Project AIR FORCE

IMPLICATIONS FOR MODEL VALIDATION

of Multiresolution, Multiperspective Modeling (MRMPM) and Exploratory Analysis

James H. Bigelow and Paul K. Davis

Prepared for the
UNITED STATES AIR FORCE

RAND

The research reported here was sponsored by the United States Air Force under Contract F49642-01-C-0003. Further information may be obtained from the Strategic Planning Division, Directorate of Plans, Hq USAF.

ISBN: 0-8330-3482-0

Published 2003 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact Distribution
Services: Telephone: (310) 451-7002; Fax: (310) 451-6915; Email: order@rand.org

Preface

This monograph is a significantly revised version of an invited presentation given at the Workshop on Foundations of Modeling and Simulation (M&S) Verification and Validation (V&V) in the 21st Century, which was held at the Applied Physics Laboratory of The Johns Hopkins University on October 22–24, 2002. The workshop had many sponsors but was of special interest to the Defense Modeling and Simulation Office (DMSO). The monograph is oriented toward working analysts and modelers concerned with defense issues, although the ideas are more general and draw upon examples from other domains.

The research reported here was sponsored by the Air Force Research Laboratory (AFRL) and conducted within the Aerospace Force Development Program of RAND's Project AIR FORCE.

Project AIR FORCE

Project AIR FORCE (PAF), a division of RAND, is the U.S. Air Force's federally funded research and development center for studies and analyses. PAF provides the Air Force with independent analyses of policy alternatives affecting the development, employment, combat readiness, and support of current and future aerospace forces. Research is performed in four programs: Aerospace Force Development; Manpower, Personnel, and Training; Resource Management; and Strategy and Doctrine.

Comments are welcome and may be addressed to either of the authors:

James H. Bigelow
bigelow@rand.org

Paul K. Davis, Project Leader
pdavis@rand.org

Contents

Preface iii

Figures vii

Tables ix

Summary xi

Acknowledgments....................xiii

Acronyms and Abbreviations xv

1. INTRODUCTION 1
   Objective 1
   Some Background Philosophy....................1
     Multiresoluton, Multiperspective Model Families and Validation....................1
     Comprehensibility and Explainability 3
     Dealing with Uncertainty 3
   Some Definitions and Distinctions....................4
     Elementary Definitions 4
     Definitions for Different Types of Low-Resolution Models 5
   Structure of This Monograph 6

2. VALIDATION-RELATED REASONS FOR MULTIPLE LEVELS OF RESOLUTION AND EXPLORATORY ANALYSIS 8
   Analysis Supporting Decisions on Forces and Weapons....................8
   Top-Down, Multiresolution Thinking in Studies and Experimentation....................11
   Broad, Exploratory Analysis Enabled by Multiresolution Modeling.... 14
   Generalizing the Concept of Model Validation 15

3. CONSISTENCY AND VALIDATION....................20
   Resolution 21
   Aggregation and Disaggregation....................22
   Consistency 25
   Right and Wrong Answers....................27

4. MOTIVATED METAMODELS, EXPLANATIONS, AND THE IMPORTANCE OF A GOOD STORY....................30

Appendix

A. Using a Simple Model to Explain, Extrapolate from, and Provide Face Validity of Complex-Model Results....................35

B. Using a Low-Resolution Calculation to Check a High-Resolution Calculation 41

C. Selecting a Good Test Set of Detailed Scenarios 47

D. Illustrations of the Use of Consistency Definitions 53

E. Basing Extrapolation on a Story.. 58

F. Motivated Metamodels.. 61

Bibliography 65

Figures

1.1. Mutual Calibration of Multiresolution Models Using All Information Available 2
1.2. A Taxonomy of Low-Resolution Models 5
2.1. A High-Level-Capabilities Comparison for Which Three Scenarios Were Analyzed in Detail 9
3.1. Aggregation Causes Loss of Information 23
3.2. Notional Network of Variables 24
3.3. The Usual Definition of Consistency Between High- and Low-Resolution Models 25
3.4. A More General Definition of Consistency Between High- and Low-Resolution Models 26
A.1. Kills per Two-Missile Salvo vs. AFVs in Footprint 37
A.2. PEM Concepts 38
B.1. Replenishment Cost Factors, Traditional (Aggregate) Method 42
B.2. Multiechelon System for Processing Recoverable Items 43
C.1. Challenges Can Be Decomposed Hierarchically 49
C.2. Some Factors Considered in EXHALT 50
C.3. Factor Combinations that Allow 500 Red Vehicles to Be Killed Within 100 km 51
C.4. Factor Combinations that Allow 4,000 Red Vehicles to Be Killed Within 300 km 52
E.1. Calibration and Structural Uncertainty 60

Tables

2.1. Types of Validation Possible in Different Circumstances 17
E.1. Sample Dataset... 58
F.1. A Dose of Theory Improves the Metamodel.................................. 63

Summary

Objective

This monograph draws upon a number of the authors' past studies to illustrate with concrete examples how the methods of multiresolution, multiperspective modeling (MRMPM) and exploratory analysis relate to model validation. The studies themselves were not concerned with model validation, but rather with matters of strategic planning, defense planning, transformation-related experimentation campaigns, logistics, flood protection, and even models of adversary behavior. Nonetheless, the studies had implications for validation, and it is our objective to pull those implications together here.

The Domain of Relevance

Our conclusions are not particularly relevant to instances in which the models in question are solidly based in settled theory or empirical testing appropriate to the new application being envisaged. Rather, they apply when the models or their data are more afflicted with uncertainty. For example, one may not be entirely sure that the model is correct structurally, but one has a seemingly reasonable theory or a good deal of experience to support it. Or perhaps the model is solid, but the input data for the application are highly uncertain. Both of these circumstances are common in strategic planning and policy analysis. They are also common in military applications such as operations planning, acquisitions planning, and training. For example, no one has a "correct" model of war with all its notorious complications, and, even if such a model existed, it would have large numbers of uncertain inputs.

Broadening the Concept of Validation

In such cases, we believe that model validation should be construed quite differently than might be suggested by the usual definition of validity. A validation process might reasonably conclude by assessing the model and its associated databases as "valid for exploratory analysis" or "valid, subject to the principal assumptions underlying the model, for exploratory analysis." These assessments, of course, would apply only to the context in question. Our point is that it often cannot reasonably be concluded that a model and its data are “valid” but less reliably predictive. They may, however, be both useful and good in more-limited ways. Therefore, one should aspire only to more-limited assessments such as those indicated.

New Criteria for Assessing Models

We believe that when working within this troubled but common domain, it is particularly important for two criteria to be met in assessing a model (and its associated data) (page 16):

- The model should be *comprehensible* and *explainable*, often in a way conducive to explaining its workings with a credible and suitable “story.”
- The model and its data should deal effectively with uncertainty, possibly *massive uncertainty*.

Crucial Enablers

If one accepts these new criteria, then it becomes important to know how to meet them. Drawing upon our experience over the years, we argue here that crucial enabling capabilities are provided by the emerging theory and practice of (1) multiresolution, multiperspective modeling, including use of famili*es* of models and games (pp. 1–3), and (2) exploratory analysis (pp. 3–4). We call upon a number of past studies that illustrate their value for at least three important functions:

1. Extrapolating, generalizing, and abstracting from a small set of analyses accomplished with detailed models (i.e., moving from results of a few scenarios to broad conclusions, which often is facilitated by having lower-resolution models than those used for the few in-depth analyses) (pp. 8–11**)**.
2. Planning top-down, with decomposition from objectives through desired results (or effects) to tasks, which lends itself naturally to multiresolution, multiperspective methods (pp. 11–14**)**.
3. Providing a broad, synoptic assessment of a problem area, as is crucial in the capabilities-based planning emphasized by the Department of Defense (DoD)—i.e., using exploratory analysis to understand “envelopes” of capability (pp. 14–15).

In all of the studies discussed, the enablers also proved important to achieving a deep understanding of problems and communicating insights credibly to others (pp. 30–34).

Acknowledgments

The authors appreciate reviews by Dale Pace of The Johns Hopkins University's Applied Physics Laboratory and RAND colleague Richard Mesic.

Acronyms and Abbreviations

DMSO	Defense Modeling and Simulation Office
DoD	Department of Defense
M&S	modeling and simulation
MRM	multiresolution modeling
MRMPM	multiresolution, multiperspective modeling
VV&A	verification, validation, and accreditation

1. Introduction

Objective

Over the past decade, we and our colleagues have done considerable theoretical and applied work involving multiresolution, multiperspective modeling (MRMPM) and exploratory analysis,[1] in part to connect the worlds of strategic planning (e.g., development of the defense guidance and defense programs)[2] with the world of more-detailed analysis and experimentation for transformation of military forces, weapon systems, and doctrine.[3] We have also used MRMPM to model adversaries within studies of deterrence and compellence.[4] Many implications for model validation have emerged as byproducts. Our objectives in this monograph are to examine those implications and to convince readers of their practical importance.

Some Background Philosophy

Three themes run through this entire monograph: (1) the importance to validation of developing and exploiting multiresolution, multiperspective *families* of models; (2) the importance of having models and model-based analyses that are *comprehensible* and *explainable*; and (3) the importance of being able to *deal with uncertainty*. These warrant some discussion at the outset.

Multiresoluton, Multiperspective Model Families and Validation

Model validation is a multifaceted activity that draws upon empirical tests and historical experience, comparison with first-principles analysis, expert appraisal, and model comparisons. The focus here is on validation-related activities involving two or more models with different resolution. Sometimes, the high-

[1] See Davis and Huber, 1992; Davis and Hillestad, 1993; Davis and Bigelow, 1998; Davis, Bigelow, and McEver, 2001; and Davis and Bigelow, 2003. Our work has been sponsored by the Air Force Research Laboratory (AFRL), the Office of the Secretary of Defense (OSD), and the Defense Advanced Research Projects Agency (DARPA).

[2] Davis, 2002a.

[3] See Defense Science Board, 1998; Davis, Bigelow, and McEver, 1999; and Gritton, Davis, Steeb, and Matsumura, 2000.

[4] See Davis (ed.), 1994; Appendix G of National Academy of Sciences, 1996; or Davis, 2002b.

resolution model is an excellent representation of the real system, in which case it should be used to calibrate the low-resolution model. Often, however, both the high- and low-resolution models (and associated data) have both strengths and weaknesses. Indeed, the low-resolution model and its data (e.g., data on real wars rather than war games) may even be superior for some purposes. As a result, the correct image is one of *mutual* calibration, as suggested in Figure 1.1, which uses double-headed arrows in a multiresolution family.[5] The arrows may be regarded as information flow or as explanations. In some instances, the various models can be all part of a larger model: That is, within the larger model, one has knobs and switches allowing one to work at different resolutions. In other instances, the models are separate, but have a known relationship to each other (Davis and Hillestad, 1993).

We often refer also to multiresolution, multi*perspective* models (MRMPM) (Davis, Bigelow, and McEver, 2001). The multiple perspectives are valuable because information often comes in different representations and formalisms, and with different semantics. Having such alternative views and relating them to each other may be very useful for seeing "the different parts of the elephant," exploiting diverse types of data, and communicating across cultural barriers. Examples include using center-of-mass coordinates in physics; comparing both phenomenological and empirical models; and communicating between the worlds of combat operations and logistics, or of combat operations and command and control.[6] Having multiple perspectives is not just about having alternative names for variables. Rather, the different perspectives usually

RAND*MR1750-1.1*

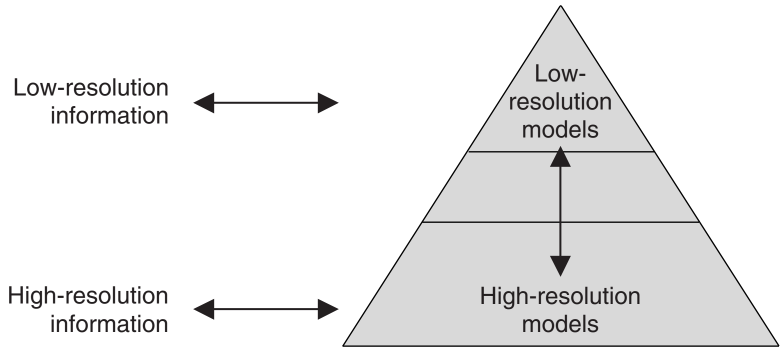

Figure 1.1—Mutual Calibration of Multiresolution Models Using All Information Available

[5] Adapted from National Research Council, 1997.

[6] Related issues are discussed under the rubrics of "hierarchical holographic modeling" in Haimes (1998) and "multifaceted modeling" in Zeigler (1984).

correspond to different decompositions of the problem or system. That is, there is no single "correct" decomposition or ontology.

Much of our past research has been devoted to learning how to turn these ideas into reality. We believe in using a range of models and games, including simple, low-resolution models that can be reduced to a formula; entity-level agent-based simulations with detailed physics; and seminar war games. For us, the family-of-models-and-games approach is not just a "nice idea," but something that can and should be more routinely adopted. It is by no means always needed, but it is needed in many instances. It is unusual, in our experience, for a complex problem to be addressed well without viewing it at different resolutions.

Comprehensibility and Explainability

Key elements of model validation often include establishing that the model's structure and behavior seem correct to subject-area experts (i.e., that the model and its behavior exhibit face validity). When models are being seriously evaluated in this way, the reviewers typically demand that the models be *comprehensible* and *their results explainable,* although the explanations needed vary greatly with context. Sometimes, in analyst-to-analyst discussions or in responding to a probing spot-check question by a client, the explanation needs to be detailed (e.g., "Well, the limiting factor is the radar's doppler filter, which is deeply embedded in old software"). At other times, and indeed in most discussions with higher-level clients, explanations need to be more abstract and lower in resolution (e.g., "It really comes down to a tradeoff between the speed of deployment versus the nature of what can be deployed. We can have either fast-response capability against small-to-moderate threats or slow-response capability against a larger threat"). It follows that we need combinations of low- and high-resolution models for explanation purposes.

Dealing with Uncertainty

The last of our themes is that dealing effectively with uncertainty is a recurrent challenge in studies, one with deep implications for the very concept of validation and validity. We emphasize the value of *exploratory analysis*, in which one studies a problem using the range of reasonable input assumptions and, indeed, assumptions about the form of the model itself. The result is often a *broad* view. For example, in Department of Defense (DoD) capabilities-based planning, the intention is to identify military capabilities that will be valuable in diverse scenarios and circumstances. Who knows the details of the future wars in which U.S. forces will be engaged?

Exploratory analysis is a more general and powerful form of sensitivity analysis (Davis, Bigelow, and McEver, 2001). It typically is best conducted with a low-resolution model, but that is most satisfactory when the low-resolution model is a reasonably valid and understood abstraction of a more-detailed view. Again, then, we see a linkage to multiresolution modeling.[7]

Some Definitions and Distinctions

Elementary Definitions

Having identified recurring themes, we must provide some definitions before we get more deeply into the discussion. Definitions are always a problem, because no source is consistently on target—whether it be the Merriam-Webster dictionary[8] for normal language or the Defense Modeling and Simulation Office (DMSO) for validation-related definitions. Nonetheless, one can try using official definitions. Drawing relevant items from the glossary on the DMSO website[9] and a recent DoD instruction,[10] we have

> *Validation*: The process of determining the degree to which a model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model.
>
> *Accuracy* of a simulation: The degree to which a parameter or variable, or a set of parameters or variables, within a model or simulation conforms exactly to reality or to some chosen standard or referent.
>
> *Fidelity* of a simulation: The degree to which a model or simulation reproduces the state and behavior of a real-world object or the perception of a real-world object, feature, condition, or chosen standard in a measurable or perceivable manner; a measure of the realism of a model or simulation; faithfulness. Fidelity should generally be described with respect to the measures, standards, or perceptions used in assessing or stating it.

[7] Multiresolution modeling can be seen as a fundamental need in many scientific domains. It has been discussed in the philosophical language of semiotics (Meystel, 1995) and in the down-to-earth context of intelligent systems (robotics) (see Meystel and Albus, 2002; and Davis and Bigelow, 2001).

[8] As an example, consider the word *sophisticated*, which we suspect most readers of this monograph use to mean something positive, as in "containing valuable detail and subtlety." The three principal definitions according to the Merriam-Webster OnLine dictionary (http://www.m-w.com), however, are all negative in connotation, as in "to alter deceptively; to deprive of genuineness, naturalness, or simplicity; and to make complicated."

[9] See www.vva.dmso.mil, the DMSO website, for a wealth of information on modeling and simulation (M&S), including a recommended practices guide for verification, validation, and accreditation (VV&A), bibliographies, and glossaries.

[10] See the May 2003 update of "DoDI 5000.61. DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)," at https://www.dmso.mil/public/library/policy/policy/i500061p.pdf. Accessed June 15, 2003.

Overall, the definitions betray their origin, i.e., in the context of hard-science models, for which it is meaningful to talk about comparing models to the real world and to precise empirical data about that real world. Usually, people have in mind that a valid model is one that makes sufficiently reliable predictions. As discussed in more detail later, however, the definitions become troublesome when applied to models and data beset with uncertainty, some of it inherent. At the end of Chapter 2, we suggest specific language to help deal with these troubles candidly.

Definitions for Different Types of Low-Resolution Models

This monograph is concerned with the role of multiresolution-model families in validation. For the sake of expositional simplicity, we imagine that we deal with models at only two levels of resolution, low and high. However, we distinguish among types of low-resolution models (Figure 1.2). In our nomenclature,

> *Empirical models* are models developed strictly from data (e.g., curve fitting).
>
> *Metamodels,* or models of models, are developed by applying statistical methods to data generated from high-resolution models (referred to as the *object models*).[11]
>
> *Theory-based low-resolution* models are models built from the viewpoint of phenomenology, but with low-resolution concepts. For example, the volume of a pond might be estimated roughly as the product of its average breadth, width, and depth.

RAND*MR1750-1.2*

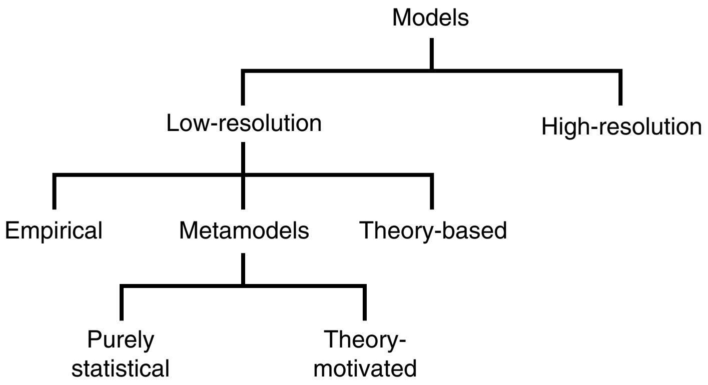

Figure 1.2—A Taxonomy of Low-Resolution Models

[11] The DMSO glossary defines a metamodel as a "model of a model," as do we, but goes on to present a relatively narrow definition relevant more to software engineering than to our purposes here.

