Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 151 (2003) 552–568
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

Decision Aiding

# Model calibration as a testing strategy for system dynamics models

Rogelio Oliva *

*Harvard Business School, Morgan Hall T87, Boston, MA 02163, USA*

Received 20 February 2001; accepted 12 August 2002

**Abstract**

System dynamics models are becoming increasingly common in the analysis of policy and managerial issues. The usefulness of these models is predicated on their ability to link observable patterns of behavior to micro-level structure and decision-making processes. This paper posits that model calibration—the process of estimating the model parameters (structure) to obtain a match between observed and simulated structures and behaviors—is a stringent test of a hypothesis linking structure to behavior, and proposes a framework to use calibration as a form of model testing. It tackles the issue at three levels: theoretical, methodological, and technical. First, it explores the nature of model testing, and suggests that the modeling process be recast as an experimental approach to gain confidence in the hypothesis articulated in the model. At the methodological level, it proposes heuristics to guide the testing strategy, and to take advantage of the strengths of automated calibration algorithms. Finally, it presents a set of techniques to support the hypothesis testing process. The paper concludes with an example and a summary of the argument for the proposed approach.
© 2002 Elsevier B.V. All rights reserved.

*Keywords:* System dynamics; Simulation; Hypotheses testing; Model calibration; Parameter estimation

## 1. Introduction

Since its inception, system dynamics (SD) has emphasized the importance of clarity of purpose for any intervention—a defined problem, issue or undesirable behavior to be corrected (Forrester, 1961). The problem behavior is usually described in a reference mode, and the purpose of the intervention is to identify how structure and decision policies generate the identified reference mode so that solutions can be generated and implemented. SD practitioners build and depend on formal simulation models to overcome the cognitive limitations to grasp the detailed complexity of the problem situation, and to make reliable behavioral inferences. Generation of problem solutions relies on using these models for policy testing (Forrester, 1961), what-if scenarios (Morecroft, 1988), or policy optimization (Kleijnen, 1995). All of these efforts, however, presume confidence that the model represents the structure of the problem situation and that his structure is responsible for the observed behavior. A theory that explicitly articulates how structure and decision policies generate

* Tel.: +1-617-495-5049; fax: +1-617-496-5265.
*E-mail address:* roliva@hbs.edu (R. Oliva).

0377-2217/$ - see front matter © 2002 Elsevier B.V. All rights reserved.
doi:10.1016/S0377-2217(02)00622-7

behavior is called a dynamic hypothesis (DH) (Richardson and Pugh, 1981), and it aims to explain "behavior as endogenous consequences of the feedback structure" (Sterman, 2000, p. 86). A good DH links observable patterns of behavior to micro-level structure and decision-making processes (Forrester, 1979; Morecroft, 1983). Arguably, the outcome of the *modeling* process should be a DH in which there is a degree of confidence that it represents the structure and observed behavior of the problem situation. Once confidence in the DH has been achieved, it is possible to proceed with the exploration of policies and scenarios, the optimization of policies, or the articulation and diffusion of insights.

Most descriptions of the SD method emphasize the iterative process required to build confidence in these models, and there is extensive literature on the kinds of tests that a model needs to pass (Barlas, 1989; Barlas and Carpenter, 1990; Barlas, 1994; Forrester and Senge, 1980; Sterman, 2000). Nevertheless, the iterative process to gain confidence in a model has been underplayed in the literature, and the advice seldom reads beyond some form of "iterate the process 'articulate DH, build model, test model' as necessary." This paper focuses on the iterative process used to develop confidence in a simulation model.

Since a DH makes a causal claim between structure and behavior, I argue that model calibration—the process of estimating the model parameters (structure) to obtain a match between observed and simulated structures and behaviors—is a stringent test of a DH. New software developments have automated the calibration process, thus making it possible to optimize the fit to historical data with a given structure. Unfortunately, because these algorithms can replicate historical behavior with relative ease, automated calibration (AC) often gives practitioners false confidence in their models. The second objective of this paper, then, is to create a framework for the use of AC within the context of model testing.

The paper tackles model calibration as a testing strategy at three levels: theoretical, methodological, and technical. The ultimate aim is to provide methodological coherence (Checkland, 1981; Eden, 1990). First, the nature of the testing process is described, and the modeling process is recast as an experimental approach to gain confidence in a DH. Caveats and limitations of the proposed approach are identified at this conceptual level. Section 3, after a brief review of parameter estimation and AC, takes the argument to a methodological level, suggesting strategies and heuristics for model calibration that assist in the testing process. Section 4 introduces a set of tools and techniques that support the calibration/testing process through analysis and interpretation of the calibration results. Although the techniques are specific steps that generate standard results, their applicability is contingent on the complexity of the model being tested, hence the need for the methodological heuristics developed in the previous section. The paper concludes with an example and a summary of the argument for the proposed approach.

