Levels of Confidence in System Dynamics Modeling: A Pragmatic Approach to Assessment of Dynamic Models

Aldo A. Zagonel and Thomas F. Corbet
CI Modeling and Simulation Department
Sandia National Laboratories[†]

Abstract

This paper provides an overview of the literature in assessment of system dynamics (SD) models to substantiate a pragmatic framework intended to guide model testing, refinement and evaluation. It recaps the predominant philosophy of science embraced in the field, and its implications for model validation. It reviews tests for building confidence in SD models. In this literature, SD is presented as a relatively uniform approach to dynamic modeling. However, surveys of the field paint a different picture, containing surprisingly diverse forms of practice. We draw upon this breadth of existing practice to develop our framework. We propose five components of practice: 1) system's mapping, 2) quantitative modeling, 3) hypothesis testing, 4) uncertainty analysis, and 5) forecasting/optimization. In light of the proposed framework, we reclassify tests for assessment of dynamic models across these five practical categories. We believe this is useful to tailor tests to specific modeling efforts, guide model testing in different phases of model development, and to help conduct partial assessments of levels of confidence.

Key words: Model testing, model evaluation, model validation, confidence building

[†] Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000. The Department of Homeland Security, Science & Technology Directorate, provided funding for this work.

Levels of Confidence in System Dynamics Modeling: A Pragmatic Approach to Assessment of Dynamic Models

Introduction

We desired guidelines for testing and building confidence in the range of models that we use to analyze interdependencies between critical infrastructures. Our intent was to draw upon the field of system dynamics (SD) for existing guidelines developed for model testing and evaluation. Forrester & Senge (1980) proposed 17 tests organized in three broad categories: tests of model structure, of model behavior, and of policy implications. Sterman (2000) revisited this work – clustering some of the tests and eliminating the categories– and produced 12 tests. Overall, there is a large degree of overlap, with the exception of tests of behavior prediction, dropped in Sterman's overview, and a test added on integration error.

However, we realized that our modeling approaches and objectives are diverse enough that not all of the tests proposed in the literature are applicable to each modeling project. It seemed that it would be helpful to bin these tests to make it easier for our modelers to identify the best subset of tests for each project. It is clear that the decision as to whether or not a particular test is appropriate should consider the assumed basis for model truth, or correctness. We therefore examined the literatures concerning the scientific basis, and the breadth of SD modeling practice, to categorize model tests. The result was a regrouping of the tests proposed by Forrester & Senge, and Sterman into five categories that are intended to represent separate components of modeling. We think that this grouping will help guide the choice of appropriate tests on a case-by-case basis.

SD model testing and evaluation literature

While we focused upon the seminal piece by Forrester & Senge, and on Sterman's modern and thorough textbook, we acknowledge that the literature on SD model testing and evaluation is wide ranging, including many specific contributions (such as Mass & Senge 1980, Peterson 1980, Tank-Nielson 1980), and other efforts to overview and discuss the tests comprehensively (such as Forrester 1961, Richardson & Pugh 1981, Barlas 1989, Ford 1999). For practical reasons and because none of these other sources are as comprehensive, we focused primarily upon the first two. In our paper we also indicate some examples of new developments in the field which are not yet incorporated into any comprehensive overview.

We perceive the organizing structure that we seek to develop here as more practically useful than the early framework adopted by Forrester & Senge (1980), also discussed in Barlas (1996), and later apparently dropped by Sterman (2000). Finding some structure to organize and discuss model testing and evaluation is an important issue, particularly to communicate the "know-how" to novice modelers, and (at least for our purposes) to develop testing protocols and standards. This effort seems useful to the SD field in general, as communicated to us by one of the leading figures in the field:

> I've always been worried that Forrester's list of 17 tests is so daunting that maybe lots of folks never use any of it, thinking that if you are going Forrester's route you have to go all the way. So your parsing of the tests is going to prove useful, I think, provided it's possible to say it all concisely and invitingly. (Professor George P. Richardson, personal communication, 11/30/2005)

Others attempted to approach this issue pragmatically. For example, Wakeland and Hoarfrost (2005) examined the portfolio of tests looking for cost/benefit measures of test usefulness. They found that contrary to its popularity, *sensitivity analysis* yielded relatively low benefit and high cost, if compared to *structure assessment, dimensional consistency and boundary adequacy*. However, they acknowledged that their measures were somewhat subjective, and that the results could vary considerably from model to model. More importantly, we fear that such an approach would induce a "check-list" type behavior, leading practitioners to focus upon checking off quickly as many as possible boxes with minimum effort. On the other hand, we would like to arrive at an organizing framework that is theoretically sound and practically useful, yet explicitly recognizes that what tests are done, and how much testing is done, is contingent upon the context of the modeling effort. While the idea of using the portfolio of tests "intelligently" is obviously not new, we find that none of the existing frameworks convey explicitly how to do this. Without it, we are left with the quandary: "I cannot give you a recipe on how to test your model, but I'll tell you if your model is valid when I look at what you've done."

Method

We followed a five-step process:

1. We grounded our research upon epistemological studies of SD theory and practice, basing our conceptual framework upon Barlas & Carpenter (1990) and Lane (2001). We tentatively established an axis of *Liberty vs. Constraint* in SD modeling and simulation. This axis serves to differentiate different components of modeling along an ordinal scale ranging left to right from loosely constrained to highly constrained modeling efforts;

2. We used expert judgment to identify *five components of modeling* based upon observed practice. While the literature was informative, we based this step essentially upon the experience of modelers of different background working within our organization. The authors, then, articulated a summary description of the essence of each component, and decided the *order of the components* in the scale established in step 1;

3. We analyzed the content of tests for building confidence in SD models. Two formally trained SD modelers, one within our organization and the other an outsider, undertook the task of specifying *24 distinct tests* working from Forrester & Senge's 17 and Sterman's 12;

4. The same two modelers discussed and agreed upon how to *cluster the 24 tests in each of the five components*, and ranked them within each component, thus also determining which tests could be considered *basic, intermediate,* and *advanced*;

5. We sought peer review of the results by sharing and discussing our findings with distinguished members of the SD Society. Their criticism, comments, and suggestions for future steps in this research are noted in this paper.