We distinguish further between purely *statistical metamodel*s (often referred to as response surfaces) and *theory-motivated metamodels*. The former are obtained by applying statistical methods to the output of experiments conducted with high-resolution models. No physical insight is necessary, and it may not even be desired by the analyst; he is doing only data analysis. Theory-motivated metamodels, to which we shall subsequently refer as *motivated metamodels* (see Davis and Bigelow, 2003) are obtained by using physical and behavioral reasoning to suggest the structure of the model, after which coefficients, exponents, and correction factors are determined using statistical analysis of high-resolution model runs.

Another term appearing in the title of this monograph is *exploratory analysis*. Exploratory analysi*s* is an analysis strategy that focuses on breadth and a synoptic view, rather than depth. It is best done with low-resolution models, with subsequent "zooming" to examine particular issues at higher resolution. Exploratory analysis is used when dealing with massive uncertainty. It is used, e.g., to develop *robust* strategies, as emphasized in Davis, Gompert, and Kugler (1996), which was written to encourage what is now called capabilities-based planning.[12] Exploratory analysis may be accomplished with parametric exploration, probabilistic exploration, or a hybrid approach.

Structure of This Monograph

With this background, the monograph proceeds as follows: Chapter 2 explains why working at multiple levels of resolution and using exploratory analysis is important to model validation. It also suggests a broadened concept of validation. Chapter 3 discusses what it means to maintain *consistency* across levels of resolution, which is closely related to model validation. Chapter 4 discusses the importance of having a low-resolution model (perhaps a metamodel) that provides a credible and insightful explanation—i.e., a story or theory about causal relationships. It also suggests that the form of metamodels should be "motivated" by drawing upon knowledge of the subject area to assure

[12] Exploratory analysis has developed over many years from a concept (see Davis and Winnefeld, 1983), through what we then called multiscenario analysis considering scores of cases (see Davis, 1988), to a more comprehensive examination over large scenario and case spaces (see Davis, Bigelow, and McEver, 2001; or Davis, 2003). Fox (2003) describes exploratory analysis with RAND's Joint Integrated Contingency Model (JICM). Progress in exploratory analysis has been made possible by a combination of technological developments and theoretical work. See also work on "exploratory modeling" in Bankes, 1993; Bankes, 2002; and Lempert, 2002. The two streams of work (exploratory analysis and exploratory modeling) began differently, but the underlying philosophies are ultimately similar.

that such a story is built in. Such considerations are crucial if models are to have face validity. Appendices provide concrete examples to illustrate points made in the main text.

2. Validation-Related Reasons for Multiple Levels of Resolution and Exploratory Analysis

Multiresolution analysis is fundamental, not merely nice to have, in at least three generic classes of analysis:

1. Extrapolating, generalizing, and abstracting from a small set of detailed analyses.
2. Planning and analyzing from a top-down perspective, perhaps with decompositions through intermediate results to be obtained (sometimes called *effects*) down to tasks to be performed. In other cases, the top-down approach may begin with a broad experimental issue and decompose into subordinate issues until narrow and well-defined experiments are specified.
3. Providing broad, synoptic assessments.

We shall consider these in turn, using specific examples related to, respectively, a choice of weapon systems, deciding how to carry out a program of transformation-related experimentation, and broadly assessing capabilities for interdiction of enemy ground forces with long-range fires.

Analysis Supporting Decisions on Forces and Weapons

To illustrate the issue of generalizing from a small set of detailed analyses, consider a military client who faces the question, "Given a specified amount of funding, should my organization buy weapon system A, B, or C?" We perform the study, which involves the heavy use of models, and find that the answer is, "Buy system B." The client will not usually be content with an unsupported recommendation, so we need to construct a persuasive argument. The legs of the argument may be the following:

- *Leg One*: System B is more combat effective than A or C in scenarios 2 and 3, and similar to A for scenario 1.
- *Leg Two*: Scenarios 1, 2, and 3 together constitute an adequate test of combat effectiveness. That is, the conclusions drawn from these can be generalized

to all scenarios for which the new capability is needed, as can be seen from an abstracted (low-resolution) depiction of results in which the scenarios appear as merely representative cases (e.g., Figure 2.1).

- *Leg Three*: System B has adequate noncombat performance (e.g., with respect to logistics, mobility, cost, and production schedule).

Consider Leg One above. Analysts typically assess the combat effectiveness of each weapon system in each of a number of scenarios, using a high-resolution simulation model. Then they construct credible but simplified explanations in

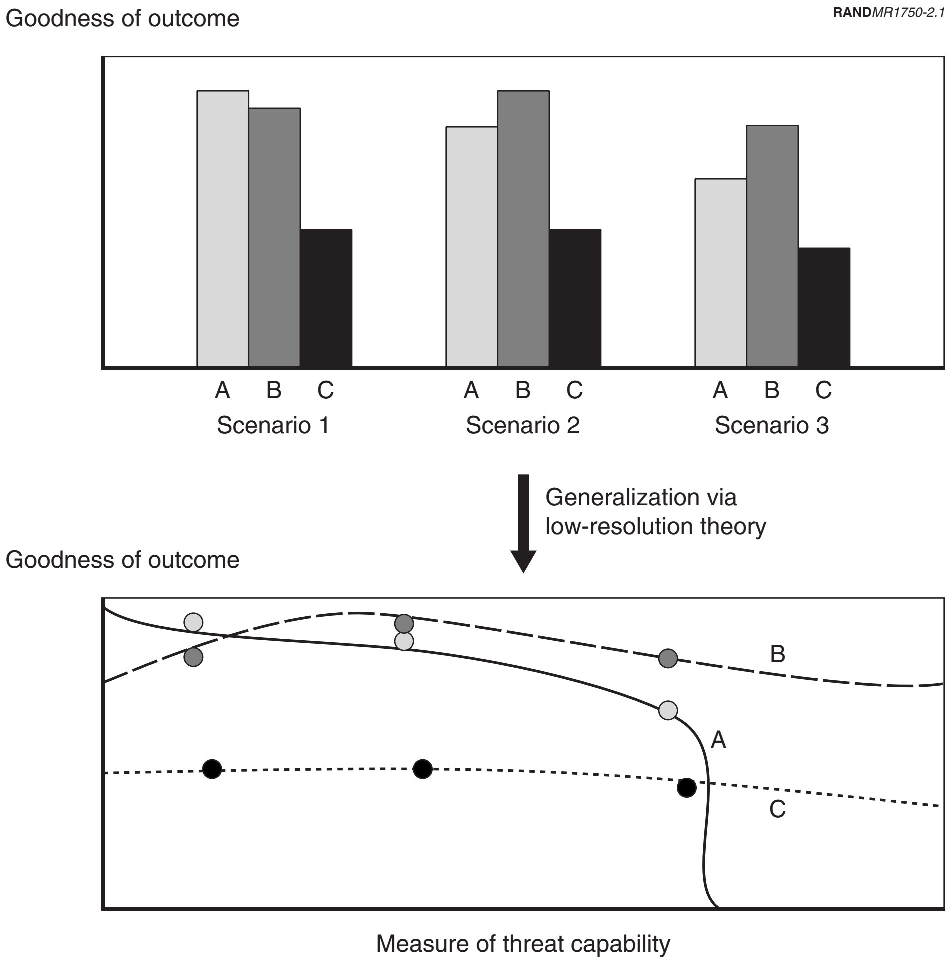

Figure 2.1—A High-Level-Capabilities Comparison for Which Three Scenarios Were Analyzed in Detail

the form of stories about why the high-resolution model behaves as it does.[1] This helps to establish the face validity of the high-resolution model and analysis. It is often a relatively small step to turn such a story into a formalized low-resolution model, but even if this step is not taken, the use of simple stories in these ways is an example of working at multiple levels of resolution.

Next, consider Leg Two. The analyst must convince the client that the results generalize and do so concisely. Unfortunately, even a modest combat simulation may have tens of megabytes of inputs and outputs. No analyst will show the client the raw inputs and outputs—at least, not more than once. Instead, he will select and summarize. He may present only two or three briefing charts, with the claim that this small amount of information captures the essence. This requires him to extrapolate beyond the relatively few high-resolution cases that were run (nine in the example of Figure 2.1). Even if the high-resolution model is considered valid, it is necessary to establish that the generalizations are correct. The analyst may then appeal to the simple stories he constructed to develop confidence in the high-resolution model and its datasets. That is, he may use the stories for himself, not merely for communication. A much better way to proceed, however, is for the analyst to be able to simultaneously discuss results with an abstracted, lower-resolution model and go into details selectively—for purposes of both credibility and explanation. Figure 2.1 shows both depictions notionally. Each point represents a case worked through in detail using the high-resolution model, but the *curves* for A, B, and C were generated by a lower-resolution model mutually calibrated with the high-resolution model. Thus, the "big picture" is seen, along with results of detailed analysis. Note, for example, that capability A is expected to be useless for a large threat, even though no such cases were run—merely extrapolating from the detailed cases is expected to be wrong.

Appendix A provides an example in which the low-resolution work not only explained important but puzzling results from high-resolution work, it also pointed out potential flaws in the high-resolution work and provided an easy mechanism for extrapolating valid results to other circumstances. The low-resolution work, however, was motivated by and dependent upon the high-resolution work. Thus there was a good deal of back-and-forth among levels. The multiresolution analysis described in Appendix A can be used to generate capability curves such as those in Figure 2.1.

[1] By *stories,* we mean simplified explanations in a narrative style rather than, say, a mathematical proof. Stories for a technical audience might use simple math, whereas stories for policymakers might depend more heavily on metaphors or on a partial description of how causes lead to effects. Obviously, the respectability of the stories will depend on the skill of the analysts.

Consider finally Leg Three, which addresses noncombat performance. Suppose a detailed computation is performed to estimate the unit cost of a proposed aircraft. Here also, multiresolution work is often necessary. Sometimes it is a mix of detailed work and back-of-the-envelope work done for validation. It may include costs for the engines, fuel system, airframe, landing gear, avionics suite, system integration, etc. The result is detailed, precise, and impressive. However, we (or competitors down the hall) may calculate the aircraft's cost per pound and compare it to historical experience for comparable technology (the back-of-the-envelope, low-resolution approach). If the estimates more-or-less agree, we have reasonable confidence in the detailed calculation. Perhaps surprisingly to some, if they don't agree, we are likely to look for errors in the *detailed* computation.

Appendix B presents an extended example of using a low-resolution computation (in this case, an empirical model) to check the results of a high-resolution calculation. Interestingly, the high-resolution results are shown to be wrong, even though they were reached with a "validated model"![2] The error occurred because the model had been validated for a very different purpose. This example again illustrates one of the points referred to in Figure 1.1, i.e., that low-resolution information and models can greatly inform judgments about the validity of a high-resolution model for a particular analysis. On the other hand, the more-detailed analysis makes it possible to consider—with caution—effects not present in historical cases.

We note that the DMSO definition of *validation* explicitly emphasizes that validation must be assessed in the context of use. Unfortunately, that principle is not always followed once models have been organizationally "accepted."

Top-Down, Multiresolution Thinking in Studies and Experimentation

We next discuss the role of multiresolution work in conceiving and executing a program of study and experimentation in a top-down manner. As an example, one can think about doing this in the context of force transformation. Here, one should worry about validating the study or experimentation campaign: Will the results be sufficient to justify decisions about acquiring a new system or adopting a new operational concept? The answer depends on the validity of the war games, models, simulations, and field experiments used and, ultimately, of the "campaign plan" for pulling things together.

[2] The term *validated model* is shown in quotation marks because the model had not in fact been validated for the application at hand. It had seemed reasonable to assume that its validity extended to the new problem, but that proved to be an erroneous judgment.

One aspect of such a campaign must be to identify the broad scenario space within which the prospective capabilities are to be assessed. A second is to identify specific cases for which the assessment would be worked through in detail, with high-resolution simulation, large-scale field experiments, or a combination of these. In developing such a plan, one should be concerned about which instrument (e.g., a simulation model or a field experiment) is appropriate for which purposes. Although military experimentation is often driven by the demands of an anticipated big event (such as a particular large-scale field experiment), that is a poor way to develop a campaign plan. As discussed in Appendix C, we recommend a top-down approach.

In the context of force transformation, we have argued that the driving impulse for initiatives should be the *important and stressful* operational challenges that might be faced in the future (Davis, Gompert, Hillestad, and Johnson, 1998; Davis, 2002a). In the case of force transformation, we dealt with projection forces and gave five examples of operational challenges for such forces:

- Early halt of a classic armored invasion, given depth (e.g., in Kuwait or Northern Saudi Arabia).
- Early shallow halt of a fast invasion on multiple axes, without depth (e.g., in Korea).
- Early offensive action without first building up a massive force (e.g., in Kosovo or Afghanistan).
- Effective low-risk interventions (e.g., in Bosnia).
- Effective peacemaking in urban environments (e.g., in Kabul).

In a sense, specifying such challenges is akin to specifying the *broad* dimensions of a name-level scenario space (*name-level* suggests that we are talking about classes of scenarios, rather than all of the many details that would apply in a specific war at a specific time). U.S. forces must be prepared for many other challenges, but the requisite capabilities may come along naturally. After all, who doubts that the Air Force will be able to achieve air-to-air superiority or that the Navy will be able to achieve control of the seas, even if the Secretary of Defense provides no specific guidance about related scenarios?

Given a set of broad challenges to work with, we recommend a hierarchical approach to fleshing matters out. Briefly (see also Appendix C), we recursively break down each challenge into the need for various building-block capabilities. For example, to accomplish an early halt given depth, we must quickly (a) establish an effective C4ISR capability[3] in the theater, *and* (b) secure bases or operating locations, *and* (c) rapidly deploy forces, *and* (d) rapidly employ those forces effectively. We break down each of these into smaller building-block capabilities, and those into still smaller blocks (hence the term *recursive*).[4] We arrive at a list of high-value building-block capabilities, which may suggest the sorts of forces and doctrinal innovations we ought to consider developing. The process of generating the building-block needs highlights the circumstances in which they have value, thus suggesting the kinds of scenarios in which they should be tested and metrics that should be used.[5] These tests can be carried out by some combination of model-based analysis and experiments in the field.

To elaborate, compare (a) and (c) in the above list of necessary building-block capabilities. There may be no reason to conduct expensive large-scale exercises to investigate deployment for diverse assumptions: Models and simulations can do well and are driven largely by the laws of physics, decision delays, and well-known logistical processes. In contrast, the ability to quickly establish an *effective* C4ISR capability is a wholly different matter. Even if current models and simulations did a good job of representing C4ISR systems in substantial technical detail, they would be utterly unreliable about major factors such as how quickly (i.e., with what spinup time) a newly assembled group of officers, brought in from diverse assignments worldwide, can understand and master their new context, learn to work together well, and develop and direct appropriate tactics, techniques, and procedures. As of 1998 (Defense Science Board, 1998; Davis, Bigelow, and McEver, 1999), the answer was that it might take weeks, but no one really knew. Since that time, OSD and US JFCOM have put great emphasis on developing standing command and control capabilities (Rumsfeld, 2001) that would greatly reduce the spinup time in conflict, but even today, much of what is being learned is “soft” information, so human experiments—and even some large-scale experiments—are still necessary.

What the low-resolution top-down analysis demonstrated for the halt problem (the first challenge in the above list) was that the spinup time was of first-order importance to the overall capability and yet was not amenable to pure modeling and simulation—laboratory and field experiments were crucial. This said, *most* of the halt problem—and the assessment of alternative programs nominally addressing that challenge—could not only be done with modeling and

[3] Command, control, communications, computing, intelligence, surveillance, and reconnaissance.

[4] This decomposition is all from the perspective of a warfighter. A logistician might construct the decomposition quite differently.

[5] See also Kelly, Davis, Bennett, Harris, Hundley, Larson, Mesic, and Miller, 2003.

simulation, but could be done with *low-resolution* modeling and simulation. In other instances, the analysis identified topics that cried out for detailed simulation and small-scale human experiments.

This top-down hierarchical approach is an example of multiresolution thinking. Although the term *top-down* is used frequently, the approach is used too sparingly. Studies typically devote space to describing the scenarios used, but few studies make more than a feeble, ad hoc attempt to show why those scenarios reflect the range of circumstances for which assured capabilities are desired.

This process can perhaps best be seen as a matter of *design*. An architect or systems engineer must understand and define his "design space" in relatively low-resolution terms so that he can see issues and address tradeoffs. At some point, he or others will need to use much more detailed descriptions to assess issues in depth, but both the broad view and the detailed view are necessary.[6]

Broad, Exploratory Analysis Enabled by Multiresolution Modeling

An important class of model-based analysis is broad, synoptic work. The approach we take is *exploratory analysis*, which is very different from the narrow and more traditional *predictive analysis*. In the predictive analysis strategy, the analyst identifies a base case and a few excursion cases. He runs his model for each one and takes the results to be predictions of what would happen if the modeled circumstances were replicated in the real world. In a concession to uncertainty, he runs a handful of sensitivity cases as well. In the exploratory analysis strategy, the analyst recognizes that he has no prior reason to single out a base case. Because of massive uncertainty or ignorance, he must explicitly examine hundreds or thousands of cases. But rigorous exploratory analysis can be done only with models that have but a handful of uncertain parameters to be varied (perhaps up to a dozen). The number of cases one must consider explodes exponentially as the number of parameters increases, and the ability to comprehend and explain both inputs and the results diminishes accordingly. Faster computers will not solve these problems.

Somewhat detailed models can also be used for exploratory analysis. We and RAND colleagues Carl Jones and Dan Fox (Fox, 2003) have done a good deal of exploratory analysis using the JICM theater-level model, and other colleagues have used experimental-design methods to explore with complex models.

[6] Such matters are discussed extensively in Davis (2002a), which was written to put more meat on the bones of "capabilities-based planning."

