Is It You or Your Model Talking? A Framework for Model Validation

James S. Hodges, James A. Dewar

RAND | PROJECT AIR FORCE
ARROYO CENTER
NATIONAL DEFENSE RESEARCH INSTITUTE

The research described in this report was sponsored by the United States Air Force, Contract No. F49620-86-C-0008; by the United States Army, Contract No. MDA903-86-C-0059; and by the Office of the Secretary of Defense (OSD), under RAND's National Defense Research Institute, an OSD-supported federally funded research and development center, Contract No. MDA903-85-C-0030.

ISBN: 0-8330-1223-1

Published 1992 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

R-4114-AF/A/OSD

Is It You or Your Model Talking? A Framework for Model Validation

James S. Hodges, James A. Dewar

Prepared for the
United States Air Force
United States Army
Office of the Secretary of Defense

RAND

PREFACE

This report had two sources: (1) the Military Operations Research Society's (MORS) First Minisymposium on Simulation Validation (SIMVAL), in Albuquerque, New Mexico, 15–18 October 1990 and (2) a paper called "Six (or so) Things You Can Do with a Bad Model."[1] The minisymposium prompted several new ideas, which were reflected in copious ad-libbing in the version of "Six (or so) Things" delivered at SIMVAL. This report incorporates those ideas in a substantially revised version of "Six (or so) Things," so that the MORS validation working group and others can use them to help DoD define validation, verification, and accreditation; to set standards; and to specify procedures.

The general sense of the minisymposium was that validation was by far the thorniest issue. While verification[2] presents some nagging problems, particularly reverifying a model that has been changed, plenty of fairly general standards and heuristics exist for verifying models. The only apparent difficulty with accreditation is how it relates to validation. Thus, this report is mainly concerned with defining validation and other terms appropriate for unvalidatable models, although accreditation is discussed briefly.

The preparation of this report was partially supported by RAND's Resource Management Department and Defense Planning and Analysis Department, using funds from the concept-formulation and research-support component of RAND's three federally funded research and development centers. Those centers are Project AIR FORCE, sponsored by the United States Air Force; the Arroyo Center, sponsored by the United States Army; and the National Defense Research Institute, sponsored by the Office of the Secretary of Defense and the Joint Staff.

---

[1]James S. Hodges, "Six (or so) Things You Can Do with a Bad Model," *Operations Research*, Vol. 39, No. 3, May–June 1991, pp. 355–365.

[2]"Determining that a model implementation accurately represents the developer's conceptual description and specifications," in the words of the MORS working group, M. L. Williams and J. Sikora, "SIMVAL Minisymposium," *Phalanx*, Vol. 24, No. 2, June 1991, p. 4.

SUMMARY

The problem of model validation is still with us. Why? Defense modelers seem to agree that validation has something to do with comparing models to reality, but disagree about how to go about it. The problem flows from a presumption that all models—be they of fuze activation or theater-level combat—can be validated and that a single validation standard and procedure can and should be defined. We believe that this presumption is a fundamental error: Some models can be validated and used to predict, while others cannot be validated and may only be put to nonpredictive uses. With this distinction, it is straightforward to define validation for validatable models and to define standards of *evaluation* for nonpredictive uses of unvalidatable or unvalidated models. Without this distinction, people modeling fuzes and theater-level combat will talk past each other forever, because their problems are fundamentally different.

This report lays out a conceptual framework for validation. The framework assumes that the standard of quality for a model should be based on the model's intended uses. The framework begins by defining a "prediction" as a statement about an observable thing, a statement about how accurate the prediction is, and a particular kind of argument about why someone else should believe those two statements. We contrast this with a common notion of prediction, which is merely a statement about an observable thing. Our definition is more restrictive because that cleaves the problem of quality assurance in the right place: Models intended to predict (according to our definition) must carry a certain kind of warranty, while under the looser definition, no clear definition of quality standards is possible.

Validation of predictions is then defined strictly—consistent with traditional usage—as the activity that supplies the statement about how accurate the prediction is and the supporting argument. Thus, only models intended to predict need to be validated. Before a model can be validated, by this standard, it must be *possible* to validate it. It is possible to validate a model when the situation being modeled satisfies four prerequisites:

1. The situation must be observable and measurable.
2. The situation must exhibit constancy of structure in time.

3. The situation must exhibit constancy across variations in conditions not specified in the model.
4. The situation must permit the collection of ample data.

Prerequisites 1 and 4 are straightforward. Prerequisite 2 is needed to ensure that a model is predictive for the same conditions as those in the validation tests. Prerequisite 3 is needed to ensure that a model is predictive for conditions that differ from those in the validation tests. Although we have stated the prerequisites as if one would specify the situation of interest and then ascertain whether the prerequisites were satisfied, it is equally possible to ascertain the range of conditions under which the prerequisites are satisfied and then see whether that range covers the situation for which a prediction is needed.

Models that can be validated *accrue validity* as they pass more varied and exacting predictive tests. The range of conditions for which Prerequisites 2 and 3 are satisfied is determined by the collection of predictive tests a model passes. If a model fails a predictive test, it can either be remade and start again at zero validity, or it can remain the same and be a candidate to satisfy Prerequisites 2 and 3 for a smaller range of conditions, excluding those of the failed test.

The notion of accruing validity may appear to conflict with our key assertion that validatability of models does not range along a continuum. But no conflict exists: Models that can be validated are different from models that cannot be; validity can accrue for the former but not for the latter. For *validatable* models, validity is not binary but accrues along a continuum between "not valid" and "valid."

If a model is unvalidated or unvalidatable, it may not be used to predict, but that does not mean it is useless. There are at least seven distinct nonpredictive uses for models:

1. As a bookkeeping device, to condense masses of data or to provide a means or incentive to improve data quality;
2. As an aid in selling an idea of which the model is but an illustration;
3. As a training aid, to induce a particular behavior;
4. As part of an automatic management system whose efficacy is not evaluated by using the model as if it were a true representation;
5. As an aid to communication, e.g., in teaching or in operating organizations;
6. As a vehicle for *a fortiori* arguments; and

7. As an aid to thinking and hypothesizing, e.g., as a stimulus to intuition in applied research or in training or as a decision aid in operating organizations.

The appropriate standard of quality for a model in a nonpredictive use depends integrally on that use. For example, evaluation of a model for Use 3, "as a training aid, to induce a particular behavior," consists of measuring whether use of the model induces the desired behavior. In this case, realism of the model is not essential; in fact, specific aspects of realism are often sacrificed to achieve particular training goals.

In cataloging nonpredictive uses of models and describing the appropriate standard of evaluation for each, this report shows that **the appropriate form of quality assurance for a model depends fundamentally on how the model is used, so any attempt to define a single validation standard and procedure for all models in all uses will surely fail.**

Our framework's strength is that it establishes specific criteria for assessing model quality based on intended uses. By doing so, it illuminates "accreditation," which becomes a formal decision about the success of a model's validation or evaluation and an official certification of such things as documentation and configuration control. While we do not claim the framework is the last word on all validation issues, it does provide a conceptual basis on which standards and procedures can be defined.

ACKNOWLEDGMENTS

In writing this report, we benefited from comments by John Adams, Steven Bankes, Carl Builder, Richard Hillestad, Mario Juncosa, Eric Larson, Adnan Rahman, Peter Stan, and Warren Walker of RAND and by Col Thomas Cardwell of USAF Studies and Analysis. Bart Bennett and Edward Harshberger of RAND made formal reviews that were extensive and very useful. Discussions with Paul Davis were also helpful, although he disagrees strongly with parts of the report. Robert Levine impressed on us the importance of "the model says X." Gene Woolsey and Hugh Miser made valuable editorial suggestions about an earlier version. Paul Steinberg saw to it that it had a beginning, as well as a middle and an end. These generous people do not necessarily agree with this report.

