A METHODOLOGICAL FRAMEWORK FOR SYSTEM DYNAMICS MODEL EVALUATION

James W. Kirchner
Resource Policy Center
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire

OVERVIEW

Modelers must choose between two fundamentally different approaches to modeling, the advocacy strategy and the method of multiple hypotheses.[1]

The advocacy strategy guides most current modeling efforts. It is a process of making the strongest possible case for a particular model or theory. It is marked by a search for confirmatory evidence. The dominant theory is modified only to make it more defensible.

The method of multiple hypotheses, by contrast, guides the most successful attempts at predictive science.[2] It is a process of selecting among a comprehensive set of credible alternative theories, through a series of tests designed to reveal the weaknesses of one or more of the competing hypotheses. It is marked by a search for disconfirmatory evidence. When one theory emerges as clearly preferable to the rest, it in turn becomes the basis for a series of competing refinements. There are at least three reasons to prefer the method of multiple hypotheses over the advocacy strategy.

1. The value of a model, whether measured in terms of "validity", "utility", or "plausibility", has no meaning without a basis for comparison.[3] Modelers using the advocacy strategy can usually gauge their results only against the existing mental and verbal models, rather than against other formal models.

2. The method of multiple hypotheses encourages the modeler to consider the types of evidence that would disprove a given theory. This aids learning.[4] The potential for disconfirmation distinguishes hypotheses from dogmas.[5]

3. Seeking to defend a favored hypothesis, under the advocacy strategy, demonstrably biases the modelers' view of the world.[6] Testing a number of models simultaneously, with disproof as the goal, diminishes the modelers' attachment to any single theory and encourages a broad, flexible view of the simuland (the real-world system being modeled).[7]

There are also drawbacks to the method of multiple hypotheses. It is more work than the advocacy strategy; choosing among models is not easy. It also may more clearly expose the weaknesses of a model, making both the model and the method less attractive to some clients.

Evaluation Techniques for the Method of Multiple Hypotheses

Many stages of modeling would remain essentially unchanged under the method of multiple hypotheses. But in adopting this method, system dynamicists would need to develop philosophies and procedures of model evaluation that emphasize disproof over verification and comparison among theories over improvement or elaboration on a single model.

For example, a model that is dimensionally incorrect, or that behaves in ways that the real world could never, is the product of either a modeling

A METHODOLOGICAL FRAMEWORK FOR SYSTEM DYNAMICS MODEL EVALUATION

James W. Kirchner
Resource Policy Center
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire

Abstract

The scientific technique known as the method of multiple hypotheses can be adapted to suit the purposes of system dynamics policy modeling. This method would allow determination of a model's value through comparison with other competing models. It would also diminish modelers' emotional attachment to any single theory. But in adopting this method, system dynamicists would need to develop a new philosophy of model evaluation, emphasizing disproof over verification and comparison among theories over improvement or elaboration on a single model.

Introduction

Virtually every modeling discipline has been subjected to criticism, both from modelers using other techniques and from potential clients.[1] One recurring objection is that neither clients nor modelers themselves can readily gauge the usefulness of a particular model for a given application; this complaint has often been lodged against system dynamics modeling.[2] To some extent the critics may misunderstand the purposes and capabilities of complex models.

But modelers themselves may be responsible for much of the controversy over whether a given model suits its purpose. In the rush to develop more powerful modeling techniques and discover new domains of application, it seems that we have largely overlooked the basic issue of what constitutes a good model--or conversely, how to tell if a model is useful--for a particular purpose. Certainly, the system dynamics literature "contains many more descriptions of models addressed to policy questions than theoretical discussions about modeling techniques."[3] This is no accident; efforts spent building models have a more direct and tangible payoff than efforts spent thinking about how to build them.

Thinking about how to build models has led to this paper. It is an attempt to outline an underlying purpose to the enterprise of modeling in general, and to lay a foundation for the design and testing of system dynamics models.

Prediction and the Purposes of Modeling

Virtually all the purposes of modeling involve some form of prediction. One kind of prediction is a forecast of the precise numerical value of some real-world quantity (e.g. population, GNP, or interest rates) at a particular point in time. System dynamics models are not well suited to such "point prediction" or "numerical prediction". Consequently, many system dynamics modelers maintain that prediction is not the goal of a model. In so doing, however, they construe prediction far more narrowly than common usage would dictate.

Most models exist to describe unobservable properties or behaviors of the real world. When policy makers can learn what they need to know through

direct observation, they do not need models or modelers. But when they need to know about events that lie in the future or characteristics of the real world that elude straightforward measurement, they turn to a model, mental or formal, to bridge the gap in their information. And any such leap beyond what can be observed can only be called a prediction. Therefore, for the purposes of this paper, I will define a prediction as "a description of an unobservable property or behavior of a real-world system."[4]

Many purposes for system dynamics models, other than point numerical forecasting, involve prediction. The role prediction plays is more obvious in some of these than in others. To wit:

* Determining which behavior modes will predominate under a variety of alternative policies, ceteris paribus, constitutes a prediction of how the system will respond to those policies.

