[21]

TIMS Studies in the Management Sciences 14 (1980) 209–228

TESTS FOR BUILDING CONFIDENCE IN SYSTEM DYNAMICS MODELS

Jay W. FORRESTER and Peter M. SENGE
Massachusetts Institute of Technology

Confidence in system dynamics models can be increased by a wide variety of tests that include tests of model structure, model behavior, and a model's policy implications. This paper describes available tests and discusses how the tests can contribute to model validation. As context for considering the tests, the paper also considers the nature of validity in system dynamics modeling and argues that validation includes transferring confidence to persons not directly involved in model construction.

1. Introduction

This paper discusses tests for building confidence in system dynamics models. Unlike earlier treatments of validation in system dynamics modeling (for example, Forrester, 1961; Ansoff and Slevin, 1968; Forrester, 1968; Nordhaus, 1973; Forrester *et al.*, 1974; Senge, 1977; Mass and Senge, 1978), the present paper concentrates on describing specific tests. Although incomplete, the tests presented here should be a useful reference for model builders, and a helpful background for developers of additional tests for system dynamics models. The breadth of tests demonstrates the variety of channels available for building confidence in a system dynamics model. There is no single test which serves to "validate" a system dynamics model. Rather, confidence in a system dynamics model accumulates gradually as the model passes more tests and as new points of correspondence between the model and empirical reality are identified.

Readers familiar with the management science literature on validation of simulation models will find the approach to validation described below to differ in several ways from more established approaches. The nature of system dynamics models permits many tests of model structure and behavior not possible with other types of models. Conversely, some widely used tests, such as standard statistical hypothesis tests, are either inappropriate or, at best, supplementary for system dynamics models. Although no attempt is made to systematically relate to the simulation validation literature, we will, wherever possible, identify major differences between confidence building in system dynamics and in other types of simulation models.

§ 2 discusses testing, validity, and the validation process that underlies the ensuing discussion of specific tests for system dynamics models. §§ 3, 4 and 5 present

the tests. § 6 summarizes the tests and considers levels of expertise required to conduct the tests.

2. The nature of validity in system dynamics models

This section clarifies the notions of testing, validation, and validity which underlie the ensuing discussion of specific tests. This section establishes an appropriate perspective for viewing the following tests. Model testing is part of a larger validation process, and this section discusses the objectives of that process for system dynamics models.

By testing, we mean the comparison of a model to empirical reality for the purpose of corroborating or refuting the model. It is important to realize that the word "empirical" means "derived from or guided by experience or experiment" (Random House Unabridged Dictionary of the English Language). Hence, empirical information for testing a model includes information in many forms other than numerical statistics. In system dynamics models, model structure can be compared directly to descriptive knowledge of real-system structure; and model behavior may be compared to observed real-system behavior. The following sections identify seventeen tests of structure and behavior suitable for system dynamics models. [1]

Validation is the process of establishing confidence in the soundness and usefulness of a model. Validation begins as the model builder accumulates confidence that a model behaves plausibly and generates problem symptoms or modes of behavior seen in the real system. Validation then extends to include persons not directly involved in constructing the model. Thus, validation includes the communication process in which the model builder (or someone else presenting a model) must communicate the bases for confidence in a model to a target audience. Unless the modeler's confidence in a model can be transferred, the potential of a model to enhance understanding and lead to more effective policies will not be realized.

Seeing validation as a process which extends beyond the model builder to encompass critics and potential model users is consistent with a view of scientific knowledge as "public knowledge," by which we do not mean merely "published knowledge," but that "The objective of Science . . . is a *consensus* of rational opinion over the widest possible field" (Ziman, 1968, p. 9).

Validation of system dynamics models is complicated by the many relevant audiences, each having its own objectives and criteria for evaluating a model. For a scientist, a model may be considered useful if it generates insight into the structure

[1] The emphasis on multiple tests in system dynamics is similar in spirit to the "multistage verification" of simulation models advocated by Thomas Naylor (Naylor *et al.*, 1966). Naylor contrasts multistage verification with three other views of the overall theory testing process – "synthetic a priorism," "ultraempiricism," and "positive economics." He argues that the multistage approach, which emphasizes equally tests of structure and behavior, is most appropriate for simulation models because it leads to the maxium variety of tests.

of real systems, makes correct predictions [2], and stimulates meaningful questions for future research. For the public and political leaders, a useful model should explain causes of important problems and provide a basis for designing policies that can improve behavior in the future. In the past, the way in which system dynamics models have been presented has often resulted in a higher degree of interest from the public and politicians than from social scientists. The salience of the problems addressed and the clarity and plausibility of model assumptions probably account for public interest in past system dynamics studies. Neither salience nor plausibility may satisfy social scientists. To progress further in validation for social scientists, system dynamicists should present the many tests available for assessing the realism of model assumptions and behavior, and for generating insights into the causes of observed phenomena.

In this paper, we take the view that the ultimate objective of validation in system dynamics is transferred confidence in a model's soundness and usefulness as a policy tool. The notion of validity as equivalent to confidence conflicts with the view many seem to hold which equates validity with absolute truth. We believe confidence is the proper criterion because there can be no proof of the absolute correctness with which a model represents reality. There is no method for proving a model to be correct. Einstein's theory of relativity has not been proven correct; it stands because it has not been disproven, and because there is shared confidence in its usefulness. Likewise, one tests a system dynamics model against a diversity of empirical evidence, seeks disproofs, and develops confidence as the model withstands tests.

