Model validation constitutes a very important step in system dynamics methodology. Yet, both published and informal evidence indicates that there has been little effort in system dynamics community explicitly devoted to model validity and validation. Validation is a prolonged and complicated process, involving both formal/quantitative tools and informal/qualitative ones. This paper focuses on the formal aspects of validation and presents a taxonomy of various aspects and steps of formal model validation. First, there is a very brief discussion of the philosophical issues involved in model validation, followed by a flowchart that describes the logical sequence in which various validation activities must be carried out. The crucial nature of structure validity in system dynamics (causal-descriptive) models is emphasized. Then examples are given of specific validity tests used in each of the three major stages of model validation: Structural tests.

Introduction

Model validation is an important, yet controversial aspect of any model-based methodology in general, and system dynamics in particular. Validity of the results in a model-based study are crucially dependent on the validity of the model. Yet, there is no single established definition of model validity and validation in the modeling literature. In particular, system dynamics has often been criticized for relying too much on informal, subjective and qualitative validation procedures (Ansoff and Slevin 1968; Nordhaus 1973; Zellner 1980). Although system dynamicists have responded to such criticisms (Forrester 1968; Forrester et al. 1974; Forrester and Senge 1980; Sterman 1984; Barlas 1989a; Barlas and Carpenter 1990), a survey of the articles published in the past 11 years indicates that there has been little effort in the system dynamics community devoted to model validity and validation: Only three of all the articles published in System Dynamics Review (between 1985 and 1995) deal with model validity/validation, which is a very low number compared to the number of papers on other methodological subjects and on applications. (Richardson 1991 provides statistics on published articles, by subject category, between 1985 and 1990.) There is no clear evidence of consistent and widespread use of even the basic established validity tools (see Peterson and Eberlein 1994 and Scholl 1995). System dynamics needs active research (not mere reactions to criticisms) to develop formal concepts and rigorous methods of model validation suitable to its nature. This article hopes to spark research in this direction. Building on the author’s past research and experience in model validation and validation, the article seeks a synthesis about the formal aspects of model validity and validation: a set of propositions as to the definition of model validity, its different aspects, the formal logic of model validation, and appropriate methods and tools.

The purpose of this article is to discuss the formal aspects of model validity and validation in system dynamics. But it is important to note that this focus of the article does not imply that model validity and validation can be cast as entirely formal/objective constructs and procedures. Model validity and validation in any discipline have to have semi-formal and subjective components for several reasons often discussed in system dynamics literature (Forrester 1961; ch. 13; Forrester 1968; Forrester and Senge 1980; Andersen 1980; Meadows 1980; Richardson and Pugh 1981, ch. 5; Sterman 1984; Barlas & Carpenter 1990; Lane 1995). An important reason has to do with the relation-
structure-oriented behavior tests and behavior pattern tests. Also discussed is if and to what extent statistical significance tests can be used in model validation. Among the three validation stages, the special importance of structure-oriented behavior tests is emphasized. These are strong behavior tests that can provide information on potential structure flaws. Since structure-oriented behavior tests combine the strength of structural orientation with the advantage of being quantifiable, they seem to be the most promising direction for research on model validation.

Yaman Barlas received his PhD degree in Industrial and Systems Engineering from Georgia Institute of Technology in 1985, and joined Miami University of Ohio. He returned to Boğaziçi University in 1993, as an associate professor of Industrial Engineering, where he is directing the system dynamics group. His interest
areas are validation of simulation models, system dynamics methodology, modeling/analysis of socio-economic problems and simulation as a learning/training platform. He is a founding member of System Dynamics Society and member of SCS International and INFORMS.

Address: Dept. of Industrial Engineering, Bogazici University, 80815 Bebek, Istanbul, Turkey; email: ybarlas@boun.edu.tr

Aspects of model validation that can be reasonably separated from the rest of the modeling activities, it becomes possible to set some standards and structures that guide the practice of model validation. It also becomes possible to carry out scientific research on model validation and potentially develop improved tools and procedures. Thus, on the one hand we acknowledge that some degree of validation takes place in every stage of modeling and that validation cannot be entirely formal and objective, and on the other hand, to make a rigorous discussion possible, we focus on those validation activities that can be separated from the rest. The purpose of the article is to discuss these “formal” aspects of model validity and model validation.

Types of models and model validity

Models can be classified in many different ways, according to different criteria, such as physical vs symbolic; dynamic vs static; deterministic vs stochastic, etc. As it relates to the notion of validity, a crucial distinction must be made between models that are “causal-descriptive” (theory-like, “white-box”) and models that are purely “correlational” (purely data-driven, “black-box”). In purely correlational (black-box) models, since there is no claim of causality in structure, what matters is the aggregate output behavior of the model; the model is assessed to be valid if its output matches the “real” output within some specified range of accuracy, without any questioning of the validity of the individual relationships that exist in the model. This type of “output” validation can often be cast as a classical statistical testing problem. Models that are built primarily for forecasting purposes (such as time-series or regression models) belong to this category. On the other hand, causal-descriptive (white-box) models are statements as to how real systems actually operate in some aspects. In this case, generating an “accurate” output behavior is not sufficient for model validity; what is crucial is the validity of the internal structure of the model. A white-box model, being a “theory” about the real system, must not only reproduce/predict its behavior, but also explain how the behavior is

Table 1. Major steps used in a typical system dynamics study

1. Problem identification
2. Model conceptualization (construction of a conceptual model)
3. Model formulation (construction of a formal model)
4. Model analysis and validation
5. Policy analysis and design
6. Implementation
generated, and possibly suggest ways of changing the existing behavior. System dynamics models—and all models that are design-oriented in general—fall in this category. Such models are built to assess the effectiveness of alternative policies or design strategies on improving the behavior of a given system. This latter is only possible, if the model has an internal structure that adequately represents those aspects of the system which are relevant to the problem behavior in hand. In short, it is often said that a system dynamics model must generate the “right output behavior for the right reasons.”