## 2. An argument for model calibration

As stated above, a DH is a theory about how structure and decision policies generate (cause) the observed behavior. The model is the conveyor of the DH, formally positing the causal link between structure—captured in terms of equations and parameters—and behavior—the simulated output generated by the interaction of the equations and initial conditions. The experimental literature outlines three criteria for inferring cause: (1) temporal precedence of the cause, (2) covariation between the presumed cause and effect, and (3) the need to rule out alternative explanations (Cook and Campbell, 1979). While the mechanics of simulation make the precedence of structure clear, the second and third criteria for inferring cause require a more thorough analysis.

For any given DH there will be many rival hypotheses—other structural explanations that might be capable of generating the observed behavior. DHs, however, are not instrumental theories, i.e., theories deemed useful for explaining certain phenomena regardless of their truth or falsehood (Flew, 1984). Rather, DHs are realistic theories of behavior (Lane, 1999). Therefore, in testing a DH, it is not enough for the model to match the observed behavior; the behavior

generated by the model has to be right for the right reasons. Clearly, no single test will be able to rule out all the possible alternative explanations. Although it is impossible to verify a hypothesis (Oreskes et al., 1994), science has refined a systematic approach to increasing the confidence of stated hypotheses: "subject your assumptions to tough tests rather than soft ones" (Bunge, 1967, p. 10). The essence of the scientific method is captured in its experimentation ethos: strive to reject the hypothesis. Thus, in order to gain confidence in the causal argument stated in a DH, the testing process has to be based on experiments that can yield a false outcome.

Recasting the modeling process—articulate DH, build model, test model—as the basic cycle of the scientific method—state hypothesis, build a laboratory, run experiments to *refute* hypothesis—clarifies the role of the formal model as an aid to experimentation. In this testing stage, the model should be regarded only as a laboratory in which to run experiments to refute the hypothesis—a role consistent with Forrester's (1971) view of the modeling process (see also Graham, 2002). Once there is confidence in the DH, the model shifts roles and can be used to develop intervention strategies. (The recasting of the SD model as an experimental setting clarifies the role of non-quantitative SD approaches, e.g., 'systems thinking' (Senge, 1990) and 'qualitative SD' (Wolstenholme, 1990), as tools for hypotheses generation without the systematic approach to falsify the hypothesis.)

The SD method is particularly well suited for this experimental approach. According to Bunge (1967), a well-formulated hypothesis should be (1) logically sound, (2) grounded in previous knowledge, and (3) empirically testable. First, the grammar of SD modeling, and the requirement for computer simulation, helps ensure that a DH's logic is sound. Formulating a model—i.e., ensuring that all equations have consistent units, that time constants are positive, that all feedback loops contain at least one stock, etc.—enforces precision, making the DH "'specific,' 'sharply defined,' and 'not vague"' (Forrester, 1961, p. 57). Second, SD model development is grounded in previous knowledge. SD is learned through apprenticeship, i.e., becoming familiar with the existing knowledge base. Furthermore, the SD academic community has endeavored to ground SD work in findings from other fields (see, for example, Morecroft, 1985; Sterman, 1989), and practitioners have a clear set of values to determine if something is well formulated (Martinez and Richardson, 2001).

Finally, since a DH will be used to generate problem solutions, it needs to be relevant to the problem situation being addressed. A model needs not only to be internally consistent, logical, and well formulated; it must also say something interesting about the real world. Although a model offers a simplified description of a problem situation, the appropriateness of that description needs to be assessed as well. Forrester (1961, p. 57) used the term "accuracy" to label "the degree of correspondence to the real world." To fully confront the model with the piece of reality that it is supposed to represent, we should capture the essence of the real system through a series of observations, measurements or facts. A DH links structure to behavior. Accordingly, observations, measurements or facts about the structure *and* the behavior of the system are needed. In a SD model both structure and behavior are, in most cases, directly observable—"model variables should correspond to those in the system being represented" (Forrester, 1961, p. 63)—and every model parameter should have a real world interpretation. The formal description of the model and the simulation results constitute a refutable causal model with multiple 'points of testing' (Bell and Bell, 1980; Bell and Senge, 1980).

Model *calibration* is the process of estimating the model parameters to obtain a match between observed and simulated behavior. Calibration explicitly attempts to link structure to behavior, which is why it is a more stringent test than solely matching structure *or* behavior. Confidence that a particular structure, with reasonable parameter values, is a valid representation increases if the structure is capable of generating the observed behavior. If the structure fails to match the observed behavior, then it can certainly be rejected; the calibration exercise constitutes a test for the DH. Examining whether the model adheres *simultaneously* to observable structure and behavior

is the toughest test to which that model can be subjected, challenging both the logic of the hypothesis and its relevance.