Validity as meaning confidence in a model's usefulness is inherently a relative concept. One must always choose between competing models. Often a model with known deficiencies may be chosen, if it inspires greater confidence than its alternatives. This is especially true when decisions must be made. Validity is also relative in the sense that it can only be properly assessed relative to a particular purpose. It is pointless to try to establish that a particular model is useful without specifying for what purpose it is to be used. Experience has repeatedly shown that debates over the relative merits of different models are often irresolvable if the purpose of the model application has not been clearly stated. [3]

3. Tests of model structure

Although all tests of a system dynamics model are aimed at establishing confidence in model structure, the tests in this section assess structure and parameters

[2] § 4 discusses the types of predictions which can and cannot be expected from a system dynamics model.

[3] Forrester (1961, Chapter 13) examines more deeply the relation between purpose and validity.

directly, without examining relationships between structure and behavior. The following discussion summarizes tests of model structure, including the purpose of each test and how the test is conducted. Examples are drawn from recent modeling studies. This section concludes with a brief discussion of statistical tests commonly used to test model structure.

3.1. Structure-verification test

Verifying structure means comparing structure of a model directly with structure of the real system that the model represents. To pass the structure-verification test, the model structure must not contradict knowledge about the structure of the real system. Structure verification may include review of model assumptions by persons highly knowledgeable about corresponding parts of the real system. Structure verification may also involve comparing model assumptions to descriptions of decision-making and organizational relationships found in relevant literature. In most instances, the structure verification test is first conducted on the basis of the model builder's personal knowledge and is then extended to include criticisms by others with direct experience from the real system.

The belief that the structure of a model should match observable goals, pressures, and constraints of real decision makers has been expressed by many economists. For example, E.H. Phelps Brown argues that many economic models rest on assumptions which cannot be observed in real decison-making and that "The remedy evidently . . . (is) to extend and deepen our observations of firms and managers as they have been and are, in the field" (Phelps Brown, 1972, p. 5). Leontief has similarly argued that what is lacking in most models ". . . is a very difficult and seldom very neat assessment and verification of . . . (model) assumptions in terms of observed facts" (Leontief, 1971, p. 2).

Verifying that model structure exists in the real system is easier and takes less skill than some other tests. Many structures can pass the structure verification test; it is easier to verify that a model structure is found in the real system than to establish that the most relevant structure for the purpose of the model has been chosen from the real system. For example, relatively few criticisms of Forrester's *Urban Dynamics* model (Forrester, 1969) contended that the structure in the model did not exist in a real city. Many more questions focused on whether the model included the most important structures for understanding urban decay. Such criticisms often asserted the need for representation of suburbs or the spatial subdivision of land area. Criticisms which ask for more of the real-life structure in the model belong to the boundary-adequacy test, discussed below.

3.2. Parameter-verification test

Model parameters (constants) can be verified against observations of real life, just as structure of a model can be compared to available knowledge. Parameter

verification means comparing model parameters to knowledge of the real system to determine if parameters correspond conceptually and numerically to real life. Conceptual correspondence means that parameters match elements of system structure. For example, an industrial model might include as a parameter a normal time to correct output inventory. Parameter verification would entail examining inventory-management decisions to determine if rapidity with which inventory imbalances are to be corrected exists as a guide in production planning. Numerical verification of the normal inventory-correction time involves determining if the value given the parameter falls within a plausible range of values for the actual correction time. [4]

Structure verification and parameter verification are interrelated. Both tests spring from the same basic objective – that system dynamics models should strive to describe real decision-making processes. If the structure of a hypothesized decision rule accurately captures the information sources underlying a real decision, the parameters in that structure should identify the relative pressures created by the information inputs. In many cases, the dividing line between parameters and variables is fluid and depends on the purpose and time horizon of the model. If a parameter is likely to change in value over the time and policy regions for which the model is to be used, then the parameter should be converted to a variable with associated structure that depends on parameters of a more enduring nature. In a model addressed to short-term issues, certain concepts can be considered constants (parameters) that for a longer-term view must be treated as variables. Therefore, structure verification, in the broadest sense, can be thought of as including parameter-verification. We distinguish the two to draw special attention to the possibility, which is often overlooked, of directly evaluating parameters from knowledge existing in the operating world.

3.3 Extreme-conditions test

Much knowledge about real systems relates to consequences of extreme conditions. If knowledge about extreme conditions is incorporated, the result is almost always an improved model in the normal operating region. As examples of extreme conditions, if in-process inventories reach zero, then output must be zero; if inventories of final goods reach zero, then shipments must be zero; if there are no houses in a city, then migration to the city will be strongly discouraged; and if pollution rises high enough, then death rate must rise.

Structure in a system dynamics model should permit extreme combinations of levels (state variables) in the system being represented. A model should be questioned if the extreme-conditions test is not met. It is not an acceptable counterargument to assert that particular extreme conditions do not occur in real life and should not occur in the model; the nonlinearities introduced by approaches to

[4] Graham (1979) discusses a variety of techniques for parameter estimation in system dynamics.

extreme conditions can have important effects in normal operating ranges. Often the nonlinearities in the transition from normal to extreme conditions are the very mechanisms that keep the extreme conditions from being reached.

To make the extreme-conditions test, one must examine each rate equation (policy) in a model, trace it back through any auxiliary equations to the level (state variables) on which the rate depends, and consider the implications of imaginary maximum and minimum (minus infinity, zero, plus infinity) values of each state variable and combinations of state variables to determine plausibility of the resulting rate equation.

The extreme-conditions test is effective for two reasons. First, it is a powerful test for discovering flaws in model structure. Many proposed formulations look plausible until considered under extreme conditions. For example, extreme conditions aid in identifying nonlinearities and asymptotes which should be incorporated into model structure. Considering extreme conditions can also reveal omitted variables. For example, Senge (1978) discusses an extreme-conditions test of the hypothesis that capital investment in a production sector of an economic model responds to imbalances in output inventory and backlog of unfilled orders for output. He notes that under normal conditions producers respond to discrepancies between desired and actual level of inventories and backlogs by adjusting employment and capacity utilization. However, if additional labor is unavailable (or available only at a prohibitively high price) or if current capacity is already fully utilized, one would expect persistantly low inventories or high order backlogs to stimulate capital investment. Hence, Senge utilizes the extreme condition of a very tight labor market to justify including inventory and backlog discrepancies as determinants of capital investment.