* Designing policies to produce desired characteristics of the system (e.g. tendency toward a particular stable equilibrium) under an array of unforseeable circumstances (that is, not ceteris paribus) involves predicting either the determinants of those characteristics or the behavior of the system under those policies and circumstances.

Of course, some goals of modeling do not necessarily require prediction. Among these are:

* communicating a set of assumptions. Per se, this does not require any kind of prediction, or any running or testing of a model. But when a model is used to enhance understanding of a real-world system, rather than of assumptions about that system, it must make predictions of the likely behavior of the simuland (the real-world system being modeled).

* determining the implications of a set of assumptions. This is a purely logical, deductive process which computer simulation can aid. But most policy purposes require that those assumptions, and thus their consequences, be representative of some real-world system. And when a model represents a real system, its behavior can only be a prediction (albeit perhaps an uncertain one) of what that system will do.

* satisfying the client. Whether this purpose is served by a model's predictive capacity depends on the client's needs.

To be useful as a guide to policy making, a model must generate conclusions that have real-world meaning and that are at least as likely to be true of the real world as those derived from other means.[5] To be significant, it must describe aspects of the system that are not already known or cannot be readily observed. In a word, it must predict.

Science and Modeling

In trying to devise models that predict well and test how well they predict, it is helpful to consider an illustrative analogue of the modeling enterprise. In other words, a model of modeling is useful. The most useful analogue I know is science.

Science and modeling have a great deal in common. Both have as their goal the prediction of unobservables (e.g. whether an as-yet-unbuilt bridge will be strong enough for its purpose, or whether a given set of tax policies will contribute to economic stability). In order to predict, both use explanations, theories, or models of the mechanisms thought to be important in determining the phenomenon of interest. Examples of these would be statics,

materials science, and a national economic model. Physical science theory is a model that is implemented through engineering.

Science also serves as a "model" for modeling in a second sense: as an ideal toward which modelers strive. We often hear that modeling should become more "scientific", which usually means more accurate, certain, precise, and objective. These properties do not arise at random, however. They are the result of the ways in which scientists (ideally) conduct their inquiries. The usefulness of a given discipline in solving a problem is a product of how its practitioners conduct their activities.

There are two fundamentally different ways to conduct such inquiries as the creation of a model or a scientific theory. I will describe these two frameworks shortly. For the moment, let me get to the point of this digression into the similarities of the scientific and modeling enterprises. It is this:

Scientists are more effective in creating useful theories if they adhere to one of these two approaches to inquiry.[6] But most modeling efforts appear to be governed by the other.

The Advocacy Strategy and the Method of Multiple Hypotheses

The many ways of attempting to create predictive models or theories can be divided into two fundamentally different approaches to modeling. I will call these the advocacy strategy and the method of multiple hypotheses.[7]

The advocacy strategy is a process of making the strongest possible case for a particular model or theory. It is marked by a search for confirmatory evidence. The dominant theory is modified only when doing so would make it clearly more defensible.

The method of multiple hypotheses, by contrast, is a process of selecting among a comprehensive set of credible alternative theories. It is marked by a search for disconfirmatory evidence, through a series of experiments or tests designed to reveal the weaknesses of one or more of the competing hypotheses. When one theory emerges as clearly preferable to the rest, it in turn becomes the basis for a series of competing refinements.

A cursory review of the literature suggests that most system dynamics modeling projects more closely resemble the advocacy strategy than the method of multiple hypotheses. In general, modelers do not describe alternative hypotheses and their relative strengths and weaknesses.[8] Often, they make no explicit comparison between their model and existing competing models. The authors usually do not enumerate the limitations they have found in their model, and may simply qualify it as being "only one of many possible theories." A glance at the literature of other modeling disciplines reveals similar trends.[9] From the published accounts of their work, it appears that most modelers follow a procedure similar to that diagrammed in Figure 1.

Examples of the advocacy strategy applied to modeling are numerous. Can the alternative also be described by reference to examples? I have been unable to find a clear case of a modeling project which has employed the method of multiple hypotheses.[10] But it is relatively straightforward to describe the characteristics such a project would have. The general procedure that such a modeling effort would follow is diagrammed in Figure 2.

A modeling project employing the method of multiple hypotheses would begin, as in the advocacy strategy, with problem definition. But here the similarity ends. It ends because the modelers would then specify a group of diverse hypotheses, each offering an alternative view of the problem at hand.

This group, which ideally would span the range of conceivable possibilities, often would contain theories that require different modeling techniques.[11] For example, one might describe the system in terms of feedback loops, while another might portray it as a group of independent linear relationships. Usually, the modelers would include "conventional wisdom" theories and those advocated by other modeling groups.

