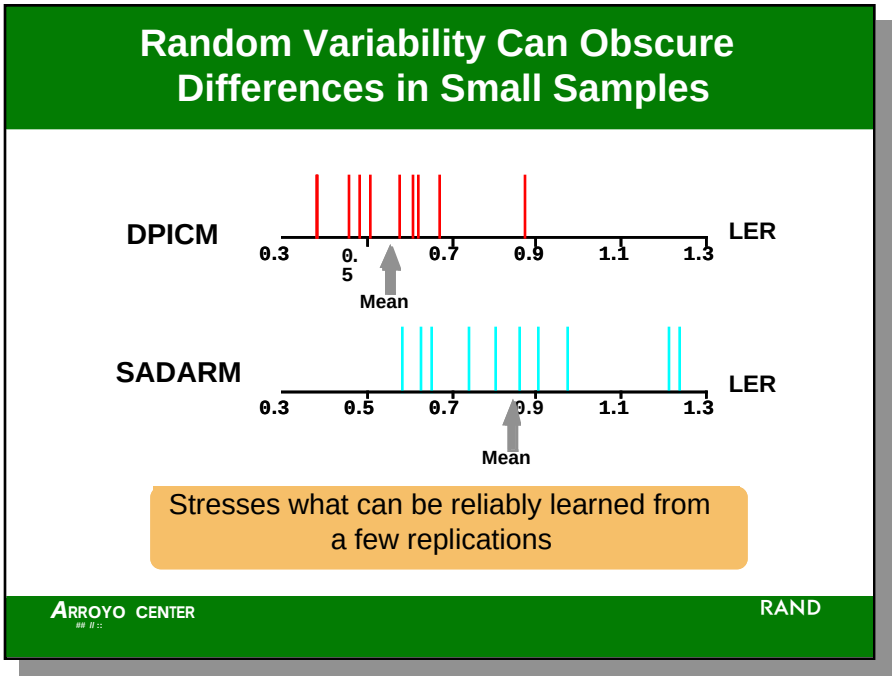
Let's postulate the following results of the experiments: All the hypotheses except number 5 (there would not be significant duplicate CFF causing a misallocation of artillery resources) are confirmed by the experiments. Remember, we have not actually done the live and virtual experiments required for hypotheses 3 and 5! Thus, the argument would hold (a decision could be credibly made) if procedures can be devised to mitigate duplicate CFF. Thus, the result of the analysis is to argue that if procedures to mitigate duplicate CFF can be devised, we should change the DTTP to allow all vehicles with IVIS to make digital CFF. Furthermore, the decision is externally traceable and the value of additional experimentation easier to determine.



If our exemplar analysis were part of the actual Focused Dispatch effort, the argument we are using would need to be strengthened. Foremost, it needs the live, DIS, and interactive JANUS runs to inform, confirm, and supplement the batch JANUS runs. This includes data to evaluate hypotheses we could not evaluate and to calibrate and otherwise inform the batch JANUS runs, especially related to human performance, as well as to make qualitative assessments about the similarity of the battles. For example, do battles in the different types of simulation "evolve" the same way in terms of maneuver, intensity, and attrition? If not, why not?