The field's epistemological standing and breadth

Barlas & Carpenter (1990) reviewed the philosophical roots of model validation, tracing the historical development of the different theories of knowledge, and found two opposing philosophies of science. One paradigm, originated with Descartes' *Rationalism* (1596-1650) and Locke's *Empiricism* (1632-1704), guided the major epistemological theories of the sixteenth through the nineteenth centuries. This traditional logical/empiricist philosophy of science assumes that *knowledge is an objective representation of reality and that theory justification can be an objective, formal process* (pp. 148-149).

A different paradigm emerged building upon criticism dating back to Hegel's *Coherence Theory* (1770-1831) and Dewey's *Pragmatism* (1859-1952), who articulated that *knowledge is socially justified belief*, rather than a product of mirroring nature and, thus, *socially, culturally, and historically dependent*. But, it was only in the second half of the 20th Century that this perspective flourished. In "The Structure of Scientific Revolutions" (1962), Kuhn argued that scientific progress is not directed toward an objective and absolute "truth" but simply toward "successful creative work." Also, Sellers (1963) proposed that knowledge acquisition is holistic rather than atomistic. These alternative perspectives led to an opposing philosophy of science denominated relativist/holistic (Barlas & Carpenter, pp. 152, and 155-157).

In their assessment, Barlas & Carpenter concluded that the field of system dynamics is more closely associated with the relativist/holistic viewpoint, than that of the logical/empiricist epistemological tradition. This is reflected both in its manifested standing, and in criticism presented by others more closely associated with a traditionalist perspective. Forrester's early work (1961, 1971) embraced the relativist approach (i) proposing that the validity of a model cannot be discussed absent of its purpose, (ii) accepting "qualitative" model validation, and (iii) interpreting "data" in a broad sense, including non-numerical and verbal information, and mental data bases. From an empiricist standpoint, his critics objected profusely and were quick to label the approach as subjective (Ansoff and Slevin 1968), and summarized it as "measurement without data" (Nordhaus 1973).

Barlas & Carpenter focused upon the core philosophy of the field, grounded in a feedback-rich, insight-driven, problem-focused approach to system dynamics modeling. They cite original methodological texts and refer to classic models such as World Dynamics (Forrester 1961, 1971; Legasto et al. 1980; Randers 1980-A; Richardson and Pugh 1981). In principle, and in practice, they concluded that system dynamics is a scientific method according to the standards placed by the relativist/holistic philosophy of science (p. 163):

> Real-life experience has taught most system dynamics practitioners that models are inherently incomplete, relative, and partly subjective, and that model validity means usefulness with respect to a purpose. But, at the same time, many

> practitioners unaware of the recent relativist philosophical developments would think that their own view of model validity is not truly scientific. Thus, many practitioners, while experiencing that validation is bound to be relative, semiformal, and conversational process, at the same time see this as a weakness of their modeling effort… System dynamics practitioners do not have to be apologetic for not meeting a utopian [logical/empiricist] criterion of scientific inquiry. [Added]

However, a recent survey of system dynamics practice (Lane 2001) demonstrated that the field is not as uniform as the core philosophical description would characterize it. Rather, Lane would denominate this characterization as "initial" or "broad" system dynamics, perhaps including the most recent text embraced by the core of the field (Sterman 2000). But, SD practice has grown in different directions. For example, "austere" SD places greater emphasis on deterministic, positivist and objectivist approaches including (p. 106): [1]

- Micro world validation, in which emphasis is placed on quantitative data to test for behavior modification (Bakken et al. 1992); and
- behavioral decision-making work (Sterman 1989; Kleinmuntz 1993)

Lane's research findings suggest that SD practice is not only growing towards the objective (logical/empiricist) end as discussed above, but also towards the subjective (relativist/holistic) viewpoint. In fact, there are more new areas of practice toward the relativist/holistic side. Some of these new forms of SD practice may be so far off towards the relativist/holistic side as to antagonize not only positivists but also mainstream system dynamicists.

This tension is clearly observed in the field's main journal (e.g., Richardson et al. 1994). It is often translated into a debate between systems "thinking" and system "dynamics" (soft and hard approaches), or between "qualitative" and "quantitative" modeling.[2] This debate was recently revisited in articles by Coyle (2000, 2001) and Homer & Oliva (2001). While there is much agreement amongst practitioners sharing Forrester's heritage, as indicated by Coyle (2001, p. 357):

---

[1] Lane's analytical framework for mapping SD practice is not the same as that of Barlas & Carpenter's. Instead, he used Burrell and Morgan's (1979) sociological approach, which in one dimension contrasts subjective and objective views of social science. However, these frameworks are *epistemologically* similar. They both describe a tension between two polar points, one which is humanistic and the other which is positivistic (Lane 2001, p. 102). Burrell and Morgan's framework deals not only with epistemology, but also with ontology, hermeneutics, and methodology:

> Epistemological issues concern the type of knowledge that can be obtained. The positivist view is that causal laws perceivable by an objective observer may be deduced, whilst the humanistic stance sees knowledge as being concerned with the significance and meaning that humans ascribe to their actions, these being drawn out via the textual interpretation… (Lane 2001, pp. 101-102)

[2] While this debate has predominantly involved soft vs. hard, and qualitative vs. quantitative approaches, it also appears in the discussion of academic vs. professional work (Graham 2002).