A series of simple models, each embodying the essence of one of the hypotheses, would then be built. These models could be tested for logical consistency and real-world meaning. In testing for these properties (which I find useful to group together as "coherence"), for example, most system dynamicists will check to see that dimensions balance, each variable has a real-world definition, and the model cannot generate physically impossible results. Failure of coherence tests indicates either that the model does not correctly portray the theory, or that the hypothesis itself is logically unsound. Thus, a hypothesis which could not be modeled in a way that satisfies the criteria of coherence would have to be abandoned.

Building these simple models would serve a second function, that of making the hypotheses explicit and thus revealing their broad similarities and differences. In doing so, the modelers are likely to discover that some of the theories can be represented by different versions of the same model. This could save work in elaborating the different models to prepare them for evaluation.

Then follows the process that defines the method of multiple hypotheses: the systematic comparison of the competing hypotheses through tests designed to uncover their weaknesses. Of course, the goal of the modeling project would guide the choice of testing techniques and criteria, as it does

in current modeling practice.[12] The essential feature of the method of multiple hypotheses is not any particular test, but rather the fact that whatever test is used is designed to reveal the limitations of a model instead of its strengths. The test should pit each model against the others in the face of potentially disconfirmatory evidence, attempting to show that it does not accurately represent some aspect of the real-world system being modeled. These tests could include those designed to answer such questions as:

* How well does each model reproduce the reference mode?

* Does each model's behavior obey constraints known to be present in the real world? For example, does the model generate populations beyond the known carrying capacities of their environments?

* Does each model exhibit dynamic characteristics, such as stability and sensitivity, that are similar to those of the simuland (the real-world system being modeled)?

* How reliable is each model in predicting the past behavior of systems similar to the simuland?

A comprehensive set of tests should reveal one model as clearly preferable to the others, or at least expose the tradeoffs that will have to be made in choosing one model over the alternatives. Once the modelers have decided on their model of choice, they can unravel its policy implications and proceed to implementation.

The procedure outlined above resembles, in some ways, the processes currently used by modelers. Any modeling effort will have attributes of both the advocacy strategy and the method of multiple hypotheses. They are not a discrete dichotomy, but instead the two ends of a continuous methodological

spectrum. For example, many modelers experiment with changes in their basic model. In so doing, they shift away from the pure advocacy strategy. But they usually do not compare a wide range of alternative models, or subject each alternative to the same comprehensive testing. Similarly, the method of multiple hypotheses may be acted out in the public arena, among modelers trying to expose the weaknesses of each others' models. But this process is at best a chaotic analogy to the procedure that each modeling group could be following in its own work.

Advantages and Disadvantages of the Method of Multiple Hypotheses

The method of multiple hypotheses offers a number of advantages over the advocacy strategy for policy modeling. Four of these are most prominent.

First, the method of multiple hypotheses allows the modeler to determine the value of a given model relative to a set of alternatives. The worth of a model, whether measured in terms of "validity", "utility", "plausibility", or some other standard, simply has no meaning without a basis for comparison.[13] Modelers using the advocacy strategy can only gauge their results against the existing mental and verbal models. These latter are so poorly suited to the criteria of policy modeling (explicitness, precision, communicability, logical consistency, etc.) that they are virtually straw men; almost any formal model would be a significant improvement over them. But that a model devised through the purest form of the advocacy strategy is preferable to other possible models is little more than an article of faith.

Second, the method of multiple hypotheses encourages the modeler to consider the types of evidence that would disprove a given hypothesis. No theory of how the world works can ever, strictly speaking, be proven--the

disconfirming facts or the better theory may lurk undiscovered--and thus understanding advances by selective disproof, clipping off one branch of possibilities after another. "Whether it is hand-waving or number-waving or equation-waving, a theory is not a theory unless it can be disproved".[14] The capacity for disconfirmation separates hypotheses from dogmas.

Third, the method of multiple hypotheses involves the modelers in developing and testing many models simultaneously, with disproof as their goal, and thus diminishes their emotional attachment to any single theory. By comparison, marshalling evidence in defense of a favored hypothesis, under the advocacy strategy, demonstrably biases the modelers' view of the world.[15] Indeed, the fact that they are doing so suggests that they already believe theirs is the most useful theory available. They are therefore less likely to detect its limitations than they would be under the method of multiple hypotheses.

Finally, the simultaneous consideration of a range of alternative theories both encourages and presupposes a broad, flexible view of the simuland. Each hypothesis suggests new ways of attacking the problem or new testing procedures and criteria. In Chamberlin's 19th century phrase, "the mind appears to become possessed of the power of simultaneous vision from different standpoints".[16]

The method of multiple hypotheses also has its drawbacks. First, it is more work than the advocacy strategy. It demands that modelers build a number of models, each of which is likely to be among those discarded later. But this apparent multiplication of effort is not as great as it seems. Modeling a number of alternative hypotheses pertaining to a single problem is far less taxing than building models to address a similar number of different problems.