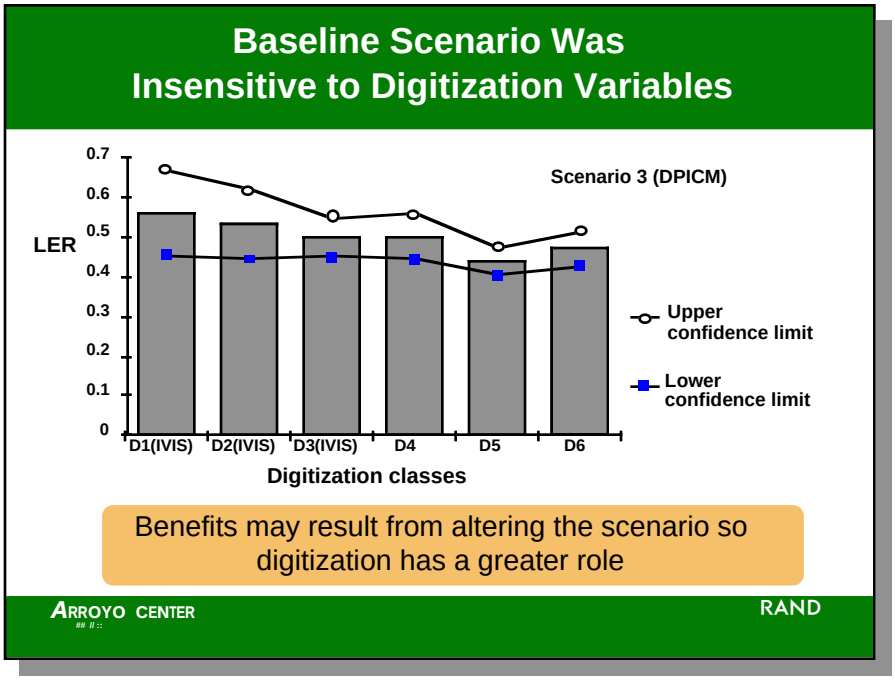
Furthermore, our analysis would benefit from looking at many more plausible cases. Those in the forefront of exploratory modeling typically run many thousands of cases (or more); see Bankes (1993). Of course, doing so requires a model that can automatically generate multiple cases. Moreover, there is a tension between model detail and the number of cases that can be examined.

Of particular interest are dramatically different scenarios, such as defensive and fluid scenarios. For the latter, the ability of Blue to maintain a coherent picture while the opposition is confused could be decisive. By contrast, scenarios like the deliberate attack in the earlier example are dependent more on firepower and terrain than the enhanced situational awareness that digitization promises.

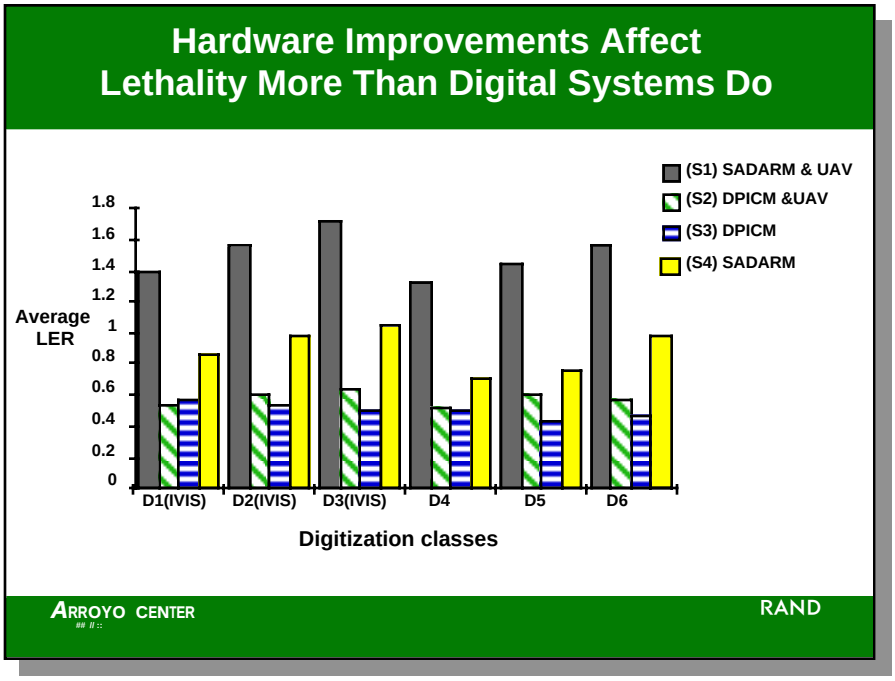


The batch JANUS runs dramatically demonstrate that there is significant variability in this scenario. This slide displays the Lethality Exchange Ratio (LER) for two different cases—the base case and DPICM replaced by SADARM. The primary difference here is a result of the choice of weapon. Over the ten replications for each combination it is clear that SADARM has a significantly higher LER. By a simple t-test we can determine that if the two weapons really are not different, then the probability of seeing such a difference by chance is less than one in a thousand. However, if only one or two replications were taken for each case, an erroneous conclusion could easily have been reached. Furthermore, the *LERs can vary by a factor of two* for both DPICM and SADARM. This substantial range in variation was typical across the cases we examined.