> Homer, Oliva and I are in agreement on several points: one cannot reliably predict dynamics by looking at a diagram, and properly quantified models are very useful in the right circumstances… that there may be cases when a diagram is all that is necessary or feasible…

It seems also to be an *agreement to disagree*. One side argues that "simulation [a more objective approach to practice] nearly always adds value, even in the face of significant uncertainties about data and the formulation of soft variables" (Homer & Oliva 2001, p. 347); while the other insists that "quantification may not 'represent value for money'… even more concerning is… the risks associated with attempting to quantify multiple and poorly understood soft relationships are likely to outweigh whatever potential benefit there might be." (Coyle 2001, p. 357) This may indicate that there is a *range* of "accepted" modeling practice, even if not everyone is on board on what those practices are.

While Barlas & Carpenter (1990) argue that the core practice of system dynamics bets fits into the relativist/holistic philosophy of scientific knowledge, other authors describe a modeling practice that sufficiently large that system dynamics modeling could also be considered to also fit into logical/empiricist tradition. Model testing should, therefore, be able to accommodate a range in belief of what constitutes a correct model.

Proposed grouping of model tests

Given the wide range of objectives and approaches using system dynamics, it is clear that a testing plan can not be specified that fits all models. Therefore we divided modeling practice into a number of components, each with associated testing requirements. A model or modeling project could consist of one or more of these components.

Components of Modeling

We chose *five* components of models or modeling projects that can be combined to represent the full range of modeling approaches and objectives: 1) system's mapping, 2) quantitative modeling, 3) hypothesis testing, 4) uncertainty analysis, and 5) forecasting/optimization. This is not to say that these components exist in isolation, while in some cases that may be possible. But in general any modeling effort will consist of a combination of these components centered upon a particular objective. Table 1 contains a summary of the characterizations that follow:

Table 1. Five components of modeling practice[3]

Loosely constrained ◄——————————————————————————► highly constrained

System's mapping	Quantitative modeling	Hypothesis-testing	Uncertainty analysis	Forecasting and optimization
Qualitative and inductive; involves drawing influence diagrams, CLDs, S&F diagrams, or any form of mapping or organization of the elements forming a system; attempts to get at the key causal interrelationships; focused upon identification of inter-organizational linkages and inter-dependencies	Quantitative and descriptive; involves formulation and simulation; largely system-focused; emphasizes S&F dynamics and the effects of delays; requires specification of the decision rules governing interrelationships; focused on representing and tracking consequences; sometimes rich in detail complexity	Quantitative and deductive; requires stating a hypothesis that explains dynamic behavior from the causal structure of the system; largely problem focused; emphasizes feedback-rich dynamics, learning, and exploration of the effect of changes in system structure; focused upon understanding and insight	Quantitative and exploratory; requires examining behavioral and quantitative sensitivity; emphasizes testing the robustness of the results produced from both quantitative modeling and hypothesis testing; focused upon uncertainty and risk, and identification of points of leverage for intervening in the system	Quantitative and predictive; within the range of the parameter space specified in the model, attempts to shed light on future behavioral patterns and the cross-sectional quantitative values of variables of interest, or to suggest optimal or robust solutions that maximize or "satisfice" particular utility functions

[3] To map these categories, we used an axis of *Liberty vs. Constraint*, which can be perceived as a common factor extracted from five dimensions: a) adoption of a divide and conquer analytical strategy (*reductionism*), whereby a complex set of facts, entities, phenomena, or structures is explained using a simpler set, such that when assembled, the small pieces explain the whole; b) knowledge is derived from reasoning (*rationalism*); c) knowledge is derived from observation and experience (*objectivism/empiricism*); d) reasoning from the general to the particular, combining thesis and antithesis in a dialectical process to produce a higher level of truth (*synthesism*); and e) amount of agreement upon problem definition and modeling objectives (*monolithism*). The first three constraining factors were derived from Barlas & Carpenter (1990), and Lane (2001). The last two were borrowed from Zagonel's (2002) distinction of models as "boundary objects" and "micro worlds." The specified "bins" are based upon a classification from observed modeling approaches and practices. The five categories fall in this ordinal scale according to a subjective assessment of the degree of constraint imposed by a combination of the above-mentioned factors.

1. System's mapping

This area of practice is qualitative and *inductive*.[4] It involves drawing influence diagrams, causal-loop diagrams (CLDs), and stock-and-flow (S&F) diagrams, or any other form of mapping or organization of the elements forming a system. Normally the map is focused upon an overriding typology or theme.

Systems mapping in and of it self can be very useful. It serves as a visual summary of a lengthier verbal or written discussion. It organizes information. If built collectively, it reflects a shared or sum of perspectives on the issue at hand. If the framework is rich enough, it may yield preliminary dynamic insights. For example, a stock-and-flow diagram helps to understand points of accumulation and intervention. Alternatively, causal-loop diagrams begin to explore reinforcing and balancing feedback. Delays can also be graphically displayed. Maps facilitate the surfacing and clarification of assumptions, and thus can help with communication.

System's mapping has always been a part of system dynamics. From the early days of Dynamo, S&F and CLDs were used in model conceptualization, and to communicate diagrammatically the structure of the computational model. Richmond et al. (1987) and Richardson (1997) articulated "principles" behind such diagrams (unit consistency, accumulation, causality, etc.). Morecroft (1982) contributed with the policy structure diagram, and Mashayeki (citation needed) with the sector diagram, implemented in iThink™ for the purpose of higher-level mapping (Richmond 1994).