The work of problem definition is shared among the models in the method of multiple hypotheses, as is much of the necessary data gathering. The bulk of the extra effort required will be that of extensively testing each model and comparing the results. The difficulties will come in making the tradeoffs involved in deciding which model will be used for making policy recommendations. Those doing single-model studies avoid this problem by tacitly selecting one model over the alternatives.

The second major disadvantage of the proposed method is that it may not build the client's confidence in the model or the modelers. It may be difficult to explain why the modelers preferred one model to the others. Further, a full enumeration of the weaknesses of a model, or those it was chosen over, is not likely to boost its credibility to many people. On the other hand, some would rather know the limitations of their planning tools than remain oblivious to them.

Evaluation Techniques for the Method of Multiple Hypotheses

To use the method of multiple hypotheses, system dynamicists would have to change their approach to modeling. In particular, because comparison of competing theories is the keystone of the proposed method, new philosophies and practices of model evaluation would be needed. The other stages of modeling would remain essentially unchanged under the process outlined above; they would simply be applied to a number of models simultaneosly. But the method of multiple hypotheses is built on an evaluation philosophy that emphasizes disproof over verification and comparison among models over improvement and elaboration of a single model. This section shows how this philosophy can be applied to changing the evaluation techniques and criteria of system dynamics.

The method of multiple hypotheses employs evaluation as an opportunity to reveal the weaknesses of a model, rather than as a means of making sure the model meets the client's standards. Nowhere is this clearer than in testing for coherence. A model that is dimensionally incorrect, or that behaves in ways that the real world could not, is the product of either careless modeling or a fundamentally flawed hypothesis. But if modelers view coherence tests as merely a hurdle to be overcome, they are likely to simply patch up the problem and move on. Replacing a constant with an asymptotic table function or redimensioning a coefficient so the units work out is no substitute for looking long and hard at the fundamental assumptions of the model.

Coherence can only be gauged by a thorough inspection of the model. It implies more than merely that the model does not happen to behave implausibly. Coherence means that the model could never generate nonsensical behavior.[17] Testing the model under extreme initial values, while it may uncover a problem overlooked in a brief examination of the model's equations, cannot be relied on to reveal every logical inconsistency in the model's assumptions.

Choosing among the competing models, as distinct from checking each for coherence, requires an evaluation strategy designed to compare them and their behavior to whatever is known about the real-world system they have been designed to represent. The bias of the method of multiple hypotheses favors tests of model "outputs" (behaviors) rather than tests of model "inputs" (assumptions). This is because the method assumes that the accuracy of assumptions is essentially unknowable, except through seeing if their consequences resemble what we can observe in the real world. But in many cases information on real world behavior may be too incomplete or vague to allow a systematic comparison of the competing models. It then becomes necessary to directly appraise the assumptions themselves.

One common way to evaluate the models' assumptions is by testing their "face validity", or plausibility to those who know the simuland. Usually the modelers explain their assumptions to someone familiar with the model's real-world counterpart, and gauge the response. They can do the same with model output as well. The opinions of authorities can make a great deal of information available to modelers in highly filtered and condensed form.[18] But three cautions are in order:

1. "Those who know," often do not know.[19] Stories of plausible but unsubstantiated assumptions being repeated in the literature until they take on the ring of fact are depressingly common.

2. If the modelers are using the experts to judge the model, and not to lend it credibility, they should be willing to let someone else pick the experts and to agree--beforehand--on the role expert opinion will play in evaluation.

3. If carried to an extreme, this process can result in domination of the modeling effort by the same assumptions that have left the problem unsolved...which is what made the model necessary in the first place.

Sensitivity testing is important in revealing how vulnerable each model is to its most uncertain assumptions. When they have insufficient data, modelers feel safe in guessing parameter values if those guesses are unlikely to affect model behavior. Under the advocacy strategy, the goal of a sensitivity test is to show that the model is insensitive to different plausible values of the uncertain parameters. A common practice, for example, is to hold all but one of the uncertain parameters at their most likely or most convenient values, and vary the remaining one as though it were the only unknown. This technique is the least likely to reveal sensitivities in the

model structure. But if the evaluation objective is to find any sensitivities in the model, rather than to demonstrate its insensitivity, all combinations of the possible parameters should be tested [20]. In practice, such a test rapidly becomes impossible as the number of uncertain parameters increases. One alternative is to test random groupings of the possible values. Another is to test each combination of the most extreme possible values of the uncertain parameters.

Replication of the reference mode constitutes, in the advocacy strategy, confirmation of the dynamic hypothesis. One of the tenets of the method of multiple hypotheses, however, is that dynamic hypotheses cannot be confirmed. Reference mode replication means only that an attempt to disconfirm the hypothesis--by showing that it cannot imply the same behavior as the real world exhibits--has failed. It does not indicate that the dynamic hypothesis is correct, because many other models could also produce behavior that mimics the reference mode.[21]

A second reason reference mode replication cannot be regarded as hypothesis confirmation is that modelers already know the behavior they are trying to re-create. As a result, they're tempted to model the reference mode instead of the system. Anyone who doubts this should eavesdrop on a homework session for any first course in system dynamics. "Let's see...It's overshoot and collapse...we'll need a positive and a negative loop..." Most modelers are smart people. If they try to duplicate the reference mode, they can hardly fail.