The variation is the result of stochastic elements in JANUS, such as detections and weapon effects. There are several random elements that JANUS does not explicitly represent—such as situational awareness, decisions, equipment and system availability, and human elements. This suggests that there might be even more variation in virtual or live experiments. The bottom line is that the potential variations due only to randomness stress what can be reliably learned from a couple of replications. The only AWE analysis tools allowing for more than a few samples are the noninteractive constructive models such as batch JANUS: thus their importance to the AWE process.

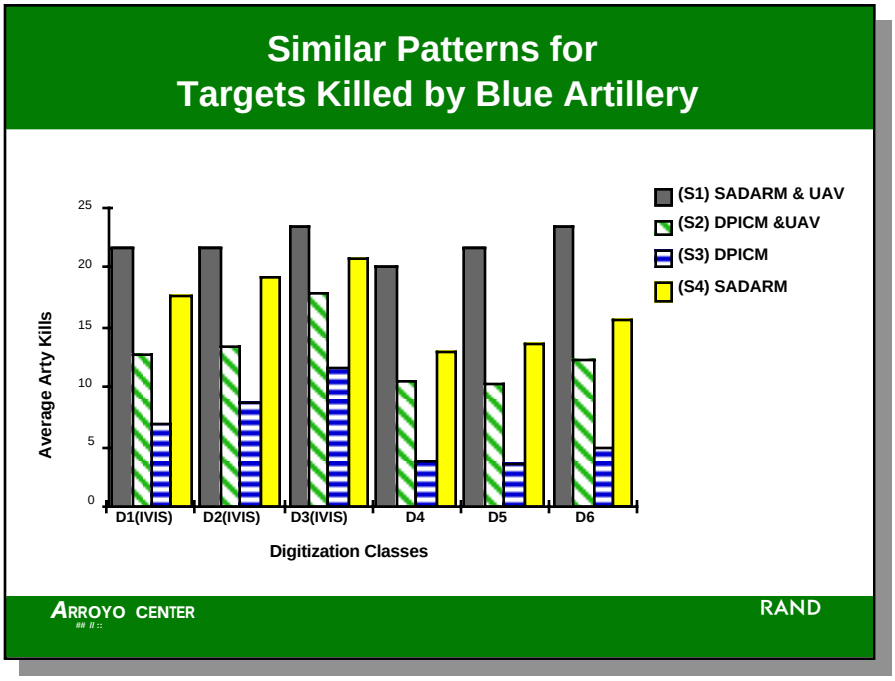


This graph displays how LER varied over the six digitization classes for scenario class 3. Scenario 3 is closest to the baseline scenario we received from TRAC-WSMR. The bars display the mean of ten runs, while the lines show the 90% confidence intervals. Over six different combinations of digitization variables there is little variation in battlefield effectiveness. This general finding held over the other scenario settings. This suggests that perhaps the given scenario does not provide maximal analytic leverage for assessing the potential benefits from digitization. Noninteractive simulations, run early in the experimental process, have the potential to identify scenarios (read DIS / live experiments) from which the most information can be obtained. *Careful design of the scenario may be the most important aspect of the overall experimental design!* If scenarios are finalized without this sort of pre-analysis early in the research plan, one runs the risk of not being able to differentiate among alternatives—not because there is no difference, but because differences are masked by the scenario.