We believe there are numerous examples of modeling practice that focus more directly upon this facet of modeling. We already mentioned Coyle's (2000, 2001) regard for this component of practice. Some new mapping approaches were developed outside of the field (Checkland 1981, Hodgson 1994, Eden 1994, Ackermann et al. 2004). They are increasing being incorporated in areas of practice focused upon the other components (Vennix et al. 1990, Lane 1993, Andersen & Richardson 1997).

2. Quantitative modeling

Quantitative modeling involves formulation and simulation. We differentiate this category from the next, hypothesis testing, because here the formulation effort is still (counter-intuitively)

---

[4] The SD method includes both inductive and deductive logic. Beveridge describes these two systems in the following manner:

> Logicians distinguish between inductive reasoning (from particular instances to general principles, from facts to theories) and deductive reasoning (from the general to the particular, applying theory to a particular case). In induction one starts from observed data and develops a generalization that explains the relationships between the objects observed. On the other hand, in deductive reasoning one starts from some general law and applies it to a particular instance. (In Babbie 1992, p. 49)

Normally the qualitative steps of the SD method emphasize the induction process of model building. This is a theory building process; the conceptual model is a representation of a theory regarding the causal relationships in the system. Alternatively, the deduction process is captured in the quantitative steps. This is a theory testing process; the formulated model is simulated, tested and evaluated in light of the hypothesized expectations, and tested against "known" aspects of the "real" system.

inductive, although it might best be characterized as "descriptive." This category is similar to the system's mapping category in this respect. While behavioral modes for key variables may be discussed both here and in system's mapping, those may not be closely tied to structural relationships embedded in the system. Quantitative modeling is focused upon understanding stock-and-flow structures and aging chains. In general, it applies very well in the representation of quantifiable systems (physical, financial or otherwise). It helps to understand mass balance, and to learn about key parameters, bottle necks, and decision points in the system. Physical delays are also closely monitored in this category of practice.

Warren's approach (2002, 2004), as well as the gist of group model building work (Richardson & Andersen 1995, Vennix 1996), appear to be centered in this modeling component. They are more akin to a hypothesis-generating approach than to a hypothesis-testing approach. But, in our view, they serve to bridge an important gap between system's mapping and hypothesis testing, often stepping in both these other components simultaneously. As stated by Warren (2005):

> Important implications from this perspective include the unavoidable causal ambiguity caused by accumulating resources, and the value of explicit and quantitative examination of resource development *prior to* investigating feedback structures. [Emphasis added]

3. Hypothesis testing

Hypothesis-testing modeling is problem focused, as opposed to system focused. It is parsimonious, as defined by the required elements needed to address the dynamic problem. The system structure is aggregated as much as possible, and detail complexity is avoided. This type of modeling tacitly assumes that there is agreement on the relevant question/issue that needs to be addressed, and sometimes even on how the system is wired. Key to hypothesis testing is a *deductive* procedure that tests if a specific feedback-rich structure is capable of explaining (in this case producing) a particular behavior (Forrester 1961). This is the so-called dynamic hypothesis (Randers 1980-B). Strictly speaking, the quantitative modeling effort does not begin until such a hypothesis is stated. Model simulations are carried out after behavioral expectations are made explicit, and serve as tests of those hypotheses (Sterman 2000).[5]

"Initial" and "broad" system dynamics –as described in Lane (2001)-- focuses upon this category of practice, drawing upon the others to the extent that they may be helpful to build, test and evaluate a model and the insights derived from the work. While using a system's map or a formulated model, the goal/product is not the map or the model, but feedback-rich insights that result from understanding dynamic complexity in the system under study. Many of the well-known works in the field of SD emphasize this component of modeling which is central to system dynamics, positing applied contributions (e.g., Ford 1990, Homer 1992, Repenning 2001), and educating on its philosophy and method (e.g., Richardson & Pugh 1981, Sterman 2000).

[5] Of course, the hypotheses evolve as things become clearer during the process of model building and testing. It can be said that the whole process is geared towards formulating a more educated hypothesis, akin to producing a more relevant problem statement or asking a better research question.

Ariza & Graham (2002) go further and discuss every aspect of the modeling process a sequence of hypothesis testing procedures, including model conceptualization, and certainly model building, testing and calibration. This is why we chose the more inclusive label of hypothesis testing for this component, as opposed to the more commonly used, *dynamic*-hypothesis testing.

4. Uncertainty analysis

In system dynamics, uncertainty analysis is often referred to as sensitivity analysis. Normally, because of the resilience of complex feedback-rich systems, *behavior* sensitivity is assumed to happen rather infrequently. However, it is necessary to build confidence in feedback-rich, problem-focused insights, by demonstrating that the behaviors of the variables of interest do not change significantly if parameters are varied within reasonable ranges, or even if marginal and justifiable changes in model boundary are made.

Other forms of sensitivity are *quantitative* and *policy* sensitivity. In system dynamics, quantitative sensitivity is rarely of concern, unless prediction or forecasting is involved, as described in the next category. In general, system dynamics is advocated for its explanatory power and as a learning instrument, rather than a predictive or forecasting tool. Still, tests for assessment of quantitative sensitivity do have the potential to identify areas for improvement in the modeling work, whether in terms of parameterization, level of aggregation or model boundary. More often discrepancies are explained away as due to model boundary decisions that treated variables exogenously, disregarded seasonality in the behaviors, or assumed away stochastic components in the system. Of course, these explanations are ultimately dependent upon problem definition and insights and recommendations. Policy sensitivity is when recommendations for system's improvement do not always hold, given the range in which parameters may vary.

Tank-Nielsen (1980) provides an overview of the objectives in sensitivity analysis, types of model changes, and interpretation of model responses. Clemson et al. (1995) discuss efficient methods for sensitivity analysis. Moxnes (2005) explores the sensitivity of policy recommendations to uncertain assumptions in fishery models.