The second reason for utilizing the extreme-conditions test is to enhance usefulness of a model for analyzing policies that may force a system to operate outside historical regions of behavior. A model which only behaves plausibly under "normal" conditions can only be used to analyze policies which do not cause the system to operate outside of those conditions. By examining model structure for extreme conditions, one develops confidence in a model's ability to behave plausibly for a wide range of conditions and thereby enhances the model's usefulness to explore policies that move the system outside of historical ranges of behavior.

The extreme-conditions test is a strong test. The test is demanding of the evaluator's time but does not impose heavy demand for system dynamics competence. It can be done by anyone who can read algebra and has extensive familiarity with the real system being modeled.

3.4. Boundary-adequacy (structure) test

Boundary adequacy appears three times in this paper — as a test in the context of structure, of behavior, and of policy. The boundary-adequacy (structure) test considers structural relationships necessary to satisfy a model's purpose. The

boundary-adequacy test asks whether or not model aggreagation is appropriate and if a model includes all relevant structure.

As a test of model structure, the boundary-adequacy test involves developing a convincing hypothesis relating proposed model structure to a particular issue addressed by a model. For example, criticism that the *Urban Dynamics* model omits city-suburb interactions might lead to a boundary-adequacy test. Such a test would begin by identifying a particular issue dealt with by the model, such as the ineffectiveness of job-training programs or the effectiveness of slum-housing demolition in reversing urban decay. It would then be necessary to identify feedback relations between city and suburb which might affect consequences of adopting the particular program. If a plausible hypothesis demonstrating importance of city-suburb interactions in assessing job training or slum-housing demolition cannot be developed, the model passes the boundary-adequacy test. If a plausible hypothesis for needing additional structure is developed, the boundary-adequacy test is not passed. To clarify the issue further, one would then need to incorporate suburban structure into the model to resolve the importance of city-suburb interactions through simulation of the expanded model. When this was done by Schroeder (in Schroder *et al.*, 1975) he found that explicit incorporation of suburbs had little impact on basic behavior or policy recommendations of the original *Urban Dynamics* model. §§ 4 and 5 discuss other aspects of boundary adequacy as tests of model behavior and policy implications.

The boundary-adequacy test requires that an evaluator be able to unify criticisms of model boundary with criticisms of model purpose. Often, criticisms of model boundary mask deeper questions about model purpose. For example, Forrester's *World Dynamics* model (Forrester, 1971) has often been criticized for failing to distinguish developed from underdeveloped countries. When one looks deeper, one sees that these criticisms generally stem from an interest in regional development rather than an interest in growth and transition for world society as a whole. Hence, they should be seen as criticism of model purpose (regional resource allocation vs. global transition) rather than boundary adequacy. Hence, the evaluator must continually distinguish questions of boundary-adequacy relative to a particular purpose from questions of model purpose. If one fails to do so, model boundary can be extended indefinitely as one incorporates into a model further aspects of real-system structure which, even if accurate, are not necessary for the particular purpose.

3.5. Dimensional-consistency test

A mundane but often revealing test, the dimensional-consistency test entails dimensional analysis of a model's rate equations. Surprisingly many models fail this simple test or pass it only by inclusion of "scaling" parameters which have little or no real-life meaning. Hence, the dimensional-consistency test is most powerful when applied in conjunction with the parameter-verification test. Failure to pass

the dimensional-consistency check, or satisfying dimensional consistency by inclusion of parameters with little or no meaning as independent structural components, often reveals faulty model structure.

3.6. Other tests

Conspicuous by their absence from the preceding tests of model structure are the statistical tests usually applied to social and economic models. For example, econometric model building relies almost completely on statistical tests which involve direct comparison of individual model equations to statistical data. However, the application of such tests to causal models has been the subject of a long-standing debate (see Keynes, 1939; Morrison and Henkel, 1970; Worswick, 1972; Phelps Brown, 1972). Although there is today fairly wide-spread agreement regarding the limitations of standard statistical tests of model structure, many modelers still rely heavily on such tests. [5]

To see why standard statistical tests should be questioned as tests of model structure, consider the widely-used t-test of statistical significance in regression analysis. The t-test tells the modeler whether or not a parameter estimate is "statistically significant," that is, that the hypothesis of a zero parameter value can be rejected with a certain probability of error. In practice, modelers frequently exclude variables from model questions when low t-statistics are obtained. It is not uncommon to see the t-test used explicitly to reject hypotheses.

Mass and Senge (1978) have specifically analyzed the application of the t-test in system dynamics modeling. They argue that statistical tests such as the t-test tell the modeler whether or not a particular hypothesis is measurable given available data, that is, whether the parameters associated with the hypothesis can be estimated with suitable precision. However, several possible causes can make a hypothesis difficult to measure, only one of which is that the hypothesis is incorrect. In particular, low t-statistics frequently result from errors in measuring the data or from "multicollinearity" of data over the period during which measurements were made. [6] To analyze measurement-error sensitivity, Mass and Senge constructed an experiment in which 10% measurement error caused a hypothesis important for the behavior of a model to be statistically insignificant, even though the data used

[5] For a representative sampling of views among simulation modelers, note that Clarkson states that "The problem of testing the mechanism employed by the model is not so simple because there is no clear way of . . . testing the functional form of the equations . . ." (Clarkson, 1962, p. 34). Naylor *et al.* (1966) call for modelers to ". . . 'verify' the postulates on which the model is based, subject to the limitations of existing statistical tests . . ." (Naylor *et al.*, 1966, pp. 314–315).