CONTENTS

1. INTRODUCTION

A half-century after the emergence of operations research, the problem of model validation lingers like an unpaid creditor. Defense modelers, at least, generally agree that validation is somehow about comparing models to reality, but disagree about how to do it. The problem is illustrated by such papers as the classic paper by C. J. Thomas[1]—considered definitive by many—which begins by presuming that validation is "an investigation of the agreement of the model with reality"[2] and concludes that such a thing cannot be defined, much less executed. The implication is that we may not be able to define it, but we must strive to do it anyway—a problematic piece of advice.

What is the origin of this impasse? All too often in serious conversation, one hears that although defense models vary greatly—from models of small-scale events like fuze activation to theater-level combat simulations—all models *can* be validated and that one procedure and standard of validation should be devised and used across the board. We believe that this presumption is a fundamental error: Some models can be validated and used to predict, while others cannot be validated and may only be put to nonpredictive uses. With this distinction, it is straightforward to define validation for validatable models and to define standards of evaluation for nonpredictive uses of unvalidatable (or merely unvalidated) models. *Without* this distinction, fuze modelers and theater-level combat modelers will continue to talk past each other fruitlessly, because their problems are fundamentally different.

The idea of model validation focuses attention on the model to the exclusion of analysts, data, and other elements of the decisionmaking context. We continue that focus here, isolating models to highlight their contribution to an analysis, thus permitting us to define assessment of a model's quality as evaluation of the efficacy with which it makes its contribution.

---

[1]C. J. Thomas, "Verification Revisited—1983," in *Military Modeling*, W. P. Hughes, Jr., ed., Military Operations Research Society, Alexandria, Virginia.

[2]Thomas, footnote to p. 293; compare the MORS validation working group's definition: "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model," M. L. Williams, "SIMVAL Minisymposium," *Phalanx*, Vol. 24, No. 2, June 1991, p. 4.

To do this, we set forth a framework based on the idea that the standard of quality for a model should depend on the model's intended use. The framework has several parts. First, Section 2 defines a "prediction" as a statement about an observable thing (along with a statement about how inaccurate the prediction can be) and a certain kind of supporting argument (summarized as "the model says X"). If an analyst wants to make predictive use of his model—to say "the model says X"—then it must be validated in a strict sense that is consistent with but broader than traditional scientific usage (Section 3). But before a model can be validated, it must be *possible* to validate it. It is not always possible to do so, and Section 3 gives four prerequisites a situation must satisfy so that a model of it can be validated. This latter discussion makes clear how a model is validated, after which we discuss other proposed ways of validating models, most of which—including validation of submodels—are insufficient for models intended to make predictions.[3]

If a model cannot be validated, it may not be used to predict. Many people infer that such a model is not useful. This does not follow, and in Section 4 we identify seven logically distinct nonpredictive uses for models and the type of evaluation appropriate for each. We say "evaluation" because we reserve "validation" for its stronger traditional sense and do not want to diminish its force by stretching it to cover unvalidatable models in nonpredictive uses.

Our central idea—that some models are validatable and others are not—departs from the fundamental presumption of the validation debate, noted above. Many models—and, in such areas as theater-level combat, maybe *all* models—cannot be validated, so it is pointless to try. But analysts are not off the hook: If a model has not been or cannot be validated, it may not be used to predict. This does not mean that analysts can ignore the quality of their models; it does mean that "quality" is not equivalent to "agreement of the model with reality." For uses other than prediction, quality and thus evaluation can be defined straightforwardly, depending on use.

---

[3]Other authors are taking different approaches to this problem. For example, RAND colleagues P. K. Davis and R. J. Hillestad define something they call "generalized validation (evaluation)" and provide a taxonomy of methods that can be brought to bear. See P. K. Davis, *Generalizing Concepts and Methods of Verification, Validation, and Accreditation (VV&A) for Military Simulations*, RAND, R-4249-DR&E, forthcoming. See also Office of the Secretary of Defense, Defense Modeling and Simulation Office, "Defense Modeling and Simulation Initiative: Appendix B," 1 May 1992, which reflects the work of MORS and a DMSO working group chaired by Davis.

One important feature of our framework is that it catalogs model uses and establishes specific criteria for assessing model quality based on use. By doing so, it illuminates accreditation, which becomes a formal decision about the success of a model's validation or evaluation and an official certification of such things as documentation or configuration control. We do not claim to have the last word on all issues related to validation. We do think that this report provides a conceptual framework suitable for defining standards and procedures.

2. PREDICTIVE USES OF MODELS, OR USES OF THE FORM "THE MODEL SAYS X"

TWO DEFINITIONS OF PREDICTION

In common usage, the term *prediction* embraces a wide class of statements ranging from forecasts of eclipses to hunches, such as "I predict the Lakers will beat the Bulls tonight." In this usage, a prediction is simply statement about an observable or potentially observable quantity or event. Following this usage, treatments inspired by policy or systems analysis might define a predictive use of a model as one in which

- A statement about an observable or potentially observable quantity or event is produced.

A prediction, in this view, is a statement about observable things that you have some reason to take seriously, presumably because it carries all the intuitive content you can give it—but it carries no warranty as to its accuracy.

We propose more restrictive definitions of prediction and predictive uses of models. Informally, the test that identifies a predictive use of a model is: Will the analysis have the form "the model says X"? For example, "the model says an eclipse of the moon will occur on Thursday at 4:23 PM," or "the model says that if a ringing bell is paired with presentation of food for r repetitions over d days in a spare environment, dogs will salivate at the sound of a bell for w weeks thereafter." If model outputs are presented this way, the model is being used predictively. More formally, a predictive use of a model is one in which

- A statement about an observable or potentially observable quantity or event is produced;
- The modeled situation is such that predictive accuracy *can* be measured; and
- The predictive accuracy of the model in the situation *has* been measured.

The addition of the last two bullets is a significant restriction of the common usage. By this second definition, a prediction is a statement about something that can be observed, a statement about how accurate the prediction is, and a particular kind of argument about why

someone else should believe those two statements. As a result, the prediction may, with known accuracy, replace a measurement of the modeled situation.

The first, more common, notion of *prediction* includes models and uses that the second does not. For example, the first permits invention of a model out of whole cloth, as long as a claim is made that the model has intuitive content. It also includes what might be described as hypothesis generation: Manipulate the model, make some hypotheses, test them elsewhere. Our definition excludes such uses: It requires that a prediction be a statement about what *will* happen with a warranted measure of its error. (Later, we classify hypothesis generation as a nonpredictive model use.)

ADVANTAGES OF THE MORE RESTRICTIVE DEFINITION

Why should our more restrictive definition be preferable? Definitions make distinctions; it is important to make useful distinctions. If validation is to be defined as a function of use—and all treatments of validation do, though most give it little emphasis—then it is crucial to define "uses" so as to distinguish cases requiring qualitatively different types of quality assurance. Our definition does this: Uses it covers require a warranty of predictive accuracy, and uses it excludes do not. The more common notion of prediction obscures this distinction.

For example, the common notion provides no way to distinguish physics from astrology. Astrologers turn inputs (birth data) into outputs expressed on observable scales (fortunes in love and war) using models full of intuitive content assembled over the millenia. Our definition does allow one to distinguish physics from astrology by introducing public statement of the accuracy of predictions (which astrologers scrupulously avoid) and the basis of the belief that that accuracy will hold in unmeasured situations.

The two definitions also differ in the light they throw on validation. Our definition requires forecasts of observable things with known accuracy, so validation consists of tests to measure accuracy and to justify claiming that it is "known." Model uses fitting the first definition but not ours are (in our terms) nonpredictive uses, and assessment of their quality is a qualitatively different task. It seems clear that a model used in hypothesis generation need not meet as high a standard as a model used to make a soft landing on Mars. The common notion of prediction mushes these uses together and provides no way to define a stiffer standard for the cases identified by our definition.