5. Forecasting and optimization

This last category has to do with predicting future patterns of behavior, changes in those patterns, and event prediction. Also, it includes research questions that are aimed at finding optimal or robust solutions.[6] As already stated, this falls outside of the main concern in system dynamics with learning, understanding, and explaining. Nevertheless, this is an area of practice that is widely used in other modeling approaches, and it has been applied using system dynamics as well (Coyle 1985, Moxnes et al. 2001, Graham & Ariza 2003). State of the art software offer capability for both sensitivity analysis and optimization (Eberlein & Peterson 1992). Forecasting

[6] Recently, practitioners focusing upon this component of modeling are looking to find "robust" solutions/policies rather than "optimal" ones. Robust solutions are preferred when the time horizon is particularly long, when situations may change significantly over time, and flexibility is needed to adjust to a changing environment, or where there is deep uncertainty. (citations needed)

is just an extension where parameters are changed within reasonable ranges, in combination with implicit or explicit definitions of utility functions (e.g. the variable of interest that is forecasted), to either examine the possible results for a particular variable in a particular time, or to attempt to have this variable seek a particular goal. Quantitative prediction is rarely an objective in system dynamics practice, but there is nothing really to prevent it from being done.

Tests clustered into the five proposed components of modeling

As a starting point, we aimed to classify the traditional tests discussed by Forrester & Senge and Sterman into one (and only one) of the components outlined above, as if there were some sort of orderly process to move from one component to another. This limitation is addressed in the discussion. We also refrained from describing the tests in detail.[7] Table 2 serves as a summary display of the results. The brief descriptions below were extracted from Sterman (2000), Table 21-4, pp. 859-861. Several of Sterman's original tests were broken down into components to accommodate to the framework.

SYSTEM'S MAPPING

1. Face validity (structural assessment through *deductive* process) – Q: Is the model structure consistent with relevant descriptive knowledge of the system?
2. Validity of decision rules (*structural* focus) – Q: Do the decision rules capture the behavior of the actors in the system?

QUANTITATIVE MODELING

3. Physical conservation – Q: Does the model conform to basic physical laws such as conservation laws?
4. Dimensional consistency – Q: Is each equation dimensionally consistent without the use of parameters having no real world meaning?
5. Integration error – Q: Are the results sensitive to the choice of time step or numerical integration method?
6. Extreme conditions tests (*equations* focus) – Q: Does each equation make sense even when its inputs take on extreme conditions?
7. Parameter assessment – Q: Do all parameters have real world counterparts? Are they consistent with relevant descriptive and numerical knowledge of the system?
8. Basic-behaviors reproduction – Q: Does the model generate the various modes of behavior observed in the system?
9. *Endogenous* behavior-reproduction tests – Q: Does the model pass behavioral reproduction tests without the aid of exogenous inputs driving the model in predetermined ways?
10. Boundary adequacy tests (*modes of behavior*) – Q: Does the behavior of the model change significantly when boundary assumptions are relaxed?

[7] More information is available in both Sterman, pp. 861-889, and Forrester & Senge (1980), pp. 212-226.

HYPOTHESIS TESTING

11. Qualitative *problem-behavior* test – Q: Does the model qualitatively reproduce the behavior(s) of interest in the system?
12. Boundary adequacy test (*problem endogeneity*) – Q: Are the important concepts for addressing the problem endogenous to the model?
13. Validity of decision rules (*policy* focus) – Q: Do the decision rules capture the behaviors of the actors in the system? (policy focus)
14. Assessment of surprise behaviors – Inspection for unusual, novel, unexpected or surprise behaviors. Q: Does the model generate previously unobserved or unrecognized behavior? Does the model successfully anticipate the response of the system to novel conditions?
15. *Behavior* sensitivity analysis – Q: Do the modes of behavior generated by the model change significantly when assumptions about parameters, boundary, and aggregation are varied over the plausible range of uncertainty?
16. Extreme conditions tests (*model behaviors* focus) – Q: Does the model respond plausibly when subjected to extreme policies, shocks, and parameters?
17. Behavior anomaly tests (changed *assumptions* tests) – Q: Do anomalous behaviors result when assumptions of the model are changed or deleted?
18. Family member (generalizability) – Ability to generalize. Q: Can the model generate the behavior observed in other instances of the same system?

UNCERTAINTY ANALYSIS

19. *Quantitative* sensitivity analysis – Q: Do the numerical values change significantly when assumptions about parameters, boundary, and aggregation are varied over the plausible range of uncertainty?
20. *Policy* sensitivity analysis – Q: Do the policy implications change significantly when assumptions about parameters are varied over the plausible range of uncertainty? Is the level of aggregation appropriate?
21. Boundary adequacy tests (*policy implications*) – Q: Do the policy recommendations change when the model boundary is extended?

FORECASTING AND OPTIMIZATION

22. Behavior correspondence – Q: Does the model quantitatively reproduce the behavior(s) of interest in the system?
23. Behavior prediction – Pattern prediction, event prediction, shifting-mode prediction
24. Changed-behavior prediction (prior to worry about number forecast; behavioral forecast)

We deliberately left out, for now, tests of system improvement. Some novel tests or approaches are listed in the discussion. The above list and Table 2 are not intended as final, but rather as a rough cut at implementing our proposed classification scheme.