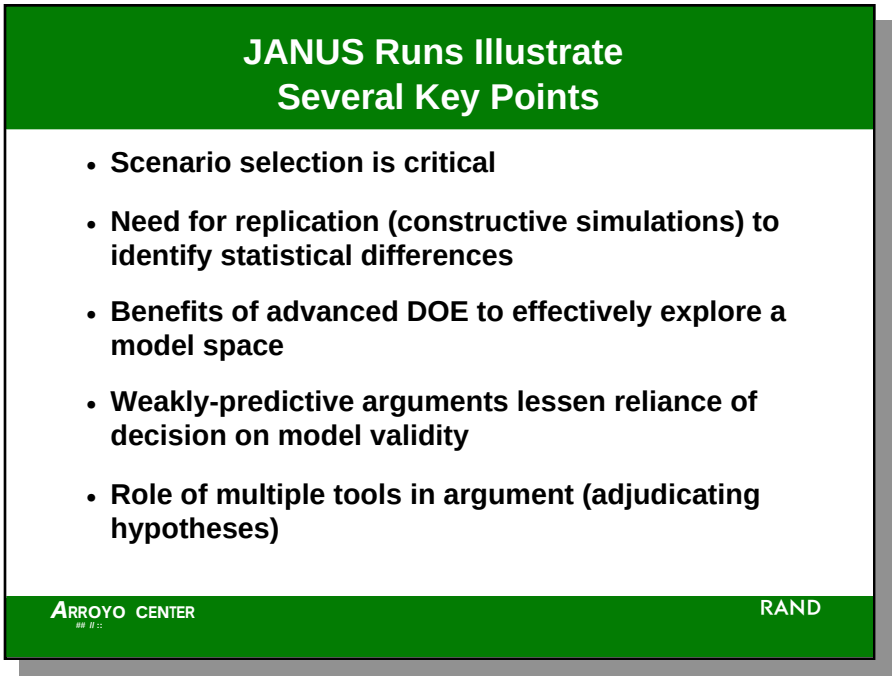


This graph displays how LER varied over the six digitization classes for all four scenario classes. The bars display the mean of ten runs. Comparisons within each digitization class show how the scenario variables effect LER. There are significant differences due to the scenario settings with enhanced performance achieved by the combination of advanced munitions (SADARM) and the use of UAVs. There is a smaller, but consistent, improvement due solely to the improved munition (SADARM). Comparisons among the six digitization classes show little effect due to the digitization classes.

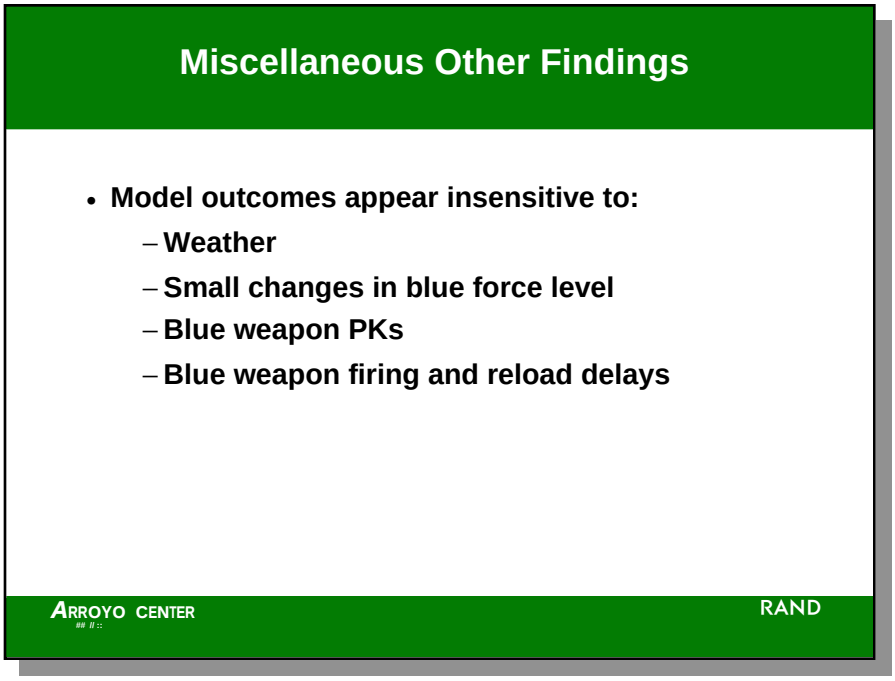
The few runs we made strongly suggested that hardware improvements, such as improved munitions and UAVs, do more for lethality in this scenario than do digitization variables—though the cases where all IVIS vehicles make CFF are generally a little higher. This too suggests that this might not be the best scenario for analyzing digitization-related DTTPs.