### 2.1. Caveats

The calibration process, however, has some limitations. First, calibration is only a partial test of the structure–behavior couple. Systemic structure in SD models is represented through equations and parameters (initial conditions for stocks have the same computational and structural functionality as system parameters). Calibration fixes the model equations and adjusts the model parameters to match observed behavior. There is a chance that a set of parameter values might be capable of replicating the observed behavior through a set of unrealistic formulations, and thus generate the right behavior for the wrong reasons. A full test of the structure–behavior couple should include an assessment of the appropriateness of model equations.

Second, rejecting a stated DH is not easy. If a discrepancy in behavior emerges when comparing the model's output to the behavior in the real world situation, a modeler's first inclination may be *not* to reject the DH in which so much time and effort has been invested. Instead, she will deflect falsification onto other assumptions used during the modeling/testing process (e.g., formulation error, measurement errors in the data). The process of protecting the core hypothesis is referred as the Quine–Duhem thesis: "because of all the background assumptions that might be wrong, any outcome can be rationally distrusted and explained away by ad hoc hypotheses that alter the background assumptions" (Cook and Campbell, 1979, p. 21). In order to perform a rigorous test of the hypothesis, the modeler must press on to gain confidence in the set of auxiliary assumptions and strive to reject the DH (Wilson, 1997). Taken to the extreme, the Quine–Duhem thesis can be equated to Lakatos' (1974) research programs. According to Lakatos, the hard-core ideas of a research program represent the defining characteristics of the research program (e.g., for SD the assumption that structure drives behavior), and are not subject to rejection. Thus, regardless of how aggressively the modeler 'attacks' the DH, the alternative DH will probably be within the context of the SD assumptions, and include explanations in terms of rates, levels, delays, and feedback loops.

Historical evidence shows that scientists (and practitioners) discount data that refutes their theory, as they prefer to work with an imperfect theory than not to have a theory at all (Kuhn, 1970; Lakatos, 1974). Under this perspective, theory validation becomes the process of building confidence in a theory, either through falsification or a functional perspective of the theory usefulness (see Gass, 1983; Miser, 1993; Mitroff, 1972; Roy, 1993; Smith, 1993 for evidence of the OR/MS community shifting to a functional perspective on model validity). Thus, validation is used as an inherently partial assessment of the degree of usefulness of the theory (van Horn, 1971). The following sections provide heuristics and tools to challenge both elements of model structure—equations and parameters—in order to extract the most information from data available, and strive for falsification. These heuristics, however, should be kept within the context of the functional utility of the theory in reference to its specific purpose.

## 3. Calibration heuristics

Parameter values for a SD model are normally estimated *a priori* from direct observations, educated guesses, and other sources of data "below the level of aggregation of model variables" (Graham, 1980, p. 144). The estimates are then revised based on an iterative process designed to match the real system's behavior. Lyneis and Pugh (1996, p. 1) describe the process as follows:

> The calibration of the model ... is typically done "by hand." In this iterative process, the modeler examines differences between simulated output and data, identifies possible reasons for those differences, adjusts model parameters in an effort to correct the discrepancy, and re-simulates the model, looping back to the first step. The entire parameter

estimation process therefore relies on the expertise and experience of the modeler.

Unsurprisingly, the process has been criticized for both its unreliability and its lack of replicable outputs (see Lyneis and Pugh, 1996 for a list of issues raised against "hand" calibration). Statistically based approaches have been adopted from other fields in an attempt to make the parameter estimation process more rigorous. Early studies (Senge, 1977), however, concluded that econometric approaches were not useful because of the propensity of SD models to violate the ordinary least squares estimation assumptions. Two approaches have been adopted to consistently estimate parameters for whole models simultaneously: full-information maximum-likelihood via optimal filtering (FIMLOF) (Peterson, 1980) based on engineering statistics (Schweppe, 1973), and model reference optimization (MRO) (Lyneis and Pugh, 1996) based on non-linear optimization algorithms that search across the parameter space. Both approaches are data- and computational-intensive; they require an error function containing all data available, and access to the set of model parameters to be adjusted. The techniques, however, yield an optimal fit with the given structure and adjusted parameters. Consequently, even novice modelers can generate good fits (Lyneis and Pugh, 1996) and, perhaps more importantly, replicable results. Operationally, the main difference between these approaches is that FIMLOF requires linearization of the system and the covariance matrix for the driving noise, while MRO can work with the model and data available 'as is'. Throughout the rest of the paper, I will use the MRO formulation of a calibration problem to illustrate some of the issues with AC, and to develop insights and heuristics for model testing. These insights and heuristics are transferable to calibration efforts with FIMLOF.