Table 2. Twenty-four tests clustered into five components of modeling

System's mapping	Quantitative modeling	Hypothesis testing	Uncertainty analysis	Forecasting & optimization
S #2a -- F&S Str #1a	1 - Face validity (structural assessment through *deductive* process)			
S #2b -- F&S Str #1b	2 - Validity of decision rules (*structural* focus)			
	S #2c -- F&S Str #1c	3 - Physical conservation		
	S #3 -- F&S Str #5	4 - Dimensional consistency		
	S #6	5 - Integration error		
	S #5a -- F&S Str #3	6 - Extreme conditions tests (*equations* focus)		
	S #4 -- F&S Str #2	7 - Parameter assessment		
	S #7a -- F&S Beh #1a	8 - Basic-behaviors reproduction		
	S #7 b-- F&S Beh #1b	9 - *Endogenous* behavior-reproduction tests		
	S #1a -- F&S Beh #7	10 - Boundary adequacy tests (*modes of behavior*)		
		S #7c -- F&S Beh #1c	11 - Qualitative *problem-behavior* test	
		S #1b -- F&S Str #4	12 - Boundary adequacy (*problem endogeneity*)	
		S #2d -- F&S Str #1d	13 - Validity of decision rules (*policy* focus)	
		S #10 -- F&S Beh #5	14 - Assessment of surprise behaviors	
		S #11a -- F&S Beh #8	15 - *Behavior* sensitivity analysis	
		S #5b -- F&S Beh #6	16 - Extreme condition tests (*model behaviors* focus)	
		S #8 -- F&S Beh #3	17 - Behavior anomaly tests (changed *assumptions* tests)	
		S #9 -- F&S Beh #4	18 - Family member (generalizability)	
		Quantitative sensitivity analysis - 19	S #11b -- F&S Beh #8	
		Policy sensititivity analysis - 20	S #s 1+11c -- F&S Pol #4	
		Boundary adequacy (*policy implications*) - 21	S #1c -- F&S Pol #3	
			Behavior correspondence - 22	S #7d -- F&S Beh #1d
			Behavior prediction - 23	F&S Beh #2
			Changed-behavior prediction - 24	F&S Pol #2
System's mapping	Quantitative modeling	Hypothesis testing	Uncertainty analysis	Forecasting & optimization

Test categories:

Basic
Intermediate
Advanced

S - Sterman (2000); F&S - Forrester and Senge (1980); Str - Structure; Beh - Behavior; Pol - Policy implications

Discussion

The proposed approach to assessment of dynamic models is aimed at guiding refinement efforts, and revealing confidence levels across areas of practice, possibly delineating a trajectory of confidence throughout model development. While we think this attempt to add structure and guidance to model testing and evaluation can be helpful and useful, it needs some refinement.

Preliminary reviews of our framework were favorable with respect to our purpose. The specific components of modeling used did not resonate well to all audiences. Specifically, the quantitative modeling and forecasting/optimization categories of practice appear to be particularly troublesome to modelers who embrace a traditionalist view of SD. The reviewers questioned the form we chose to cluster the tests, each in a single category, and proposed changes discussed below. They also suggested that we go beyond the two sources studied, and include novel tests and approaches. We address the latter two points below.

Do the components exist in isolation?

Our initial attempt to cluster the tests into the components may have produced an artificial result, in which it might be interpreted that the components exist in isolation. While this may be true, more often than not a modeling effort or project will encompass more than one component. This begs the question whether we discuss (1) how one might move from one component to another, with many possibilities for beginning and end points, and steps along the way, or (2) revisiting the results to portray the relevance of each test vis-à-vis the components.

In the former, for example, we could describe a modeling effort driven by a dynamic hypothesis, that moves into the mapping, quantitative, and uncertainty analysis components in a particular order or via iterations in the modeling process, but is centered in the hypothesis testing component, as described by Sterman (2000), Figures 3-1 and 3-2 (pages 87-105), or Richardson and Pugh (1980), Figure 1.11 (pages 15-17), among others. Alternative approaches might deemphasize hypothesis testing, focusing primarily upon the system's mapping component (Powell & Coyle 2002), or the quantitative modeling component (Richardson et al. 2004), to provide a couple of examples.

An alternative path might be to attempt to create a matrix of tests and components that shows the relative relevance of each test for each component. This solution would have to propose how important and useful each of the tests is for each component. Thus, if the overall purpose of the modeling effort is forecasting, for example, how important is testing for face validity, compared with its relevance in a modeling effort that is primarily devoted to mapping a system and showing interdependencies amongst key stakeholders.

From our organizational perspective and need, because most of our projects follow a particular path, the first alternative seems more attractive. However, for the breadth of practice in the field, the second alternative may prove more useful. We welcome other modelers' feedback and reactions to our framework, preliminary clustering, and alternative future paths, as you think about how helpful this framework could be in your daily practice.

Adding novel tests and approaches

A number of new tests not included in the early literature (Forrester & Senge 1980), and texts (Sterman 2000) exist, and would be worth including in future iterations of this framework. The list below is illustrative but not comprehensive. We try to provide at least one example of new development for each of the components:

- Soft, qualitative deductive procedures used to validate system's maps (e.g. Vennix 1990, Luna-Reyes & Andersen 2003)

Group model building has additional issues with regard to testing and confidence. For some models correctness can be defined quantitatively, but others correctness might mean including all viewpoints of stakeholders.

- Reality checks (e.g. Peterson & Eberlein 1994)

Vensim® has built-in features that automate checking model conformance against statements of "truth." We think that embedding these solidly held beliefs about the nature of reality in the model is particularly helpful to validate the quantitative model.

- Automated procedures to corroborate or refute causal stories linking structure to behavior (e.g. Mojtahedzadeh et al. 2004, Oliva 2004)

Using pathway participation metric (PPM), *Digest* detects and displays which feedback loops are most influential in explaining patterns of behavior in a model. This and other promising approaches (based upon Eigenvalue analysis) can help test a modeler's experiential understanding of the link between model structure and behaviors.