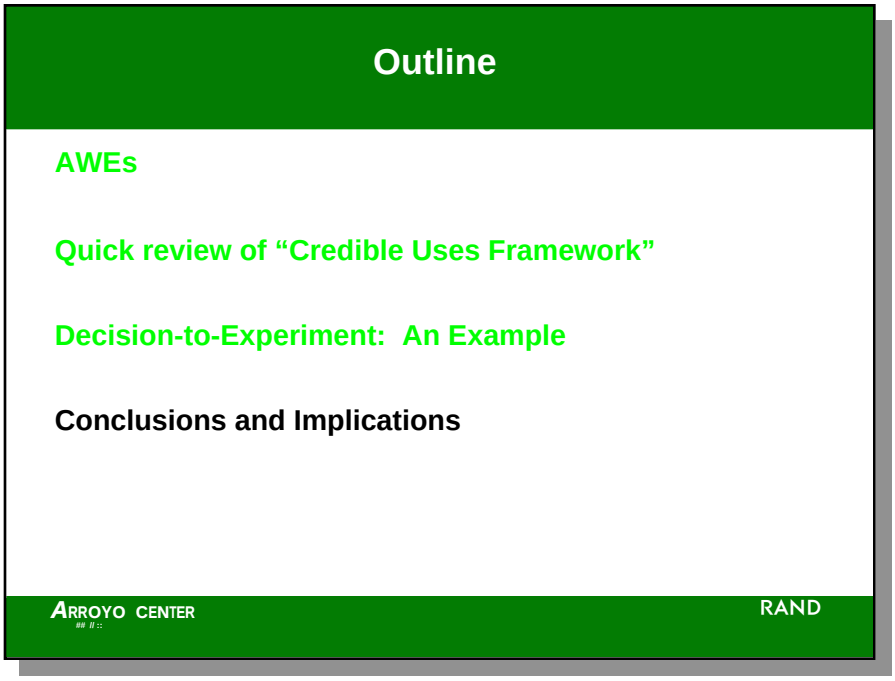
Our hypotheses are stated in terms of artillery effectiveness, while so far we have been using LER as a surrogate. It turns out that the important conclusions are consistent when the number of Red targets killed by artillery is used as the MOP.



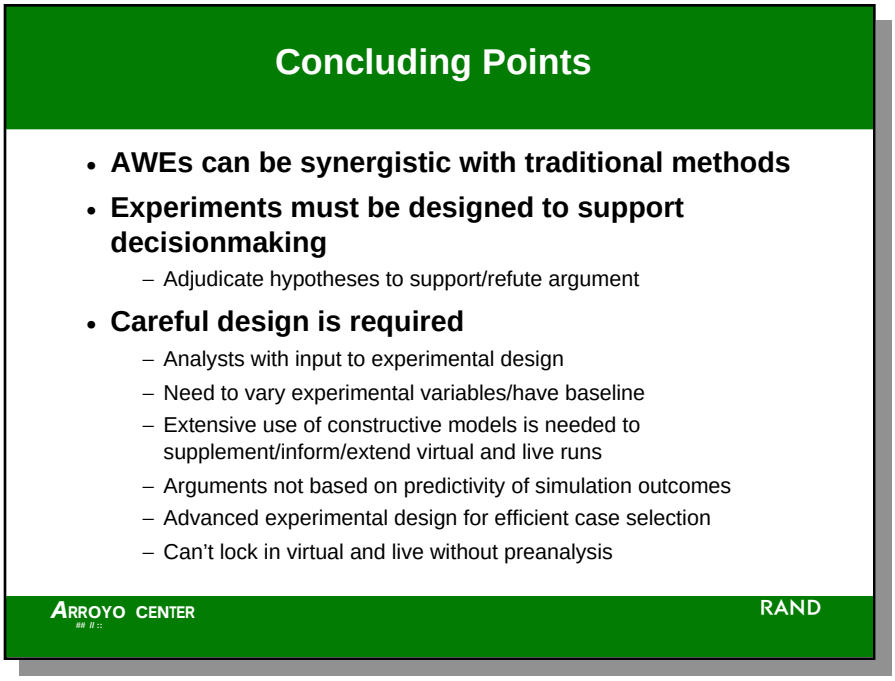
These few batch JANUS runs illustrate several interesting points, as shown above, relating to AWEs in addition to examining the DTTP decision. Here, "few batch JANUS runs" is relative to what we believe is typically necessary and possible for effective exploratory modeling.