[6] Multicollinearity means that two time series of data are highly correlated. If one is testing the hypothesis that X is a determinant of changes in Y, and the data series for X is highly correlated with the time series for some other variable Z also being used to explain movements in Y, the hypothesized effect of X can be difficult or impossible to measure.

for the statistical test *were generated by the model itself.*

The above experiment illustrates that conventional statistical tests of model structure are not sufficient grounds for rejecting the causal hypotheses in a system dynamics model. Such tests may be useful for discovering possible flaws in model structure, but they should be buttressed by the tests described above and in the following sections before model assumptions are altered. [7]

4. Tests of model behavior

Tests of model behavior evaluate adequacy of model structure through analysis of behavior generated by the structure. Tests of model behavior include behavior reproduction, behavior prediction, behavior anomaly, family member, surprise behavior, extreme policy, boundary-adequacy (behavior), and behavior sensitivity.

4.1. Behavior-reproduction tests

The familiy of behavior-reproduction tests examines how well model-generated behavior matches observed behavior of the real system. Behavior-reproduction tests include: symptom generation, frequency generation, relative phasing, multiple mode, and behavior characteristic.

The *symptom-generation* test examines whether or not a model recreates the symptoms of difficulty that motivated construction of the model. Presumably the model was made to show how a particular kind of undesirable situation arises, so it can be alleviated. Unless one can show how internal policies and structure cause the symptoms, one is in a poor position to alter those causes.

For example, in a corporate model to deal with loss of market share, the model should show how known policies and structure lead to loss of market share. If the corporate problem is instability of employment, the model should persuade one that, for the right reasons, it is generating the observed kind of employment fluctuation. If the objective is to understand and correct policies that cause unemployment and a faltering economy in an older American city, the appropriate model should show the internal mechanism of transition from urban growth to stagnation.

The *frequency-generation* and *relative-phasing* tests focus on periodicities of fluctuation and phase relationships between variables. [8] For example, Senge (1978) employs the frequency-generation test in comparing two investment functions. When the two investment functions are embedded in a model of a production sector, Senge shows that only one is able to generate the longer term, 10-to-25 year

[7] Senge (1978) provides one example of how common statistical tests can be combined with the behavior tests described in § 4.

[8] Similar tests have been proposed by Cohen and Cyert (1961) and Fishman and Kiviat (1967).

fluctuations seen in industry data for capital investment. The study by Mass of alternative business-cycle theories (Mass, 1975) provides examples of the relative-phasing test by showing that relative timing of production, inventory, backlog, and employment generated by a production-sector like that in the System Dynamics National Model now being developed at MIT matches the relative timing of those variables in the real economy.

In the literature on modeling and simulation, there are a wide range of tests involving point-by-point comparisons of model-generated and observed behavior (for example see Orcutt *et al.*, 1961; Holt 1965; Cohen and Cyert, 1961; Naylor and Finger, 1967). Despite widespread acceptance, such tests involving point-by-point measures of goodness of fit are generally less appropriate for system dynamics models than the symptom-generation, frequency-generation, and multiple-mode tests outlined above. Forrester has shown that predicting the exact future values of a real system or replicating the point-by-point values of past data is unsound as a basis for evaluating assumptions in a system dynamics model. The problem with point-by-point measures of fit and point predictions stems from the sensitivity of the exact timing of variables to random noise. To illustrate, Forrester conducted an experiment in which a model was resimulated several times with only the exact sequences of random inputs altered (the statistical characteristics of the random inputs were the same in all simulations). The same variables from different simulations were very poor point-by-point predictors of one another. However, random noise did not inhibit the ability to make correct choices between alternative policies in the search for system improvement. The results argue against point-by-point measures of fit and point prediction as being effective in model validation (see Forrester, 1961, Appendix K).

The *multiple-mode-test* considers whether or not a model is able to generate more than one mode of observed behavior. The usefulness of such models for policy analysis makes the multiple-mode test an important test of model behavior. A model able to generate two distinct periodicities of fluctuation observed in a real system provides the possibility for studying possible interaction of the modes and how policies differentially affect each mode. For example, the production-sector model developed by Mass (1975) generates two periodicities – a 3-to-7-year fluctuation and an approximately 18-year fluctuation. Both modes of behavior are observed in the real economy. Short-term fluctuations in production, employment, inventories, and prices generated by the model closely match observed business-cycle fluctuations of the variables, both in period and relative phasing. The 18-year fluctuations in investment and capital stock likewise correspond to so-called Kuznets-cycle fluctuations which have been observed. Through analysis of model behavior, Mass reaches a significant conclusion for economic policy: capital investment is not a major cause of the short-term business cycle, and consequently, policies designed to influence capital spending may provide relatively little leverage for influencing business cycles. Conversely, investment policies may have considerable impact on longer term economic cycles.

Alternatively, the multiple-mode test might be applied to a model that explains why one mode of historical behavior gives way to another. An example is the sudden reversal in Figure 3-1b of *Urban Dynamics* of the underemployed/housing and underemployed/jobs ratios that occurs at the time the growth curve reaches its peak. The rapid transition from low unemployment and tight housing to high unemployment and excess housing has characterized the shift from growth to stagnation in many American cities.

Lastly, *behavior-characteristic tests* are included as a miscellaneous category for other behavior-reproduction tests. Aspects of behavior such as a peculiar shape of a fluctuating time series (e.g., sharp peaks and long troughs) may be the focus of a behavior-characteristic test. Unusual events, such as a great depression or "oil crisis" are also features of behavior which a model might be intended to reproduce; one would expect a model to show the pattern of circumstances and behavior leading to the event rather than the exact time predicted for the event.