Validation of a system dynamics model is much more complicated than that of a black-box model, because judging the validity of the internal structure of a model is very problematic, both philosophically and technically. It is philosophically difficult, because, as we shall briefly review in the next section, the problem is directly related to the unresolved philosophical issue of verifying the truth of a (scientific) statement. And the problem is technically difficult, because there are no established formal tests (such as statistical hypothesis tests) that one can use in deciding if the structure of a given model is close enough to the “real” structure. Furthermore, standard statistical tests can not even be used in validating the behavior of a system dynamics model, because of problems of autocorrelations and multicollinearity. (This topic will be discussed further in the section Role of statistical significance testing.) A review of model validation literature in general shows that a large majority of technical research deals only with what we call behavior validation. (System dynamics literature does offer a few articles that discuss the structural aspects of model validity. See for instance, Forrester 1961, ch. 13; Forrester and Senge 1980; Richardson and Pugh 1981, ch. 5; Barlas 1989b; Peterson and Eberlein 1994. Also, an exception is the special issue of European Journal of Operational Research (EJOR), 1993). As a matter of fact, even in the conceptual articles that discuss the nature of validity, structure validity is at best barely mentioned. And most tools that are suggested for structure validation are informal and qualitative in nature (such as expert reviews, inspections, walkthroughs, data flow and control flow analyses, consistency checking, etc.; see Balci 1994). There may be two main reasons why structure validity has been ignored for so long in modeling literature. The first one stems from a lack of recognition of the philosophical importance of structure validity in white-box modeling (as opposed to black-box modeling). The second reason has to do with the technical difficulty of designing formal/statistical tools that address structure validity. In what follows, we first provide a brief overview of the philosophical aspect of the problem, then discuss the formal logic of model validation and suggest some specific tests that are appropriate.
Philosophical aspects of model validity

In some fundamental ways, the issue of validating causal-descriptive (e.g. system dynamics) models has strong ties with philosophy of science issues. A system dynamics model is refuted if a critic can show that a relationship in the model conflicts with an established “real relationship”, even if the output behavior of the model matches the observed system behavior. For such models, validity primarily means validity of the internal structure of the model, not its output behavior (“right behavior for the right reason”). It can be said that a valid system dynamics model embodies a theory about how a system actually works in some respect. Therefore, there has to be a strong connection between how theories are justified in the sciences (a major and unresolved philosophy of science question) and how such models are validated. This means that our conception of model validity depends on our philosophy (implicit or explicit) of how knowledge is obtained and confirmed. This connection between various philosophies and model validity is discussed by several authors in modeling literature (most significantly, Naylor and Finger 1968; Mitroff 1969; Bell and Bell 1980; Forrester and Senge 1980; Barlas & Carpenter 1990; Carson and Flood 1990; the special issue of EJOR 1993). These articles discuss the two main epistemologies, namely empiricism and rationalism, and the major philosophies of science, namely verificationism, logical positivism, falsificationism, instrumentalism, and different versions of relativism (such as Kuhnian “moderate” relativism and Feyerabend’s extreme relativism). Discussion of these philosophies is beyond the scope of this article. Barlas and Carpenter (1990) summarize and compare these philosophies and conclude that, for our purpose, it is possible to classify them into two opposing philosophical camps. The traditional reductionist/logical positivist school (including empiricism, rationalism, verificationism and the “strong” falsificationism) would see a valid model as an objective representation of a real system. The model can be either “correct” or “incorrect”; once the model confronts the empirical facts, its truth or falsehood would be automatically revealed. In this philosophy, validity is seen as a matter of accuracy, rather than usefulness (Barlas and Carpenter 1990). The opposing school (including more recent relativistic, holistic and pragmatist philosophies), in contrast, would see a valid model as one of many possible ways of describing a real situation. “No particular representation is superior to others in any absolute sense, although one could prove to be more effective. No model can claim absolute objectivity, for every model carries in it the modeler's worldview. Models are not true or false, but lie on a continuum of usefulness.” (Barlas and Carpenter 1990). The authors, by citing major system dynamics articles dealing with validity (Forrester 1961, ch. 13; Forrester 1968; Forrester and
Senge 1980; Sterman 1984), show that the philosophy of system dynamics is in agreement with the relativist/holistic philosophy of science. Accordingly, model validity is not absolute and validation cannot be entirely objective and formal. Since validity means "adequacy with respect to a purpose", model validation has to have informal, subjective and qualitative components. Model validation is a gradual process of "confidence building", rather than a binary "accept/reject" division; see also Radzicki (1990) for a discussion of a similar philosophical division that exists among the different schools of economics.

This brief philosophical overview shows that

1. validity of a system dynamics model primarily means validity of its internal structure and
2. the recent relativist/holistic philosophy argues that validation of the internal structure cannot be made entirely objective, formal and quantitative (in the sense that even scientific theory confirmation has informal and subjective aspects).

At this point it is crucial to note that relativist/holistic philosophy does not reject the role of formal/quantitative tests in model validation. On the contrary, since this philosophy claims that validity is gradually established as a result of a "controversial" (rather than confrontational) process, collecting, organizing, interpreting and effectively communicating information on model validity would play a major role in the process. Formal/quantitative tests provide crucial inputs to the larger validation process, which is gradual, semi-formal and conversational. The challenge is to design formal/quantitative validation procedures and tests suitable for system dynamics models, while keeping the above philosophical perspective.

**Aspects of formal model validation**

As discussed above, the ultimate objective of system dynamics model validation is to establish the validity of the *structure* of the model. Accuracy of the model behavior's reproduction of real behavior is also evaluated, but this is meaningful only if we already have sufficient confidence in the structure of the model. Thus, the general logical order of validation is, first to test the validity of the structure, and then start testing the behavior accuracy, only after the structure of the model is perceived adequate. This logical sequence is depicted in Figure 1.
Fig. 1. Overall nature and selected tests of formal model validation

**Direct structure tests**
Observe in Figure 1 that we distinguish between two types of structure tests: Direct structure tests and Structure-oriented behavior tests. Direct structure tests assess the validity of the model structure, by direct comparison with knowledge about real system structure. This involves taking each relationship
(mathematical equation or any form of logical relationship) individually and comparing it with available knowledge about the real system. There is no simulation involved. Forrester and Senge (1980) give examples of direct structure tests, such as, structure and parameter confirmation tests, direct extreme-conditions test and dimensional consistency test. Direct structure tests, as shown in Figure 1, can be classified as empirical or theoretical. Empirical structure tests involve comparing the model structure with information (quantitative or qualitative) obtained directly from the real system being modeled. Theoretical structure tests involve comparing the model structure with generalized knowledge about the system that exists in the literature.