However, the dimensionality of input datasets will depend on the resolution of the model. Suppose, for example, that the analyst is using high-resolution simulation models for entity-level analysis of force-on-force encounters. For each case, the analyst must specify the *detailed* scenarios, including, e.g., terrain, build-up schedules of Blue and Red forces, Blue and Red tactics, and so forth. He must also estimate a host of parameters describing the force's capabilities relative to capabilities of weapons that may be arrayed against it. There will be dozens or even thousands of parameters. Varying them all, and in combinations, can rapidly generate astronomical numbers of cases. Yes, clever experimental design may reduce the number of cases needed, but not to single digits. The alternative, i.e., varying only a small number of parameters while leaving others constant, is reasonable if and only if the model is well understood (as, for example, in RAND's JICM work). In our experience, that often is not the case, and deeply buried data treated as "known" are actually quite uncertain, having a big impact on results.[7]

Theory suggests that the way to avoid this "curse of dimensionality" is to work at multiple levels of resolution. Assume that it is possible to build a low-resolution model that is reasonably consistent with behaviors of the analyst's high-resolution model (see Appendix B) and that only a small number of parameters are needed to do so. The space of input datasets will have relatively few dimensions (5 to 12 in our studies). We can generate and analyze a large enough sample of input datasets to be reasonably sure that we have identified all the tough tests. We can then select one of the low-resolution model's input datasets and build a number of corresponding input datasets for the high-resolution model. This is an extension of the top-down process outlined earlier: Start with a broad, low-resolution view; identify areas where adding detail will add value and not just increase the workload; and drill on down to pay dirt.

Generalizing the Concept of Model Validation

Our extensive experience with exploratory analysis has convinced us that the current concept of model validation needs to be generalized. As noted earlier, model *validation* is usually defined as the process of determining the degree to which a model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model. Under this

[7] As examples, strategy and tactics are often represented as constant data, even though real-world commanders would change them if circumstances proved inauspicious. Another example that we have seen frequently is "requirements analysis," in which one option proved "superior" only because the threshold stated in the requirement was, say, 0.80 rather than 0.79. See Davis, Bigelow, and McEver, 2000.

definition, a black-box model may be completely valid even if we do not understand what is inside that black box: All that matters is the accuracy of predictions. Such a definition is perhaps adequate for models that can reasonably be expected to make reliable predictions and that can be checked for their ability to do so. That is feasible when the models rely on *settled theory* and have *reliable data for cases appropriate to the application*. We are all familiar with models of this type, and we may not have thought much about their internal workings for many years—e.g., since our school days or a time when we did specialized work in the area treated by the models.

In practice, however, a great deal of model-based military analysis (and analysis in other domains) depends instead on models that are incomplete, tentative, less than rigorous, and accompanied by highly uncertain data. This is reality, however much the point is obscured by organizations treating their mature models as solid. In such instances, we believe that good model-based analysis should typically have two underemphasized features:

- The models and model-based analyses should be reviewably *comprehensible* and, as part of that, should lend themselves to explanations in the form of credible, coherent, and insightful stories.
- The models and model-based analyses should deal effectively with *massive uncertainty.*

These features are enabled by MRMPM, by the kind of "stories" we emphasize in this monograph, and by the theory and methods of exploratory analysis. Typically, these are especially important when models or their input data are uncertain, in which case it makes little sense to assess their validity in the usual way. There is a dilemma here: It may be clear that using a particular model will be useful and part of a sound analysis, but the model simply cannot be deemed "valid" by the customary definition. One might consider introducing yet a new term and a new process for evaluating such a model. However, the world typically does not need more processes, and the process of validation is a well-established part of quality control in organizations such as DoD. It therefore makes sense, in our view, to generalize the concept of the validation process so that the assessments it reaches are meaningful.

Table 2.1 provides a concrete image of our suggested approach. For simplicity, the model undergoing a validation process is based either on settled theory or on clearly explained but unproven assumptions (e.g., the assumptions underlying a

Table 2.1

Types of Validation Possible in Different Circumstances

Basis of Model	Empirical Validation for Comparable Cases	Quality of Input Data for Application	Best Possible Assessment	Examples
Settled theory	Yes or no	Good	"Valid"	Bomb-damage calculations against an understood target
Settled theory	Yes or no	Uncertain	"Valid for exploratory analysis"	Bomb-damage calculations if the target's character is poorly known Mobility calculation for a scenario with unknown warning time, access rights, etc.
None	Yes	Good	"Valid"	Empirical cost model ($ per pound) for a new product comparable to earlier ones Medical statistical research showing correlations between using X and developing cancer
None	No	Good	"Not validated"	
None	Yes	Uncertain	"Valid for exploratory analysis"	Empirical cost model ($ per pound) for a new product comparable to earlier ones, but with factor-of-two uncertainty in final weight
None	No	Uncertain	"Not validated"	
Assumptions and hypotheses	Yes	Good	"Valid"	Doctrinal estimates of standard movement times
Assumptions and hypotheses	No	Good	"Valid subject to assumptions"	Cost calculation assuming a linear model
Assumptions and hypotheses	Yes	Uncertain	"Valid for exploratory analysis"	Cost calculation with large uncertainties about eventual scale
Assumptions and hypotheses	No	Uncertain	"Valid, subject to assumptions, for exploratory analysis"	Cost calculation assuming a linear model, with large uncertainties about eventual scale

plausible explanatory “story”), or it has no theoretical basis. The model either has or has not been tested empirically against data relevant to the application at hand, and the input data for the application are either good or highly uncertain. The result is a set of twelve logical cases. In our generalized approach to the validation process, the best result of an assessment would be either “valid,” “valid for exploratory analysis,” “not validated,” or “valid, subject to assumptions, for exploratory analysis.” These would be the output judgments of a successful validation process. The last column of the table provides examples.

In summary,

- We recommend that the validation process be generalized to recognize the value and importance of models that cannot be validated in the customary way, but to characterize good examples of such models (and data) with new phrases, notably “valid for exploratory analysis” and “valid, subject to assumptions, for exploratory analysis.”

A cynic might complain that the phrase “subject to assumptions” introduces circularity, as in “The model is valid if it is correct.” That is obviously not what we intend, but we acknowledge that the issue is inherently somewhat uncomfortable. Some semi-notional examples may help demonstrate that the point we are trying to confront cannot easily be avoided.

> “The many battles will be very complex, but if on average the campaign is described reasonably well by a Lanchester square law, we see that …[the point being illustrated by the model may relate to the power of concentrating forces or how certain tactics enable doing so].”
>
> “Assuming that the dose-response relationship is linear, we can extrapolate from the clinical tests to conclude that a dose of 20 mg/day should be deemed safe.”
>
> “Assuming that both sides use reserves intelligently, that neither side breaks off battle, and…, then if attrition is akin to that of WW II, battle will proceed more or less as described by …[some theater model].”
>
> “For the sake of minimaxing, let us assume that both sides use optimal tactics in employing their air forces. In that case, the analysis suggests that….”
>
> “If our model of unit cohesion captures the key factors at play, then we should expect a quick victory tomorrow because we know that (a) the enemy’s morale is low, (b) the enemy is attempting to defend something that both soldiers and officers understand is not of fundamental importance to their country, (c) both officers and

> men know that for them to continue fighting would be to pursue a lost cause, and (d) there is no particular zealotry factor at work."[8]

> "The interdiction model assumes that on each day, enemy armored vehicles are destroyed in proportion to the number of shots taken against them and the average kills per shot, with the units in question halting if attrition exceeds a 'break point.' All of the parameters are highly uncertain, but the model can be used to assess, roughly, how much is enough, as a function of assumptions about those parameters. The result can be used to help establish requirements (e.g., sortie rates)."

All of these examples approximate statements made in real-world studies with which we are familiar, and which we respect. Note that in these cases there is no really good way to compare the models being used to "reality," at least not as yet. Nonetheless, they may be good models for the purposes at hand.

We suggest, then, that validation be understood in the exploratory analysis context as follows:

- A model and its case space (databases) are valid for exploratory analysis if the case space represents well the degree of uncertainty in inputs and, within that case space, the model is either structurally valid, behaves in ways adequately consistent with what is known empirically, or is based on what appear to be reasonable assumptions. As always in a discussion of validation, adequacy must be judged in the context of the particular application.

This is more subtle than it may at first appear. A low-resolution model necessarily leaves out details and is unlikely to be highly accurate in all point cases of interest. However, if these imperfections are addressed by varying model parameters adequately, then all may be well.[9]

In summary, we have discussed in this chapter three generic reasons why multiresolution modeling is often crucial. These involve extrapolating from a few detailed analyses to more general conclusions, doing top-down analysis, and doing broad synoptic analysis.

[8] Such a discussion sketches a "qualitative model" as described in Davis (2001).

[9] A reviewer noted that somewhat-related issues have recently been discussed in DMSO's VV&A technical working group. Thus, the issue we raise appears timely. See also Harmon and Youngblood, 2003.

3. Consistency and Validation

So far, we have discussed why multiresolution modeling is needed. In doing so, we have relied on an intuitive notion of consistency between models with different levels of resolution. One can conclude intuitively that a multiresolution family displays consistency if employing its family members at different levels of resolution generates no contradictions. In this chapter, we develop a more precise notion.

We do this for two reasons. First, there is a close relationship between validation and consistency. To be valid, according to the DMSO definition, a model must be an accurate representation of the real world for the purposes of its intended uses. We don't expect it to provide a complete, fully detailed picture of the real world. Rather, we interpret its "accurate representation" as one that is "consistent with"—i.e., does not contradict—the real world in the context of the model's use. More mundanely, suppose we have a high-resolution model considered to be valid. If we build a metamodel of it, then demonstrating that the two are consistent will presumably validate the metamodel. That is, a transitivity principle applies, if the contexts of use for the high-resolution model and metamodel are those for which the high-resolution model has been validated.

Second, low-resolution models are often deterministic, while high-resolution models frequently (but not always, to be sure) have stochastic components.[1] But, as an examination of the notion of consistency will show, one ought to expect low-resolution models to be more stochastic than high-resolution models. Each model has a representation of the thing modeled that leaves out details. Thus, an ensemble of many states of the thing modeled map into each state of the model. The stochastic elements in the model must represent not only the stochastic elements in the thing modeled, but the variation of outcomes over the ensemble as well. Failure to fully represent this variability is a major reason for the bad name that low-resolution models have gotten in some circles. This bad name is unfair, because low-resolution models are often used in connection with sensitivity analysis or even exploratory analysis, in which case, having a

[1] There are more general challenges here involving stochastic effects. For example, real-world data are often relatively sparse and difficult to interpret precisely because the phenomena represented are stochastic.

stochastic model may be unimportant, but in any case, many analysts distrust and dislike low-resolution models.

Resolution

To discuss consistency, we first need to elaborate on *resolution*. A model's resolution is the degree of detail with which it represents something.[2] We measure the resolution of an optical system by the smallest (angular) separation two points must have before their images are distinct. By analogy, every model has a representation of the system it models. We measure the model's resolution by the separation two states of the modeled system must have before they are represented differently in the model.

Of course, resolution is a much more complex notion in models than in optics. An optical image is a straightforward transformation of the scene viewed. By contrast, a model can use representations that are complex, obscure, and abstract. Let us look at some examples.

Terrain is often represented digitally, with the sides of grid cells ranging from, say, 10 km (low resolution) to 100 m (relatively high resolution). Similarly, time can be measured in days (low resolution) or seconds (relatively high resolution). The lowest resolution might be adequate for representing the placement and movement of divisions in the theater. The highest might be needed for the simulation of target-seeking by a precision weapon.

Entities in the simulation could be sorted into many categories or lumped into only a few categories. Each kind of aircraft and missile could be represented as a different class of object, or they could all be lumped together as "long-range shooters."

Processes can be represented in detail as consisting of many different steps or activities, or they can be represented by one or two factors. For example, resupply of ammunition can be modeled as

1. Communicating the need for ammunition to a storage site.
2. Loading bulk ammunition on large trucks.
3. Driving the ammunition to a transshipment point.

[2] The most relevant definition from the Merriam-Webster OnLine dictionary, is "the process or capability of making distinguishable the individual parts of an object, closely adjacent optical images, or sources of light."

4. Breaking the bulk truckloads into "combat-configured loads" (i.e., consisting of the precise mix of ammunition types needed by the recipient).
5. Delivering the ammunition to the consumer.

More simply, however, one could represent ammunition resupply by a delay time (or one could simply ignore the problem, as do most combat models).

Low-resolution representations can be statistical. To use a social-policy example, air quality is measured by the concentrations of pollutants at measuring stations in an air basin. The raw data will show the concentration as a function of time. But an air-quality standard is expressed in terms of an averaging time for a pollutant, a threshold concentration, and a frequency. For example, in 1970, the air-quality standard for oxidants was exceeded if the oxidant level, averaged over one hour, exceeded 0.08 parts per million during more than one hour per year. So it is convenient to calculate hourly averages from the raw data and then to describe the set of those averages as a frequency distribution, without regard for the times at which a particular concentration was observed. Further, Larsen and Zimmer (1965) observed that this frequency distribution could be fit relatively well by a log-normal distribution, and hence described by two parameters. The statistical distribution, of course, is a lower-resolution representation of air quality than the original time series of measurements, because there are many time series that have the same statistical representation.

Finally, abstraction plays a role in resolution. In Davis, Bigelow, and McEver (2000) (see also Appendix A), the low-resolution model used the concept of a "clearing," in this case, an open stretch of roadway where munitions could acquire targets unobstructed by foliage. The only attribute a clearing had was its width. In the high-resolution model, the terrain was modeled as an array of cells 100 m on a side. Each cell either contained trees or did not contain trees. To identify clearings in this terrain, we had to decide about ambiguous cases. Was a stretch of road a clearing if trees lined the road closely but the road itself was clear? Was an open stretch of road one clearing or two if it was interrupted only by a very short stand of trees? Did the treeless area need to be large enough that a distant reconnaissance platform could identify it, or only large enough so that trees did not interfere with the terminal search algorithm of a munition?

Aggregation and Disaggregation

Aggregation and *disaggregation* are terms often used in discussing changes of resolution. With *aggregation* one moves from high to low on some dimension of resolution. With *disaggregation* one does the reverse, moving from low resolution to high. When one aggregates, one loses information, and it is not retrieved by disaggregating. This is inherent in the definition of resolution given earlier, namely that it involves the ability to distinguish among states of the modeled system. Many high-resolution states correspond to each low-resolution state. One can move from a high-resolution state to a unique corresponding low-resolution state, but when one tries to move back—when one disaggregates—one must be content to deal with an ensemble of high-resolution states, all of which will aggregate back into the same low-resolution state. Here are some examples.

Consider terrain represented as a rectangular array of cells 100 m on a side (Figure 3.1). Each cell has an attribute with one of three values. The cell can be (1) open, (2) covered with trees (T), or (3) covered by buildings (C, for city). Now we aggregate this description to an array of cells 1 km on a side. Each of the new cells consists of 100 old cells. We can characterize each new cell by three numbers: (1) the percentage of the old cells that are open; (2) the percentage that are covered with trees; and (3) the percentage that are covered with buildings. If we now try to disaggregate, we cannot recover the original, high-resolution description. As shown in Figure 3.1, there are many 10×10 arrays of small cells that would aggregate to the same large cell.

We conceive the aggregation of a process as truncating a network of variables such as that depicted in Figure 3.2. Describing the process in detail requires many variables, but an aggregated description requires only a few. In the

RAND*MR1750-3.1*

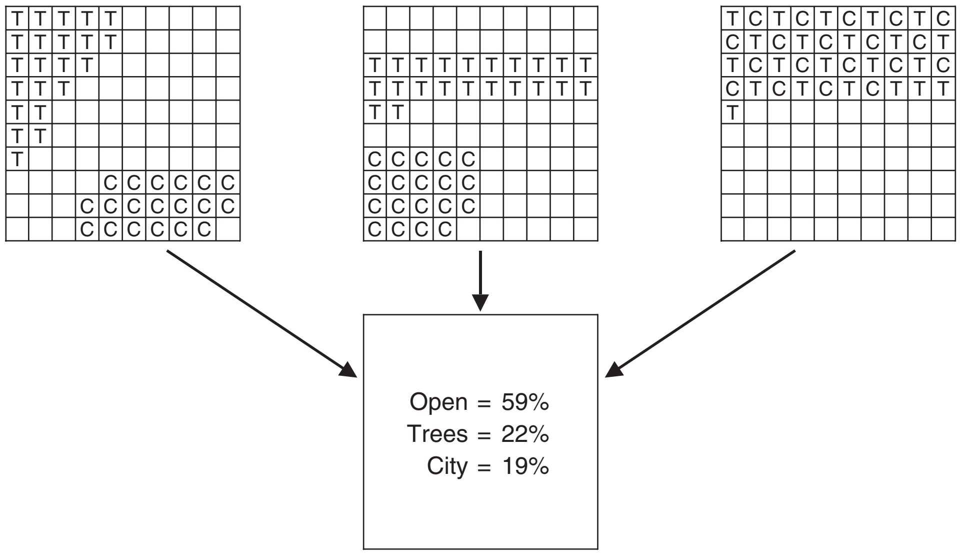

Figure 3.1—Aggregation Causes Loss of Information

RANDMR1750-3.2

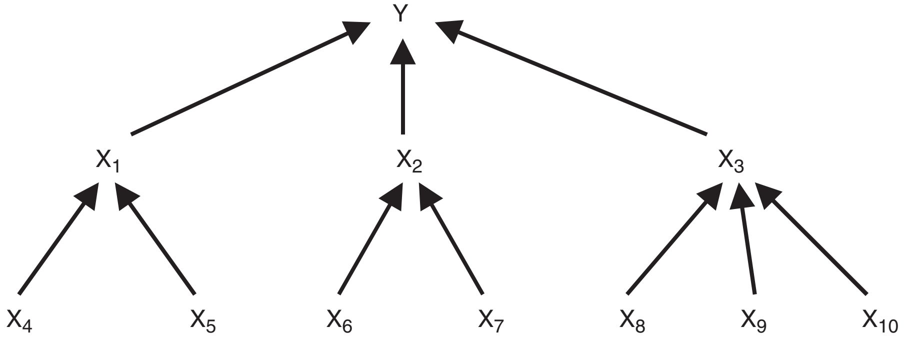

Figure 3.2—Notional Network of Variables

figure, Y represents an outcome of the process, the variables X_1, X_2, X_3 constitute the low-resolution description, and the variables X_4,…, X_{10} constitute the high-resolution description.

Algebraically, we write the low-resolution description as

$$Y = F(X_1, X_2, X_3) \tag{1}$$

We write the high-resolution description as

$$Y = F(G_1(X_4, X_5), G_2(X_6, X_7), G_3(X_8, X_9, X_{10})) \tag{2}$$

The links between them are the definitions of the low-resolution variables in terms of the high-resolution variables:

$$X_1 = G_1(X_4, X_5)$$

$$X_2 = G_2(X_6, X_7)$$

$$X_3 = G_3(X_8, X_9, X_{10}) \tag{3}$$

That is, we can substitute Equations 3 into the low-resolution description (Equation 1) to obtain the high-resolution description (Equation 2).