- New methods designed to conduct sensitivity analysis (e.g. Ford & Flynn 2005)

In the latest issue of the *System Dynamics Review*, a statistical screening approach is proposed to learn which of the many uncertain inputs to a model stand out as most influential.

- New approaches to model calibration (e.g. Oliva 2003)

Model structure analysis through graph theory uses partition heuristics and feedback structure decomposition to increase model confidence through careful calibration. This development may be viewed as a stepping stone toward testing dynamic hypotheses. But, in and of it self, it delineates a path toward automation of the calibration process, which is crucial to forecasting and optimization.

Epilogue

We believe this framework to be theoretically sound and practically useful to guide model testing in different phases of model development, and to conduct partial assessments of levels of confidence. It can help negotiate project deliverables, reconciling client and modeler

expectations regarding level of effort involved to achieve a certain product, depending upon where the model stands. It also creates a terminology to differentiate modeling products. However, we've only taken a first cut at reviewing and organizing the whole spectrum of model testing and evaluation procedures. Future steps will include assessing the practical usefulness of adopting this framework in our daily work, as well as thinking about how to move forward. Your reactions and suggestions will help us refine and expand the proposed framework.

Acknowledgements

The authors acknowledge David Andersen, Steve Conrad, Sharon Deland, Andy Ford, Mohammad Mojtahedzadeh, George Richardson, and Silvia Ulli-Beer for their contributions to this research effort.

References

Ackermann F, S Howick and DF Andersen. 2004. Stirling revisited: Practical approaches to merging two systems thinking streams. Proceedings of the 22nd International Conference of the System Dynamics Society. Oxford, England (July 25-29).

Andersen DF and GP Richardson. 1997. Scripts for group model building. *System Dynamics Review* 13(2): 107-129.

Ansoff HI and DP Slevin. 1968. An appreciation of industrial dynamics. *Management Science* 14: 383-397.

Ariza CA and AK Graham. 2002. Quick and rigorous, strategic and participative: 12 ways to improve on the expected tradeoffs. *Proceedings of the 20th International Conference of the System Dynamics Society*. Palermo, Italy (July 28-August 1).

Babbie E. 1992. *The Practice of Social Research*. Belmond, CA: Wadsworth Publishing Company.

Bakken B, J Gould and D Kim. 1992. Experimentation in learning organizations: A management flight simulator approach. *European Journal of Operational Research* 59(1): 167-182.

Barlas Y. 1989. Multiple tests for validation of system dynamics type of simulation models. *European Journal of Operations Research* 42(1): 59-87.

Barlas Y. 1996. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review* 12 (3): 183-210.

Barlas Y and S Carpenter. 1990. Philosophical roots of model validation: Two paradigms. *System Dynamics Review* 6(2): 148-166.

Burrell G and G Morgan. 1979. *Sociological Paradigms and Organizational Analysis: Elements of the Sociology of Corporate Life*. Gower: Aldershot.

Checkland P. 1981. *Systems Thinking, Systems Practice*. Chichester, England: John Wiley & Sons.

Clemson B., Y Tang, J Pyne and R Unal. 1995. Efficient methods for sensitivity analysis. *System Dynamics Review* 11(1): 31-49.

Coyle RG. 1985. The use of optimization methods for policy design in a system dynamics model. *System Dynamics Review* 1 (1): 81-91.

Coyle RG. 2000. Qualitative and quantitative modeling in system dynamics: Some research questions. *System Dynamics Review* 16(3): 225-244.

Coyle RG. 2001. Rejoinder to Homer and Oliva. *System Dynamics Review* 17(4): 357-363.

Eberlein RL and DW Peterson. 1992. Understanding models with Vensim™. *European Journal of Operations Research* 59(1): 216-219.

Eden C. 1994. Cognitive mapping and problem structuring for system dynamics model building. *System Dynamics Review* 10(2-3): 257-276.

Ford A. 1990. Estimating the impact of efficiency standards on uncertainty of the northwest electrical system. *Operations Research* (July-August)

Ford A. 1999. *Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems*. Washington, DC: Island Press.

Ford A and H Flynn. 2005. Statistical screening of system dynamics models. *System Dynamics Review* 21(4): 273-303.

Forrester JW. 1961. *Industrial Dynamics*. Portland, Oregon: Productivity Press.

Forrester JW. 1971. *World Dynamics*. Waltham, MA: Pegasus Communications.

Forrester JW and PM Senge. 1980. Tests for building confidence in system dynamics models. In AA Legasto Jr, JW Forrester and JM Lyneis (eds.). *System Dynamics. TIMS Studies in the Management Sciences* 14. New York: North-Holland: 209-228.

Graham A. 2002. On positioning system dynamics as an applied science of strategy. Or: SD is scientific. We haven't said so explicitly, and we should. *Proceedings of the 20th International Conference of the System Dynamics Society*. Palermo, Italy (July 28-August 1).

Graham AK and CA Ariza. 2003. Dynamic, hard and strategic questions: Using optimization to answer a marketing resource allocation question. *System Dynamics Review* 19(1): 27-46.

Hodgson AM. 1994. Hexagons for systems thinking. In JDW Morecroft and JD Sterman (eds.). *Modeling for Learning Organizations*. Portland, OR: Productivity Press: 359-374.

Homer JB. 1992. A system dynamics model of national cocaine prevalence. *System Dynamics Review* 9(1): 49-78.

Homer J and R Oliva. 2001. Maps and models in system dynamics: A response to Coyle. *System Dynamics Review* 17(4): 347-355.

Kleinmuntz DN. 1993. Information processing and misperceptions of the implications of feedback in dynamic decision making. *System Dynamics Review* 9(3): 223-237.