The structure confirmation test, applied as an empirical one, means comparing the form of the equations of the model with the relationships that exist in the real system (Forrester and Senge 1980). It may also be carried out as a theoretical structure test, by comparing the model equations with generalized knowledge in the literature. Structure confirmation tests are perhaps most difficult to formalize and quantify, as they attempt to compare the form of the equations of the model, directly with the form of the relationships that exist in the real system. The information needed for this type of comparison is highly qualitative in nature; it cannot be captured simply by a set of numerical data. There are nevertheless some semi-formal tools that can be used to carry out direct structure confirmation tests. Examples are formal inspections, reviews, walkthroughs, and data flow analysis, typically used in the verification and validation of computer models (see Balci 1994). The second direct structure test, parameter confirmation means evaluating the constant parameters against knowledge of the real system, both conceptually and numerically (Forrester and Senge 1980). Conceptual confirmation means being able to identify elements in the real system that correspond to the parameters of the model. Numerical confirmation consists of estimating the numerical value of the parameter with enough accuracy. Parameter confirmation test, as seen in Figure 1, may be applied both as an empirical test and as a theoretical one.

Another important direct structure test, direct extreme-condition testing, involves evaluating the validity of model equations under extreme conditions, by assessing the plausibility of the resulting values against the knowledge/anticipation of what would happen under a similar condition in real life (Forrester and Senge 1980). Unlike normal operating conditions, it is relatively easy to anticipate what values the variables of the real system would (asymptotically) take under extreme conditions (that is why this test is classified as a theoretical one in Figure 1). For instance, if the population is 0, then there can be no births, no workers, no consumption, etc. If the pollution level is extremely high, then death rates must rise and migration to the city must decline. Each model equation can thus be tested by assigning
extreme values to its input variables, and comparing the value of the output variable to what would logically happen in the real system under the same extreme condition. (Note that direct extreme-condition testing does not involve dynamic simulation; it is applied to each equation in isolation, statically.) Finally, dimensional consistency test entails checking the right-hand side and left-hand side of each equation for dimensional consistency (it is classified as a theoretical test in the sense that it is an internal consistency test). To be meaningful, the test requires that the model has no dummy “scaling” parameters that have no meaning in real life, i.e. the model has already passed the conceptual parameter-confirmation test (Forrester and Senge 1980).

Structure-oriented behavior tests

The second general category of structural tests, structure-oriented behavior tests, assess the validity of the structure indirectly, by applying certain behavior tests on model-generated behavior patterns (Barlas 1989b; Forrester and Senge). These tests involve simulation, and can be applied to the entire model, as well as to isolated sub-models of it. These are “strong” behavior tests that can help the modeler uncover potential structural flaws. Figure 1 illustrates several such tests: Extreme-condition (indirect) test involves assigning extreme values to selected parameters and comparing the model-generated behavior to the observed (or anticipated) behavior of the real system under the same extreme condition. (This is also called stress testing; see Balci 1994.) Behavior sensitivity test consists of determining those parameters to which the model is highly sensitive, and asking if the real system would exhibit similar high sensitivity to the corresponding parameters. Modified-behavior prediction can be done if it is possible to find data about the behavior of a modified version of the real system. The model passes this test if it can generate similar modified behavior, when simulated with structural modifications that reflect the structure of the “modified” real system (see Forrester and Senge 1980 and Barlas 1989b for more details). Phase-relationship test uses the phase relationships between pairs of variables in the model, obtained as a result of simulations. If certain phase relationships obtained from the model contradict the phase relationships that are observed/expected from the real system, this may indicate a structural flaw in the model (Forrester and Senge 1980; Barlas 1989b).

An interesting implementation of behavior sensitivity and extreme-condition testing is provided by Carson and Flood (1990). The authors present an application of what they call “Qualitative Features Analysis” to a model of fluid/electrolyte balance in the human body. The test consists of specifying
the major qualitative features of the "expected" behavior under specific test conditions, and then comparing them with actual simulation results. Examples of "qualitative features" (behavior patterns) used in the article are: a fall, a rise, a fall followed by a rise, a delayed fall, a delayed rise, and oscillatory. Then, a summary measure is obtained by counting the total number of runs in which the expected and simulated qualitative features match. The higher the number, the higher the structure validity of the model with respect to the test cases. More importantly, each test case, if failed, may help the modeler uncover a structural flaw. Finally, an interesting test that can be used in structure-oriented behavior evaluation is the Turing test. In this test, experts are presented with a shuffled collection of real and simulated output behavior patterns, and asked if they can distinguish between these two types of patterns. If experiments show that the two types of behavior patterns are statistically indistinguishable, then the model passes the particular Turing test (Schruben 1980). However, if experts do detect significant differences, then they are interviewed with the purpose of uncovering the structural flaws in the model that may have caused the differences.

Structure-oriented behavior tests are strong behavior tests that can provide information on potential structural flaws (see Barlas 1989b for illustrations). Their main advantage over direct structure tests is that they are much more suitable to formalize and quantify. Direct structure tests, although powerful in concept, have the disadvantage of being too qualitative and informal by their nature. Since structure-oriented behavior tests combined the strength of structural orientation with the advantage of being quantifiable, they seem to be the most promising direction for research on model validation. A nice implementation of extreme-condition testing is provided by the "Reality Check" feature of VENSIM simulation software (Peterson and Eberlein 1994). Reality Check consists of statements of the form: "if input A is imposed on the system, then behavior B should result." Then, the software performs simulations and tests the conformance of the model to the anticipated behavior. Reality Check is thus a version of the general concept of assertion checking typically used in program verification (Balci 1994). Further research is needed to quantify and formalize such procedures, which are somewhat qualitative and semi-formal at this stage.