Clearly, the high-resolution description contains more information than the low-resolution description. If we specify values for all the high-resolution variables, we can calculate unique values for the low-resolution variables using Equations 3. But if we are given values for only the low-resolution variables, we can find many high-resolution values that satisfy Equations 3. What, then, does consistency mean?

Consistency

Figure 3.3 depicts the usual definition of consistency (e.g., see Davis and Bigelow, 1998). It assumes that the models are being used for the same application. Starting in the upper left-hand corner with high-resolution inputs, one can follow either of two paths. Going first to the right and then down, one uses the high-resolution model to produce high-resolution outputs, and then aggregates them to the lower level of resolution. Going first down and then to the right, one aggregates the high-resolution inputs to the lower level of resolution and then uses the low-resolution model to produce low-resolution outputs. The two models are consistent if the results are the same (or nearly so), regardless of path.

However, as stated earlier, one loses information when one aggregates. Aggregation moves from a high-resolution state to a unique corresponding low-resolution state, but when one tries to move back—when one disaggregates—one obtains an ensemble of high-resolution states, all of which will aggregate back into the same low-resolution state. In some instances, it may be adequate to select a single, typical high-resolution state from the ensemble, in which case, the consistency definition just outlined is adequate (and there is a lot of redundancy in the high-resolution representation). In some cases, additional information is

RANDMR1750-3.3

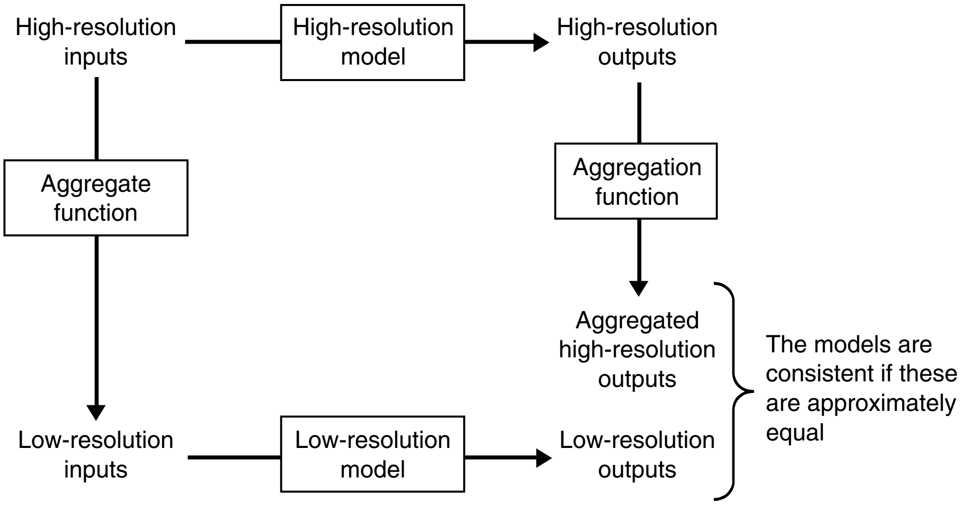

Figure 3.3—The Usual Definition of Consistency Between High- and Low-Resolution Models

necessary to specify the mapping (e.g., doctrine may dictate how a unit would configure itself in a situation, without regard to the unit's history through a road march to that situation). In other instances, it may be necessary to acknowledge the variation within the ensemble.

In such cases, we need the definition depicted in Figure 3.4. In this definition, one starts in the *lower* left-hand corner, with low-resolution inputs. Going first upwards, one disaggregates those inputs into an ensemble of high-resolution inputs. The high-resolution model is used to generate high-resolution outputs for all the inputs in the ensemble, and then to aggregate each set of outputs. Taking the other path, one uses the low-resolution model to generate low-resolution outputs from the low-resolution inputs. The two models are consistent in this new sense if the ensemble of aggregated high-resolution outputs is "comparable" to the low-resolution outputs.

In principle, we should use an even more general definition of consistency. Since resolution is multidimensional, one model of a pair could have higher resolution in one dimension, and the other could have higher resolution in a second dimension.

Often, though, this extra complication will not be necessary. Perhaps the most common example of models with mixed resolutions is that of models with different scope. For example, Model 1 may represent long-range missile strikes

RANDMR1750-3.4

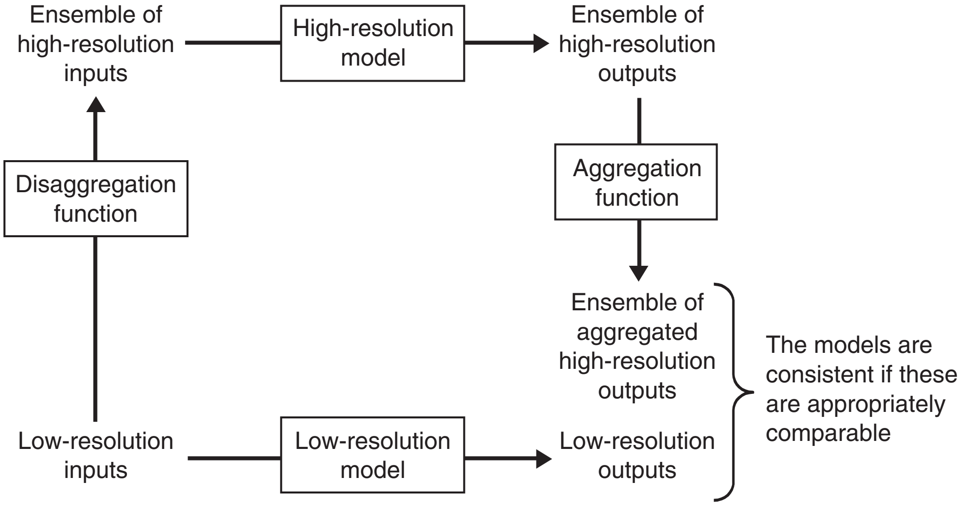
Ensemble of high-resolution inputs
High-resolution model
Ensemble of high-resolution outputs
Disaggregation function
Aggregation function
Ensemble of aggregated high-resolution outputs
The models are consistent if these are appropriately comparable
Low-resolution inputs
Low-resolution model
Low-resolution outputs

Figure 3.4—A More General Definition of Consistency Between High- and Low-Resolution Models

in detail, taking as inputs the positions and velocities of target vehicles within a specified distance from the impact point (i.e., in the missile's footprint). Model 2 might represent missile strikes as a simple kill probability multiplied by the number of targets in the missile's footprint. But it might calculate vehicle movements in detail, including movements of vehicles far from any particular missile's footprint. So Model 1 has greater resolution than Model 2 in the immediate neighborhood of a missile strike, but lower resolution (in fact, no information at all) at a distance.

Finally, let us mention briefly another complication. Not uncommonly, one has outputs generated from low- and high-resolution models but does not know the aggregation and disaggregation functions referred to in Figures 3.3 and 3.4. What would then constitute "consistency"? We find the question (raised by a reviewer) to be uncomfortable because we advocate seeking to understand model relationships (i.e., estimating those functions). Nonetheless, if one is faced with the question, it may be possible to do enough sensitivity analysis with the high-resolution model to determine whether the low-resolution model's results are at least within the range of those of the high-resolution model for the same problem and situation. The appropriate language, then, is cautious, as in "The results are not obviously inconsistent with what can be understood from…."

Right and Wrong Answers

The key, of course, is what we mean by "comparable" in Figures 3.3 and 3.4. Following the figure, let

y = high-resolution output

$\overrightarrow{xhr}$ = vector of high-resolution input variables

D = domain of high-resolution inputs

$HR(\cdot)$ = high-resolution model

$\overrightarrow{xlr}$ = vector of low-resolution inputs

$AggIn(\cdot)$ = function for aggregating high-resolution inputs to low-resolution inputs

$LR(\cdot)$ = low-resolution model

For simplicity, we will assume that the high-resolution model is deterministic. For any low-resolution input vector, the ensemble of high-resolution outputs is a set of values for *y*, and the output of the low-resolution model should be one or more attributes of that set, i.e.,

$$LR\left(\overrightarrow{xlr}\right) = Attribs\left\{ y \middle| y = HR\left(\overrightarrow{xhr}\right), \overrightarrow{xhr} \in D, AggIn\left(\overrightarrow{xhr}\right) = \overrightarrow{xlr} \right\} \tag{4}$$

But what attributes of this set should the low-resolution model estimate? If the set is a small interval—i.e., if there is little variation in the ensemble of high-resolution outputs—it is enough to estimate any *y* in the set. The first example in Appendix D provides an illustration.

The situation becomes so simple when the ensemble is a small interval (or even a single point) that it can be tempting to make the ensemble such an interval. The usual way to do this is to limit the domain D of high-resolution input vectors to be considered. For example, the analyst may assume certain of the inputs to the high-resolution model are constant. This undoubtedly simplifies the low-resolution model, but it greatly limits its applicability. And when one conveniently forgets the limitation and applies it anyway, one obtains spuriously precise results. We suspect this is one of the major reasons for the bad reputation that low-resolution models have in some circles.

There are, of course, "honest" ways to limit the domain D. For example, the object model may contain five or ten different kinds of Blue shooters. The analyst may have information that constrains the mix, either because a particular mix is infeasible or because it would be silly to choose it. Some shooters may require airfields with long runways, and there may not be many such airfields in the theater. There may be limited numbers of other shooters in the force structure. Some shooters may not be rapidly deployable. Such constraints may reduce the size of the ensemble of high-resolution cases that corresponds to each low-resolution case.

Even if the ensemble of high-resolution outputs is not a small interval, one can sometimes make do with a low-resolution model that estimates a single output. The second example in Appendix D outlines a low-resolution model that estimates the maximum size an algae bloom could attain under certain conditions. In the notation above, this corresponds to estimating the maximum value of *y* in the ensemble, i.e.,

$$LR\left(\overrightarrow{xlr}\right) = MAX\left\{ y \middle| y = HR\left(\overrightarrow{xhr}\right), \overrightarrow{xhr} \in D, AggIn\left(\overrightarrow{xhr}\right) = \overrightarrow{xlr} \right\} \tag{5}$$

Much, perhaps most, of the time, the ensemble of high-resolution outputs is not a small interval. Appendix D provides an illustration of this. In these instance, all the information needed about the ensemble cannot be captured in a single output, so the low-resolution model will have to estimate two or more outputs. The obvious candidates are a mean and a standard deviation, or the upper and lower bounds of a confidence interval, but these make no sense unless a probability distribution can be assigned to the high-resolution input vectors. Without such a distribution, the most one can do is estimate the range of y in the ensemble.

But such a distribution will generally not describe the probabilities that various high-resolution cases would occur in the real world. Rather, the analyst will specify a distribution that captures the analytic importance of various regions of the domain D, or equivalently, he will design a sample of high-resolution cases. This problem is covered under the topic of experimental design in numerous statistics texts (see, e.g., Saltelli, Chan, and Scott, 2000). Given such a sample (or distribution), one can readily calculate means, standard deviations, and confidence intervals for the low-resolution results.

4. Motivated Metamodels, Explanations, and the Importance of a Good Story

Nobody denies the importance of basing a model on—or calibrating it to—data. In our view, however, a good story is often equally important. By *story*, we mean what is often referred to as *explanation*. Explanations are not always needed. Analysts are often willing to use a model that they do not personally understand, but that is known to be reliable—because of either prior comparisons with relevant data or rigorous development from well-established theory. In other cases, however, that is not sufficient. Perhaps one is about to make an important decision on a complex subject and the model in question cannot be considered to be based on settled theory and reliable data. In such cases, one cannot treat the model as a mere "black box"; to be useful, the model must tell a credible story about how things work in the relevant portion of the world. It must express a set of logical relations, cause-and-effect mechanisms on which to base inferences. The term *story* suggests that the explanation may be an ad hoc invention, and one might therefore prefer to use another term such as *theory* or *phenomenological explanation*. Sometimes the story can be quite speculative, however, and then the alternative terms suggest that the modeler knows more than he actually does. But no matter how skimpy one's knowledge, we consider it important to support a model with an explanatory and motivational story. Again, we are referring here to instances in which the model and its data cannot be treated as an answer machine.

Not everyone agrees with our approach. Some consider it to be enough to fit equations to data, e.g., by statistical methods such as regression. In such models, an independent variable is said to "explain" some part of the variation of the dependent variable, but this does not mean that a change in the independent variable causes the dependent variable to change. Rather, it means that the two variables are correlated, with possibly no causal relation at all. Before we would find such a model useful, we would need to identify the cause-and-effect relations—i.e., we would need to discover or construct the story.

There are several reasons why a good story is important. First, as mentioned earlier, it is necessary to explain to the client why the model yielded the results it did and why those results are generally true, not just true in the specific cases run. A persuasive story is invaluable for this task. The cause-and-effect aspect of the story is essential here, because the client (presumably a decisionmaker) wants to take actions that will cause desired consequences.

Second, one step in validating the model is establishing its face validity. Face validation is the process of persuading subject-matter experts that the model behaves reasonably, i.e., that for reasonable inputs it produces reasonable outputs. How will they judge that the model is reasonable if not by determining that it conforms to a good story or theory?

Third, a model is used to estimate things that cannot be observed directly.[1] This means that the model will be used to extrapolate beyond the available data. We are acutely aware that extrapolation has a bad name. However, as Law and Kelton (1991) observe:

> The greater the commonality between the existing and proposed systems, the greater our confidence in the model of the proposed system. There is no completely definitive approach for validating the model of the proposed system. If there were, there might be no need for a simulation model in the first place.

So extrapolation is where the action is. Somebody had to extrapolate beyond historical experience to suppose that precision-guided munitions (PGMs) would increase the capability of the force that used them. Today, it is necessary to extrapolate beyond past experience to suggest that information technology will revolutionize warfare. We agree with both of these extrapolations, although the jury is still out on the second.

Extrapolation is never based on data alone; it is based on a function that one fits to the data. Extrapolated results depend crucially on the form of the function chosen. Appendix E illustrates this with a simple example. The form of the function, in turn, is suggested by plausible stories (or theories) about how the real world behaves. Thus the story influences how one builds the model in the first place.

A story is just as important for a metamodel—a model of a model—as for any other model. Ideally, the story would be enough by itself. The low-resolution models in a family could be derived or inferred from physical considerations, derived explicitly from the theory embodied in valid high-resolution models, or obtained by algebraically reducing or simplifying the high-resolution model. In practice, however, one must augment the story by applying statistical methods to data from high-resolution experiments.

[1] It might be possible in principle to observe these things directly, but it might be costly, dangerous, impossible to do quickly, or otherwise inconvenient.

A common method of building a metamodel is to run the high-resolution object model many times, collect the results in a dataset, and simply fit a response surface to the data. Appendix F describes an experiment in which we examined ways of constructing a metamodel of a given high-resolution object model, and the first method we tested was this "most common" approach. By judiciously selecting independent variables, we were able to achieve a fairly good overall fit (e.g., as measured by root-mean-square error).

However, we found ways that are much more effective than blindly regressing the outcome(s) on the inputs. We defined transformations of the inputs and various functions of them to use as additional independent variables in the regression model. When these additional variables were chosen on the basis of knowledge of the object model's underlying theory, they vastly improved the regression results. This knowledge could be a description of how the object model works, perhaps from documentation. Or it could be a coherent story about how the model works. Or the knowledge could be a thought experiment about how the target model "ought" to work, i.e., how it would work if it had been built. We coined the term *motivated metamodel* to describe a low-resolution model whose structure is based on knowledge of these kinds (Davis and Bigelow, 2003).

Appendix A provides another example of constructing a metamodel, this time a metamodel to estimate the effectiveness of missile salvos fired from long range at groups of armored vehicles. The object model was extremely large, and the analysis dataset was constructed from only about a dozen cases. But each case provided data on hundreds of missile salvos, so our analysis dataset contained plenty of observations.

Each case from a large model provides thousands of data elements. To build a metamodel, one must select the data elements that should be included among the independent variables. In the example of Appendix A, we assumed that the number of vehicles killed by a missile salvo depended on the vehicles that were within a specified distance from the impact point (i.e., in the missile's footprint) and not obscured by foliage at the time of impact. It seemed only reasonable to exclude such variables as the number of vehicles very far from the impact point and the locations of vehicles at times other than the impact time. We offer these exclusions as examples of using knowledge of how a large model works (or how it should work) to help structure the metamodel. Reducing the set of candidate independent variables in this way vastly simplifies the task of building the metamodel.

In our example, having a story was also important because the high-resolution model we used has parameters that remained constant within a case but that we wanted to be able to vary when we employed the metamodel. All cases assumed the same failure probabilities of the missiles and their warheads. Some parameters, such as the transparency of the foliage and the delay time between selecting an aim point and a salvo's impact, did vary, but few combinations of these parameters could be explored within the dozen or so cases we had available. To estimate the effects of these unvaried or hardly varied parameters required that we extrapolate and interpolate in ways that the data did not support.

Motivating the metamodel can be strategically important. For example, suppose one is dealing with a system that could fail if any of several critical components failed. Naïve (unmotivated) metamodels may fail to reflect the individual criticality of such components and may therefore be quite misleading if used for policy analysis. Naïve metamodels may be correct "on average" but give misleading results on the relative importance of inputs, thereby skewing resource-allocation decisions. Naïve metamodels may also fail in "corners" that seem to be obscure but that can actually be focused upon by adversaries. Motivated metamodels can ameliorate such problems.

All in all, a good story (or theory, or phenomenological explanation) is an essential element of any model. It helps immensely in

- Building the model.
- Establishing its face validity.
- Generalizing and extrapolating.
- Explaining and justifying results to a client.

To those who argue that data alone provide a sufficient basis for modeling, we offer this quotation. It was written about world affairs and international relations, but we believe it applies in the present context as well:

> [I]nevitably we operate with some kind of theory. It is sheer myth to believe that we need merely observe the circumstances of a situation in order to understand them. Facts do not speak for themselves; observers give them voice by sorting out those that are relevant from those that are irrelevant and, in doing so, they bring theoretical perspective to bear. Whether it be realism, liberalism, or pragmatism, analysts and policy makers alike must have some theoretical orientation if they are to know anything. Theory provides guidelines; it sensitizes observers to alternative possibilities; it highlights where levers might be pulled and influence wielded; it links ends to means and strategies to resources; and perhaps

> most of all, it infuses context and pattern into a welter of seemingly disarrayed and unrelated phenomena. (Rosenau, 1998, p. 92)

In summary, we—like Rosenau—believe that theory is badly needed; it is only sometimes sufficient to treat models as answer machines. Sometimes our theories can be well established by empirical data or rigorous comparison with more fundamental settled theory, but sometimes we must make do with something less rigorous. Also, when reaching conclusions or supporting decisions made by policymakers, we often must abstract what we believe to be the essence of the problem—discarding many details in the process. The result, then, may be more of a story than a robust theory, but if we have done our job well, the story will be insightful and accurate enough for the purpose at hand.