It is important that a model pass the behavior-reproduction tests without the aid of exogenous time-series inputs driving the model in a predetermined way. Unless the model shows how internal policies generate observed behavior, the model fails to provide a persuasive basis for improving behavior. Excluding exogenous input variables also aids in analyzing the causes of model behavior. Behavior-reproduction tests become much more convincing when one can show why the tests are passed. When one can show that a particular feature of observed behavior is a necessary consequence of model structure, one has much greater confidence in the significance of a behavior-reproduction test. (Forrester (1961, Chapter 12) discusses this issue of exogenous inputs in more depth.)

To pass the behavior-reproduction tests a model may need to be excited by a simple test input. The type of input required depends on the nature of model behavior. Random disturbances are important for systems whose most significant characteristic is damped oscillation. Random disturbances trigger the irregular fluctuations that are conspicuous when such systems appear in real life. For models focusing on "life-cycle" dynamics, random disturbances are less important because principal interest is in one-time phenomena.

4.2. Behavior-prediction tests

Behavior-prediction tests are analogous to behavior-reproduction tests. Whereas behavior-reproduction tests focus on reproducing historical behavior, behavior-prediction tests focus on future behavior. System dynamics model-builders have often stressed that their models do not strive for prediction of future values of system variables – that is, for "point prediction" (see Forrester, 1961, pp. 123–128). However, system dynamics models should tell certain things about behavior in the future. The possible range of predictive objectives for a system dynamics model are illustrated by the pattern-prediction and event-prediction tests.

The *pattern-prediction test* examines whether or not a model generates qualita-

tively correct patterns of future behavior. Conduct of the pattern-prediction test may entail evaluation of periods, phase relationships, shape, or other characteristics of behavior predicted by the model.

The *event-prediction test* focuses on a particular change in circumstances, such as a sharp drop in market share or a rapid upsurge in a commodity price, which is found likely on the basis of analysis of model behavior. As in the other predictive tests, evaluation of the event-prediction test should hinge on the dynamic nature of an event and identification of conditions leading to it rather than on the exact time when an event will occur. A good example of the event-prediction test can be seen in the studies of natural-resource depletion by Behrens (1973) and Naill (1973). Both studies showed that the price of a resource could rise precipitously even after a long period of steady or falling prices. The model showed the nature and ultimate inevitability of an event, but not necessarily the timing. Since the papers were written, such a sharp rise has occurred in the world price of oil and natural gas.

One other predictive test, the changed-behavior-prediction test, is presented in § 4.

4.3. Behavior-anomaly test

The behavior-anomaly test frequently arises with system dynamics models. In constructing and analyzing a system dynamics model, one expects it to behave like the real system under study; but frequently the model-builder discovers anomalous features of model behavior which sharply conflict with behavior of the real system. Once the behavioral anomaly is traced to the elements of model structure responsible for the behavior, one often finds obvious flaws in model assumptions.

Although the behavior-anomaly test is used extensively in model development, it can also play a broader role in validation. For example, one can often defend particular model assumptions by showing how implausible behavior arises if the assumption is altered. The investment-function study cited above (Senge, 1978) offers numerous examples of behavior-anomaly tests of the investment function developed for the System Dynamics National Model (Forrester *et al.*, 1976). The tests show how behavior anomalies arise when components of the investment function are eliminated.

4.4. Family-member test

System dynamics models usually represent a family of social systems. In other words, *when possible* a model should be a general model of the class of system to which belongs the particular member of interest. One should usually be interested in why a particular member of the class differs from the various other members. How did different policies produce the different behaviors? An important step in validation is to show that the model takes on the characteristics of different members of the class when policies are altered in accordance with the known decision-

making differences between the members. The model is a general theory; its structure is the structure of the entire class. For example, a corporate model of loss of market share should show the different behaviors of loss or gain of market share as its parameters are changed to represent policies followed in contrasting companies. Likewise, the *Urban Dynamics* model should be interpreted as a general model of urban growth and equilibrium. With appropriate choice of parameters, it should behave like cities as different as New York, Dallas, West Berlin, and Calcutta. To behave in such diverse ways, the parameters and tables of the urban model must be changed to represent the appropriate geographical, cultural, climatic, sociological, and economic conditions.

The family-member test permits a repeat of the other tests of the model in the context of different special cases that fall within the general theory covered by the model. The general theory is embodied in the structure of the model. The special cases are embodied in the parameters. To make the test, one uses the particular member of the general family for picking parameter values. Then one examines the newly parameterized model in terms of the various model tests to see if the model has withstood transplantation to the special case.

4.5. Surprise-behavior test

The better and more comprehensive a system dynamics model, the more likely it is to exhibit behavior that is present in the real system but which has gone unrecognized. Often such behavior emerges to the surprise of the model builder. When unexpected behavior appears, the model builder must first understand causes of the unexpected behavior within the model, then compare the behavior and its causes to those of the real system. When this procedure leads to identification of previously unrecognized behavior in the real system, the surprise-behavior test contributes to confidence in a model's usefulness.

For example, Lyneis *et al.* (1977) describe a corporate model dealing with instability of employment which showed loss of market share as had been anticipated, but which also showed that the drop in market share was occurring at the time of business downturns, which had not been realized. In the model, the product was less available during declining business than during times of high demand. The model showed a steeper reduction in production than in demand and inability to deliver during times when more sales could have been made. A review of the data showed the same timing had been occurring in the actual system, though the behavior had gone unnoticed in formulating company policies.