In the method of multiple hypotheses, reference mode replication is a useful means of testing a theory rather than of confirming it. To scrutinize a theory, modelers should not test only the behavior of the variables that are

central to the dynamic hypothesis and the upcoming forecasts. They should compare the behavior of every variable in the model against whatever data--time series, point value, or qualitative--they can find to describe the history of its real-world counterpart. In so doing, they will give the model every opportunity to show them that it is generating the right output behavior in the wrong way.

Of course, if the goal of modeling is to be able to predict, the goal of model evaluation should be to assess how well each model predicts. As mentioned above, because the reference mode is known in advance its replication is not a prediction. So, too, with the natural sciences; showing a theory (or model) can explain a previously observed phenomenon is reassuring, while demonstrating that it can predict what was not known at the time is impressive.[22] That is the difference between mimicry and prediction.

Modelers may be able to directly measure the predictive capacity of their models. All that they need do is compare the model's behavior to system behavior they were previously unaware of. If they know the behavior of the simuland too well, other similar systems may exhibit a host of behaviors that the system under study could have, but did not, reveal.

In the process of developing and testing a model, most modelers learn the conditions under which its different behavior modes arise. They can then test its predictive power by finding the extent to which similar conditions produce similar behaviors in the model and in real-world systems resembling the simuland.

A complete, unambiguous comparison with concurrent predictions will be impossible. Some conditions--those implied by unprecedented policies, for example--have never arisen before. Some systems may more closely resemble the

simuland than others, or every system available may show the same behavior. Data may be incomplete or imprecise.

Nonetheless, any information gained in this way can be useful if it is conscientiously applied to testing the hypotheses. For example, a model may predict that a given range of different conditions will produce a wide array of behaviors, while the real systems under these conditions exhibit the same behavior modes. This may result if the hypothesis includes a causal agent that is more influential in the model than in the real world. Or different behaviors may arise out of indistinguishably similar circumstances. In this case the hypothesis is probably inadequate to capture the effects of the differences between these seemingly "indistinguishable" conditions.

This kind of testing is another opportunity for modelers to try to disprove their hypotheses. It focuses attention on the system of causal mechanisms and a wide range of potential behaviors rather than on the particular behavior dominating a single reference mode. This is necessary if modelers are to capture the determinants of the system's spectrum of possible future behaviors.

This process is not straightforward. In particular, the attitude of the modeler can transform it from a technique for hypothesis testing to an excercise in "hypothesis confirmation,"--a fiction that dies hard. The modelers' approach is critical. This is equally true of every other stage of model building.

If modelers simply want a product that sells, the advocacy strategy may serve them well. The single-hypothesis study, culminating in a well-defended model, has been successful in pleasing some clients. Shepherding a model through the necessary "validation" exercises is relatively straightforward.

But if modelers are striving instead toward a science of prediction, they would do well to abandon their attachment to any single model of a simuland. A model has little value, except as one of a number of alternatives which have been subjected to the same rigorous, comprehensive testing. Perhaps the most reasonable attitude toward a theory is one of conscientious skepticism. "Models should be tested with a vengeance, not simply estimated and admired."[23] What better reason to have confidence in a model than to know its weaknesses, and to know that despite them, it is the best policy design tool available?

NOTES

1. See, for example, Sheridan (1975), p. 196.

2. See Armstrong (1978), pp. 239-240, and Sharp (1972), p. 1222.

3. Meadows (1976), p. 14.

4. This definition is at odds with that used by some systems dynamics authors. However, "prediction" and its synonyms are the only terms available to express this concept. This definition also agrees with the general usage of the term. See Armstrong (1978), pp. 481 and 484, and Sullivan and Claycombe (1977), p. 1.

5. Other valuable attributes of policy models, such as clarity and simplicity, may have to be traded off against predictive capacity.

6. Platt (1964), pp. 347 and 352.

7. These terms and their definitions are borrowed from Armstrong (1978), pp. 406-407, and Chamberlin (1890), p. 756.

8. While almost all modelers test the behavior of the model under different conditions, they usually do not test different theories of how the system actually works. One exception is the effort of the World 3 modelers to expand various hypotheses in subsector models. See Meadows (1973).

9. Armstrong (1978), p. 407.

10. Some modelers claim to test different hypotheses, for example by trying a number of regression model specifications or statistically rejecting the null hypothesis. However, the mathematical constraints on regression estimation techniques allow only a very narrow range of alternative hypotheses to be estimated. Further, the fact that the model is being estimated a posteriori calls into question whether such a model can be said to express a hypothesis at all. Similarly, rejecting the null hypothesis does not reject any of the innumerable alternatives to it.