Parenthetically, the two definitions are also relevant to aspects of model quality other than validation, such as parsimony. Our definition implies a natural standard of parsimony: If a factor's exclusion does not create systematic prediction errors that are "too large," then it can be omitted. The common notion of prediction actually militates against parsimony: Since you can't test whether a factor is important, you typically end up including everything that you think matters. Other principles must be added to engender parsimony or to justify models omitting factors that are "known to matter"—arguments like "you can't do systems analysis with a detailed model, because there are too many things to vary."

EXAMPLES OF PREDICTIVE USES OF MODELS

Our notion of prediction is quite expansive. To suggest its breadth, we give some examples of predictive uses of models.

Predictions, in our sense, can be specific or weak. One example of a specific prediction is a forecast of the location of a planet: "Mars will be at such-and-such a position at such-and-such a time." Highly accurate specific predictions are needed to make a soft landing on Mars. In contrast, Aristarchus's outmoded model of circular planetary motion makes specific predictions with known accuracy (up to 15 degrees of error for Mars), good enough for backyard astronomers.

An example of a weak prediction is a comparison, such as "weapon A will survive more often than weapon B." This is a weak prediction because it only forecasts that A will survive more often than B, not how often either will survive. Nonetheless, it is a prediction and thus is subject to the standard to be introduced in Section 3. Another kind of weak prediction is an *a fortiori* argument, discussed in Section 4. Probabilistic predictions are also weak predictions in that they do not forecast specific events, but some have been shown to satisfy all three aspects of our definition. For example, weather forecasts like "the chance of rain tomorrow is 70 percent" have been shown to be accurate in the sense that about 70 percent of such predictions are followed by rain.[1]

Finally, predictions in our sense are not limited to the physical sciences. One such prediction, Pavlovian conditioning, was mentioned earlier, and others will be mentioned below.

---

[1]A. H. Murphy and R. L. Winkler, "Reliability of subjective probability forecasts of precipitation and temperature," *Applied Statistics*, Vol 26, 1974, pp. 41–47.

Our definition does exclude one large class of predictions, which we will now describe. The essential distinction is given by Mayr,[2] who distinguishes *temporal* and *logical* predictions. A temporal prediction is an inference from the present to the future, such as a prediction of the point of impact of an artillery shell, made just before the shell is fired. A logical prediction forecasts the conformance of individual observations with a theory or scientific law. For example, based on a theory of evolution and of a species' origin, a biologist predicts that other species will be found in the fossil record to have evolved by the same mechanism. Some temporal predictions are logical predictions, but many logical predictions are not temporal predictions, the evolutionary example being an instance.

Our notion of predictive uses of models refers primarily to temporal predictions, because they are relevant to decisionmaking. Logical predictions may be part of building the science that is used, in turn, to build models to aid decisions, but the models themselves make temporal predictions.

---

[2]E. Mayr, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*, Belknap Press, Harvard University Press, Cambridge, Mass., 1982.

3. VALIDATABILITY AND VALIDATION

If an analysis will produce results in the predictive form "the model says X," then it must make a statement about how accurate the prediction is and make a particular kind of argument about why somebody else should believe that statement. In our framework, validation supplies the statement of predictive accuracy and the argument about believability. That is, we define "validation" to apply only to models in predictive uses. This is deliberate and in keeping with scientific usage for temporal predictions, as defined by Mayr. The tendency in the validation debate is to lump under "validation" all quality-assurance activity other than verification. We will use "validation" for models in predictive uses and "evaluation" for models in nonpredictive uses.

If an analysis is to take the form "the model says X," that model must be validated, but before that can be done, it must be *possible* to validate it. This section argues that the situation being modeled determines whether a model of it can be validated, and that some situations do not allow the possibility of validatable models.

The cleavage between validatable and unvalidatable models is important, because if a situation does not admit validatable models, no model of the situation may be used to make predictions. Models of the situation *may* be put to nonpredictive uses without being subjected to the rigors of an attempted but inevitably unsuccessful validation, although such models must be evaluated in the manner appropriate to the nonpredictive use.

We reiterate that a model that cannot be validated in this sense is not necessarily useless; it simply may not be used to make sentences like "the model says X." Section 4 discusses seven logically distinct nonpredictive uses.

A DEFINITION OF VALIDATABILITY

We do not claim to have a definitive test of validatability, but the four prerequisites that follow have withstood considerable scrutiny.[1] Before discussing the prerequisites, we make two general points. First, the prerequisites apply not to models but to the situation being

[1]At this point we are willing to say, as Justice Potter Stewart once said of obscenity, that we know it when we see it.

modeled. A model inherits validatability from the situation it models. Second, a situation may not satisfy the prerequisites at a given time, but technological or scientific progress can change that. For example, some models based on physical laws were once unvalidatable, but many of these models now are validatable. Examples given below will make these points clearer.

Prerequisite 1 (P1): It must be possible to *observe and measure* the situation being modeled.

It must be possible to make specific predictions from the model, to take measurements corresponding to the predictions, and to compare the predictions and measurements without adjusting model inputs or outputs. P1 makes it possible to compare predictions to actual measurements, a requirement of scientific validation. If the real thing cannot be measured, one cannot judge whether a model's predictions are any good. An example of a situation in which P1 is not satisfied is a notional weapon system going against notional countermeasures: Neither exists, so neither can be measured—even if "the physics of the situation" are thought to be well-known—so no model of the situation can be validated.

Prerequisite 2 (P2): The situation being modeled must exhibit a *constancy of structure in time.*

Consider a situation in which water flows turbulently through some system. Turbulent flow still eludes physicists, but a model of the system could be constructed by measuring the flow and fitting curves to the data. In this instance, one could presume that the model would predict the system's future behavior under the same conditions, even without specific theory. That confidence would derive from a presumed constancy of the as-yet-unknown physical laws underlying the behavior of the system. This is *constancy of structure in time*: One must have reason to believe that the causal structure of the situation is sufficiently constant that measurements taken at one time can be reproduced under the same conditions at a later time. This would not be true, for example, of the specific path of a small object through our turbulent system, which would not be expected to be the same in two trials.

For an example where P1 holds but P2 might not, consider small-unit combat. If an infantry platoon is measured at a given time, do those measurements predict future measurements of the platoon? Even in a controlled test, soldiers learn from one trial to the next, so mea-

surements of one trial need not predict measurements of other trials. But some measurements may predict. One possibility is Marshall's famous assertion that only a small fraction of infantrymen actually fire their weapons in combat.[2] Treating Marshall's assertion as a model, and assuming it were accepted as true, it would be an example of constancy of structure that was not accepted at one point—so that no validatable model could be built then—but that came to be accepted with the accumulation of experience, thus permitting a validatable model. This is an example not only of constancy of structure becoming accepted over time, but also of it being established outside the hard sciences.

Prerequisite 3 (P3): The situation being modeled must exhibit a *constancy across variations in conditions not specified in the model.*

Return to the example of water flowing turbulently through a system. We said above that a presumed physical constancy allows an assumption that measurements taken under given conditions predict the system's behavior under the same conditions. One might systematically vary conditions—rate of flow or pressure, say—and model empirically how the conditions relate to output measures. But if a change were then made in some other condition not specified in the empirical model, all bets would be off about predictive power. Without further justification, the system could not be presumed to be constant across variations in conditions not specified in the model. Thus, the model could not be validated for predictions for situations in which conditions were not known to be identical to those used to build the model empirically.

As with P2, physics provides many situations satisfying P3, because outcomes are often driven by a few factors. In contrast, outcomes in evolutionary biology can be driven by countless interventions, and predictions are for the foolhardy. But, again, this does not mean that only physical models can be validated. For example, Ehrenberg has gathered data sets measuring children of both genders and a great variety of ages, nationalities, and social classes and found that within each data set, log (average weight) is close to 0.4 + 0.8 (average height) with stunning consistency.[3] Some systematic deviations can be found—e.g., French boys tend to be heavier than other boys as they

---

[2]S.L.A. Marshall, *Men Against Fire*, Gloucester, Mass., Peter Smith, 1978.