Appendix

A. Using a Simple Model to Explain, Extrapolate from, and Provide Face Validity of Complex-Model Results

Background

RAND has a suite of high-resolution simulation models for high-fidelity analysis of force-on-force encounters. Our colleagues have used this suite of simulation models to assess the combat effectiveness of numerous weapon systems, including the Army Tactical Missile System (ATACMS) with the Brilliant Anti-Tank (BAT) munition. In a 1996 study for the Defense Science Board (Matsumura, Steeb, Herbert, Lees, Eisenhard, and Stich, 1997), four cases were simulated in which a total of 88 ATACMS were fired, killing 283 enemy tanks, or about three tanks per missile. Over time, this figure of three kills per BAT warhead gained status as a reasonable estimate of the effectiveness one should expect from this weapon. "Three tanks killed per BAT" had graduated from a mere statistic to a rule of thumb (a particularly simple metamodel).

The transition to rule of thumb came about in this way. People were disappointed that BAT did not perform better in the 1996 study. Each BAT has 13 submunitions, which spread out over a huge footprint (4 km radius) and independently use acoustic sensors to home in on targets. People had hoped that substantially more than a quarter of the submunitions would score hits. So the 1996 results came to be interpreted as an upper bound: "You can't count on more than three tanks killed per BAT." It seemed natural to begin using it as an estimate of BAT's effectiveness in general.

Clearly, granting rule-of-thumb status to this statistic was unjustified. Yet such extrapolations occur often—in part, because clients seeing the results for a handful of scenarios demand estimates of how the results generalize. An analyst might say, "I think three kills per BAT is a reasonable general estimate," or he could say, "I can't generalize from these results, and I strongly recommend that you don't either." Although more appropriate scientifically, the latter answer would not be responsive and would not be appreciated.

With this background, a 1998 study for the Defense Science Board (Matsumura, Steeb, Isensee, Herbert, Eisehnard, and Gordon, 1999) simulated six new cases in which 324 missiles were fired, killing 142 armored vehicles (tanks and BMPs)—only about 0.44 kill per missile. At first, it seemed that the differences could easily be rationalized. In DSB ’96, the terrain was entirely open, while in DSB ’98, the terrain had considerable tree cover. In DSB ’96, almost all the Red vehicles were armored fighting vehicles (AFVs), whereas in DSB ’98, less than 20 percent of the Red vehicles were AFVs. And in DSB ’96, the Red vehicles were in dense formations (50 to 100 m spacing), while in DSB ’98, vehicles were much more dispersed (150 to 600 m).

But for each of these explanations, there was a plausible counterargument. Thus in DSB ‘98, missiles were aimed only at clearings. One could argue that this should reduce the number of shots but not necessarily the effectiveness per shot. The BAT submunition preferentially homes in on AFVs, so the presence of trucks should have made little difference. And the fact that ATACMS/BAT has such a huge footprint should have negated the large separations between vehicles. So what were the real reasons that ATACMS/BAT performed so much more poorly in DSB ‘98 than in DSB ‘96? And how should one extrapolate? Although the rationalizations were presented at the time (*rationalization* is the term for *story* when the story is not well supported by analysis), the questions were troublesome.

Multiresolution Analysis

Over the next year, as part of a small project for OSD, we used this puzzling experience as an opportunity to demonstrate what could be accomplished with multiresolution analysis of the phenomena. We examined the performance of BAT in every missile salvo simulated in the two studies. We found that BAT’s performance differed in the two studies because, as Figure A.1 shows, there were generally many fewer AFVs in BAT’s footprint in DSB ’98 than there were in DSB ’96. It was a rare salvo (of two missiles) in DSB ’96 that had as few as 20 AFVs in its footprint, and often there were 40 to 100, while there were rarely more than 15 AFVs in the footprint of a DSB ’98 salvo. As one would expect, salvos killed fewer AFVs when there were fewer AFVs in the footprint.

This may seem like an obvious conclusion, but it was not easy to reach. The RAND suite of models, like most high-resolution models, consumes a great deal of input and generates an enormous amount of output. Producing a graph like Figure A.1 is a major undertaking, almost as great an undertaking as producing a similar graph from field experiments. One must manipulate a lot of data. One must define concepts such as “AFV” and “footprint” in terms of the entities in the high-resolution model. Thus several different kinds of vehicles fall into the

RANDMR1750-A.1

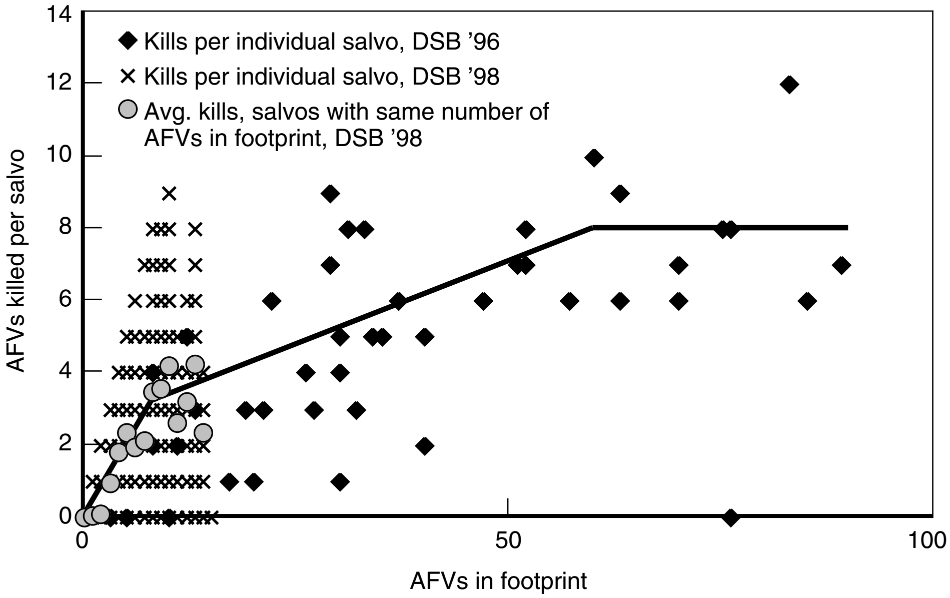

Figure A.1—Kills per Two-Missile Salvo vs. AFVs in Footprint

class "AFV," and while almost all kills occurred within 4 km of the impact point, there were a few scattered kills at distances of 6 or 7 km. In our experience, it is rare for an analyst to devote this level of effort to the examination of the inputs and outputs from a high-resolution model of combat. [1] Yet without such an effort, why should one have confidence that the high-resolution suite was treating BAT in a reasonable way?

Further, the analysis was not yet complete. After DSB '98, we knew that "three tanks killed per BAT" was not a good rule of thumb, and after the data analysis, we knew that the line in Figure A.1 was better—a kind of simple statistical metamodel. But to use it, we had to be able to estimate the number of AFVs in the footprint from variables under the control of Blue and Red. That created a new challenge.

With this in mind, we built a low-resolution theory-based model called PEM (PGM Effectiveness Model) (Davis, Bigelow, and McEver, 2000). We structured it to be consistent with the functions of RAND's high-fidelity suite of models that are involved in interdiction and calibrated it to the high-fidelity results. PEM is

[1] In other domains, such detailed analysis is more routine. A reviewer noted that about 970,000 runs of the highest-resolution fly-out models were made to provide adequate information about capabilities of two competing missile designs in the selection of the PATRIOT PAC-3 missile to ensure that adequate information was available over the complete spectrum of threat types, intercept conditions, and other operational considerations to support a correct decision about which missile design to use.

not quite small enough to be written on the back of an envelope, but it is nonetheless small and simple. We implemented it in Analytica™, a very flexible visual modeling tool.

PEM assumes that a column of Red vehicles is traveling along a road and moves through a clearing of width W. Rather than being uniformly spaced, the Red vehicles are grouped into packets, perhaps representing platoons. Each packet has N AFVs separated from one another by a distance S. Successive packets are separated by a distance P, which is larger than S. This column of vehicles moves through the clearing at a velocity V.

Blue attacks the column by firing a salvo of one or more missiles at the clearing, timed to arrive when a selected packet is expected to be in the center of the clearing (see Figure A.2). But there is a random error in the arrival time (TOA_error) whose mean is proportional to the time since the missile last received information about the position of the target packet (Time_of_last_update). If TOA_error is too large, the target packet may have passed completely through the clearing, or beyond the weapon's footprint, F, whichever is larger; or, if the missile arrives early, the target packet may not have entered the clearing or the footprint. A smaller error will find the target packet not centered in the clearing, and part of

RANDMR1750-A.2

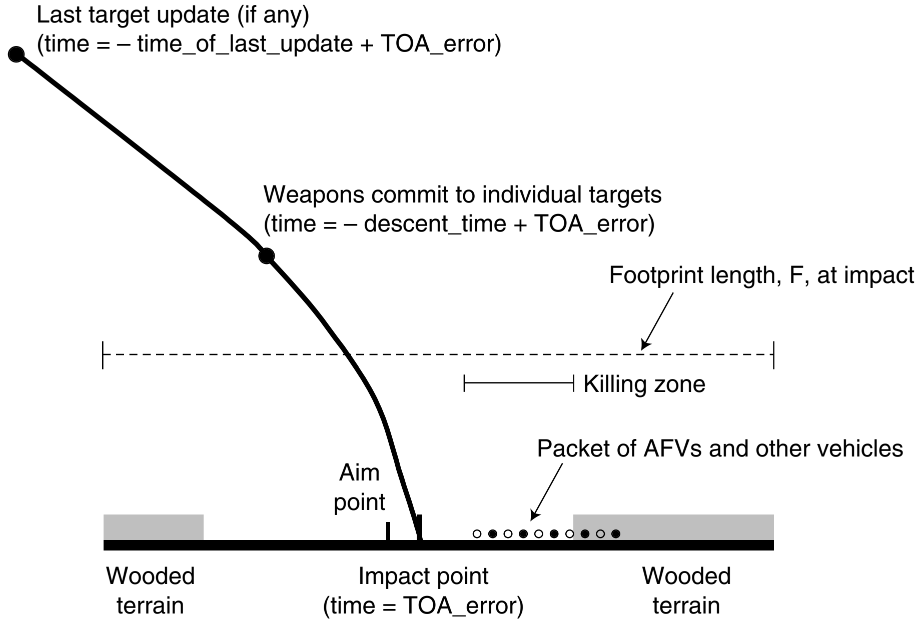
Last target update (if any)
(time = – time_of_last_update + TOA_error)
Weapons commit to individual targets
(time = – descent_time + TOA_error)
Footprint length, F, at impact
Killing zone
Packet of AFVs and other vehicles
Aim
point
Wooded
terrain
Impact point
(time = TOA_error)
Wooded
terrain

Figure A.2—PEM Concepts

it may be hidden in the trees at either end of the clearing. Depending on the various parameters, parts of the packets just forward and rearward of the target packet may be in the killing zone.

Except for a few details, this determines how many Red AFVs are in the killing zone of the weapon at its time of impact. We can then use a function like the line from Figure A.1 to estimate the number of AFVs actually killed.

One can think of PEM as a scaling function that adjusts the effect of long-range precision fires for the influence of a variety of factors. These factors include the time of last update, which operates through the error in the missile arrival time; the footprint of the weapon; the openness of the terrain; and the formation (including the dispersion) of Red's vehicles. We have used the model for exploratory analysis to identify both positive and negative interactions among the factors. We have also developed an even simpler version of the model that could be used as a subroutine to incorporate these factors in other models. Of course, we could use "three tanks killed per BAT" in all the same ways. But PEM is richer, more plausible, and ultimately more nearly valid.

Returning to Figure A.1, note that there is a large variation in number of kills per salvo, even when the number of AFVs in the footprint is held constant. Some of this variation is due to the fact that the high-resolution model performs Monte Carlo trials to determine which BAT submunitions fail at one point or another in the process of acquiring and killing a target. But some of the variation may have been systematically related to variables we had not considered. Further analysis of the high-resolution inputs and outputs might suggest ways to increase the average kills per salvo, even as AFVs in the footprint are held constant.

What does all this have to do with validation? First, by developing PEM and using it to reason about the data from high-resolution simulation, we constructed a satisfying explanation for the high-resolution results, which had previously been troubling. Again, that explanation was at a much lower resolution than that of the original models and could not easily have been inferred from those models. Had we been unable to construct PEM, we would have been left to argue that the result must be right "because the high-resolution model says so, and it must be correct because of all the careful work that has gone into it." The fact that we could construct a low-resolution explanation thus tended to validate the high-resolution model. Second, having implemented the low-resolution explanation as a simple model (PEM), we could then use it to explore a scenario space looking for the kinds of scenarios that would challenge BAT most severely (Davis, Bigelow, and McEver, 2000).

In the course of the work, we also found instances in which the correctness of the high-resolution models was questionable, or at least not well understood. For example, we tentatively concluded that large-area acoustic noise due to AFVs was apparently having a substantial effect on results. It was questionable whether the estimates of this effect had been realistically calculated for mixed terrain. Also, with the benefit of hindsight, we concluded that the tactics used in the DSB '98 study had probably been unrealistically extreme, thereby further emphasizing the importance of being able to scale results to other cases, as was possible once PEM was created.

The point here is not, of course, to criticize high-resolution models, much less to suggest that lower-resolution models are better. To the contrary, as discussed in Chapter 1, we believe strongly in a family-of-models approach that seeks to use all available theory and data to develop the best possible integrated understanding of phenomena. The high-resolution work referred to here was critical in illuminating many fundamental features of the interdiction problem being studied, particularly features related to the subtle interactions of terrain, line of sight, finite weapon speeds, and weapon characteristics. Had we tried to work the problem with only a simple model such as PEM, we very likely would have made serious mistakes. Going back and forth between models, however, allowed us to do better than we could have done with either low- or high-resolution models alone.

B. Using a Low-Resolution Calculation to Check a High-Resolution Calculation

Sometimes low-resolution work turns out to be more accurate than much-richer, detailed analysis. An example of this arose in RAND work on logistics requirements of the Army's Light Helicopter, Experimental (LHX), which has since been redesignated the RAH-66 Comanche (Smith, Acker, Bigelow, Dreyfuss, La Forge, Pei, Resetar, and Petruschell, 1988). The requirements we estimated were the cost per flying hour of replenishment spares and the cost per flying hour of replenishment repair parts. Spares are parts that can be removed from the aircraft, repaired, and reused. Repair parts are used once and thrown away. These cost coefficients are labeled *replenishment costs* to distinguish them from *initial costs*. When any service buys an item of equipment, they buy initial spares and repair parts to stock the supply system. Subsequent operation of the equipment will break spares and repair parts, and those stocks will need to be replenished.

Traditionally, these cost coefficients are estimated from historical data (Figure B.1). The resulting lines correspond to linear empirical models. Such empirical models are a special case of low-resolution models. They are not metamodels as we have used the term in this monograph; they are based on empirical data, not on some more-detailed model.

Smith, Acker, Bigelow, Dreyfuss, La Forge, Pei, Resetar, and Petruschell (1988) used data for six existing helicopters (the OH-6 Cayuse, the OH-58 Kiowa, the UH-1 Iroquois—better known as the "Huey," the AH-1 Cobra, the UH-60 Black Hawk, and the AH-64 Apache). These were the helicopters whose missions the LHX was intended to perform. The costs per flying hour were plotted against the investment cost per helicopter, and it was found that the six historical points lay reasonably close to straight lines. When the LHX was plotted at its projected investment cost ($10.6 million), the straight lines implied that the cost elements should be over $700 and $350 per flying hour, respectively.

The Army was unwilling to use this simple, aggregate model for estimating these cost factors. Army analysts had adopted a reliability-improvement program that was intended to reduce these costs, and the traditional method has no "hooks"

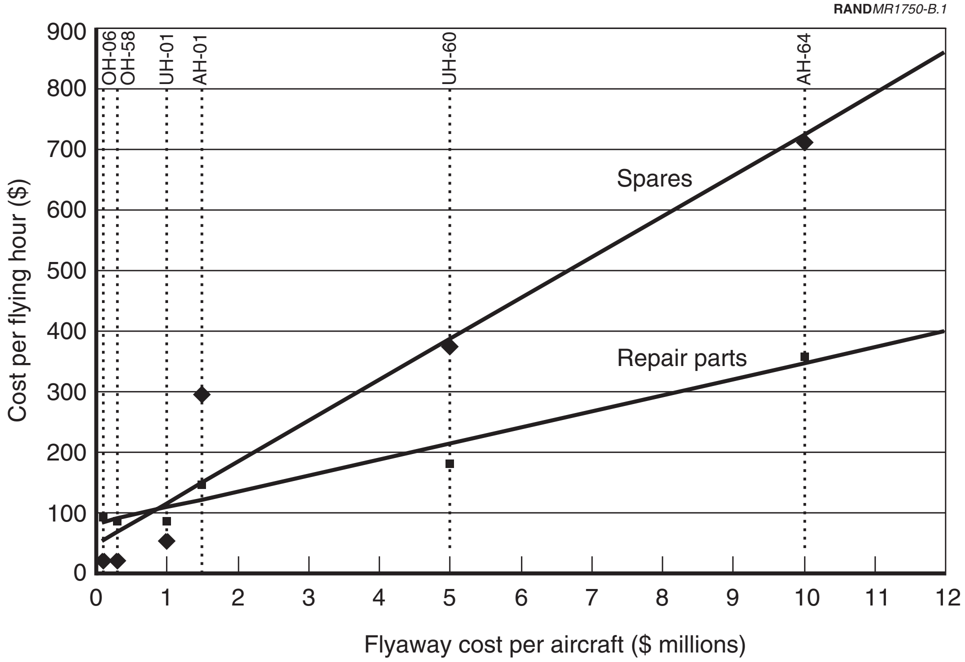

Figure B.1—Replenishment Cost Factors, Traditional (Aggregate) Method

for including its effects. So in their baseline cost estimate (BCE) for the LHX, the Army analysts estimated these two cost elements one spare and repair part at a time. They estimated a unit price for each item, a mean time between failures, and for spares the fraction of the item price that would be spent on repairs and the fraction of items that could not be repaired (the condemnation fraction). The Army analysts were optimistic; they found a close analog for each item on their existing helicopters and used the one with the lowest cost in their LHX estimate. They also adjusted the costs downward to account for the reliability-improvement effort in the LHX program. The result put the two cost factors at about $35 and $16 per flying hour for spares and repair parts, respectively, or about one-twentieth of the estimates made using the empirical model.