4.6. Extreme-policy test

The extreme-policy test involves altering a policy statement (rate equation) in an extreme way and running the model to determine dynamic consequences. Does the model behave as we might expect for the real system under the same extreme pol-

icy circumstances? For example, one could ask for the above-mentioned National Model, "What would happen if further capital investment were not possible?" Does the remainder of the system (for example, employment and personal savings flows) respond as might be expected under conditions of zero new-capital acquisition?

The extreme-policy test is important because one may be quite sure what would happen under the extreme circumstances even if no real-life example has been observed. The test shows the resilience of a model to major policy changes. The better a model passes a multiplicity of extreme-policy tests, the greater can be confidence over the range of normal policy analysis and design.

4.7. Boundary-adequcay (behavior) test

The boundary-adequacy (structure) test discussed in § 3 often must be extended to include analysis of model behavior. The boundary-adequacy (behavior) test considers whether or not a model includes the structure necessary to address the issues for which it is designed. The test involves conceptualizing additional structure that might influence behavior of the model. When conducted as a behavior test, the boundary-adequacy test includes analysis of behavior with and without the additional structure. Conduct of the boundary-adequacy test requires modeling skill, both in conceptualizing model structure and in analyzing the behavior generated by alternative structures.

In his business-cycle study Mass (1975) provides an example of the boundary-adequacy test conducted as a test of model behavior. First, he showed that basic inventory-management and backlog-management policies in conjunction with delays in adjusting workforce were capable of generating business-cycle fluctuations. He then extended the model boundary to incorporate endogenous consumer demand, thereby incorporating another possible cause of business cycles. He found that the consumption multiplier link had little influence on model behavior; the model generated essentially the same modes of behavior with and without endogenous consumption. By this test, Mass strengthened confidence that the original model boundary which excluded endogenous consumption was adequate for understanding the primary causes of short-term business cycles.

4.8. Behavior-sensitivity test

The behavior-sensitivity test focuses on sensitivity of model behavior to changes in parameter values. The behavior-sensitivity test ascertains whether or not plausible shifts in model parameters can cause a model to fail behavior tests previously passed. To the extent that such alternative parameter values are not found, confidence in the model is enhanced. For example, does there exist another equally plausible set of parameter values that can lead the model to fail to generate observed patterns of behavior or to behave implausibly under conditions where plausible behavior was previously exhibited?

The behavior-sensitivity test is typically conducted by experimenting with different parameter values and analyzing their impact on behavior. Frequently, after extensive model analysis, the system dynamics modeler has a good idea where sensitive parameters might lie and this understanding effectively guides sensitivity analysis. However, the behavior-sensitivity test can be formalized. Numerous studies (see, for example, Vermeulen and DeJongh, 1977) have applied mathematical and formalized computational procedures to sensitivity testing in system dynamics models. For example, such studies often compute time-varying partial derivatives of state variables with respect to changes in parameters. Although formal procedures can clearly aid sensitivity analysis, many formal sensitivity studies fall into the trap of losing sight of the other confidence-building tests for system dynamics models. In particular, they often stop at identifying "sensitive parameters" and fail to establish whether or not plausible shifts in those parameters cause the model to fail behavior tests which were previously passed. Unless formal sensitivity analyses are related to the other confidence-building tests, their implications for validity usually remain unclear.

Typically, the behavior of system dynamics models is insensitive to plausible changes in most parameter values. It appears that real systems are likewise insensitive. For example, behavior of many corporations continues with characteristic successes and failures over several changes of presidents and under changing external conditions. Likewise, all the older Northeastern cities in the United States show the same symptoms of aging and unemployment whether they be seacoast cities, manufacturing centers, or the nation's capital. On the other hand, both real systems and models or real systems shown behavior sensitive to a few parameters. Finding a sensitive parameter does not necessarily invalidate the model. Even though it has a substantial effect on behavior, plausible variations in the parameter may not lead to failure of other behavior tests. Moreover, one should attempt to ascertain by comparing different members of the class of systems, whether or not the real system is likewise sensitive to the parameter in question. If it is, the sensitive parameter may be an important input for policy analysis.

4.9. Other tests

In the preceding section, we commented on the use in system dynamics of standard statistical tests such as are common in econometrics. A more promising approach to statistical testing in system dynamics models may lie with a newer set of statistical tools using the Kalman filter developed in the field of engineering. Statistical tests based on the Kalman filter differ from conventional econometric tests as behavior tests differ from structure tests. Conventional statistical tests attempt to compare model structure directly to data; Kalman filter tests compare model behavior to data. The difference in approach permits tests based on the Kalman filter to separate out the effects of measurement error when testing hypotheses and may prove significant for applications in system dynamics. Peterson (1979) has

shown the potential usefulness for system dynamics models of statistical tests based on the Kalman filter; broader acceptance of such methods awaits further evidence of their benefits in practice.

5. Tests of policy implications

Tests can be conducted to build confidence in a model's implications for policy. Although all tests of system dynamics models aim at usefulness of a model as a policy-analysis tool, tests of policy implications differ from other tests in their explicit focus on comparing policy changes in a model and in the corresponding reality. Policy implication tests attempt to verify that response of a real system to a policy change would correspond to the response predicted by a model. The tests also examine how robust are policy implications when changes are made in boundaries or parameters.

5.1. System-improvement test

The ultimate test of a system dynamics model lies in identifying policies that lead to improved performance of the real system. The system-improvement test considers whether or not policies found beneficial after working with a model, when implemented, also improve real-system behavior.

Although it is the ultimate real-life test, the system-improvement test presents many difficulties. First, it will not be tried until the model from which the new policies come enjoys enough confidence for the implementation experiment to be made. Second, if the real-life experiment is made and results are as predicted, the test is often clouded by the assertion that the beneficial results came from causes other than the new policies. No matter what the outcome, interpretation of actual policy implementation is invariably subject to uncertainty as to whether or not other conditions were adequately constant to permit attributing the results to the policies. Third, the very long time for reaction in most social systems (running to months or years for a corporation, and to decades for a national economy) mean that results of the system-improvement test accumulate slowly.