[3]A.S.C. Ehrenberg, "The elements of law-like relationships," *Journal of the Royal Statistical Society, Series A*, Vol. 131, pp. 280–302.

get older—but these deviations are small. In the complete absence of contradictory data sets, one could presume that height and weight have this relationship across unspecified conditions.

P3 must hold so that test measurements are relevant to future situations under which unspecified conditions vary from those that held in the test. The great fear of testers is that they will do a well-designed test, varying many conditions, and the system will fail anyway because the test missed a condition that was more important than those in the test. This is the point of P3: If the model is validated with some conditions fixed, then it is valid only for future situations in which those conditions hold, unless the situation being modeled is known to exhibit constancy across variation of unspecified conditions.

One further issue must be addressed: The difference between known and unknown unspecified conditions. The history of science is full of examples of unknown unspecified conditions popping up to confound a "validated" law. Newton's laws, for example, were thought to be valid across all unspecified conditions until Einstein pointed out that they break down at significant fractions of the speed of light. Cases like these are sometimes used by those who would debunk the notion of validity altogether. We think this undesirable, for theoretical and practical reasons. The theoretical reason is that such a new unspecified condition does not make the earlier successful tests irrelevant; the conditions under which P3 can be assumed to hold have simply been narrowed by the discovery of the new condition. The practical reason is that the possible existence of unknown important conditions should not relieve modelers of the obligation to worry about specified and known unspecified conditions, if they want to claim that their model predicts.

Prerequisite 4 (P4): It must be possible to collect ample data with which to make predictive tests of the model.

It must be possible to gather enough data to conduct predictive tests and to permit models to accrue validity. For a case in which P1 through P3 might be argued to hold but P4 does not, consider theater-level combat. One can imagine having data from millions of actual wars from which to ascertain the range of situations across which P2 and P3 hold. Should we then consider models of theater-level combat validatable, given that the millions of wars do not exist and never will? Models of such a situation might be said to be potentially validatable, but the potential cannot be realized, because the data to do the validation cannot be collected.

"Ample data" is deliberately vague. We do not think it is possible to be more specific at this point, except to say that it must be possible to gather vastly more observations than the number of adjustable constants in the model.

THE CENTRAL ROLE OF THE CONSTANCY PREREQUISITES, P2 AND P3

The *role* of the two constancy prerequisites, P2 and P3, can be summarized simply: P2 is necessary if you want to validate a model for the same conditions as those in your tests, and P3 is necessary if you want to validate a model for a wider range of conditions. The *content* of P2 and P3 is conceptually difficult, and given their importance, we will discuss them further.

We have defined P2 and P3 as if one would specify a range of situations and then see whether P2 and P3 hold. But one can work in the other direction, by determining a range of situations across which P2 and P3 hold, and then seeing whether that range covers the needed cases. P2 and P3 define the range of cases for which a model can be argued to predict; if that range covers the situation for which a prediction is needed—and the model is validated—then the model may be used to make predictions. Otherwise, it may not.

As an example, consider a validation done for a model of a Vietnam-era air-to-air missile.[4] A model was built of the missile's performance against North Vietnamese aircraft, an extensive validation was done in the United States, and the model was used to compute the missile's kill probability. When the missile was deployed, its kill probability was much lower, because the behavior of American and North Vietnamese pilots created conditions different from those of the validation. In our view, the model was validatable and validated—all four prerequisites could be supported *for conditions like those in the test*—but that range of conditions did not include wartime conditions. The issue for the modelers was (or should have been) whether wartime conditions were included in the range of cases for which P2 and, in this case particularly, P3 held. If so, the model should have predicted; if not, the model could not be expected to predict.

This and similar examples have been presented in support of arguments that it is necessarily a subjective judgment whether P2 and P3 are satisfied. We consider this the most serious challenge to the framework presented here, so we discuss it further. The argument

[4]We omit the details because they are unnecessary for our purpose.

against the utility of P2 and P3 goes as follows: In validating and using a model, one must make judgments about system stability and hidden variables that go beyond anything empirical. That is, one must invoke an implicit or explicit metamodel of the situation under study—*meta* in the sense of more comprehensive—which is necessarily subjective and error-prone, so any validation is necessarily subjective, error-prone, and a matter of degree.

This argument has three parts:

- One must make assertions about system stability and hidden variables;
- Such assertions constitute a metamodel; and
- The metamodel goes beyond anything empirical and is necessarily subjective and error-prone.

The first bullet restates P2 and P3, and the second bullet says that P2 and P3 require a model more comprehensive than the one in hand. These bullets create no problems. The third bullet is about the difficulty of making a case that P2 and P3 hold, and here we disagree. To say that arguments for P2 and P3 are necessarily subjective and extraempirical is to deny the history of the physical sciences: Stoichiometry and Newtonian mechanics satisfy P2 and P3 for well-defined ranges of cases; is there any shortage of empirical evidence? The second bullet does suggest an infinite regress in which the four prerequisites must be satisfied for the metamodel, which, in turn, requires a metametamodel, and so on. But indisputable cases, such as stoichiometry, demonstrate that infinite regress is not inevitable. Moreover, P2 and P3 can be founded firmly for cases outside the physical sciences, as the example of children's heights and weights illustrates.

That does not mean that it is *easy* to support P2 and P3, or that it is even possible for many situations of interest to defense modelers. But this is a cruel feature of life, not an argument to dispose of P2 and P3. If you want to be able to predict and to give a tested accuracy with your prediction, you need P2 and P3. If you do not have P2 and P3, you cannot make predictions.

Besides, P2 and P3 force a modeler to ask the right questions about the breadth across which a model can be validated. If one asks "should this model be used in this situation?" and interprets that as a question about predictive validity, then the questions that need to be asked and answered are summarized by P2 and P3.

IMPLICATIONS OF THE PREREQUISITES FOR MILITARY COMBAT MODELS

The full implications of these prerequisites will be explored incompletely in the rest of this report, so we need to be explicit now about their implications for combat models. As stated earlier, we were led to our framework as a means of breaking the validation impasse between predictive and nonpredictive kinds of models. This framework has led some to believe that we have drawn the line between predictive and nonpredictive so as to put all combat models in the nonpredictive category. However, we do not assert that all combat models are unvalidatable.

It is true that the prerequisites make it difficult to model combat validatably. Satisfying P1 is not a problem: Many aspects of combat are measurable and have been measured. Prerequisites P2 and P3 are more problematic. It is difficult to show that human behavior is constant in time or across unspecified conditions, although the example of the children's heights and weights and Ehrenberg's examples from marketing[5] show that some measures of human activity exhibit P2 and P3 constancy. The weather-forecasting example suggests further that some aspects of the idiosyncracies of combat might be modeled and validated. The most problematic prerequisite, however, may be P4. Detailed combat models typically have hundreds of adjustable constants, and historical examples or analogs are typically limited to a few dozen at most (with no incentive to gather more). This mismatch in number of adjustable constants and number of data points makes is unlikely that large-scale combat can be modeled validatably.

It is not necessarily impossible to model combat validatably. At the other end of the combat spectrum, there is some reason for hope. Small-unit combat involves many fewer adjustable constants than large-scale combat, and arenas such as the National Training Center at Fort Irwin provide a venue for taking many measurements on some aspects of small-unit combat. It is conceivable that a persuasive argument could be made that small-unit combat would satisfy all four of our prerequisites.

ACCRUING VALIDITY

Given that a situation admits a validatable model, how does a model of it attain validity? We use a notion of validity given by Miser and

[5]A.S.C. Ehrenberg, *Repeat Buying*, 2d ed., Charles Griffin, London, 1988.