There were several reasons for the discrepancy. A major reason, of course, was the unbridled optimism of the Army analysts. A new weapon system is always cheaper and more reliable, not to say more effective, than anything you already have.

But most of the discrepancy was due to the fact that the detailed model—the model that describes what happens to an individual item—was an unrealistic idealization. A high-level view of this multiechelon inventory model, used by all the services, is represented by the network shown in Figure B.2.

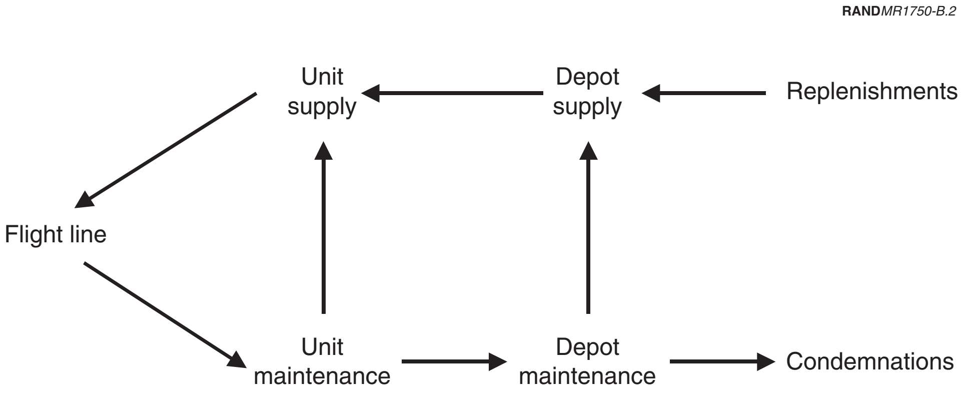

Figure B.2—Multiechelon System for Processing Recoverable Items

When helicopters fly, radars, fuel pumps, or engine parts may fail. When a part fails, it is removed at the flight line and sent to unit maintenance for repair. A replacement is drawn from unit supply, if one is available. Unit maintenance tries to repair the item and, if successful, returns it to unit supply to replenish the stock. If the repair is too difficult for unit maintenance, personnel there return the part to the depot for repair and order a replacement from depot supply. Depot maintenance repairs the item and sends it to depot supply.[1] As the system operates, the inventory of each item is distributed among the transportation and repair pipelines. Items in transit between the unit and the depot in both directions are in transportation pipelines; items in repair at both the unit and depot are in repair pipelines. The number of items in a pipeline will depend on the transportation or repair time and will equal the number that have entered during a period that equals the pipeline time.

In theory (i.e., assumed in the model), initial procurement provides just enough items to fill the pipelines. Over time, the system will lose items due to condemnations. Replenishment spares and replenishment repair parts are intended to replace these condemned items through procurement from manufacturers. Consistent with this theory and model, the Army BCE calculated the costs per flying hour of replenishment spares and repair parts as the costs of replacing condemned items.

[1] The network for repair parts is simpler. Repair parts are discarded when they are removed from a helicopter, and they are replaced from intermediate supply. Intermediate supply is replenished from depot supply. Repair parts are not cycled through maintenance facilities for reuse. But the model described here can represent repair parts as recoverable items with condemnation fractions of one. So our comments about the deficiencies of this model for estimating the cost per flying hour of replenishment spares also apply to estimating the cost per flying hour of replenishment repair parts.

The theory, however, has many shortcomings. For example, initial procurement may provide incorrect inventories of some parts. If more items are provided to the system initially than are needed to fill the pipelines, the excess will migrate to intermediate and depot supply. There will be no need to replenish these items until the excess is used up. On the other hand, if too few items are provided initially to fill the pipelines, parts will migrate out of war reserve stocks or, in extreme cases, be made up by leaving "holes" in helicopters (i.e., not replacing items that have been removed). In this case, requirements for replenishment may vastly exceed condemnations.

Nor is misestimation of initial procurement the only reason replenishments may differ from condemnations. Removal rates of items at the flight line vary for many reasons (Crawford, 1988; Hodges, 1980). The military frequently modifies its aircraft, either to improve safety or to increase capability. These changes can cause removal rates to increase or decrease. If an expensive recoverable item has a high removal rate, engineers will redesign or reinforce it, or they will design modifications to other portions of the aircraft to relieve stress on the item. If aircraft are redeployed to a location with extreme environmental factors (e.g., temperature extremes or sand storms), removal rates can rise. In all such cases, there are procurement demands that do not result from condemnations. Indeed, since condemnations tend to be very low, purchases to fill pipelines dominate replenishment.

Perhaps some individuals at the time expected the law of large numbers to rescue the Army cost estimates. That is, the demand rates could be considered to be random numbers, and this would have made the contribution of each item to its cost factor a random number as well. Since there are thousands of items per weapon system, why shouldn't the sum of all these contributions converge on the right answer?

The answer is one we all know well. The expected value of a function of a random variable x is generally not equal to the function applied to the expected value of x. That is:

$$E(f(x)) \neq f(E(x))$$

We do find equality, of course, if *f(x)* is a linear function, and it may seem at first sight that we have a linear function in the present case. In fact, if we hold all the other parameters constant (i.e., pipeline times, fraction repairable at unit maintenance, and condemnation fraction), then the total required inventory of an item is proportional to the demand rate. But it isn't the total inventory that contributes to the cost factor; it is the incremental inventory. It is the amount that must be bought *this year*, not the total amount that must be kept on hand. For this year's buy, it is necessary to calculate the total inventory required and then subtract the inventory already on hand. And then the result must be *truncated*, for if it happens that the increment is negative, then the contribution of this item to the cost factor is zero. None of the current inventory can be sold.

In addition, the other parameters do not in fact remain constant. If a spare (a reparable item) cannot be repaired at unit maintenance, the engineers will try to redesign it, or to design appropriate tools so that the item will become reparable at unit maintenance. The engineers will try to fix parts with high condemnation fractions as well.

So why did the Army use this model to estimate the replenishment cost factors? Inventory models based on the multiechelon structure of Figure B.2 are used by all the services and the Defense Logistics Agency (DLA) to manage items in their supply systems. These models work very well for most items most of the time, and they signal the item manager when an item begins to misbehave. At that point, the item manager takes action, for example, by assigning a high repair or transportation priority to an item (reducing repair and transportation times) or by redistributing the items in the system (Hodges, 1980).

Nor does the fact that the item manager must intervene from time to time mean that the model is not good for its management purpose. Simon (1982a, p. 197) describes a study that devised rules for inventory and workflow smoothing. He later commented that he had simplified the description of the real world to make his model mathematically tractable, and he never expected that the optimal solution to the idealization would be optimal in the real world (Simon, 1982b, p. 434). He was relying on the people actually implementing the policy to intervene from time to time when the policy went off track.

In short, the detailed model meets the DMSO definition of validity in regard to its use as an item-management tool. But it is not valid for the use the Army made of it in their cost estimating.

In a related effort, the validity of the theory was tested in another application, one using Air Force data for selected aircraft (Bigelow, 1984).[2] The model calculated a very large requirement to buy spares in the first year, and much smaller requirements (about 10 percent of the first year's requirement) in each subsequent year. The first year's requirement sufficed to redress the inventory

[2] The main purpose of this study was to build a metamodel of the item-level model described here, so that OSD could quickly estimate the impact of changes in flying programs on elements of the Air Force budget. The invalidation of the item-level model to predict the requirement for replenishment-spares dollars was an unanticipated byproduct.

discrepancies, due to the difference between historical and current demand rates—the same problem that derailed the Army helicopter-cost estimates. The model assumed that once the discrepancies were redressed, they would never recur. Actual requirements have never shown this drop after the first year, suggesting that these discrepancies are a permanent fact of life. Empirical models, then, however simple, reflect real-world considerations that often do not appear in theory—even in meticulously developed high-resolution models.

Strictly speaking, this is not an example of metamodeling or multiresolution modeling (MRM). But it does suggest that working at multiple levels of aggregation can reveal errors that would remain hidden if one worked at only a single level.

To cast this result in MRM terms, these methods for estimating the cost factors are both based on widely used models. The low-resolution empirical method uses statistics (or less formal means) to fit historical data. Cost analysts are very comfortable with this method; they have been using it since the beginning of cost analysis and systems analysis in the 1950s. It is based on the supposition that the future will look like the past or (more generally) can be extrapolated from past trends.

The higher-resolution method is based on an idealized description of the process that generates the demands for spares and repair parts. Because it looks explicitly at the process, it has the "hooks" for estimating the effects of changing aspects of the process. But because it is an idealization, not everything that can be estimated with this method is correct.

The hooks provide the means for extrapolating from historical practices to new practices. In this case, they are needed to represent an attempt to design reliability into the LHX. However, there is nothing in the historical records that represents attempts to do this for previous helicopters, so no variables can be included in a statistical model based on historical data. The hooks must be based on thought experiments (that is, theory). This raises the question of just how effectively a method with these hooks can be validated. The whole point of using the detailed method that has hooks was to allow extrapolation into an unobserved realm.

Unfortunately, the two methods are not mutually consistent. The exercise of making them consistent—which was never done—would presumably have yielded more defensible estimates of the annual cost of replenishment spares and repair parts, one that would provide a plausible explanation of deviations of the estimates from the historical data.

C. Selecting a Good Test Set of Detailed Scenarios

One of the generic problems in combat analysis is that of picking appropriate test cases for study with high-resolution simulations. Another problem is how to use a combination of field experiments and constructive modeling and simulation. We addressed both such issues in a 1999 study advising OSD on the use of analyses within efforts to "transform the force."[1] The motivations involve both opportunities and necessity. By exploiting modern technology and new operational concepts, DoD expects that U.S. forces can greatly increase their capabilities and, in some cases, can do so while reducing costs. At the same time, major changes will also be *necessary* to mitigate difficulties that can be posed by even midlevel rogue states. These include short-warning attacks and other so-called "asymmetric" strategies involving weapons of mass destruction (WMD), missiles, mining, high-lethality conventional weapons, exploitation of urban sprawl and innocent civilians, and coercion of regional states resulting in access constraints.

A key mission of the newly created U.S. Joint Forces Command was experimentation in support of force transformation. It was often observed that such experimentation should employ a combination of models, simulations, war games, and field experiments. This observation, however, was rather abstract, and many interpreted it to mean only that in both the work-up and follow-up phases of a major field experiment, models and simulations should be used. We had, and continue to have, a sharply different view, in which field experiments are seen as merely one part of experimentation, and not even the most scientifically important part. Further, in our view, field experiments are rare and valuable opportunities that should be used to collect information that cannot be obtained more readily in other ways.[2]

In Davis, Bigelow, and McEver (1999), we outlined an approach to thinking about ways in which analysis could guide and supplement research and experimentation. We recommended a hierarchical analysis approach, combining

[1] DoD first highlighted the transformation issue in its 1997 Quadrennial Defense Review. See also Davis, Bigelow, and McEver (1999), which documented suggestions made in late 1996.

[2] To be sure, field experiments have other important objectives as well, notably, demonstrating to high-level officers and officials that certain capabilities can be realized. This can be crucial for achieving support and acceptance.

top-down and bottom-up methods. *Top-down* means starting with a broad, low-resolution view of the issue, and identifying a succession of higher-resolution formulations of narrower problems. *Bottom-up* refers to combining solutions to high-resolution problems into an integrated response to the top-level issue. (The common use of these terms is yet more evidence that analysts routinely work at multiple levels of resolution.)

At the top of our illustrative application of the top-down approach was the operational challenge of bringing about the *early* halt of an invading armored column, even under stressful circumstances related to asymmetric strategies.[3] Working down, we developed variations of the challenges that stressed U. S. forces in different ways, and from the variations we identified high-value building-block operational capabilities, adaptive integration capabilities, and cross-cutting functional capabilities. The process of identifying these capabilities also highlights the circumstances in which they are valuable and suggests the scenarios in which they should be tested. In some cases, detailed analysis or even experiments would be needed to do the testing.

In our methodology, any challenge problem is the outer shell of a multilayer structure. To show on a diagram that the United States should be able to bring about an early halt is straightforward, but accomplishing the challenge requires many building-block capabilities, as shown in Figure C.1 (an incomplete depiction). For example, at the left of the figure, we see the requirement to establish quick and effective command and control and theater missile defense. That in itself is an extraordinary operational challenge, but in this depiction the related capability is a subordinate building block. In many cases, success of the whole requires success of all the parts. That is, we have a complex *system problem*.

Certain crucial features of the overall operational capability are cross-cutting and therefore do not appear in any single branch of the decomposition tree. In particular, *all* of the subordinate operations are, in our view, likely to depend on network-centric command and control, long-range fires, effective operations with allies, forward presence, forward leaning during crisis, and mobility (Bigelow, Bolten, and DeHaven, 1977).

[3] This problem was worked in much more detail in the context of asymmetric strategies in Davis, McEver, and Wilson (2002). That study also emphasized that the same interdiction capability measured by a "halt distance" was quite relevant to countermaneuver strategies that might be important in, e.g., an invasion of Iraq.

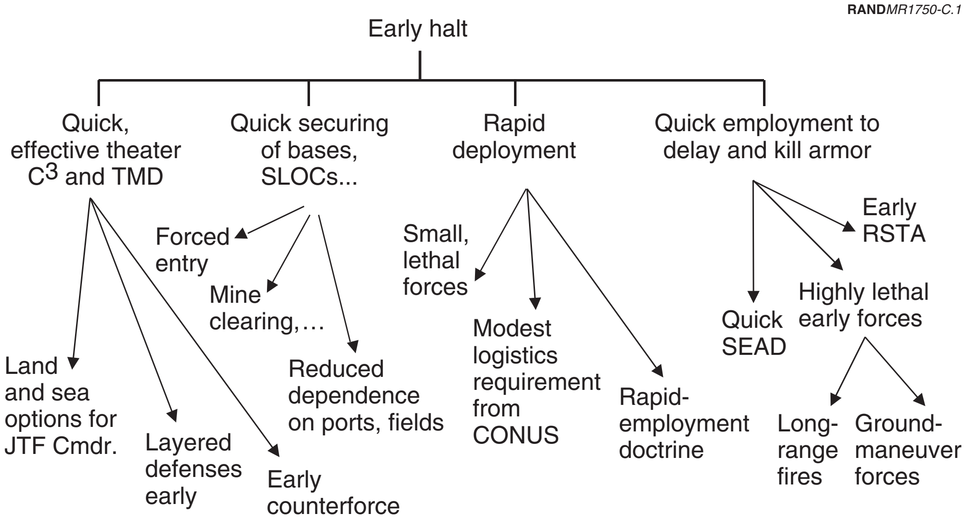
RANDMR1750-C.1
Early halt
Quick, effective theater C3 and TMD
Quick securing of bases, SLOCs...
Rapid deployment
Quick employment to delay and kill armor
Land and sea options for JTF Cmdr.
Layered defenses early
Early counterforce
Forced entry
Mine clearing,…
Reduced dependence on ports, fields
Small, lethal forces
Modest logistics requirement from CONUS
Rapid-employment doctrine
Quick SEAD
Early RSTA
Highly lethal early forces
Long-range fires
Ground-maneuver forces

Figure C.1—Challenges Can Be Decomposed Hierarchically

We can drill down further. We implemented a multiresolution version of a halt-phase model called EXHALT (*ex*ploring the *halt* problem) in the Analytica™ modeling system (various versions are documented in Davis and Bigelow, 1998; McEver, Davis, and Bigelow, 2000; and Davis, McEver, and Wilson, 2002). Figure C.2 shows some of the factors EXHALT considers to determine the effectiveness of long-range fires (a building block on the "quick employment to delay and kill armor" branch) in halting the invading armored column. We measure effectiveness by how far the Red column penetrates. This depends on the time it takes to halt Red and its speed of advance. Moving down the left-hand branch, the halt time will depend on kills per day as a function of time, which in turn depends on the number of shooters of all types available to oppose the column and the effectiveness of each type of shooter (the underlining of these factors indicates that they are vectors). As shown in the figure, these factors can be decomposed further.

It is useful to express the factors in a multiresolution hierarchy in this way because one can visualize the combinations of factors necessary to bring about a halt. To illustrate, let us truncate Figure C.2 at the level of "kills per day" on the left branch from Thalt, and at "Red vehicles to kill" on the right branch. If we make the approximation that KPD(t) is zero until air defenses are suppressed and constant thereafter, we can then express the halt distance as

$$Dhalt = V \times \left(\left(T_s - T_{delay} \right) + \frac{\xi}{KPD} \right) \tag{C.1}$$

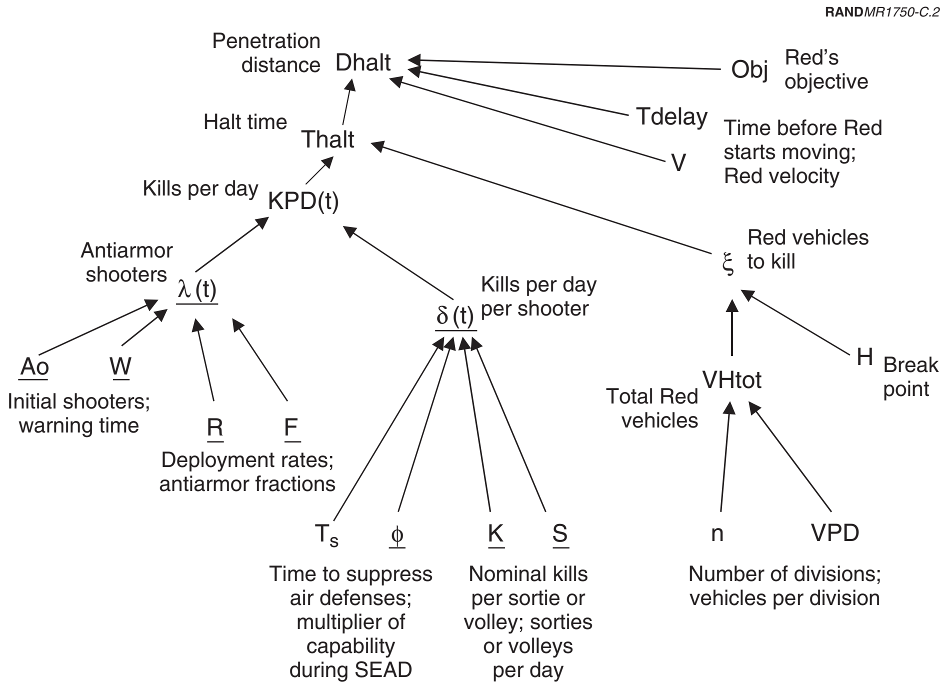
RANDMR1750-C.2
Penetration distance
Dhalt
Obj
Red's objective
Halt time
Thalt
Tdelay
Time before Red starts moving; Red velocity
V
Kills per day
KPD(t)
Red vehicles to kill
ξ
Antiarmor shooters
λ (t)
δ (t)
Kills per day per shooter
Ao
W
Initial shooters; warning time
R
F
Deployment rates; antiarmor fractions
H
Break point
VHtot
Total Red vehicles
Ts
ϕ
K
S
n
VPD
Time to suppress air defenses; multiplier of capability during SEAD
Nominal kills per sortie or volley; sorties or volleys per day
Number of divisions; vehicles per division

Figure C.2—Some Factors Considered in EXHALT

Figure C.3 shows the combinations of Red velocity, wait time (the difference $T_s - T_{delay}$), and kills per day that are sufficient, under these assumptions, to kill 500 Red vehicles by the time the column has traveled 100 km. Points on or below the surface will do the job; points above the surface will fail. Clearly, to succeed at this task, Blue must concentrate on measures that will either slow Red to a crawl or reduce the wait time to a day or less. In this figure, the wait time is the period between the time the Red column begins to move and the time Blue begins to attack it. Thus wait time may include warning time as well. Unless the wait time can be made small, the number of kills per day the Blue force can achieve hardly matters.