While not central to our analysis, it is interesting to note that the results remained pretty consistent over the combinations of weather, blue force level, probabilities of kill (PKs), and weapon firing and reload delays.



We now summarize our conclusions and some implications for future AWEs.

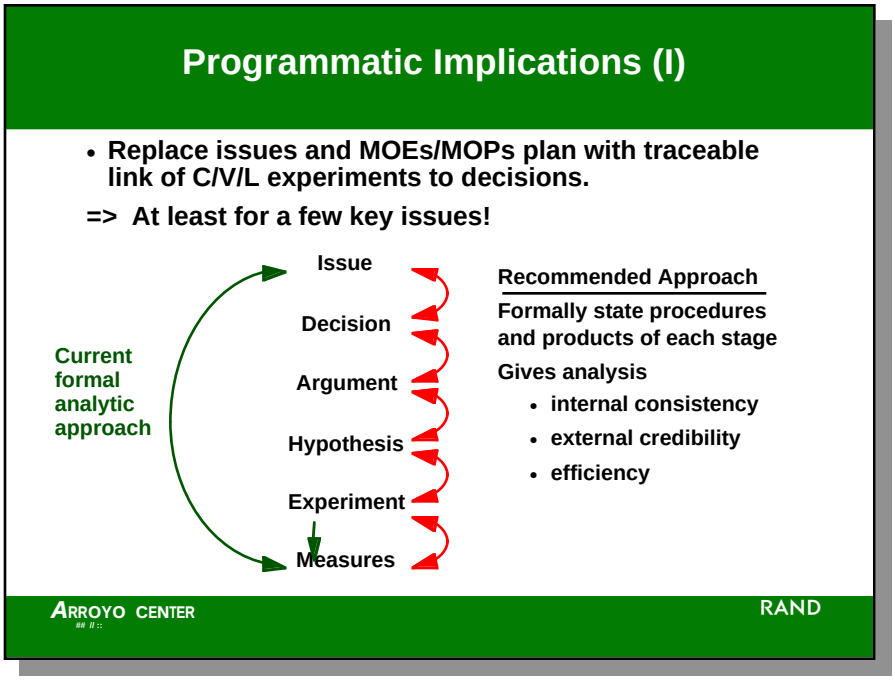


AWEs are a new way of doing analysis. While our briefing has focused on methods to improve AWEs, we believe that the live and virtual components can be synergistic with traditional approaches. The strengths of live and virtual simulations are that they use the real equipment and provide better data about the behavior of human beings on the battlefield. This is critical when looking at the effects of information systems—like digitization. It is something that closed constructive models do not do well. By themselves, however, virtual and live simulations are error prone, subject to simulation or network failure, and very expensive; they have numerous uncontrollable factors and allow only extremely small sample sizes. The promise for AWEs rests on research strategies that can effectively integrate the different types of tools.

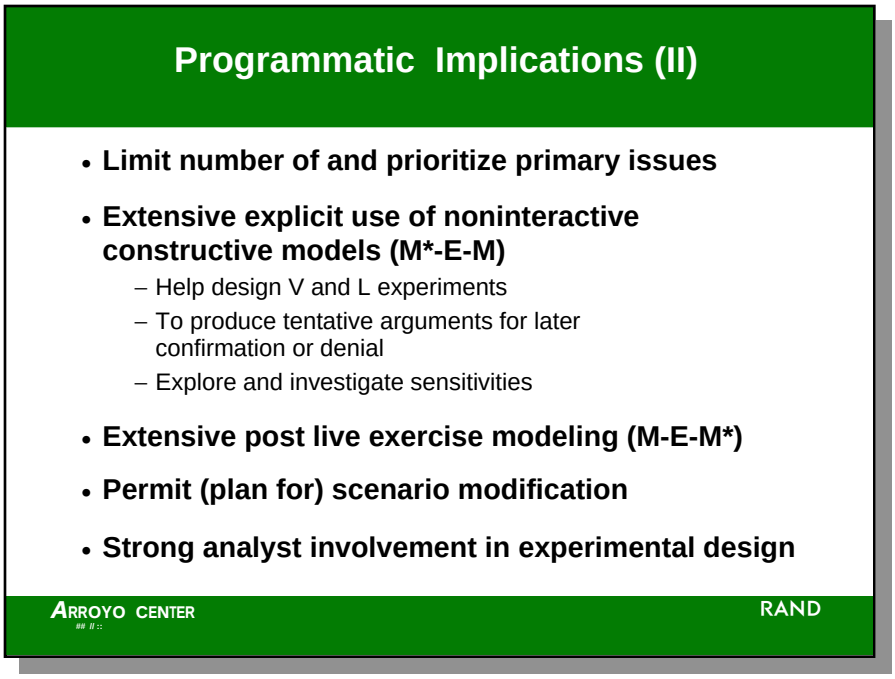
The potential benefits of AWEs cannot fully be realized if the analysis is based solely on data gleaned from training exercises. There must be an objective and traceable link from experimental results to decisions on important issues. At least a portion of the experiments must be designed to adjudicate hypotheses that inform arguments that will make the decisions.