Quade,[6] which was intended for predictive uses. They argue that validity is not binary—i.e., valid or not valid—but a degree of credibility that accrues as a model survives more varied and exacting predictive tests. (A predictive test consists of using the model to predict specific actual measurements, making the measurements, and comparing them without fiddling with model parameters or inputs.) In our terms, if you want to say "the model says X," it must have passed predictive tests of enough variety and difficulty that an honest argument can be made that what "the model says" is actually going to happen. The tests should, as noted, cover the entire area circumscribed by the P2 and P3 arguments.

The notion of accruing validity may appear to conflict with our key assertion that validatability of models does not range along a continuum. But no conflict exists. *Validatability* does not range along a continuum: A clear distinction can be made between models that can be validated and models that cannot be, the former being *models for which validity can accrue* and the latter *models for which validity cannot accrue*. For validatable models, *validity* is not binary, but accrues along a continuum between "not valid" and "valid." Validatable models that have accrued no validity are at the same end of the continuum as are unvalidatable models; what makes them different is that unvalidatable models cannot move off zero, while validatable models can.

What happens if a model fails a validity test? In simplistic terms, it goes onto the junk heap of failed models. In reality, failing a validity test happens all the time. A model is tested, it fails, the model is changed, and it is retested. This process continues until the model stops failing validity tests. Validity then continues to accrue up to some level of confidence.

To make this point more explicit, consider a model that has just failed a validity test. There are generally two possible explanations. One, there is something wrong with the model, and it should be fixed. After it is fixed, it must *reaccrue* validity (from zero) in all areas affected by the test failure. Two, there is nothing wrong with the model, but something wrong with the P2 and/or P3 arguments. In this case, the model does not lose any of its validity, but it *does* lose in the collection of conditions over which it is predictive. This suggests another way to view the prerequisites P2 and P3: The collection of

[6]H. J. Miser and E. S. Quade, "Validation," in *Handbook of Systems Analysis: Craft Issues and Procedural Choices*, North-Holland, New York, 1988, pp. 527–565.

successful predictive tests conducted to date determines the range of conditions for which P2 and P3 can be argued.

VARIANTS AND SUBSTITUTES FOR PREDICTIVE TESTS

The notions of predictive uses of models and of validity accrual are demanding—by design—and many alternatives less exacting than predictive tests have been proposed. We now examine these alternatives to evaluate them and to illuminate our preferred notion of validation. We have found only two to be acceptable.

The first acceptable alternative is a transitivity argument: Use the outputs from validated model B to validate a simpler model A. B must actually be valid—there *is* no free lunch. Moreover, if this method is used, model A cannot be shown to be any more accurate than model B. If B is known to be accurate to within Z units and if A reproduces B exactly, then A is at best accurate to within Z. However, if A reproduces B with an error of W units, A is accurate to within

$$\sqrt{Z^2 + W^2} \ .$$

(This result actually holds for standard deviations when the errors of B and A are stochastically independent, but it is not misleading as a rule of thumb.)

The other acceptable substitute for predictive tests is tests based on observational data, e.g., data from administrative files or data on historical battles. Such data have the disadvantage that there were no controls on experimental conditions or assignment of conditions to individuals. Of course, some sets of observational data (e.g., observations of Mars' orbit) may be uncontrollable in an experimental sense, yet may completely describe the system if governed adequately. In general, however, it is hazardous to draw causal inferences from such data unless it is clear how they relate to the requirements of an adequate notional experimental design. In any case, if a model cannot reproduce a result in observational data, the model's owner has something to explain, and if the result cannot be attributed to defects in the data, the model loses credibility as a predictor.

Other proposed substitutes do not make the grade. We will discuss four: validating a model by validating its submodels, validating a model against data used to construct it, checking the model's math and logic, and peer review.

Some suggest that it should be "legal" to validate a large model by validating its submodels, for example, when it is possible to get output measurements for the submodels but not for the large model. In general this will not do: Showing that the submodels predict intermediate measurements with known accuracy is not the same as showing that the total model predicts ultimate outputs with known accuracy. If an analysis only uses one submodel of the larger model, the larger model has been tautologically validated for that use. Otherwise, validatability of submodels does not carry over to the larger model. Consideration of P2 and P3 for the submodels establishes the range of situations in which each is validatable, and the range of situations for which the overall model is validatable is no greater than the intersection of the ranges for the submodels. That intersection could be empty; if so, the submodels could all be validatable while the larger model was not.

Another alternative often suggested is a kind of circular reasoning in which a combat model (say) is run; the outputs are compared to a new war and differ greatly; the model's inputs or parameters are adjusted until the outputs are "close enough" to the actual war; and then increased credibility is claimed. The circularity is clear: We alter the model until it reproduces the wars in our data set; therefore, the model is validated because we have shown that it reproduces the wars in our data set. This is particularly egregious in models with many adjustable constants, as is clear to anyone familiar with linear regression: If enough terms are added to the right-hand side of a regression, it *fits* perfectly, but that does not mean that it *predicts* perfectly.

This kind of circular reasoning is healthy and useful at the right time, e.g., during hypothesis generation. If a hypothesis cannot explain existing data, it is in trouble; if it can, it may be a candidate for a predictive test. In this way, making models fit existing data can be a pruning device and a mental exercise, but it does not bolster a model's validity.

One variant of this reasoning *is* acceptable for accruing validity. It is not uncommon to leave aside part of a data set, fit a model to the bulk, and predict the left-out data as a test of validity. This *is* a legitimate step in a progression of ever-more-convincing predictive tests, the next step being prediction of a distinct data set, and the most convincing being a series of successful forecasts of data sets of varied origins.

Some suggest that checking a model's math and logic should increase its validity. To the contrary, these activities are irrelevant to validity.

If a model is unvalidatable, nothing can make it "more valid." If a model is validatable *and* valid, checking its math and logic will increase its chances of passing a predictive test, but cannot show the model's validity: Only predictive tests can do that. The third logical possibility is that a situation permits validatable models but that a given model of it is not valid. Then, checking its math and logic cannot make the model valid. Looking through algorithms for counterintuitive things may help in evaluation for nonpredictive uses or in verification, but by itself provides no extra confidence in a model's predictive accuracy.

Similarly, peer review is not an adequate substitute for predictive tests. In this approach, a piece of work is given to parties with no interest or with a competing interest. They try to discredit it, and if they fail, the model is supposed to be "more valid" than before the peer review. While this can be useful in evaluating models for nonpredictive uses (Section 4), it does not contribute to validity unless it measures predictive accuracy or strengthens an argument for a measure of predictive accuracy. If the peer review consists of a predictive test, yes; if it consists of bouncing model outputs or algorithms off an antagonist's intuition, no.

Our concept of validation is strict and demanding, but consistent with traditional scientific usage. It also makes clear the central importance of distinguishing models intended for predictive uses from models intended for other uses: The two kinds of usage call for different standards of quality. The next section is about quality standards for nonpredictive uses.

4. SEVEN USES OF UNVALIDATED (INCLUDING UNVALIDATABLE) MODELS, AND THE QUALITY STANDARD RELEVANT TO EACH USE

If an analysis is to have the form "the model says X," the notions of quality given in Sections 2 and 3 must apply. However, it might be unnecessary or impossible to satisfy them. If so, something is gained and something is lost: The appropriate standard of quality is less exacting than validation, but the model cannot be used to predict. This does not mean it is useless; it may, however, only be put to nonpredictive uses. We distinguish seven such uses (listed in decreasing order of tangibility of the model's contribution):

1. As a bookkeeping device, to condense masses of data or to provide a means or incentive to improve data quality
2. As an aid in selling an idea of which the model is but an illustration
3. As a training aid, to induce a particular behavior
4. As part of an automatic management system whose efficacy is not evaluated by using the model as if it were a true representation
5. As an aid to communication, e.g., in purely intellectual explorations or in operating organizations
6. As a vehicle for *a fortiori* arguments
7. As an aid to thinking and hypothesizing, e.g., as a stimulus to intuition in applied research or in training or as a decision aid in operating organizations.