Next, consider the challenge of killing a much larger force (4,000 vehicles) by the time Red has traveled 300 km. Now, it pays to improve any of the factors (Figure C.4).

We find that it is important to include in our test set of scenarios both a scenario that stresses the rapid accomplishment of a relatively small task and a scenario that stresses the slower accomplishment of a very large task. Note that a single worst-case scenario, i.e., one that requires halting a very large and fast force within a day or two, will not suffice. True, if the worst-case scenario can be

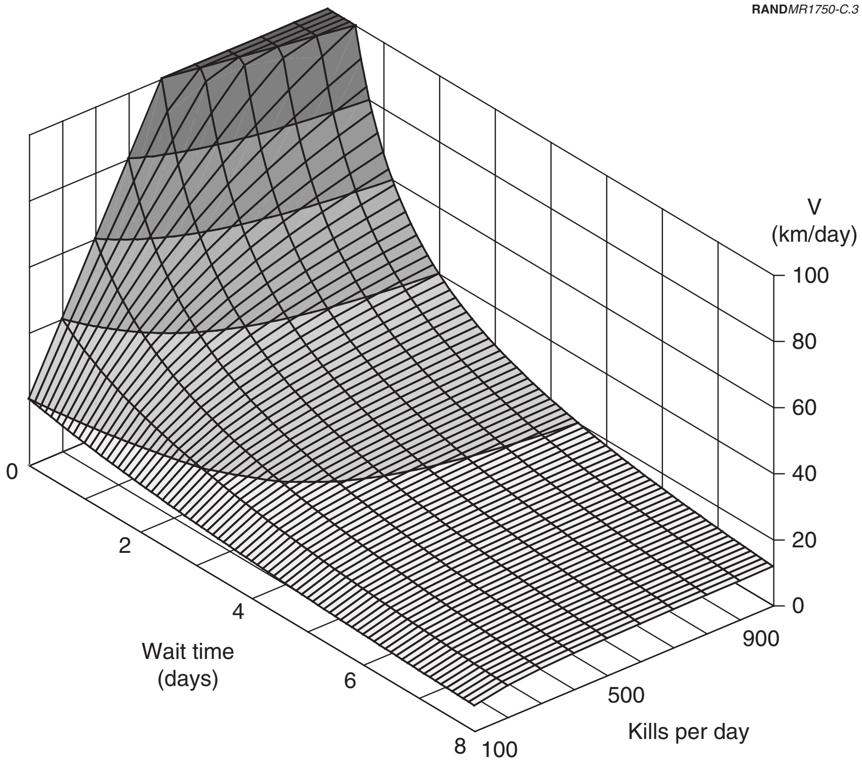

Figure C.3—Factor Combinations that Allow 500 Red Vehicles to Be Killed Within 100 km

accomplished, the two lesser scenarios can be accomplished. But the cost of a worst-case capability may be prohibitive. For example, it might require buying a force consisting entirely of very rapidly deployable, highly lethal weapons capable of operating from austere bases. Weapons with all of these capabilities will have very high unit costs. By contrast, both lesser scenarios might be accomplished by buying a couple of squadrons of not-so-lethal but highly deployable weapons that are capable of operating from austere bases. These weapons would be first on the scene and could accomplish the scenario that stresses the rapid accomplishment of a relatively small task. They would be supplemented in the longer, bigger scenario with more lethal weapons that need more time to deploy and can be operated only from better-prepared bases. This mixed force would doubtless be considerably less costly than the worst-case force would be.

Of course, this discussion does not complete the top-down development of scenario and building blocks. Figure C.2 expands on the parameters explored in Figures C.3 and C.4, and the other branches and building blocks in Figure C.1

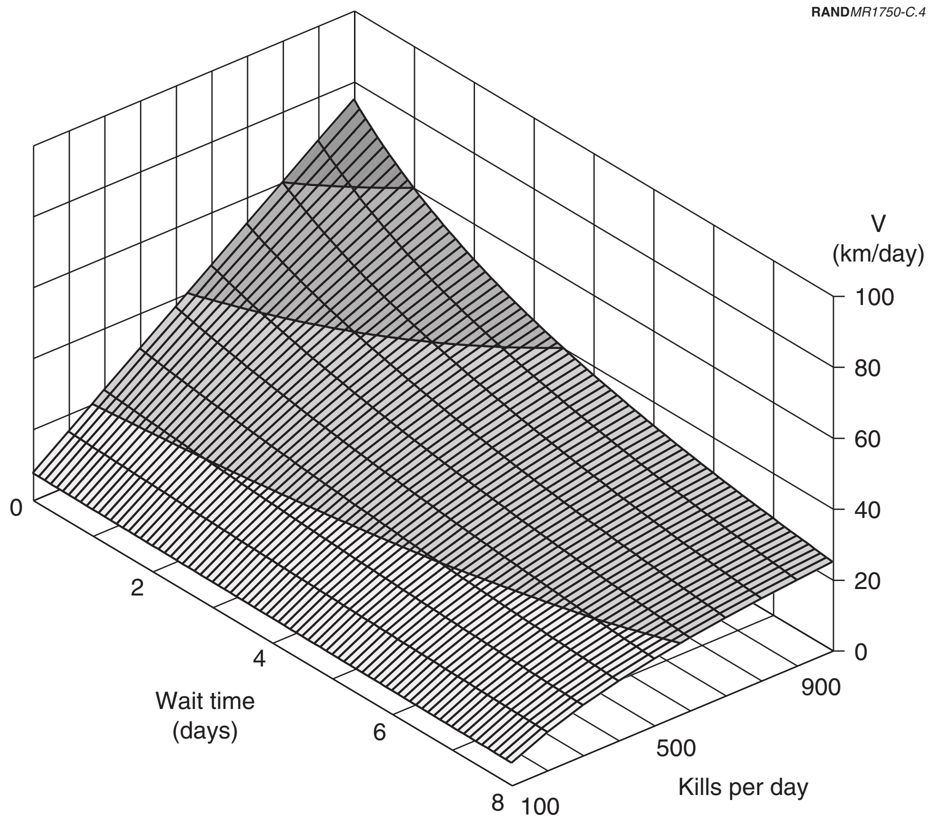

Figure C.4—Factor Combinations that Allow 4,000 Red Vehicles to Be Killed Within 300 km

remain to be explored. The result of this analysis will be a list of high-value building blocks, each described in terms of a combination of capabilities. The process of generating them also highlights the circumstances in which they are valuable and suggests the scenarios in which they should be tested. Detailed analysis would be needed to do the testing.

D. Illustrations of the Use of Consistency Definitions

The Ensemble of High-Resolution Outputs Can Be a Narrow Interval

In 1975, RAND conducted the Policy Analysis of the Oosterschelde (POLANO)[1] project for the government of the Netherlands. In 1953, a flood caused by a very large storm sweeping down the North Sea had devastated the Delta region of the Netherlands. After recovering from the immediate effects, the Dutch government decided to build engineering works in the Delta region to protect the area and its population and resources from any future flood.

The Oosterschelde was the last estuary whose protection was undertaken. The original plan called for the Oosterschelde to be dammed off from the sea, creating a freshwater lake. By 1974, however, controversy had arisen over the damage this would do to the ecology of the Oosterschelde and to its thriving oyster and mussel fishing industry. So the Dutch Cabinet directed the Rijkswaterstaat (RWS), the government agency responsible for water control and public works, to assess alternative protection plans. The RWS turned to RAND for help in this assessment.

Ultimately, three plans were considered. The original plan was to close off the Oosterschelde completely. The first alternative was to leave it open and build massive new dikes around its perimeter. The second alternative, and the most challenging to assess, was to build a storm-surge barrier—a dam with gates that could be left open most of the time but closed if a large storm threatened. Different variants of the storm-surge barrier had different aperture sizes, each size producing a different reduction in the tide and hence a potentially different effect on the Oosterschelde’s ecology.

It was necessary to estimate peak water levels at each dike section around the Oosterschelde for a wide range of storm scenarios and barrier apertures. (Additional estimating relations were created to account for the incremental

[1] This description of the project has been paraphrased from the Preface of Bigelow, Bolten, and DeHaven (1977).

effect of wind and waves on peak water levels, by dike section.) By comparing the peak water level at a dike section with the height of that section, we could estimate the likelihood and severity of flooding for different scenarios. In the follow-on Barrier Control (BARCON) study,[2] it was necessary to test barrier-control strategies, i.e., to determine how quickly the barrier should be closed, what the target water level inside the basin should be while the barrier was closed, and what the relation of inside and outside water levels should be when the barrier was opened.

The RWS had developed a model called IMPLIC that calculated water levels as a function of time throughout the Oosterschelde, using an implicit finite difference scheme to integrate partial differential equations. Its calculated water levels and flows throughout the basin matched measured water levels very closely.

We constructed a model we called SIMPLIC (Simple IMPLIC) to estimate the same water levels, but at much less cost (see Abrahamse, Bigelow, Gladstone, Goeller, Kirkwood, and Petruschell, 1977; and Catlett, Stanton, and Yildiz, 1979). SIMPLIC consisted of a single ordinary differential equation which could be integrated much more economically than IMPLIC's partial differential equations. Where IMPLIC required the user to input water levels over time at many points on a section across the mouth of the estuary, SIMPLIC required only the average water level over time at the mouth. Similarly, IMPLIC required a detailed description of the basin, including depths at every point relative to a standard elevation, whereas SIMPLIC merely needed the surface area of the basin as a function of average water level. It was thus much easier to specify cases for SIMPLIC, and SIMPLIC was much less costly to run.

Needless to say, SIMPLIC did not reproduce IMPLIC's water levels exactly. While IMPLIC successfully represented basin resonance phenomena (i.e., the phenomena that make tidal amplitude different at different locations) and the phase shift of the tide at different points within the basin, SIMPLIC predicted the water level at only a single reference point in the basin, and a calibration curve to adjust the SIMPLIC results for other locations around the basin had to be introduced. We actually did not care about phase shifts. We were interested in estimating the peak water level achieved at each point in the basin during each tidal cycle, and that had almost zero dependence on the phase. Once we calibrated SIMPLIC for the resonances of the basin, SIMPLIC's estimates of peak

[2] Once the Dutch Cabinet had determined to protect the Oosterschelde estuary by building the storm-surge barrier, they commissioned RAND to investigate what conditions should trigger the closing and reopening of the barrier.

levels per tidal cycle matched IMPLIC's estimates within two or three centimeters.

In this example, the ensemble of high-resolution results would consist of all scenarios that produce a given peak water level at the reference point used in SIMPLIC. But even after accounting for resonances and phase shifts, IMPLIC's predictions of peak water levels at different locations are very highly correlated. Once the peak water level at one location is known, the peak levels at all other points can be predicted with very little error. This is equivalent to saying that if we were to look at all cases in the ensemble, we would find little variation in peak levels at any given location. Hence SIMPLIC is consistent with IMPLIC according to the first definition of consistency given in the main text of this monograph.

Selecting the Maximum of the Ensemble of High-Resolution Outputs

The POLANO project developed a model to predict the severity of algae blooms and the effectiveness of methods for controlling them. That model was also used in a subsequent project, PAWN (Policy Analysis of Water Management for the Netherlands). F. J. Los later developed the model further at the Delft Hydraulics Laboratory (Bigelow, Bolten, and DeHaven, 1977; Los, 1991).

Algae—otherwise known as phytoplankton—are single-celled, waterborne organisms that consume nutrients (nitrogen, phosphorus, and, for some species, silicon) plus energy from sunlight in order to grow. Given enough of these resources, a population of algae can grow large enough to become a nuisance. They can cause the death of plants on the bottoms of water basins by shielding them from sunlight. They often produce substances that are toxic to fish and shellfish. They can clog filters in water systems. And when the algae population dies off, the process of decomposition can exhaust the dissolved oxygen in the water, causing fish kills and bad odors.

Most published models of algae growth simulate the population of algae over time, as it grows from a small to a large concentration. For our purposes, however, the peak concentration is a good measure of the severity of an algae bloom. Therefore we built a model that estimates the peak directly, without trying to map the trajectory by which the bloom arrives at the peak. We formulated the model as a linear program in which the variables were the concentrations of different species of algae, and the constraints ensured that the algae population did not outgrow the available nutrients and solar energy. The model calculated the mix of species that maximized total biomass.

We were aware that the model could overestimate the peak of a bloom. It takes time for algae to grow, and the conditions most favorable for algae might not persist long enough. But we argued that sometimes overestimating the bloom was acceptable for policy purposes, as long as blooms reached or nearly reached their "theoretical" peaks with reasonable probability.

An Example That Requires Ensembles of High-Resolution Cases: Force Allocation

Consider the rule of thumb that in order to succeed, an attacking force must have a superiority of 3:1 over the defending force. If this rule is valid for a small sector along the line of battle (i.e., in detail), what can we say about its applicability to the theater as a whole (i.e., in the aggregate) (Davis, 1995)?

Suppose that Blue has one unit of force in the whole theater, and Red has two units. Suppose there are two sectors on the battle line, and the 3:1 rule is true in each sector individually. Clearly, if Blue and Red both decide to allocate their forces equally between the sectors, the ratio in both sectors will be only 2:1 in favor of Red. Since this is smaller than 3:1, a Blue defense will be successful. But if Blue allocates its forces equally, and Red "loads up" on Sector 1, Red will be able to create a 3:1 or greater advantage there while maintaining a safe 1:3 or better ratio on Sector 2. Now Red can attack successfully on one sector and, having defeated half of the Blue force, presumably can fall upon the other half and finish the job. On the other hand, if Blue can detect Red's concentration of forces on Sector 1, it can reallocate its own forces to counter Red's move.

The situation becomes even less clear if we divide the line of battle into more sectors—ten sectors, for example. Let Red allocate its two force units however it chooses. Let Blue then allocate two-thirds of its one force unit to mirror the Red allocation. This is enough to defend successfully on every sector. Now let Blue put its remaining one-third force units on the sector where Red is weakest. Red must have no more than one-tenth of its force, or 0.2 force units, on the weakest sector, so Blue will enjoy a superiority on that sector greater than 3:1. It appears Blue can attack successfully and can defeat the Red force in detail.

In this example, the low-resolution case is characterized by the theaterwide force ratio of 2:1 in Red's favor. The corresponding ensemble of high-resolution cases consists of all allocations of Blue and Red forces over the sectors that adhere to the 2:1 constraint. That is, if B_j and R_j are the Blue and Red forces allocated to sector j, then the high-resolution ensemble consists of all solutions to the constraints

$$\begin{aligned}&\sum_j B_j = 1, B_j \geq 0\\&\sum_j R_j = 2, R_j \geq 0\end{aligned}\qquad \text{(D.1)}$$

Clearly, the fact is that the 2:1 theater force ratio does not constrain the ratio on any given sector. There are many solutions to the constraints of Equation D.1 in which different sectors have force ratios that deviate from 2:1. Without additional constraints, we cannot predict the theater outcome from the theater force ratio.

Unless the force allocation to sectors were an input, a high-resolution model would have to include a force-allocation algorithm. This might be modeled as an OODA[3] loop. Each side would use its intelligence assets to observe the opponent's force allocation and would estimate the opponent's intentions concerning changes to the allocation. Each side would then decide how to change its own allocation and would proceed to do so. If one side had substantially better intelligence and mobility, that side might be able to largely dictate the sector-by-sector force ratios. If each side could quickly observe what was happening but only slowly change its own allocation, neither side could readily change the initial allocations. If each side could move its forces very rapidly but was unable to detect the opponent's movements, the theater campaign would be a crapshoot. In none of these cases would it be appropriate to apply the 3:1 rule at the theater level.

[3] Observe, orient, decide, act.

E. Basing Extrapolation on a Story

In this appendix, we illustrate the need to use a story or theory as a basis for extrapolating beyond the available data. We generated a dataset consisting of 10 points (y_i, x_i), using the following model:

$$y = f(x) + \varepsilon \tag{E.1}$$

The residual ε is a normal random variable with mean zero and standard deviation 6. We will withhold the function $f(x)$ for the time being. Table E.1 shows the values of the independent variable x and the dependent variable y. Because we generated the dataset, we can provide the "true" value of the dependent variable, $f(x)$, as well.

Given these data, we wish to estimate the value of y that corresponds to x = 15. In addition, we want to state how uncertain we ought to be about our estimate. The standard method is to use statistical regression to fit a straight line to these data. When we regress y on x, the least-squares fit is

$$\hat{y} = -10.1 + 7.12 \times x \tag{E.2}$$

Substituting x = 15 into this equation yields our estimate, y = 96.76.

Table E.1

Sample Dataset

x	y	$f(x)$
3.0	9.72	10.98
3.5	20.13	15.00
4.0	12.01	19.28
4.5	19.76	23.62
5.0	25.35	27.89
5.5	43.58	31.95
6.0	25.46	35.70
6.5	31.86	39.07
7.0	42.32	42.00
7.5	42.82	44.47

There are three sources of uncertainty in our estimate (we prefer the term *uncertainty* to *error*). First, there is a *residual uncertainty*. We know that Model E.1 has a residual term, which by construction is a normal random variable with mean zero and standard deviation 6. If this were the only source of uncertainty, the 90 percent confidence interval for our estimate would be 96.76 ± 1.645σ, or (86.83 – 106.63).