Behavior pattern tests

The two categories of tests discussed above are designed to evaluate the validity of the model structure. As a result of these tests, once enough confidence has been built in the validity of the model structure, one can start applying certain tests designed to measure how accurately the model can reproduce the
major behavior patterns exhibited by the real system. It is crucial to note that the emphasis is on *pattern* prediction (periods, frequencies, trends, phase lags, amplitudes, ...), rather than point (event) prediction. This is a logical result of the long-term policy orientation of system dynamics models. Furthermore, since such models, starting with a set of initial conditions, create the dynamic behavior patterns endogenously (not dictated by external input functions), it can be shown that even "perfect" structures may not yield accurate point prediction (Forrester 1961, ch. 13 and Appendix K; Forrester 1968; Wright 1972; Forrester and Senge 1980; Barlas 1985; 1989a). Among the behavior pattern tests are the multiple-test procedure by Barlas (1989a) and modifications of it (Barlas and Erdem 1994), an overall summary statistic proposed by Sterman (1984) and several tests discussed in Forrester and Senge (1980).

Figure 1 summarizes our discussion of the three stages of model validation. Observe that, in Figure 1, all three stages are dependent on "model purposes", which is determined in the problem identification step (the very first step) of system dynamics methodology. Thus, as discussed in the philosophical overview section, no validity test can be carried out in the absolute sense, without reference to the specific purpose of the model. Figure 1 also emphasizes the principle that, to be meaningful, direct structure tests must be carried out first, then structure-oriented behavior tests and then finally behavior pattern tests. Figure 1 should not give the wrong impression that the entire model validation consists of single sequential process, starting with the initial model and ending with behavior tests. In reality, depending on the results of the specific tests, there may be numerous loops and model revisions in the process. This feedback nature and its logic is depicted in Figure 2. The tests are carried out in some logical sequence, and it makes sense to move to the next step, only if we are able to establish sufficient confidence in the current step. Otherwise, we return back to work on necessary model revisions (typically *structural* revisions, not *ad hoc* parameter changes). Once the model has been through all the structural tests, we can start assessing the pattern prediction ability of the model by applying a series of behavior tests. Note the two dashed lines from "Model construction and revisions" to "Structure-oriented behavior tests" and "Behavior pattern tests." These lines denote the possible legitimacy of skipping direct structure and structure-oriented behavior tests, but only when the model revisions are of non-structural nature (made as a result of failing behavior pattern tests). Thus, for instance, if the model passes all the direct-structure and structure-oriented behavior tests but fails the behavior pattern tests, then, after certain parameter and/or input function reestimation, it would be logical to skip the structure tests and apply behavior pattern tests only. Once we have reached the final step of behavior pattern
testing, the emphasis is on the accuracy of pattern predictions, and is essentially done for communication of results and implementation purposes. Behavior pattern tests are "weak" tests that provide no information on the validity of the structure of the model.

Behavior pattern tests must themselves be carried out in some logical order too. Figure 3 suggests a behavior pattern validation procedure. There are two fundamentally different types of behavior patterns that call for two different types of behavior tests. If the problem involves a transient, highly non-stationary behavior (such as a truncated S-shaped growth, or a single boom-then-bust pattern), then it is impossible to apply any standard statistical measure. The problem is of no statistical nature to start with and therefore no general statistical tests can be offered in this case. The best approach is to compare graphical/visual measures of the most typical behavior-pattern characteristics, such as the amplitude of a peak, time between two peaks, minimum value, slope, number of inflection points, time to settle, etc. (Although statistical testing is
not suitable in this case, it may still be possible to design situation-specific formulas that can estimate these—and some other—behavior pattern characteristics.) Carson and Flood (1990) illustrate an application of this type of testing which they call "Quantitative features" validation (see also Forrester & Senge 1980; Barlas 1985). If, on the other hand, the problem involves a long-term steady-state simulation, then, it is possible to apply certain standard statistical measures and tests.

Figure 3 includes the multi-step behavior validation procedure developed by Barlas (1989a, 1990). Formulas and some more detailed information are provided in the appendix. (See also Barlas 1994 for some enhancements in step 4, comparing amplitude variations.) Note that, if a model is judged to fail the behavior pattern tests, we return once again back to work on “model revisions.” But in this case, since confidence in the model structure must have been already established, model revisions involve parameter/input changes, rather than structural revisions.
The role of "statistical significance" testing in model validation is a controversial issue. In general, statistical significance testing consists of advancing a "null hypothesis" and then rejecting it if the discrepancy between what the hypothesis predicts and real data is "statistically significant." This latter means (assumes) that the discrepancy cannot be explained by mere chance. In order to carry out such a test, one must assume a fixed level of discrepancy above which the hypothesis would be rejected, which in turn implies admitting a certain probability of error (type I error, or "significance level"). For instance, a fixed significance level of 0.05 means that one is willing to reject a "true" null hypothesis by mistake 5% of the time. Significance testing is frequently used in validating models in social, behavioral and economic sciences. In system dynamics validation however, there is very little use of statistical significance testing and system dynamics has often been criticized for it. System dynamicists have responded by arguing that statistical significance can contribute very little to model validation in system dynamics. The controversy exists not just between system dynamicists and their critics, but also among social, behavioral and social scientists. The issue is not likely to be easily resolved by logical/rational debate, because statistical significance testing has gradually evolved into being a "norm" in some of the social/behavioral sciences (Morrison and Henkel 1970).