In time, the system-improvement test becomes the decisive test, but only as repeated real-life applications of a model lead overwhelmingly to the conclusion that models pointed the way to improved policies. In the meantime, confidence in policy implications of models must be achieved through other tests.

5.2. Changed-behavior-prediction test

The changed-behavior-prediction test asks if a model correctly predicts how behavior of the system will change if a governing policy is changed. The test can be made in several ways. Initially, the test can be made by changing policies in a model

and verifying plausibility of resulting behavioral changes. Alternatively, one can examine response of a model to policies which have been pursued in the real system to see if the model responds to a policy change as the real system responded. If the model represents a family of systems, some of those systems will probably be operating under different policies, and the policies of the model can be altered to see if its behavior takes on the different behaviors that distinguish members of the family.

Several examples of the changed-behavior-prediction test can be drawn from research connected with the *Urban Dynamics* model. In the book presenting the original model, Forrester (1969) examined the response of the model to several policies that had been tried in real cities. He examined model response to job-creation and job-training programs, a low-income-housing program, and financial aid, and found that, in each case, a set of pressures arose within the model that combined to defeat the intended positive results of the program. Although Forrester did not discuss in any detail real-life examples of failure of the same programs, there appears to be ample evidence of such real-city failures. To complete the changed-behavior prediction test, one would need to examine the pressures which arose to defeat the real programs and compare those pressures to reasons for failure in the model. (Some of the evidence was examined in subsequent urban dynamics research – see Mass (1974) and Schroeder *et al.* (1975).)

5.3. Boundary-adequacy (policy) test

The boundary-adequacy test, when viewed as a test of the policy implications of a model, examines how modifying the model boundary would alter policy recommendations arrived at by using the model. The boundary-adequacy test requires conceptualization of additional structure and analysis of the effects of the additional structure on model behavior. One repeats the simulations involving a particular policy recommendation to determine why the additional structure does or does not alter the recommendation.

One example of the boundary-adequacy test of policy implications was Schroeder's response to a boundary issue in the *Urban Dynamics* model, cited above. To respond to the issue that the original model was inadequate for policy testing because it omitted city-suburb interactions, Schroeder (in Schroeder *et al.*, 1975) constructed a suburb model and merged it with the original *Urban Dynamics* model. Analysis of the revised model showed no significant shifts in policy recommendations from the original *Urban Dynamics* model. By conducting the test, Schroeder demonstrated that the geographical boundary assumed in the original model was adequate for the original set of policy issues considered.

5.4. Policy-sensitivity test

Parameter sensitivity testing can, in addition to revealing the degree of robustness of model behavior, indicate the degree to which policy recommendations

might be influenced by uncertainty in parameter values. Such testing can help to show the risk involved in adopting a model for policy making. If the same policies would be recommended, regardless of paramater values within a plausible range, risk in using the model will be less than if two plausible sets of parameters lead to opposite policy recommendations. Exploration of parameter-sensitivity testing as related to policy is illustrated in Appendix B of *Urban Dynamics* (Forrester, 1969). Illustrated there is the one parameter change known to the author that could invalidate the recommended policies that were given. The parameter change requires the assumption that people are almost totally indifferent to the availability of housing – indifferent to the extent that removing most of the housing in a city would have negligible effect on decisions by people moving to and from the city. In this case, Forrester viewed as implausible the only parameter change he found capable of invalidating the model's policy recommendations. Hence, the policy-sensitivity test suggested that the policy recommendations were not likely to be affected by uncertainties in parameters.

6. Conclusions

Table 1 summarizes the tests of model structure, behavior, and policy implications which have been presented above. The table identifies 17 tests, illustrating the breadth of channels for building confidence in system dynamics models.

With so many tests available, one naturally asks whether, in fact, all tests must be carried out in all modeling applications and, in particular, if there isn't a subset of tests which might be considered "core tests for system dynamics." One can identify such a subset, based on the tests that accomplished systems dynamicists generally rely on. Included in this set of core tests would be all the structure tests, because they are intrinsically part of constructing a system dynamics model. Accomplished modelers appear to rely particularly heavily on the extreme-condition test as a means of identifying faulty hypotheses. Of the behavior tests, the most utilized are probably the behavior-reproduction tests, the behavior-anomaly test, and the behavior-sensitivity test. Almost all system dynamics models should be capable of reproducing certain "target" modes of observed behavior and responding plausibly to a wide range of test conditions. Changed-behavior prediction and policy-sensitivity are essential tests of a model's policy implications. This set of core tests are indicated by a superscript "a" in table 1.

Despite the fact that it might not always be possible or cost-effective to conduct all the confidence-building tests, the existence of a wide variety of tests increases the likelihood that more tests will be conducted and that more people can be involved in the overall validation process. In fact, one of the key features of the tests discussed above is the extent to which they can be readily carried out by many types of evaluators. Virtually all tests can be either conducted or understood by an interested nontechnical model user. None of the tests require mathematical or com-

Table 1
Confidence-building tests.

Tests of Model Structure	
[a]	1. Structure Verification
[a]	2. Parameter Verification
[a]	3. Extreme Conditions
[a]	4. Boundary Adequacy
[a]	5. Dimensional Consistency
Test of Model Behavior	
[a]	1. Behavior Reproduction (symptom generation, frequency generation relative phasing, multiple mode, behavior characteristic)
	2. Behavior Prediction (pattern prediction, event prediction, shifting-mode prediction)
[a]	3. Behavior Anomaly
	4. Family Member
	5. Surprise Behavior
	6. Extreme Policy
	7. Boundary Adequacy
[a]	8. Behavior Sensitivity
Tests of Policy Implications	
	1. System Improvement
[a]	2. Changed-Behavior Prediction
	3. Boundary Adequacy
[a]	4. Policy Sensitivity

[a] Core tests.

putational techniques which cannot be easily explained. Technical evaluators can conduct all the tests except where limited competence in system dynamics might preclude conceptualizing new model structure (boundary-adequacy tests).