This section discusses the seven uses and gives examples of each. Once the use is specified, it is fairly simple to define how to evaluate the model for that use, and we do so below. As it turns out, the seven uses are listed above in decreasing order of straightforwardness of evaluation, as well as in decreasing order of tangibility of the model's contribution. This section will also show that evaluation of an unvalidated model need not involve comparing the model to reality and that for some uses unrealism is deliberately introduced.

USE 1: AS A BOOKKEEPING DEVICE

Condensing Masses of Data

Some models digest great volumes of inputs and produce handy numbers or pictures. For example, a team from the Regional Forces Division of the Studies and Analysis group of the U.S. Air Force operated the C3ISIM model in Saudi Arabia in support of planning before Operation Desert Storm. After the bombing campaign began, the model's inputs could not be altered quickly enough to keep up with changes in the target array and the Air Tasking Order, but before and after the war, the model's graphical displays were used to acquaint new mission commanders with air traffic information and the military layout of the area.[1]

Providing a Means or Incentive to Improve Data Quality

Analysts and managers are often inconvenienced by low-quality data. Sometimes quality is low because data collectors have no incentive to record the data carefully. But a model—one that data collectors care about and that needs good data—can create the right incentives. For example, Air Force base commanders and maintenance officers are becoming aware of the potential usefulness of the DRIVE model—which schedules component repair and distribution for Air Force depots—within theaters. Quality problems have plagued the data DRIVE uses, but interest in DRIVE has stimulated interest in the data it uses as inputs. While the original stimulus may have been provided by DRIVE, unrelated uses have since been found for the improved data.[2]

Evaluation

An unvalidated model used to condense masses of data is evaluated by ensuring that it reads the right input numbers and then summarizes them without error.[3] An unvalidated model used to induce higher data quality is evaluated by comparing the accuracy of the data before and after the use of the model.

---

[1]This example is from Major Frederic Case, USAF.

[2]This use and the example were contributed by Warren Walker.

[3]This is identical to verification, as that term was defined in the Preface.

USE 2: AS AN AID IN SELLING AN IDEA OF WHICH THE MODEL IS BUT AN ILLUSTRATION

Architects build scale models of new projects so that developers, financiers, and city officials can see how the projects will look. A scale model may be an unvalidated model in the sense used here: It does not and cannot predict how the plumbing will work. Nonetheless, it can do a good job of selling the idea—the project—of which it is but an illustration by conveying aspects of the idea concretely.

Mathematical models can serve the same function. A project the first author works on is conceiving a spreadsheet-like model of the Navy aviation logistics system. This model would allow a user to trade off expenditures for physical distribution, say, against expenditures on stocks of spare parts. In our armed forces, this idea is unfamiliar, and the first thing we will need to do is sell it. Even a crude version of such a model would be a good marketing tool, and once the idea is sold, it will be up to the Services to take care of the details.

The last paragraph is fine except for its last clause, which illustrates the danger of this use of an unvalidated model: It almost begs to be used disingenuously. A model may be fine as a descriptive tool ("here are some things in your logistics system that you are trading off whether you know it or not") but poor as a predictive tool ("you will save this much if you make the trade-off this other way"). The requirements of the two uses are different but not always distinguished. If we succeed in building our Navy model, it will most likely be unvalidatable. It will sell the idea of trade-offs, and we would like it to be a predictive tool, but we will need some other logic to justify that use. That logic has not been devised—because we have not yet built the model, and we may not be able to devise the logic or build the model. If not, our model may only be used to sell the idea of trade-offs, not to make them.

Evaluation

An unvalidated model used to sell an idea need only represent the idea and display benefits. Evaluation consists of ensuring that the model does both things, and this is all that can be asked of a model in this role. The analyst-cum-salesman is not off the hook: He must have a good idea about how to produce the benefits and must accompany presentations with appropriate caveats. In the case of the architect's model, these warnings would sound like, "this model is only

intended to show how the project looks; do not attempt to flush the toilets." This sounds silly but is in the correct form for caveats that must accompany unvalidated models used to sell ideas.

USE 3: AS A TRAINING AID, TO INDUCE A PARTICULAR BEHAVIOR

Railroad engineers are trained partly in railroad engine simulators. These simulators are usually realistic, but not always. For example, if presented as a moving picture, the movement of near-field telegraph poles past the side windows of the engine can "strobe" at certain simulated speeds, and this is very distracting for trainees. Thus, instead of simulating the apparent movement of telegraph poles, some simulators run past the side windows a black-and-white zebra pattern that is not distracting at any speed. In this case, a deliberately unrealistic model is used because the avoidance of distraction is more important than realism in the engineer's peripheral vision.

In the U.S. Army, brigades train at the National Training Center (NTC), a 1000 square mile piece of desert with a home team (the OPFOR) trained to fight using Soviet tactics. The OPFOR is extremely proficient and has other advantages, such as familiarity with the terrain. This deliberately unrealistic aspect of the NTC is maintained because the trainers do not want Blue units to make mistakes without paying for them.

While this use is related to Use 6, the *a fortiori* argument, it is distinct and may be antagonistic to that use. Use 6 is analytic—drawing analytical conclusions from an unvalidated model—while the present use has a training purpose. The ability to use the NTC analytically is diminished by the OPFOR's skill, because some apparent outcomes can plausibly be caused by the OPFOR's unrealistic advantage and not by anything inherent in U.S. Army doctrine, tactics, or equipment.

Evaluation

The use of a deliberately unrealistic model to induce particular behavior is evaluated by determining whether it induces the desired behavior. The mechanics of this are familiar and will not be discussed. Not only is realism not inherently important, specific aspects of it are sacrificed to produce particular effects.

USE 4: AS PART OF AN AUTOMATIC MANAGEMENT SYSTEM WHOSE EFFICACY IS NOT EVALUATED BY USING THE MODEL AS IF IT WERE A TRUE REPRESENTATION

Some models can be viewed reductively as algorithms that turn input numbers into output numbers. As such, a model can be inserted into a management system in which the outputs drive more or less automatic functions. For example, Kalman filters and other time-series models are used to process data from sensors in freeway road surfaces, and in turn to run metered on-ramps. There is little reason to take the Kalman filter model at face value as a representation of traffic flow, but its performance—as part of the system of metered on-ramps—can be judged easily enough.

A management system driven by an unvalidated model may not be tested by using the model as if it were true. By presumption, the model is a suspect representation of the problem the management system faces. It is particularly dangerous to use the model in a test, because it is the model's picture of the world that the management system is designed to handle. Using the model as if it were true amounts to rigging the test. But an unvalidated model can be used as a vehicle for *a fortiori* arguments in a test of a system of which it is a part. (See Use 6.)

Evaluation

An unvalidated model used as part of a management system is evaluated by measuring the efficacy of the management system. If it works, it works, and that's all that matters. This points to an important truth: Cost-effectiveness and realistic appearance have no necessary connection. It might even be cost-effective to use a less realistic model if it were a lot cheaper to run and at worst only a bit less effective.

Contrary impressions notwithstanding, this is not a defense of "black boxes." It is a defense of simple, dumb-looking, transparent boxes that do the job.

USE 5: AS AN AID TO COMMUNICATION

A model can be a systematic description of belief and knowledge about a situation. This can aid communication in two ways: by serving as a basis for purely intellectual explorations and, in operating

organizations, by naming things so that different groups of people can speak a common language.