The problems involved in using statistical significance tests in validating system dynamics (and other socio-economic) models are both technical and philosophical. The technical reasons why statistical significance has little relevance in model validation have to do with some fundamental assumptions that must hold for statistical tests to be valid. Most statistical tests assume at least that data are (i) serially independent (not autocorrelated); (ii) not cross-correlated; (iii) normally distributed. The first two of these assumptions are almost never met by a system dynamics model. Data generated by system dynamics models are autocorrelated and cross-correlated by their very nature. Applying statistical tests to auto/cross-correlated data requires extensive model simplification, and/or data transformation, frequently a complex problem in itself, sometimes with no satisfactory solution at all (Senge 1977; Mass and Senge 1980; Sterman 1984; Barlas 1985). There are some other technical difficulties. For instance, it can be shown that the results of statistical significance tests will be ambiguous, even misleading, if data are corrupted with "measurement errors", which is common in socio-economic studies (Wright 1972; Senge 1977; Mass and Senge 1980; Barlas 1985). Finally, in system dynamics models, there is no single "output variable" that one can focus on in validity testing; there are typically many variables of importance for the pur-
pose of the study. Thus, model validation, if cast as statistical hypothesis testing, would constitute a simultaneous multiple-hypothesis problem (Miller 1981). Computing the overall effective significance level of this type of test is a very difficult procedure, because the variables are cross-correlated in system dynamics models. (See Anderson 1971, section 6.7 for the m-dimensional joint distribution estimation problem.) Although there are some simpler approximate methods (Miller 1981), the problem of knowing/estimating the significance level of simultaneous multiple-hypothesis tests does not have a solution general enough to be used in an automated validation procedure. In summary, there are many technical/statistical problems that render significance testing inapplicable in system dynamics model validation. (See Morrison and Henkel 1970 for many other statistical and technical problems with statistical significance testing.)

There are also philosophical problems associated with statistical significance testing. A major problem has to do with the common practice of arbitrarily fixing the significance level (typically at 0.05), which, as explained above, determines the “rejection” region of the test. In hypothesis testing, in addition to type I error explained above, there is a second type of erroneous decision, which is the probability of accepting a null hypothesis that is in reality false. There is an important trade-off between type I and type II errors. Selecting a very small significance level (probability of rejecting a “true” hypothesis type I error) would mean allowing a relatively large type II error, while a relatively large significance level would mean a reduced probability of type II error. In an actual decision problem, each type of error would result in a certain cost and, if we knew these costs, the proper approach would be to seek the “optimum” significance level that minimizes the long-run expected total cost of committing type I and type II errors. Mathematical difficulties aside, this approach cannot be implemented in model validation, because the costs involved in the two types of errors are too complicated to estimate, or even to define. What would be the cost of “rejecting a valid model”? And the cost of “accepting an invalid one”? Faced with these problems, some socio-economic disciplines seem to have avoided the question altogether. In many modeling studies, a significance level of 0.05 is chosen arbitrarily, and models are rejected or accepted as a result of the related tests (Forrester 1961, ch. 13; 1973; Barlas 1985). This practice assumes that fixing the significance level a priori and using precise numerical bounds to either reject or fail to reject the model constitutes an “objective and scientific” method. The alternative would be to simply report the obtained validity statistics; but no significance testing, no “reject” or “fail to reject” decisions. It is clear that these two approaches belong to the two different (opposing) philosophical schools discussed earlier in the article. The first approach, under the influence of logical positivism,
seems to assume that models are either “true” or “false”, which would be revealed once model predictions confront real data. Since the validity of the model is seen as absolute and objective, there is no attempt to relate validity and purpose; no acknowledgement that the level of significance must in principle depend on the context of the study. As a result, an arbitrary level of significance is chosen (the myth of 0.05), as if dictated by the “objective world” and models get rejected at once, depending where the test statistic falls (Forrester 1961, ch. 13 and 1973 discusses why this method is no less subjective than the informal/qualitative methods that do not use fixed significance levels.) This type of binary reject/not reject decision in validity testing is very much against the relativist/holistic philosophy of science, which would argue that since validity depends on the purpose of the model, significance level must also be context-dependent. In particular, our reject/accept decision must weigh the costs involved in accepting an invalid model and rejecting a valid one. And if we cannot formulate this as a formal validity test, then it is better (and not less scientific) to let the modeler/user incorporate these factors at least in a qualitative/informal way. In the end, the modeler/user must accept the model as valid, if (s)he thinks it is more beneficial to use it than not to use it. The model is thus “judged” to be valid, invalid, or typically somewhere in between, as a result of a gradual, rich, yet semi-formal, semi-quantitative process, to which many quantitative statistics serve as inputs.

Finally, a major conceptual problem in applying statistical significance to validation, has to do with the very nature of statistical significance testing. In such a test, the null hypothesis is of general form “$X_m = X_r$” where $X_m$ represents some measure of the model and $X_r$ corresponds to the same measure of the real system. The problem is that, this type of statistical test is useful only if the final decision is to reject the null hypothesis (which would mean “the model not valid”). If, on the other hand, we fail to reject the null hypothesis (which would be our ultimate goal in validation), the result is a very weak one. That is why, in standard statistical tests, the hypothesis that one seeks to establish is typically placed in the “alternative” hypothesis. Thus, if one tries to show that a certain “treatment” does actually improve a given process, the null hypothesis would state “treatment makes no difference,” and if, at the end, the null hypothesis is rejected, then the alternative hypothesis would be considered valid. This idea cannot be applied to model validation, where we seek to show “no significant difference” between the model and the real system. The problem has no solution, since according to statistical theory, it would make no sense to have a null hypothesis of the form $X_m = X_r$ (see Wright 1972, and especially Morrison and Henkel 1970 for more on this).

In this section, we have given several reasons why statistical significance is
Barlas: Formal Aspects of Model Validity

not suitable for model validity testing. (Note that our argument is specifically against the use of statistical significance testing, not against the use of statistical measures. It is perfectly legitimate and indeed desirable to use appropriate statistical measures in all stages of model validation). Is there no potential role for statistical significance in system dynamics methodology? There are indeed some legitimate and useful applications of statistical significance testing in system dynamics studies. Statistical significance is appropriate in testing the behavior prediction accuracy of a model, given that the structure validity of the model has already been established. At this final stage, statistical significance is suitable, since the purpose is to establish that the pattern-prediction ability of the model is high enough, which is a question of accuracy and calibration, rather than validity. (Kleijnen 1995 discusses several statistical procedures to be used in behavior validation, as well as in program verification.) But note that statistical tests suitable for system dynamics behavior evaluation are different in nature than classical statistical measures and tests. As discussed early in the article, these statistics and tests must be pattern-oriented rather than point-oriented (see Figure 2 and Barlas 1989a). Statistical significance testing also has a proper role in other steps of system dynamics methodology, most notably in the analysis of simulation experiments after model validity has been established (Wright 1972). In this type of analysis, the null hypothesis would state $X_0 = X_1$, where $X_0$ and $X_1$ represent a certain behavior pattern component (say amplitudes of oscillations), before and after the design change respectively. Let us assume that the alternative hypothesis states that $X_0 > X_1$. Thus, if we can reject the null hypothesis, this result would establish that the design change does indeed reduce the oscillation amplitude significantly (a strong conclusion). Another appropriate application of statistical significance is in the analysis of experiments run on interactive simulation games (Sterman 1987; 1989). Results obtained from different subjects running simulation models interactively would have random components by nature and are as such suitable for statistical analysis. Differences between different subjects' game results, or differences between policies of subjects and those hypothesized by the simulation formulation, can be tested for statistical significance (see Sterman 1987 and 1989 for examples).