Maximizing the analysis yields from AWEs requires significant up-front analysis input. Careful design is required in choosing the scenarios and variables for study. The virtual and live scenarios should not be fixed without some preanalysis to determine their analytic potential. Because combat models are not validated in any strong sense, i.e., they are at best weakly predictive of real outcomes, careful argumentation is needed to credibly affect decisions. Typically, a large number of plausible cases need to be examined. Constraints on other tools suggest that batch constructive models are the only way to do this.

Therefore, AWEs need to formally include batch simulations in their analysis toolkit. Even then, advanced experimental design will be necessary to obtain the necessary information from the constructive runs.



The number-one programmatic recommendation for future AWEs is that they formally link the experiments to potential decisions on important issues, as illustrated in this briefing—at least for a few key issues. Current plans jump directly from issues to measures that inform on the issue. Analysts are informally doing parts of this process now. However, explicitly stating the procedures and products of each stage ensures internal consistency, provides external traceability and credibility, and promotes efficiency.



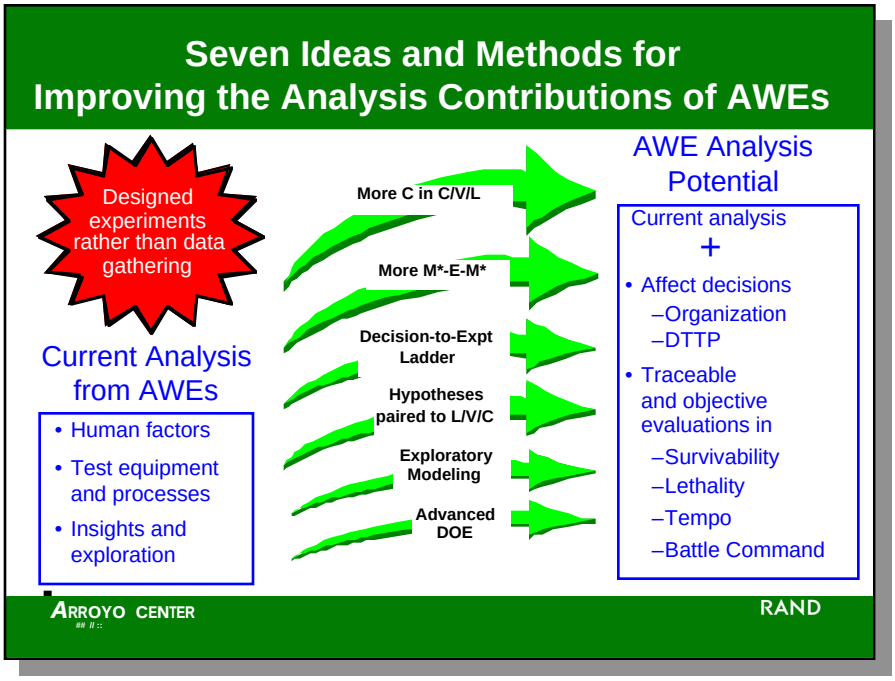
There are several other changes that we believe, based on our observations, previous research, and fundamental principles, could strengthen AWEs. It is important to limit the number of primary issues that the experiments will be designed to address. How many can be addressed, and how forcefully, falls naturally out of the approach we recommend (i.e., the decision-to-experiment ladder). Issues of lesser importance can still be addressed much as they are today, that is, by looking at the MOEs / MOPs that relate to them.