When used as a basis for purely intellectual explorations, a model is a strawman, a group of null hypotheses. It provides a collection of questions that need to be answered. To the extent that it organizes what is believed and known, it structures data, debate, and teaching—although it also constrains those activities. This argument has been used by a RAND researcher to defend unvalidated models generally and combat models in particular. As presented, it is unassailable as a defense of *building* unvalidated models. But if an unvalidated model, once built, is used to draw conclusions or advice about something in the world, then it is no longer being used merely to organize beliefs and knowledge, and its use must be justified by some other argument. We have yet to see a policy model built without some intent to influence decisions. Thus, although it is legitimate to use an unvalidated policy model as a basis for purely intellectual explorations, this use is probably not important in practice.

In operating organizations, however, a model can be a useful communication aid. The Combat Analysis Group of the U.S. Central Command used three models in its support of planning and operations during Operations Desert Shield and Desert Storm. The modules and entities of these models provided a language in which analysts could discuss their results concisely with the various staff elements they served.[4]

Evaluation

An unvalidated model used to aid communication is evaluated by testing whether its introduction improves communication. Formal tests of this are straightforward for either of the variants discussed here. For practical purposes, if a model's language is used voluntarily in an operating organization, then it works.

If an unvalidated model is set forth merely as a hypothetical statement as the basis for intellectual exploration, we have no complaint, but if it slips into use as a statement about what will happen in some future situation, then it must be evaluated by ensuring that it satisfies the logic of this more ambitious use. We discuss this under Use 7, below.

---

[4]This example is from Colonel Gary R. Ware, USAF, Combat Analysis Group, U.S. Central Command.

USE 6: AS A VEHICLE FOR *A FORTIORI* ARGUMENTS

An *a fortiori* argument can work like this: If condition Z were true, then policy A would be preferable to the other candidates. But the actual situation deviates from Z in ways that favor A even more. Thus, *a fortiori*, A is preferable.

An unvalidated model may be used in an *a fortiori* argument. For example, a RAND colleague who does Army research on new types of weapon systems has defended the JANUS combat simulation model on the grounds that it "limits the bull——" of advocates of new technologies. This is an *a fortiori* argument: Actual combat would tax exotic systems more than JANUS does; JANUS finds exotic system A to be wanting; therefore, JANUS has "limited the bull——" by rejecting exotic system A.[5] This particular argument, if correct, can be used to reject exotic systems; an analogous argument using a different model might be used to show that an exotic system *is* cost-effective.

Evaluation

An *a fortiori* argument has three parts: (1) Condition Z implies policy A is preferable; (2) Z represents a boundary on the actual situation; and (3) reality's deviations from Z favor A. Evaluation of an unvalidated model for use in an *a fortiori* argument depends on how the model is used in each of the three parts. This discussion will cover evaluation for the example given above, because we know of no logically distinct *a fortiori* arguments using unvalidated models. If others exist, they might require different forms of evaluation.

The first part of the argument, "Z implies A is preferable," takes Z as true and draws an implication from it. In the example, Z is "the unvalidated model accurately represents the actual situation." The burden of this part of the argument is on drawing the implication correctly, and evaluation, as distinct from verification, plays no role. Indeed, the argument presumes that Z is false in a specific way, so there is no reason to "validate" Z in the usual sense.[6]

The second part of the argument, "Z represents a boundary on the actual situation," does require that an assertion in the model be related

[5]This observation is from Dick Salter.

[6]In formal logic, implications are statements of the form "if A, then B." They impose relations between the truth values of statements A and B. The obvious relation is that if A is true, B is also, but if A is false B can be true or false. The first part of the *a fortiori* argument uses the property that A can be false and B true without creating a contradiction.

to a fact in the world, although it is a descriptive assertion, not a prediction. In the JANUS model, the assumption is that a notional weapon system works as advertised, is not subject to Murphy's Law, and so on. This makes JANUS a more benign environment than real combat, so the assumption that JANUS accurately represents the actual situation forms a boundary on the actual situation.

The third part of the *a fortiori* argument, "reality's deviations from Z favor A," is likewise subject to evaluation. If deviations from Z "go in one direction" in a sense meaningful in context, they must be shown to favor A. It is tempting to try to evaluate this part of the argument by using the unvalidated model itself, as in "If we push parameter values away from those given by Z in the direction we know to be true, the model produces outcomes more favorable to A." This is fine as long as the model is deficient only because its parameter values are unknown. It is not acceptable if the model is unvalidated in some other way.

In the *a fortiori* argument used as an example here, the third part of the argument is evaluated by an appeal to common sense or folk wisdom. We presume, based on experience, that unexpected—and hence unmodeled—problems will make an exotic system less effective, so that the results from JANUS must overstate actual effectiveness. As long as the common sense behind such arguments is common sense about the world *and not about the model,* appealing to it is acceptable, although not squeaky clean.

The third part of the *a fortiori* argument is a very weak prediction, an assertion that deviation from certain conditions will favor policy A. As such, it could be subjected to the strictures of Sections 2 through 4, although we believe that it is consistent with typical usage to classify it with nonpredictive uses and to loosen up on evaluation of the third part of the argument.

USE 7: AS AN AID TO THINKING AND HYPOTHESIZING

An unvalidated model is a combination of assertions, some factual, some approximate, some conjectural, and some plainly false but convenient. What, then, are we to make of the ubiquitous claim that unvalidated models provide insight? *Webster's Ninth New Collegiate Dictionary* defines *insight* as "The power or act of seeing into a situation: Penetration" (p. 626). By definition, an unvalidated model does not give power to see into the actual situation, only into the assertions embodied in the model. Thus, if the use of an unvalidated model provides insight, it does so not by revealing truth about the world but

by revealing key features of its own assumptions and thereby causing its user to go learn whether those key assumptions are true. The model does not provide insight: It helps its user formulate questions that might be insightful or that might be utterly ridiculous. Two instances are sufficiently distinct to deserve separate discussion:

- As a stimulus to intuition in applied research or in training
- As a decision aid in operating organizations.

Stimulus to Intuition in Applied Research or in Training

Sometimes it can be useful to draw implications from the assertions in an unvalidated model. Many are intuitively obvious, and nothing is learned. (As one reviewer pointed out, people often take this as a confirmation of their intuition and "learn" to have greater confidence than they should.) On the other hand, some implications are not obvious, or rather, they conflict with a prior belief. Of these, most turn out to be errors in data or computer code, or artifacts of a specific assumption representing a vague belief. These implications disappear on examination. Some striking implications, however, do not disappear. At this point, the model's user has only learned something about the assumptions in the model: This is arithmetic, not science. If the user is then moved to go learn something about the world, the model may be said to have provided an insight by poking him to go look at something out there. However, it cannot be overemphasized that the model only tells the user about its own assumptions and *not necessarily about the world*: An unvalidated model can suggest but cannot reveal truth. That must be found elsewhere, if it can be found at all.

As an aid in understanding this use of models, consider a slightly facetious alternative to combat simulations, in effect a model that adds no value to an analysis. A team of analysts could mock up several sets of notional briefing charts, with blanks on the charts in place of numbers that the model might produce. They could then hire the first author's 15-year-old Nephew Mike to write numbers in the blanks. For each set of charts, the analysts would try to devise a plausible explanation for Mike's numbers—i.e., to tell a story—and if none were found, the set would be discarded. If a plausible explanation were devised, the analysts would be in the same logical position as if they had gotten the numbers from a combat model: Mike's numbers had suggested, but the truth must be found elsewhere.

This "Nephew Mike process" illustrates several points. First, models in this use generate stories, as does Nephew Mike. A potential difference between a model and Mike is that Mike plainly adds no value, because his stories have no inherent plausibility. Second, a model is used as a story generator in the following way: The model generates a story, the analysts try to find fault with it ("break it") by examining it for errors or absurdities, iterating between running the model and examining its output until they arrive at stories they cannot break, i.e., with which they cannot find fault. One difference between a model and Mike is that Mike does not produce a "paper trail"[7] to aid the story-breaking; he just writes numbers in the blanks.