Model validity and different uses of models

In the preceding paragraph (in particular in note 7), we mentioned two different uses of system dynamics modeling:

1. modeling/analysis of a real system in order to improve some undesirable performance patterns;
2. modeling of an existing theory in order to evaluate/test the theory (for instance, Sterman 1985)

The first modeling mode is a typical application, while the second one is theoretical research. Although the same general philosophy of model validation applies in both modes, there are some practical differences in the validity tests involved. The first difference is that behavior accuracy testing (Figure 3) has very little place in theory testing. While in both modes the emphasis is on structural validity through direct-structure tests and structure-oriented behavior tests, in an application there is the third stage of testing the behavior (pattern) prediction accuracy of the model against real data. This last step is typically neither possible nor necessary in theory testing mode. The second main difference is that, in an application model, the ultimate goal is to design policies that improve the behavior of the system. Therefore, policy analysis experiments, derivation of improved policies and implementation issues are all crucial in such studies. These policy-oriented steps are non-existent in studies involving models of theories. (See Wittenberg 1992; Sterman 1992; Barlas 1992; and Radzicki 1992 for an interesting debate on models of theories vs models of real systems.)

Another use of system dynamics modeling that has become increasingly popular in recent years is interactive simulation gaming ("management flight simulators"—see Graham et al. 1994). In this mode, some selected decisions are made by a player, interactively during the course of the simulation. Thus, an interactive simulator consists of two components: the simulation model (the "engine") and the interactive interface between the player and the model. Validation of the model itself would involve the same general steps discussed in this article. (For this to be possible, the analyst must first build a decisions-in-the-loops version of the model, do the validation tests, and then turn the model into an interactive game.) In an interactive simulation study, validity of the stand-alone version of the model obviously does not imply validity of the game. Therefore, there needs to be an added step of validating the interactive game. The issues involved in validating dynamic simulation games are quite complex and interdisciplinary in nature, requiring help from different disciplines like cognitive psychology, information systems design and computer technology. As such, these issues are beyond the scope of this article (Meadows 1989; Andersen et al. 1990; Graham et al. 1994).

Finally, a much more recent usage mode of system dynamics is called "modeling for learning." According to this approach,

...
makers, not by technical experts. They are created in a group process. The policy insights from models are disseminated throughout an organization in hands-on workshops, not presentations. Such workshops are designed to mimic the learning process of the original team. (Morecroft and Sterman 1994: xvii–xviii).

Note that such models differ from traditional system dynamics models in several aspects: they are typically smaller, they are built in a group process (including non-technical participants), and their primary use is to enhance (organizational) learning. These differences may imply that the nature of validity and validation of such models would also be different than that of traditional models. Since modeling as learning is a rather new trend, there is limited conceptual or applied literature dealing with the unique aspects of validity and validation of such models. (A recent exception is Lane 1995, who discusses four different modes of application of system dynamics and how they each relate to different aspects of model validity. See also note 3.) That these models tend to be smaller may allow (and require) more extensive use of formal validation procedures, both in direct structural and in structure-oriented behavior testing. As a result, one may speculate that, if guided with proper scientific rigor and expertise, these smaller models built in groups may actually yield higher quality (thoroughly tested) models. On the other hand, much like models of scientific theories, the models built for learning may not necessitate as much behavior accuracy testing as the traditional applications do. (In this sense, these models are like pedagogical models, or “generic” models.) These are open questions to be explored with further research. It is however safe to state that these models too, being system dynamics models, should obey the fundamental philosophy and logic of validation of system dynamics discussed in this article.

Implementation issues and suggestions

One of the difficulties faced by practitioners is that the number and diversity of validity tests offered in the literature is too large, so that selecting and applying a sub-set is an overwhelming problem. With this in mind, the chart offered in Figure 1 includes only those tests that are believed to be most important (see Forrester and Senge 1980; Richardson and Pugh 1981; and Lane 1995 for more tests). Yet, it may be argued that even the narrow selection of tests suggested in Figure 1 may be too demanding in practice. Surely, all the tests included in the given chart are not of equal importance and relevance. For instance, in many cases, modified-behavior prediction, boundary-adequacy and Turing tests may be too demanding and difficult to formalize
and apply in practice. Therefore, it is desirable to define a “minimum” most crucial set of formalizable validity tests, necessary (but not always sufficient) for any system dynamics study, regardless of the usage mode (discussed in the previous section). Such a set may include the following tests in each of the three stages of model validation: In direct-structure testing; parameter-confirmation, extreme-condition (direct) and dimensional consistency tests. In structure-oriented behavior testing; extreme condition, behavior sensitivity, and phase relationship tests. And in behavior testing; the complete flowchart offered in Figure 3.

Current simulation software offers limited support for model validation. The strongest one in this sense is VENSIM, which offers a nice extreme-condition testing environment (“Reality Check”). It also supports direct-structure testing, via its “Causal Tracing” and documentation features. Finally, VENSIM offers full dimensional-consistency testing and parameter-confirmation support (Eberlein and Peterson 1994; Peterson and Eberlein 1994). The only other software that supports similar automatic dimensional-consistency testing is DYSMAP2 (Dangerfield 1994). DYNAMO offers some support for direct-structure testing, by providing strong semantic analysis, cross-references and other documentation techniques (Lyneis et al. 1994). What is currently missing in simulation software is a complete validation environment that guides the user through the entire process of model validation. Our suggested research direction is toward the development of an environment that implements the logic of the flowchart offered in Figure 2. The environment should ultimately offer all the validity tests suggested in Figure 1, but since the development of such a complete environment is quite a formidable task, our short-term research recommendation is to focus on the “minimum” subset of tests defined above. And more urgently, research must focus on the most crucial sub-step of the validation phase, namely, behavior-oriented structure testing. Although the current software provides some support for extreme-condition and sensitivity testing, the process is still quite informal. We need to design formal ways of carrying out these tests, interpreting the test results (defining a “failure”), counting, summarizing and reporting them. (See Carson and Flood 1990 for a step in this direction.) Since we cannot hope to report hundreds or thousands, of test results individually, there is a crucial need for summary measures that capture, process and summarize the results of the structure-oriented behavior tests and report them to the audience in the most informative way.