Kuhn T. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lane 1993. The road not taken: observing a process of issue selection and model conceptualization. *System Dynamics Review* 9 (3): 239-264.

Lane DC. 2001. Rerum cognoscere causas: Part I – How do the ideas of system dynamics relate to traditional social theories and the voluntarism/determinism debate? *System Dynamics Review* 17(2): 97-118.

Legasto Jr AA, JW Forrester and JM Lyneis (eds.). 1980. *System Dynamics*. *TIMS Studies in the Management Sciences* 14. New York: North-Holland.

Luna-Reyes L and DL Andersen. 2003. Collecting and analyzing qualitative data for system dynamics: methods and models. *System Dynamics Review* 19(4): 271-296.

Mass NJ and PM Senge. 1980. Alternative tests for selecting model variables. In J Randers (ed.). 1980. *Elements of the System Dynamics Method*. Cambridge, MA: Productivity Press: 205-225.

Mojtahedzadeh M, D Andersen and GP Richardson. 2004. Using Digest to implement the pathway participation method for detecting influential system structure. *System Dynamics Review* 20(1): 1-20.

Morecroft J. 1982. A critical review of diagramming tools for conceptualizing feedback system models. *Dynamica* 8(1): 20-29.

Moxnes E. 2005. Policy sensitivity analysis: Simple versus complex fishery models. *System Dynamics Review* 21(2): 123-145.

Moxnes E, O Danell, E Gaare and J Kumpula. 2001. Optimal strategies for the use of reindeer rangelands. *Ecological Modelling* 145(2-3): 225-241.

Nordhaus B. 1973. World dynamics: Measurement without data. *Economic Journal* (December).

Oliva R. 2003. Model calibration as a testing strategy for system dynamics models. *European Journal of Operational Research* 151(3): 525-568.

Oliva R. 2004. Model structure analysis through graph theory: Partition heuristics and feedback structure decomposition. *System Dynamics Review* 20(4): 313-336.

Peterson DW. 1980. Statistical tools for system dynamics. In J Randers (ed.). 1980. *Elements of the System Dynamics Method*. Cambridge, MA: Productivity Press: 226-242.

Peterson DW and RL Eberlein. 1994. Reality check: A bridge between systems thinking and system dynamics. *System Dynamics Review* 10(2-3): 159-174.

Powell JH and RG Coyle. 2002. Setting strategic agendas: The use of qualitative methods in highly politicized contexts. *Proceedings of the 20th International Conference of the System Dynamics Society*. Palermo, Italy (July 28-August 1).

Randers J (Ed.). 1980-A. *Elements of the System Dynamics Method*. Cambridge, MA: Productivity Press.

Randers J. 1980-B. Guidelines for model conceptualization. In J Randers (ed.). *Elements of the System Dynamics Method*. Cambridge, MA: Productivity Press: 117-139.

Repenning NP. 2001. Understanding fire fighting in new product development. *Journal of Product Innovation Management* 18(5): 285-300.

Richardson GP. 1997. Problems in causal loop diagrams revisited. *System Dynamics Review* 13(3): 247-252.

Richardson GP and DF Andersen. 1995. Teamwork in group model building. *System Dynamics Review* 11(2): 113-137.

Richardson GP, DF Andersen and L Luna-Reyes. 2004. Joining minds: Group modeling to link people, process, analysis and policy design. 26th Annual Association for Public Policy Analysis and Management Research Conference. Atlanta, GA (October 28-30).

Richardson GP and AL Pugh III. 1981. *Introduction to System Dynamics Modeling with DYNAMO*. Cambridge, MA: Productivity Press.

Richardson GP, EF Wolstenholme and JDW Morecroft (eds.). 1994. *Systems Thinkers, Systems Thinking*. *System Dynamics Review* 10(2-3).

Richmond B, S Peterson and P Vescuso. 1987. *An Academic User's Guide to STELLA*. Lyme, NH: High Performance Systems, Inc.

Sellers W. 1963. Empiricism and the philosophy of mind. In *Science, Perception and Reality*. New York: Humanities Press.

Sterman JD. 1989. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science* 35(3): 321-339.

Sterman JD. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston, MA: The McGraw-Hill Companies.

Tank-Nielsen C. 1980. Sensitivity analysis in system dynamics. In J Randers (ed.). 1980. *Elements of the System Dynamics Method*. Cambridge, MA: Productivity Press: 226-242.

Vennix JAM. 1996. *Group Model Building: Facilitating Team Learning Using System Dynamics*. London: John Wiley & Sons.

Vennix JAM, JW Gubbels, D Post and HJ Poppen. 1990. A structured approach to knowledge elicitation in conceptual model building. *System Dynamics Review* 6(2): 194-208.

Wakeland W and M Hoarfrost. 2005. The case for thoroughly testing complex system dynamics models. *Proceedings of the 23rd International Conference of the System Dynamics Society*. Boston (July 17-21).

Warren K. 2002. *Competitive Strategy Dynamics*. United Kingdom: Wiley.

Warren K. 2004. Why has feedback systems thinking struggled to influence strategy and policy formulation? *Systems Research and Behavioral Science* 21(4): 331-347.

Warren K. 2005. Improving strategic management with the fundamental principles of system dynamics. *System Dynamics Review* 21(4): 329-350.

Wolstenholme EF. 1994. A Systematic approach to model creation. In JDW Morecroft and JD Sterman (eds.). *Modeling for Learning Organizations*. Portland, Oregon: Productivity Press: 175-194.

Zagonel AA. 2002. Model conceptualization in Group Model Building: A review of the literature exploring the tension between representing reality and negotiating a social order. *Proceedings of the 20th International Conference of the System Dynamics Society*. Palermo, Italy (July 28-August 1).