11. In the natural sciences, the method calls for hypotheses that encompass all logical possibilities. To specify these in the social sciences would be exhausting. Nonetheless, modelers can gain efficiency by cleverly specifying alternatives that "cleave the universe" into mutually exclusive possibilities, thus using a branching technique rather than a random search to find a useful theory.

12. Randers (1974), p. 16-17.

13. Forrester (1968), pp. 3-3 and 3-4, Armstrong (1978), pp. 274-276.

14. Platt (1964), p. 350.

15. See Geller and Pitz (1968), Fischoff and Beyth (1975), and Pruitt (1961), cited in Armstrong (1978), pp. 354 and 407. Chamberlin said so much earlier, with only anecdotal evidence.[Chamberlin (1890), pp. 754-755]

16. Chamberlin (1890), p. 756.

17. The coherency test does not ask, "Is the model representative of the real world?" but rather, "Could the model represent the real world?" The concept employed here is the philosopher's division of statements into three classes: truth, falsehood, and nonsense. Logically inconsistent or vacuous statements are neither true or false; they are simply babble. The goal of a coherency test is to separate meaningful theories, which may or may not be true, from the nonsensical. Many verbal and mental models fail the test of coherence.

18. Randers (1974), p. 35.

19. For this reason, face validity is known to some as "faith validity." See Armstrong (1978), pp. 297-299.

20. Attempts in this direction can yield surprises. Clark and Cole (1975, p. 67) report that, with less than 5% change in any of the initialization parameters, "World 3 gave totally different patterns of 'results'."

21. This argument has been made at least as long ago as 1848. See Mill (1848), pp. 356-358, Ravetz (1971), pp. 150-151, Randers (1974), p. 35, and Godet (1979), p. 11. However, talk of "proving" hypotheses still abounds in the literature.

22. For example, the theory of general relativity attracted interest but little acclaim for some time after Einstein first announced it in 1916. But three years later, a team of astronomers confirmed the predicted deflection of starlight by the sun's gravitational field. Einstein was famous overnight.

23. Larkey and Sproull (1981), p. 241.

REFERENCES

Amara, Roy 1981, "The Futures Field: How to Tell Good Work from Bad", The Futurist, April 1981 pp. 63-71.

Armstrong, J. Scott 1978, Long Range Forecasting, Wiley and Sons, New York

Ascher, William 1981, "The Forecasting Potential of Complex Models", Policy Sciences, 13 pp. 247-267.

Boulding, Kenneth E. 1974, "Reflections on Planning: The Value of Uncertainty", Technology Review, Oct./Nov. 1974

Brewer, Garry D. 1974, "The Policy Sciences Emerge: To Nurture and Structure a Discipline", Policy Sciences, 5 pp. 239-244.

Chamberlin, Thomas C. 1890, "The Method of Multiple Working Hypotheses", reprinted in Science, 148, pp. 754-759.

Clark, John and Sam Cole 1975, Global Simulation Models: A Comparative Study, Wiley and Sons, London.

Fischoff, Baruch and Ruth Beyth 1975, "I Knew It Would Happen: Remembered Probabilities of Once-Future Things", Organizational Behavior and Human Performance, 13, pp. 1-16.

Forrester, Jay W. 1968, Priciples of Systems, Wright-Allen Press, Cambridge.

Geller, E. Scott and G. F. Pitz 1968, "Confidence and Decision Speed in the Revision of Opinion", Organizational Behavior and Human Performance, 3, pp. 190-201.

Godet, Michel 1979, The Crisis in Forecasting and the Emergence of the "Prospective" Approach, Pergamon Press, New York.

Gould, L., E. Kuh, F. Schweppe, and R. Welsch 1977, "Discussion on Model Validation", Prepared for Workshop on Model Validation, Cambridge, Mass., May 23-24, 1977.

Larkey, Patrick D. and Lee S. Sproull 1981, "Models in Theory and Practice: Some Examples, Problems, and Prospects", Policy Sciences, 13, pp. 233-246.

Meadows, Dennis L. and Donnella H. Meadows 1973, Toward Global Equilibrium: Collected Papers, Wright-Allen Press, Cambridge.

Meadows, Donnella H. 1976, The Unavoidable A Priori, DSD #71, Available from Resource Policy Center, Dartmouth College, Hanover, N.H. 03755.

Mill, John Stuart 1848, A System of Logic, Harper and Brothers, New York, eighth edition, 1900.

Platt, John R. 1964, "Strong Inference", Science, 146, pp. 347-353.

Pruitt, D. G. 1961, "Informational Requirements in Making Decisions", American Journal of Psychology, 74, pp. 433-439.

Randers, Jørgen 1974, "Conceptualizing Models of Social Systems", in Methodological Aspects of Social System Simulation, DSD #14, available from Resource Policy Center, Dartmouth College, Hanover, N.H. 03755.