Conclusions

The purpose of this paper was to provide an overview of model validity and analyze the formal aspects of system dynamics validation. We first summarize
the philosophical issues involved in validation, and conclude that a relativist, holistic, conversationalist philosophy is appropriate (rather than an absolutist, reductionist, confrontational one). However, it is crucial to note that relativist/holistic philosophy does not reject the role of formal/quantitative tests in model validation. On the contrary, formal/quantitative tests provide crucial inputs to the larger validation process, which is gradual, semi-formal and conversational. The challenge is to design formal/quantitative validation procedures and tests suitable for system dynamics models, while keeping the above philosophical perspective.

In analyzing the major aspects of system dynamics validation, we emphasize the crucial nature of structure validity in system dynamics. Among the three major steps of model validation (direct structure testing, structure-oriented behavior testing, behavior pattern accuracy testing), the one that deserves most research attention is structure-oriented behavior testing. These tests have the special feature of revealing information on structure (in)adequacy, while having the great potential of being quantifiable and formalizable. There is some work in this direction in the existing literature, but much more concentrated research is needed. The system dynamics field must emphasize high-quality modeling by both publishing the best examples of high quality, compact models and doing methodological research into how to achieve "high quality." ("Quality" is indeed a better term than validity, as it fits much better with the relativistic/holistic philosophy that emphasizes "degree" of validity with respect to a "purpose.") We are living in an age in which many different systemic methodologies are on the rise, which implies stronger competition. The single most important and unique expertise that system dynamics has to offer to the market is, I submit, how to build good models. Although this has been the defining feature of the field since its origins, the field has not placed sufficient emphasis on formalizing (hence being able to transfer to others) the art and science of producing high-quality models. One way of starting this process is by focusing on the major aspects of model validity and validation described in this article. Formal/quantitative structure-oriented tests must be designed, tested and, finally, turned into user-friendly validation environments, following the formal logic of model validation provided in this article.

Notes
1. Note that the "distributed" nature of model validation is further complicated by the fact that the steps shown in Table 1 are not applied in a strict sequential order. System dynamics methodology has a cyclical (iterative) nature, so that
each step is executed many times, as needed in practice, which is a dynamic feedback process itself (Roberts et al. 1983; Richardson and Pugh 1981). Yet, there is a good reason for placing the "formal" validation activities right after the "formal model construction" step. Most formal validity tests, by their nature, necessitate the existence of some sort of "formal" model (i.e. a set mathematical/logical expressions that can be executed on a computer). And the logic of placing the formal validation step right before the "policy analysis/design" step is that it would be meaningless to carry out analysis and design experiments on a model in which there is no confidence.

2: There is an interesting analogy here between the "quality" of a product and the "validity" of a model. (The term quality may actually be a much better term than validity, as the Philosophical review section will later indicate.) In traditional quality control, quality of the product is "controlled" by heavy inspection and statistical process control. In Total Quality practice however, the principle is that the quality of the product must be "built into" the product in the design phases (both product design and process design); the quality cannot be "inspected in." Could the same principle apply in modeling? Could it be that modelers must focus their effort so as to build a high-quality model at the outset and seek to minimize (ideally eliminate?) the formal stage of model validation in the methodology? This question does not have an obvious answer. There is certainly a value in having a high-quality conceptual model and initial mathematical model, but this is not free and the question is how much time and effort to spend to assure such quality in these early phases. There could be a law of decreasing marginal returns, thus a potential "optimum" somewhere.

3. The "dual" nature of model validation can also be seen by comparing our paper and that of Lane (1995), which in a sense complement each other. While our paper focuses on the narrower topic of "formal" model validation, Lane discusses the many facets of model validity and how each is related to the different types of system dynamics activities. Lane's paper focuses on the richness and diversity of model validation in general. Our paper, while acknowledging this diversity and its importance, seeks to derive a formal "core" of model validation that would be useful for most typical system dynamics applications (called "ardent SD" by Lane). The proper approach is to have both perspectives simultaneously: designing/applying formal, rigorous validation procedures suitable for system dynamics models, while being aware of the limitations inherent in such formal procedures, as compared to the diversity of information used in ultimately establishing the validity/quality of models. With this approach, it would be possible to acquire formal validation procedures that are in a continuous state of improvement.

4. In the modeling literature in general, different authors use different terms for what we call "structure" validity. For instance, Carson and Flood (1990) call it "representational" validity, Schlesinger et al. (1979) call it "model qualification", yet others call it "internal" validity, or "face" validity.

5. Forrester and Senge (1980) call these tests "structure verification" and "parameter verification" tests. We prefer "confirmation" to "verification", because verification has a different established meaning in the simulation literature, namely denoting the process of making sure that the computer program is free of logical errors, that it does what the programmer intends to do.
6. For instance, *inspections* consist of five structured processes (overview, preparation, inspection, rework and follow-up), carried out by an inspection team, designed to uncover modeling flaws and programming faults. Inspections are very comprehensive, time-consuming processes that ultimately include the resolution of the flaws and faults detected (Balci 1994). *Walkthroughs* are simpler versions of inspections that require less effort and time; their purpose is to uncover and document flaws and faults. *Reviews* are “high-level” versions of walkthroughs that involve managers and sponsors of the study. As such, they are concerned with whether the model meets the project objectives, development specifications, standards and guidelines, rather than with the technical validity and correctness of the model (Balci 1994). In *semantic analysis*, the simulation software provides information about the source code, such as “symbol tables” and “cross-reference tables”, that describe the variables used in the model, their type, their relationships, where they are used, where they are altered, etc. (Balci 1994). Such information can be very valuable in direct structure confirmation, especially when the model is large.