Third, the result of an exercise with a model-as-story-generator is—like the result of an exercise with Nephew Mike—a collection of statements of the following form:

- The model said [a substantive result];
- It did so because [an explanation in the model's terms];
- We attempted to break the story by [list of attempts to break the story], none of which succeeded because [list of explanations in substantive and model terms].

This is often considered a prediction, with the third step presumed to establish the predictive power of the model. This presumption is mistaken: The model has not been shown to predict, because its predictions have not been tested. The analysis has merely produced statements in which no fault has yet been found.

The fourth point is that the Nephew Mike process provides a standard of comparison for the contribution and cost of models in this use. We develop this more below, under "evaluation."

The logic of the foregoing discussion applies almost without change to the use of unvalidated models for training. A retired Air Force pilot at RAND says that the JANUS model gave him his first opportunity to attack a column of tanks. The model permits low-risk, relatively cheap trial-and-error learning. The question is whether the pilot learns about an actual attack on real tanks or only about some modeler's rendition of it. Again, the model can only suggest.

---

[7]We use "paper trail" figuratively to refer to the entire record of models and cases considered, as well as sequences of steps within particular model runs. The problem of how to set up computing environments to promote ease in the use of such a paper trail is discussed in S. C. Bankes, *Exploratory Modeling and the Use of Simulation for Policy Analysis*, RAND, N-3093-A, 1991.

Decision Aid in Operational Settings

An unvalidated model can serve as a decision aid to someone in, for example, a military staff position. A RAND project has developed a spreadsheet model intended to help Air Force theater staffs consider wartime redistributions of logistics assets. The model's developers make no pretense that it can be validated; it is not supposed to think for staff planners, only to help. As above, the model is a story generator and might suggest things that the planner would never think of, because of its speed and thoroughness. It might also suggest things the planner would never *want* to think of. Evaluation of the model's suggestions must be done outside the model, because the planner knows things that it cannot know, and this is inevitable.

Evaluation

This discussion will apply primarily to models used to stimulate thinking and hypothesizing in applied research. Decision aids will be treated briefly at the end of this section.

The correct criterion for evaluating models in this use is cost-effectiveness: how many "good" stories a model yields for a given cost. The problem is to isolate the model's contribution to cost-effectiveness because, as the foregoing suggests, this use of models is just one element of an iterative process.

The "Nephew Mike process" helps clarify the model's role by distinguishing what the analysts do and what the model does. The model fills two roles: It provides a device on which analysts can set "knobs" to produce stories reflecting the knob-settings, and it provides a paper trail that analysts can use to unravel *how* the model produced a story for given knob-settings. The model may be able to provide more "effectiveness" through either of these roles: More knobs or wider ranges of settings on existing knobs permit more stories to be generated, and a better paper trail allows analysts to examine in greater depth and detail how the model produced the stories. For this, the model's owner pays in man-hours and in hardware and software costs.

In this sense, a given model can be made more effective if it is subjected to peer review and to checks of its math and logic, because these activities can remove obvious problems that would, if discovered

during analyses, cause stories to break.[8] Such activities are investments; savings are recouped during analyses that arrive at their final stories more cheaply because they do not have to discover problems that were removed by peer review and math or logic checks.

However, incurring these costs and reaping these benefits do not imply better predictive power or even insight into the situation being modeled. The Gulf War of 1991 provides a cautionary example. Before the war, the nearly universal judgment of combat modelers (and everyone else) was that U.S. forces would sustain thousands of casualties. As is well known, actual U.S. casualties were in the low hundreds. Some modelers argued that their models were not the problem: Had they known the Iraqis would not fight, they could have changed some parameters and predicted the actual outcome. For the present purpose, we have three comments. First, as noted in Section 3, a model with enough adjustable parameters can fit any outcome; the ability to do so does not imply predictive power. Second, many of the models in question had been through dozens of cycles of story generation and breaking. No amount of story breaking ensures that an unvalidated model will not be grossly wrong. Third, consider adding second-order effects to the portrayals of soldier motivation in the combat models used for the Gulf War. Any "excursions" using those second-order effects would amount to minor variants on what is, in the final analysis, a gross mistake. The quandary of most modelers is that they cannot know whether their model is so mistaken until it is too late.

This points to an important truth: A model's potential utility has an upper limit, and no amount of extra knobs or paper trail will yield utility above that limit. For some situations, Nephew Mike *is* the most cost-effective model, because time spent with an ordinary model, turning knobs and sifting a paper trail, is time wasted. Unfortunately, it is not always easy to see what the upper limit is. In some cases, the upper bound is provided by alternative activities that, like the unvalidated model, also stimulate thinking and hypothesizing. The first author recently built a model of repair parts consumption in Naval aviation depots and put it to Use 7. With some hypotheses in hand, the model can be elaborated or a field test can be run in actual depots. The field test is not reality, but taken as a model, it permits

[8]These activities can involve subtle problems of evaluating how well a model represents vague beliefs or beliefs about one phenomenon given that related phenomena are not represented in the model. See the discussion of evaluation of Use 4 in J. S. Hodges, "Six (or so) Things You Can Do with a Bad Model," *Operations Research*, May–June 1991, Vol. 39, No. 3, pp. 355–365.

observation of effects that a computer model can never capture, particularly effects involving human interaction. Little can be learned about how the proposed policies will work by tinkering with the model, but much can be learned by seeing how they work in real depots.

However, some situations allow few alternative methods to stimulate thinking and hypotheses. Theater-level combat is one: Even when REFORGER exercises were run regularly, they modeled only a part of the action in a hypothetical European theater war. This lack of alternatives does not imply that more knobs and paper trails add value without bound, but it does make the upper bound difficult to specify. Clearly, an infinitely large model would take infinitely long to run and thus be infinitely costly. The point at which returns to added detail become negative, then, is finite, although it is difficult to say exactly when it is reached.

All of the foregoing applies to models used as decision aids. A decision aid that produces many ultimately useless suggestions, thus consuming a lot of the user's time, is less cost-effective than a decision aid that is easier to use or that produces fewer suggestions that the user discards. Indeed, it might even be possible to compare decision aids with formal tests of the effectiveness of their users, with the aid that allows its users to be more effective being the winner.

5. CONCLUSION

This report's objective was to isolate the model in an analysis and ask two questions about it. The first question was, "what does the model do in the analysis?" with the key distinction being between prediction and other uses of models. The second question was "how well does the model do what it does?" where "how well" was defined in terms of "what the model does." As Sections 3 and 4 made clear, **the appropriate form of quality assurance depends fundamentally on how the model is used, so any attempt to define a single validation standard and procedure for all models in all uses will surely fail.**

Some other implications follow from our analysis. **Few military models or models of human decisionmaking can be validated, and it is counterproductive to demand as a matter of policy that users and institutional parents attempt to validate them.** Asking people to do the impossible is an invitation to cynicism and corruption. The foregoing does not mean that such models are useless, although it does restrict their uses. This restriction focuses attention on what models actually do, on what standards should be for models in those roles, and on whether and how various activities aid quality assurance. The focus on what models do provides a way to think about the costs and benefits of models in given uses. The key here is a baseline, such as the "Nephew Mike process" (Section 4), that adds no inherent value to the analysts' activities: To judge a given model in a given use, compare its contribution and cost to that of Nephew Mike.[1]

The problem of validation is actually two problems, one intellectual and the other bureaucratic or political. Our framework addresses the intellectual problem, but it seems clear that the bureaucratic problem cannot be solved until the intellectual problem is. Our framework's strength is that it establishes specific criteria for assessing model quality based on intended uses; without some such basis, "validation" procedures will be eyewash. While we do not claim the framework is the last word on all validation issues, it does provide a conceptual basis on which appropriate standards and procedures can be defined.

[1]Indeed, we would suggest that consumers of analyses begin today to ask analysts what their models provide that Nephew Mike could not, and at what cost.

RAND/R-4114-AF/A/OSD