Getting the most analysis out of the AWEs requires careful use of the analysis tools. More explicit use of noninteractive constructive models would buttress some of the weakest points of the other tools, such as sample size, control, and lack of repeatability. Constructive runs can help design the scarce HIL runs, adjudicate some hypotheses by themselves, and be used for exploration or sensitivity analysis.

Many of the AWEs culminate at (or shortly after) the live training exercise (analysis experiment). Since the live experiment can provide important information on both human factors and real equipment performance, it is desirable to use this information for extensive post live exercise modeling—mainly with constructive models. The constructive models will be more credible when they can incorporate information and data obtained from the live simulations. Additionally, they can be used to explore issues (insights) that are obtained from the live simulation. Furthermore, they can be used for extensive sensitivity analysis. Planning for several months of post live simulation analysis could greatly enhance the final analysis products. The selection of scenario(s) can have a great impact on the analytic leverage of an AWE. Thus, the use of

initial analysis to shape the scenario could add to the information obtained from the experiments.

Finally, credible analysis of battlefield effectiveness requires strong analyst participation in all phases of the experiments. This is essential if the AWEs are to help make decisions on the organization and employment of forces, as well as demonstrate improvements in survivability, lethality, etc.



Contained in the preceding slides are seven ideas and methods that we believe can improve the analysis contribution of AWEs. These also relate to other "few event" HIL experiments, such as SIMNET experiments.

(1) Foremost among the ideas is the overarching theme that the experiments (at least some of them) must be carefully designed. Design includes scenario and variable selection, as well as what data to extract.

(2) Constraints imposed by humans-in-loop simulations (i.e., virtual and live) imply that closed constructive models are the only simulation tool that can vary many factors and give analysts the control required for reliable and repeatable conclusions. Thus, there should be expanded use of traditional methods (i.e., constructive simulations) in AWEs.

(3) Constructive simulations can be effectively used by applying the principles in the M-E-M process, that is, (a) more up-front modeling to establish preliminary conclusions and focus the live and virtual runs, and (b) more post–live experiment modeling with the constructive models informed by the virtual and live simulations.

(4) Designing the experiments that will best affect decisions can be done efficiently by using the decision-to-experiment ladder. Specifically, the experiments are run to adjudicate hypotheses which provide the foundation of arguments that will make the decisions.

(5) The different classes of simulations (i.e., C / V / L) have different strengths. Evaluating hypotheses may require information from one or several types of

simulation. Which simulations or combination of simulations are used, and their ordering, is determined by the hypothesis to be tested.

(6) Exploratory Modeling is a research methodology that uses computational experiments to analyze complex and uncertain systems. It is applicable to AWEs, which rely heavily on weakly predictive combat simulations.

(7) Exploratory modeling typically requires that a large number of “plausible” cases be examined. Even with constructive models, the combinatorics constrain what can practically be varied. Advanced DOE methods can assist in efficiently using these models.

BIBLIOGRAPHY

Bankes, S. C., "Exploratory Modeling for Policy Analysis," *Operations Research*, Vol. 41, No. 3, May–June 1993.

Bankes, S. C., "Exploratory Modeling," in *Encyclopedia of Operations Research and Management Science*, Saul I. Gass and Carl M. Harris (eds.), Boston: Kluwer Academic Publishers, 1996, pp. 203–205.

Dewar, J., J. Gillogly, and M. Juncosa, *Non-Monotonicity, Chaos, and Combat Models*, Santa Monica, CA: RAND, R-3995-RC, 1991.

Dewar, J., S. Bankes, J. Hodges, T. Lucas, D. Saunders-Newton, and P. Vye, *Credible Uses of the Distributed Interactive Simulation (DIS) System*, Santa Monica, CA: RAND, MR-607-A, 1996.

Hodges, J. S. and J. A. Dewar, *Is It You or Your Model Talking? A Framework for Model Validation*, Santa Monica, CA: RAND, R-4114-A/AF/OSD, 1992.

TRAC-WSMR, *Draft Focused Dispatch Experiment Assessment Plan*, White Sands, NM, January 1995.