Ravetz, Jerome R. 1971, Scientific Knowledge and its Social Problems, Oxford University Press, New York.

Rothkopf, Michael H. 1975, "Limits to Models", Proceedings of the International Conference on Cybernetics and Society, p. 465.

Sharp, J.A. 1972, "Problems of Systems Dynamics Methodology", in Advances in Cybernetics and Systems, (J. Rose, ed.), Gordon and Breach, London, v. 3, pp. 1221-1227.

Sheridan, Thomas B. 1975, "On Interfacing Models and Decision-Makers", Proceedings of the International Conference on Cybernetics and Society, pp. 196-199.

Sullivan, William G. and W. Wayne Claycombe 1977, Fundamentals of Forecasting, Reston Publishing Co., Reston, Va.

Fig. 1. Schematic diagram of the advocacy strategy as applied to modeling.

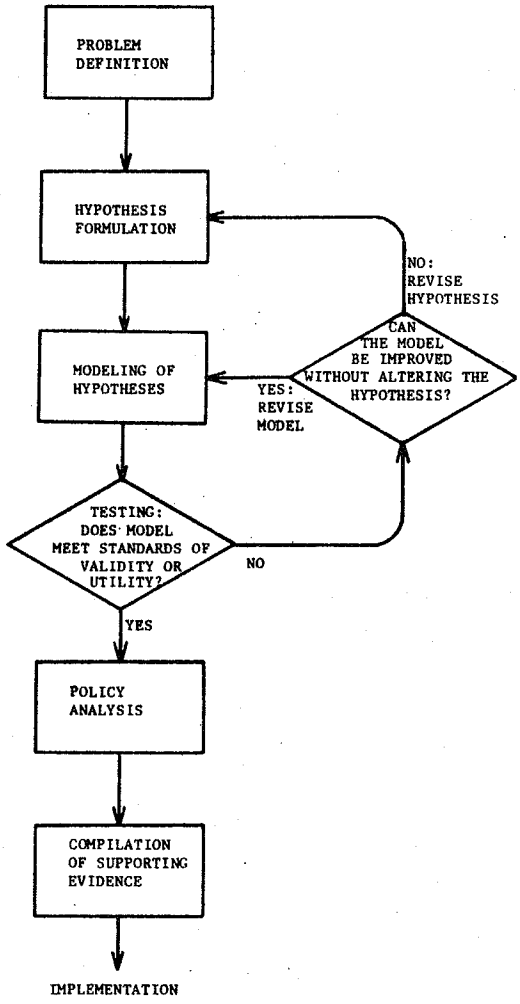
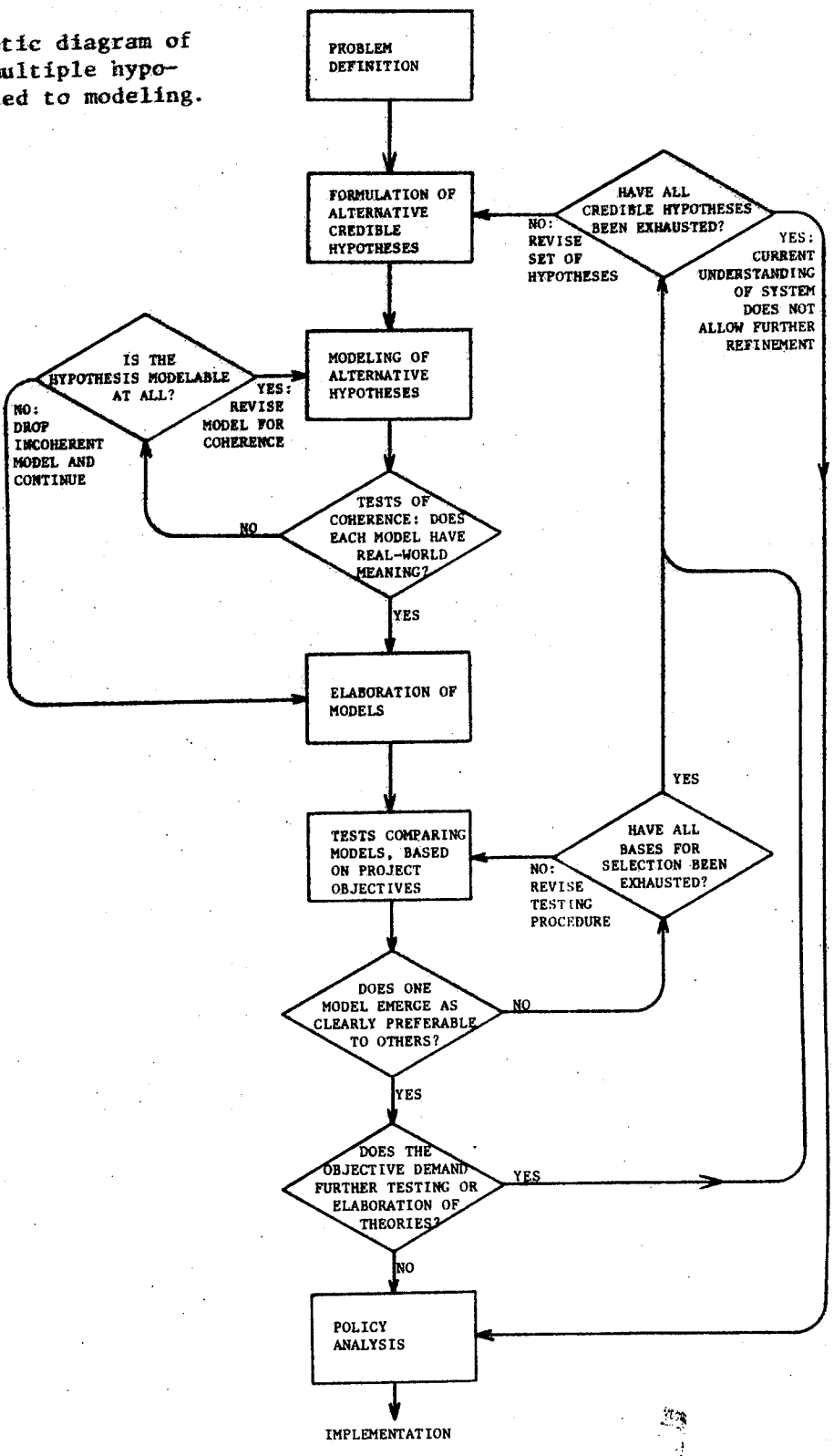