The accessibility of the whole testing process is crucial to possibilities for success in system dynamics modeling. If fully exploited, the large variety of tests available to a multiplicity of evaluators should enable the development of useful models in which there is widely-shared confidence.

References

Ansoff, H.I. and D.P. Slevin, 1968, An appreciation of industrial dynamics, Management Science 14, 383–397.

Behrens, W.B., 1973, The dynamics of natural resource utilization, in: Toward Global Equilibrium: Collected Papers (MIT Press, Cambridge, Mass.) 141–164.

Clarkson, Geoffrey, P.E., 1962, Portfolio Selection: A Simulation of Trust Investment (Prentice-Hall, Englewood Cliffs, N.J.).

Cohen, K.J. and R.M. Cyert, 1961, Computer models in dynamic economics, The Quarterly Journal of Economics 75, 112–122.

Fishman, G.S. and O.J. Kiviat, 1967, The analysis of simulation-generated time series, Management Science 13, 525–557.

Forrester, J.W., 1961, Industrial Dynamics (MIT Press, Cambridge, Mass.).

Forrester, J.W., 1968, Industrial dynamics – a response to Ansoff and Slevin, Management Science 14, 601–618.

Forrrster, J.W., 1969, Urban Dynamics (MIT Press, Cambridge, Mass.).
Forrester, J.W., 1971, World Dynamics (MIT Press, Cambridge, Mass.).
Forrester, J.W., G.W. Low and N.J. Mass, 1974, The debate on world dynamics – a response to Nordhaus, Policy Sciences 5, 169–190.
Forrester, J.W., 1976, Business structure, economic cycles, and national policy, Futures 9, 195–214.
Forrester, J.W., N.J. Mass and C.J. Ryan, 1976, The system dynamics national model: understanding socio-economic change and policy alternatives, Technological Forecasting and Social Change 9, 51–68.
Graham, A.K., 1979, Parameter estimation in system dynamics modeling, this volume.
Holt, C.C., 1965, Validation and application of macroeconomic model using computer simulation, in Duesenberry *et al.*, The Brookings Quarterly Economic Model of the U.S. Economy (Rand McNally & Co., Chicago, Ill.).
Keynes, J.M., 1939, Professor Tinbergen's method, Economic Journal 49, 567–569.
Leontief, W., 1971. Theoretical assumptions and nonobserved facts, American Economic Review 61, 1–7.
Lyneis, J.M., D.W. Peterson and B.E. Tuttle, 1977, Implementing the results of computer-based models – lessons from a case study, Systems Dynamics Group Working Paper D-2674, Alfred P. Sloan School of Management (MIT, Cambridge, Mass.).
Mass, N.J. ed., 1974, Readings in Urban Dynamics: Vol. 1 (MIT Press, Cambridge, Mass.).
Mass, N.J., 1975, Economic Cycles: An Analysis of Underlying Causes (MIT Press, Cambridge, Mass.).
Mass, N.J. and P.M. Senge, 1978, Alternative tests for the selection of model variables, IEEE Systems, Man and Cybernetics 8, 450–459.
Morrison, D.E. and R.E. Henkel, 1970, The Significance Test Controversy (Aldine, Chicago, Ill.).
Nail, R.F., 1973, The discovery life-cycle of a natural resource: a case study of U.S. natural gas, in: D.L. Meadows and D.H. Meadows, eds., Towards Global Equilibrium: Collected papers (MIT Press, Cambridge, Mass.) 213–256.
Naylor, Thomas H., J.L. Balintty, D.S. Burdick and K. Chu, 1966, Computer Simulation Techniques (Wiley, New York).
Naylor, T.H. and J.M. Finger, 1967, Verification of computer simulation models, Management Science 14, B92–B101.
Nordhaus, W.D., 1973, World dynamics: measurement without data, Economic Journal 83, 1156–1183.
Orcutt, G.H., M. Greenberger, J. Korbel and A.M. Rivlin, 1961, Microanalysis of Socioeconomic Systems: A Simulation Study (Harper and Brothers, New York).
Peterson, D.W., 1979, Statistical tools for system dynamics, in: J. Randers, eds., Elements of the System Dynamics Method (MIT Press, Cambridge, Mass.).
Phelps Brown, E.H, 1972, The underdevelopment of economics, Economic Journal 82, 1–10.
Randers, J., ed., 1979, Elements of the System Dynamics Method (MIT Press, Cambridge, Mass.).
Schroeder, W.W. III, R.E. Sweeney and L.E. Alfeld, eds., 1975, Readings in Urban Dynamics: Vol. 2 (MIT Press, Cambridge, Mass.).
Senge, P.M., 1977, Statistical estimation of feedback models, Simulation (June) 177–184.
Senge, P.M., 1978, The System Dynamics National Model Investment Function: A Comparison to the Neoclassical Investment Function, Ph.D. dissertation, Alfred P. Sloan School of Management (MIT, Cambridge, Mass.).
Vermeulen, P.J. and D.C. DeJongh, 1977, Growth in a finite world – a comprehensive sensitivity analysis, Automatica 13, 77–84.
Worswick, G.D.N., 1972, Is progress in economic science possible? Economic Journal 82, 73–86.
Ziman, J.M., 1968, Public Knowledge: An Essay Concerning the Social Dimension of Science (Cambridge University Press, London).