Second, there is a *calibration uncertainty*. Calibration uncertainty arises from the uncertainty in the values of the intercept and slope of the regression equation. To assess the size of this uncertainty, we generated 20 sets of residuals using the Excel RAND() and NORMINV() functions, calculated the least-squares linear function for each, and used them to make 20 estimates of the value of y at $x = 15$. We estimated the 90 percent confidence interval for calibration uncertainty by dropping the minimum and maximum estimates. The range of the remaining 90 percent of the estimates was from 88.81 to 123.29. This interval is about 1.8 times the width of the 90 percent confidence interval for the residuals.

Third, there is *structural uncertainty*. We chose to fit a linear model to the data, and there is no clear indication from the data that this is a bad choice. But that is no guarantee that it is a good choice. In fact, we generated the data using a function:

$$f(x) = x^3 e^{-0.3x} \qquad \text{(E.3)}$$

As shown in Figure E.1, this function looks very nearly linear for $x \in [3, 7.5]$, but it extrapolates to larger x in a very nonlinear fashion. Indeed, $f(15) = 37.49$, a much smaller value than any of the linear fits could suggest. Figure E.1 also shows the minimum, maximum, and nominal linear fits, and the data from Table E.1.

The problem, of course, is that we are extrapolating beyond the range of the data. The data cannot reliably guide us in this task; something more is needed. We suggest that the "something more" is a good story, an explanation for how these particular data came to be used. This explanation can then be used to suggest what hypothetical data might look like in ranges not covered by the actual data. For example, the physical meaning of the variable y might preclude its being negative, which would imply that a linear function with a negative intercept is a poor fit. Similarly, one might reason that y cannot become arbitrarily large. A more powerful story might describe the process that generates y from x as involving two factors. One causes y to increase with increasing x, while the other has the opposite effect. The former factor might be expected to dominate when x

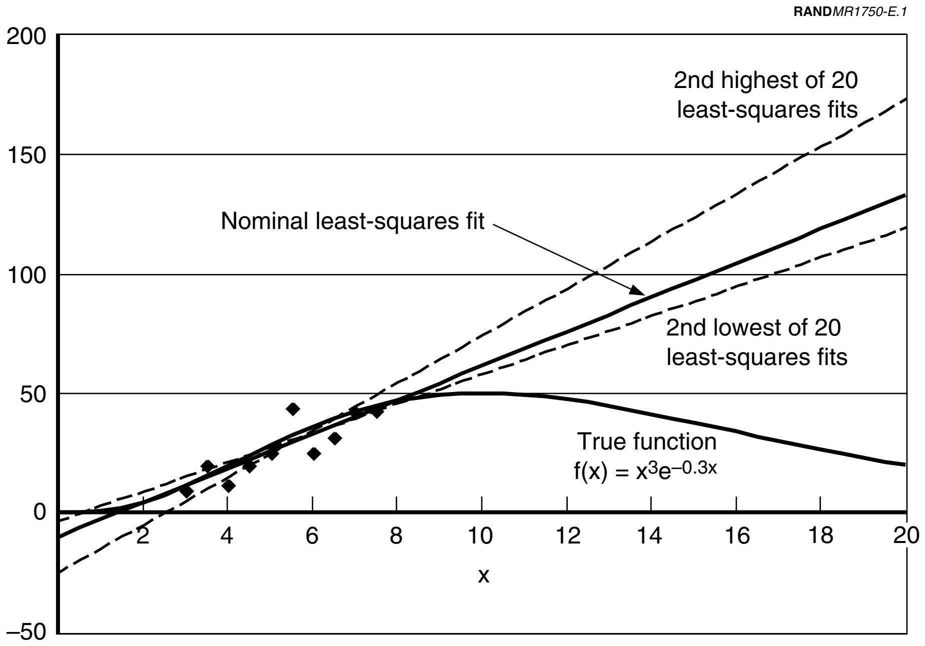

Figure E.1—Calibration and Structural Uncertainty

is small, the latter when *x* is large. In the present example, of course, the first factor is modeled as x^3, the second as $e^{-0.3x}$, but other functional forms could be proposed. To be less abstract, if x represents time, x^3 might represent the volume of an expanding cloud, e.g., from a rocket, and e^{-bx} might represent the decay of some chemical process.

A good story may be needed for interpolation as well as for extrapolation. Indeed, when we use a linear or quadratic function to interpolate between points in a dataset, we are implicitly assuming that the function is smooth. When fitting a sample from a waveform with sine and cosine function, we acknowledge that the result will match the actual waveform (as opposed to the data sample) only if the waveform contains no components with wavelengths much shorter than the distance between successive data points. "The function is smooth at this scale" isn't a very complicated story, but it is a story nonetheless.

This is not to say that "the function is smooth" is the only one that can be used for interpolation. Nor, of course, is it necessarily the right story. In particular, the sparser one's data, the more someone might question it.

F. Motivated Metamodels

In Davis and Bigelow (2003), we describe an experiment to test approaches to building metamodels. For an object model we used EXHALT-CF (Davis, McEver, and Wilson, 2002), a model that treats the halt phase of a military operation. In its simplest version, the halt phase is a mere race. An attacking force (Red) is advancing on an objective, while the defenders (Blue) interdict its armored vehicles with long-range fires. Red will halt when it reaches its objective (a Red win) or when Blue has killed a specified number of vehicles (a Blue win), whichever comes first. EXHALT-CF, however, adds many embellishments relevant to current strategic concerns about real-world military operations, especially in the Persian Gulf.

First, the model must represent Blue deployments. Some number of shooters may be stationed in the theater in peacetime. Depending on strategic warning, diplomatic relations, Red's deceptiveness, and Red's ability to threaten bases in the theater (e.g., with weapons of mass destruction (WMD)), Blue may or may not be able to augment this number before Red begins its advance. Once Red's advance begins, Blue will deploy more shooters into the theater, up to a theater capacity, which reflects logistical shortcomings.

The effectiveness of Blue's shooters is measured by kills per shooter-day. Early in the campaign, Blue may be unable or unwilling to attack the Red column because of Red air defenses. After a period of air-defense suppression, Blue's attacks will start. Even then, however, sortie rates may be reduced because of a continued threat of attack with WMD, which would force Blue personnel to work in protective gear or would force Blue to operate from more-distant and more-poorly-prepared bases.

The weapons and strategy Blue selects will also influence Blue shooter effectiveness. Blue may select an area weapon capable of killing several Red armored vehicles per shot. To counter this, Red may space its vehicles more widely. Or Blue may select a point weapon that kills no more than one vehicle per shot and is thus unaffected by Red's vehicle spacing. Also, Blue will likely have limited supplies of its best weapons and will revert to lesser weapons when its best are exhausted. Blue may attack the entire Red column in depth or focus its attack on the leading edge. If Blue does the latter, its attack may slow Red, but each sortie may be less effective than the previous one, due to deconfliction problems.

We set out to build a metamodel that would estimate Red's halt distance as a simple function of EXHALT-CF's inputs. We generated a dataset of 1,000 cases. At first, we used a purely statistical approach, blindly regressing the outcome (the Red halt distance) against everything in sight. Even adding cross-product terms as independent variables did little to improve matters. Of course, had we been fitting a response surface in a small neighborhood of a base case, this approach might have worked well. Assuming the model is smooth (twice differentiable when seen as a function), Taylor's Approximation from elementary calculus guarantees success if the neighborhood is small enough. But we were looking for a metamodel that performed well over a much larger region of the input parameters.

We constructed a series of four metamodels. For the first, we (almost) blindly regressed Red's penetration distance on the 25 input variables that we had varied to create the analysis dataset. For the second, we introduced some composite variables that actually appear as intermediate variables in the object model. For example, three of the variables in the analysis dataset are the number of Red divisions, the armored vehicles per division, and the fraction of Red vehicles that must be killed to effect a halt. These variables influence Red's penetration distance exclusively though their product, which we duly included in the second metamodel as a replacement for the original three.

We based the third and fourth metamodels on an explicit story, hardly more complicated than this. Red will start moving at a time T_{delay} and thereafter will move at a velocity V until it reaches its objective or Blue has killed enough vehicles, whichever comes first. Blue will begin shooting at a time T and thereafter will kill vehicles at a rate proportional to the number of shooters and the effectiveness of each shooter. The time to kill enough vehicles is the ratio of required kills to the kills per day. This story is very similar to the one described in Equation B.1 in Appendix B. In metamodel 3, we used a simple estimate for the average number of shooters in the theater. In metamodel 4, we used a more complicated expression for average shooters that accounted for the fact that more shooters would deploy if the campaign were longer. As shown in Table F.1, the more "theory" we added, the simpler the metamodels became and the better they fit the data.[1]

[1] A reviewer noted that we provide no guidance here on how to construct the theory to be used in a metamodel and no guidance as to how to assess alternative proposed theories. That is true. Modeling is a mix of art and science.

Table F.1

A Dose of Theory Improves the Metamodel

Metamodel Number	Number of Calibration Coefficients	Number of Aggregate Variables	RMS Error (km)
1	15	14	140
2	11	10	84
3	9	5	30
4	9	5	8

Bibliography

Abrahamse, Allan, James H. Bigelow, R. J. Gladstone, Bruce F. Goeller, Thomas F. Kirkwood, and Robert L. Petruschell (1977), *Protecting an Estuary from Floods—A Policy Analysis for the Oosterschelde*, Vol. II, *Assessment of Security from Flooding,* Santa Monica, CA: RAND.

Bankes, Stephen C. (1993), "Exploratory Modeling for Policy Analysis," *Operations Research,* Vol. 41, No. 3, May-June.

_____ (2002), "Tools and Techniques for Developing Policies for Complex and Uncertain Systems," *Proceedings of the National Academy of Sciences, Colloquium,* Vol. 99, Suppl. 3.

Bigelow, James H. (1984), "Managing Recoverable Aircraft Components in the PPB and Related Processes, Executive Summary," Santa Monica, CA: RAND.

Bigelow, James H., Joseph H. Bolten, and James C. DeHaven (1977), *Protecting an Esturary from Floods—A Policy Analysis for the Oosterschelde*, Santa Monica, CA: RAND.

Catlett, Louis, Richard Stanton, and Orphan Yildiz (1979), *Controlling the Oosterschelde Storm-Surge Barrier—A Policy Analysis of Alternative Strategies*, Vol. IV, *Basin Response to North Sea Water Levels: The BarconSIMPLIC Model*, Santa Monica, CA: RAND.

Crawford, Gordon B. (1988), *Variability in the Demands for Aircraft Spare Parts: Its Magnitude and Implications*, Santa Monica, CA: RAND.

Davis, Paul K. (1988), *The Role of Uncertainty in Assessing the NATO/Pact Central-Region Balance*, N-2839-RC, Santa Monica, CA: RAND.

_____ (ed.) (1994), *New Challenges in Defense Planning: Rethinking How Much is Enough*, Santa Monica, CA: RAND.

_____ (1995), *Aggregation, Disaggregation, and the 3:1 Rule in Ground Combat,* Santa Monica, CA: RAND.

_____ (2001), *Effects-Based Operations: A Grand Challenge for the Analytical Community,* Santa Monica, CA: RAND.

_____ (2002a), *Analytic Architecture for Capabilities-Based Planning, Mission-System Analysis, and Transformation,* Santa Monica, CA: RAND.

_____ (2002b), "Synthetic Cognitive Modeling of Adversaries for Effects-Based Planning," *Proceedings of the SPIE,* 4716 (27).

_____ (2003), “Exploratory Analysis and Implications for Modeling,” in Stuart E. Johnson, Martin C. Libicki, and Gregory F. Treverton (eds.), *New Challenges, New Tools for Defense Decisionmaking*, Santa Monica, CA: RAND, pp. 255–283.

Davis, Paul K., and James H. Bigelow (1998), *Experiments in Multiresolution Modeling,* Santa Monica, CA: RAND.

_____ (2001), “Metamodels to Aid Planning by Intelligent Machines,” *Proceedings of PERMIS 2001*.

_____ (2003), *Motivated Metamodels: Synthesis of Cause-Effect Reasoning and Statistical Metamodeling*, Santa Monica, CA: RAND.

Davis, Paul K., James H. Bigelow, and Jimmie McEver (1999), *Analytical Methods for Studies and Experiments on “Transforming the Force,”* Santa Monica, CA: RAND, DB-278-OSD.

_____ (2000), *Effects of Terrain, Maneuver Tactics, and C4ISR on the Effectiveness of Long-Range Precision Fires*, Santa Monica, CA: RAND.

_____ (2001), *Exploratory Analysis and a Case History of Multiresolution, Multiperspective Modeling*, Santa Monica, CA: RAND, RP-925.

Davis, Paul K., David Gompert, Richard Hillestad, and Stuart Johnson (1998), *Transforming U.S. Forces: Suggestions for DoD Strategy*, Santa Monica, CA: RAND, IP 179.

Davis, Paul K., David Gompert, and Richard Kugler (1996), *Adaptiveness in National Defense: The Basis of a New Framework,* Santa Monica, CA: RAND.

Davis, Paul K., and Richard Hillestad (1993), “Families of Models That Cross Levels of Resolution: Issues for Design, Calibration, and Management,” *Proceedings of the 1993 Winter Simulation Conference*, December.

Davis, Paul K., and Reiner Huber (1992), *Variable Resoltion Modeling: Issues, Principles and Challenges*, Santa Monica, CA: RAND, N-3400-DARPA.

Davis, Paul K., Jimmie McEver, and Barry Wilson (2002), *Measuring Interdiction Capabilities in the Presence of Anti-Access Strategies: Exploratory Analysis to Inform Adaptive Strategies for the Persian Gulf*, Santa Monica, CA: RAND, MR-1513-OSD.

Davis, Paul K., and James A. Winnefeld (1983), *The RAND Strategy Assessment Center*, *An Overview and Interim Conclusions About Utility and Development Options,* Santa Monica, CA: RAND, R-2945-DNA.

Defense Science Board (1998), *Joint Operations Superiority in the 21st Century: Integrating Capabilities Underwriting Joint Vision 2010 and Beyond,* Washington, DC: Office of the Under Secretary of Defense for Acquisition and Technology.

Fox, Daniel (2003), “Using Exploratory Modeling,” in Stuart E. Johnson, Martin C. Libicki, and Gregory F. Treverton (eds.), *New Challenges, New Tools for Defense Decisionmaking,* Santa Monica, CA: RAND, pp. 258–298.

Gritton, Eugene, Paul K. Davis, Randall Steeb, and John Matsumura (2000), *Ground Forces for a Rapidly Employable Joint Task Force*, Santa Monica, CA: RAND.

Haimes, Yacov (1998), *Risk Modeling, Assessment, and Management*, New York: John Wiley & Sons.

Harmon, S. Y., and S. M. Youngblood (2003), *Leveraging Fidelity to Achieve Substantive Interoperability*, Washington, DC: U.S. Department of Defense, Defense Modeling and Simulation Office.

Hodges, James (1980), *Onward Through the Fog: Uncertainty and Management Adaptation in Systems Analysis and Design*, Santa Monica, CA: RAND.

Kelley, Charles, Paul K. Davis, Bruce Bennett, Elwyn Harris, Richard Hundley, Eric Larson, Richard Mesic, and Michael Miller (2003), *Metrics for the Quadrennial Defense Review's Operational Goals*, Santa Monica, CA: RAND, DB-402-OSD.

Larsen, Ralph I., and C.E. Zimmer (1965), "Calculating Air Quality and Its Control," *APCA Journal*, Vol. 15, No. 12, pp. 565–562.

Law, Averill, and David W. Kelton (1991), *Simulation Modeling and Analysis,* 2d ed., new York: McGraw-Hill.

Lempert, Robert J. (2002), "A New Decision Science for Complex Systems," *Proceedings of the National Academy of Sciences Colloquium*, Vol. 99, Suppl. e.

Los, F. J. (1991), *Mathematical Simulation of Algae Blooms by the Model Bloom 1i,* Version 2.

Matsumura, John, Randall Steeb, Thomas J. Herbert, Mark R. Lees, Scot Eisenhard, and Angela B. Stich (1997), *Analytic Support to the Defense Science Board: Tactics and Technology for 21st Century Military Superiority,* Santa Monica, CA: RAND.

Matsumura, John, Randall Steeb, Ernest Isensee, Thomas J. Herbert, Scot Eisenhard, and John Gordon (1999), *Joint Operations Superiority in the 21st Century: Analytic Support to the 1998 Defense Science Board,* Santa Monica, CA: RAND.

McEver, Jimmie, Paul K. Davis, and James H. Bigelow (2000), *EXHALT*, Santa Monica, CA: RAND.

Meystel, Alex (1995), *Semiotic Modeling and Situation Analysis: An Introduction,* Bala Cynwyd, PA: AdRem, Inc.

Meystel, Alex, and James Albus (2002), *Intelligent Systems: Architecture, Design and Control,* New York: Wiley.

National Academy of Sciences (1996), *Post-Cold War Conflict Deterrence,* Washington, DC: National Academy of Science.

National Research Council (1997), *Modeling and Simulation,* Vol. 9 of *Technology for the United States Navy and Marine Corps: 2000-2035*, Washington, DC: National Academy Press.

Rosenau, James N. (1998), "Many Damn Things Simultaneously: Complexity Theory Theory and World Affairs," in David S. Alberts and Thomas J. Czerwinski (eds.), *Complexity, Global Politics, and National Security,* Washington, DC: National Defense University.

Rumsfeld, Donald (2001), *Report of the Quadrennial Defense Review,* Washington, DC: Department of Defense.

Saltelli, Andrea, Karen Chan, and Marian E. Scott (eds.) (2000), *Sensitivity Analysis*, Chichester, UK: John Wiley & Sons, Ltd.

Simon, Herbert (1982a), *Models of Bounded Rationality,* Vol. 1, Cambridge, MA: MIT Press.

Simon, Herbert (1982b), *Models of Bounded Rationality,* Vol. 2, Cambridge, MA: MIT Press.

Smith, Giles K., Gordon F. Acker, James H. Bigelow, David J. Dreyfuss, Richard La Forge, Richard Y. Pei, Susan A. Resetar, and Robert Petruschell (1988), *Design, Performance, and Cost of Alternative LHX Configurations*, Santa Monica, CA: RAND.

Zeigler, Bernard (1984), *Multifacetted Modelling and Discrete Event Simulation*, Ontario, CA: Academic Press.