Fig. 2. Schematic diagram of the method of multiple hypotheses as applied to modeling.



error or a fundamentally flawed hypothesis. But if modelers view tests of logical consistency merely as hurdles to be overcome, they are likely to just patch up the problem and move on. Replacing a constant with an asymptotic function or redimensioning a coefficient so that the units balance is no substitute for looking long and hard at the assumptions of the model.

Similarly, under the advocacy strategy the goal of a sensitivity test is to show that the model is insensitive to different plausible values of its uncertain parameters. A common practice is to vary one parameter (as though it were the only unknown) while holding the others at their most convenient values. This technique is least likely to reveal sensitivities in the model structure. If, on the other hand, the objective is to find the model's sensitivities, modelers will test random combinations of values of the uncertain parameters. Or they may test each combination of the uncertain parameters' most extreme possible values.[8]

In the advocacy strategy, replication of the reference mode is taken as confirmation of the dynamic hypothesis. The method of multiple hypotheses, on the other hand, assumes that hypotheses cannot be confirmed. Reference mode replication means that an attempt to disconfirm the hypothesis--by showing that it cannot imply the behavior that the real world exhibits--has failed. Many other models could also produce behavior that mimics the reference mode.

To scrutinize a theory, modelers should not look only at the behavior of the "output" variables. Ideally, they should compare the behavior of every variable in the model against whatever data they can find to describe its real-world counterpart. In so doing, they maximize the chance of discovering that the model is generating the right output behavior in the wrong way.

Similarly, modelers should look for opportunities to compare the models' behaviors to system behaviors they were previously unaware of. Because they know the reference mode in advance, reproducing it does not demonstrate any capacity of the model to forecast behavior. But this capability can be gauged by finding the extent which similar conditions produce similar behaviors in the model and in real-world systems resembling the simuland.

NOTES

1. These terms and their definitions are borrowed from Armstrong, J. Scott 1978, Long Range Forecasting, Wiley & Sons, pp. 406-407, and Chamberlin, T.C. 1890, "The Method of Multiple Working Hypotheses", reprinted in Science, 148, 754-759, p. 756.

2. Platt, John R. 1964, "Strong Inference", Science, 146, pp. 347 & 352.

3. Forrester, Jay W. 1968, Principles of Systems, Wright-Allen Press, Cambridge, pp. 3-3 and 3-4, and Armstrong (1978), pp. 274-276.

4. Armstrong (1978), p. 354.

5. Platt (1964), p. 350.

6. See Geller, E. Scott and G. F. Pitz 1968, "Confidence and Decision Speed in the Revision of Opinion", Organizational Behavior and Human Performance, 3, pp. 190-201, Fischoff, Baruch and Ruth Beyth 1975, "I Knew It Would Happen: Remembered Probabilities of Once-Future Things", Organizational Behavior and Human Performance, 13, pp. 1-16, and Pruitt, D. G. 1961, "Informational Requirements in Making Decisions", American Journal of Psychology, 74, pp. 433-439.

7. Chamberlin (1890), pp. 754-756 and Platt (1964), p. 348.

8. Clark, John and Sam Cole 1975, Global Simulation Models: A Comparative Study, Wiley and Sons, London, p. 67, report that with less than 5% change in some groups of parameters, "World 3 gave totally different patterns of 'results'."