7. Some system dynamics models are built not for “users,” but as platforms to test social/economic theories. There is a popular view which argues that it is only acceptable to tie the accept/reject decisions about model validity to the estimated costs of type I and type II errors if the model is a potential application. According to this view, in scientific theory testing mode, the “truth” of the theory must be tested objectively; such a decision cannot/must not depend on any “costs” involved. And thus, the argument goes, when a model is tested as a scientific theory, it is proper and indeed necessary to use formal reject/not reject decisions of statistical significance testing. The relativist/holistic philosophy outlined above would naturally reject this argument, since scientific theory confirmation is described as a gradual, semi-formal, conversational process by its very nature. Such a philosophy is against the notion of absolute truth and objectivity even in the natural sciences. Indeed, a curious fact is that significance testing is very rarely used in natural sciences, but extensively used in socio-economic/behavioral sciences, as if to give an “appearance” of objectivity. Morrison and Henkel (1970) provide a very comprehensive critical discussion of this extensive use of statistical significance in social/behavioral sciences. Their book is an excellent treatment of both statistical and philosophical aspects of the issue, in particular how statistical and scientific inference (or statistical and scientific “significance”) differ. There are very strong arguments in the book against using significance testing in scientific inquiry, even if all the technical problems were hypothetically solved. (These arguments are outside the scope of our article; see Morrison and Henkel 1970.) These authors conclude that significance testing in social/behavioral sciences characterizes “a false search for empirical associations, rather than a search for hypotheses that explain” and that this contributes to the “impoverished state of our theory” in these sciences (Morrison and Henkel 1970: 309).
Appendix: The Multi-step validation procedure (Barlas 1989b)

**Trend comparison and removal**

A linear trend can be estimated by \( Y = b_0 + b_1 t \), a quadratic trend by \( Y = b_0 + b_1 t + b_2 t^2 \), or an exponential one by \( Y = b_0 e^{bt} \). Then, if \( Y_i \) are the data, the trend component can be removed by: \( Z_i = Y_i - b_1 t \), or \( Y_i - b_1 t - b_2 t^2 \) or \( Y_i - b_0 (e^{bt} - 1) \).

**Comparing the periods:**

An autocorrelation function test is able to detect significant errors in the periods. The “sample autocovariance function” of a time series \( X_i \) is given by

\[
\text{Cov}(k) = \frac{1}{N} \sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x})
\]

for lag \( k = 0, 1, 2, \ldots < N \). Then, the “sample autocorrelation function” is obtained by dividing Cov\((k)\) by Cov\((0)\):

\[
r(k) = \frac{\text{Cov}(k)}{\text{Cov}(0)} = \frac{\text{Cov}(k)}{\text{Var}(X_i)}
\]

for lag \( k = 0, 1, 2, \ldots , N \). We use the following \( \text{Var}(r(k)) \) provided by Anderson (1982):

\[
\text{Var}(r(k)) = \frac{1}{N(N+2)} \sum_{i=1}^{N-1} (N-1)(r(k-i)+r(k+i)-2r(k)r(i))^2.
\]

If \( r_s(k) \) belongs to the simulation output and \( r_A(k) \) to the actual (observed) one, then the null hypothesis is:

\[ H_0 : r_s(1) - r_A(1) = 0, r_s(2) - r_A(2) = 0, \ldots , r_s(M) - r_A(M) = 0. \]

and the alternative hypothesis is,

\[ H_1 : r_s(k) - r_A(k) \neq 0 \text{ for at least one } k. \]

Now consider the difference \( d_1 = r_s(1) - r_A(1) \). The standard error of \( d_1 \) is

\[
\text{Se}(d_1) = \sqrt{\text{Var}(r_s(1)) + \text{Var}(r_A(1))}.
\]

Since \( d_1 = 0 \) under \( H_0 \), we construct the interval \([-2\text{Se}(d_1), \text{Se}(d_1)]\) and reject \( H_0 \) if \( d_1 \) falls outside the interval.
Comparing the means

“Percent error in the means” $E_1$ can be defined as:

$$E_1 = \frac{|\bar{S} - \bar{A}|}{\bar{A}}$$

where

$$\bar{S} = \frac{1}{N} \sum_{i=1}^{N} S_i, \quad \bar{A} = \frac{1}{N} \sum_{i=1}^{N} A_i$$

Comparing the variations

“Percent error in the variations” $E_2$ is defined as:

$$E_2 = \frac{|s_S - s_A|}{s_A}$$

where

$$s_S = \sqrt{\frac{1}{N} \sum (S_i - \bar{S})^2}, \quad s_A = \sqrt{\frac{1}{N} \sum (A_i - \bar{A})^2}.$$ 

Testing the phase lag

The cross-correlation function provides an estimate of a potential phase lag. The cross-correlation function between the simulated ($S$) and the actual ($A$) time patterns is given by:

$$C_{SA}(k) = \frac{(1/N) \sum_{i=k}^{N} (S_i - \bar{S})(A_{i-k} - \bar{A})}{s_Ss_A} \quad \text{for } k = 0, 1, 2, \ldots$$

$$C_{SA}(k) = \frac{(1/N) \sum_{i=k}^{N} (A_i - \bar{A})(S_{i+k} - \bar{S})}{s_Ss_A} \quad \text{for } k = 0, -1, -2, \ldots$$

An overall summary measure

As a final step, only after all validity tests have been passed, can the discrepancy coefficient $U$ be computed as a single summary measure:
\[ U = \frac{\sqrt{\sum (S_i - \bar{S} - A_i + \bar{A})^2}}{\sqrt{\sum (A_i - \bar{A})^2} + \sqrt{\sum (S_i - \bar{S})^2}} \]

\[ = \frac{\sqrt{\sum (E_i - \bar{E})^2}}{\sqrt{\sum (A_i - \bar{A})^2} + \sqrt{\sum (S_i - \bar{S})^2}} \]

\[ = \frac{s_E}{s_A + s_S} \]

**References**