### 3.1. Automated calibration

Under MRO, the calibration problem from longitudinal data is specified as an optimization problem, adjusting system parameters ($\boldsymbol{p}$), to minimize a function of the differences between the available data series ($\boldsymbol{d}_t$) and the corresponding model variable ($\boldsymbol{y}_t$). Since multiple data series might be available, the objective function must specify the relative weighting ($\boldsymbol{w}$) for each series. Model outcome variables are a function of the model's state variables ($\boldsymbol{s}_t$), parameters ($\boldsymbol{p}$), and known inputs ($\boldsymbol{u}_t$). Lastly, the values of system parameters can be limited to a feasible range [**ll**, **ul**]. Formally, the calibration problem is stated as

$$\underset{\boldsymbol{p}}{\text{Min}} \quad \sum_{i=1}^{n} w_i \sum_{t=T_0}^{T_f} f(y_{it} - d_{it}),$$

$$\text{Subject to} \quad \boldsymbol{y}_t = c(\boldsymbol{s}_t, \boldsymbol{p}, \boldsymbol{u}_t), \quad \mathbf{ll} \leqslant \boldsymbol{p} \leqslant \mathbf{ul}$$

where

$w_i$ = weight of $i$th error series,
$y_{it}$ = model variable $i$ at time $t$,
$d_{it}$ = data for variable $i$ at time $t$,
$\boldsymbol{s}_t$ = model state variables,
$\boldsymbol{p}$ = model parameters,
$\boldsymbol{u}_t$ = known inputs (data series),
**ll** = lower limit of parameter feasible range,
**ul** = upper limit of parameter feasible range,
$T_0$ = initial simulation time,
$T_f$ = final simulation time, and
$n$ = the number of variable − data pairs in error function.

There is a range of options for defining the error function $f$ and the relative weight of each error series $w_i$ (Kleijnen and Sargent, 2000; see also Reichelt et al., 1996 for a detailed comparison of error functions and their relative attributes when used to estimate cyclical SD models). The constraint function $c$, however, is directly determined by the model equations, and in most cases will not be linear. A variety of available optimization algorithms are suitable to search the parameter space to minimize the deviation between model outcome and historical data. Good algorithms for this task should be capable of searching over large parameter spaces while confronting noise, discontinuities, and pervasive non-linearity (Miller, 1998); algorithms typically implemented include grid-search, hill-climbing, simulated annealing, genetic algorithms, and gradient estimation techniques (see Miller, 1998 for a brief discussion and references on these optimization techniques).

While AC techniques are capable of generating an optimal fit to historical data with a given structure and calibration parameters, a number of critiques have been offered against this approach to inferring parameter values. These critiques can be grouped under three main headings: the source of the estimated parameter values, the tractability of mismatches and model diagnosis, and the nature of the implied testing process.

(i) *Source of the estimated parameter values*. AC estimates parameter values from model equations and aggregate data—collected statistics corresponding to model variables. This process, however, assumes that the model structure (equations) is known, and that all uncertainty resides in the parameter values. Using aggregated data to determine parameter values "forces" the proposed structure, through the estimated parameters, to match historical behavior. Graham (1980, p. 144) argues that most parameters in SD models should be estimated from 'unaggregate data'—"information [that is] more detailed than data that corresponds (sic) directly to model variables."

Graham (1980) provides two strong reasons for estimating parameters with data below the level of aggregation. First, most factual knowledge about the system falls into the category of unaggregate data, and parameters can be estimated independently of model equations. Second, parameters that are directly observable, or that can be estimated from unaggregate data—records, interviews, etc.—increase the model's ability to anticipate events outside of the historical experience, and are intrinsically more robust than parameters that are inferred from observed behavior.

(ii) *Tractability of mismatches and model diagnosis*. One of the main benefits of the AC techniques is that it is possible to specify the calibration problem as a single optimization problem with an error function that contains all data available and allows for adjustment of all model parameters. By providing total flexibility to the model structure to adapt to the existing data, such an approach generates the best possible fit to all data available. From an operational perspective, however, having a complex error function and multiple parameters to adjust makes the tractability of the errors and the diagnosis of mismatches more difficult.

Since not all model parameters affect all output variables in the model, as the number of data series in the error function increases, individual parameters become less significant; variations in individual parameter values have a small impact in a complex error function, thus resulting in wider confidence intervals for the estimated parameters. Similarly, increasing the number of parameters to be estimated through an error function reduces the degrees of freedom in the estimation problem, thus resulting in wider confidence intervals, i.e., less efficient estimators.

The most serious difficulty with a large number of calibration parameters, however, is the increased difficulty in detecting formulation errors. In an effort to match historical data, the calibration process 'fixes' the model structure to cover formulation errors. Since these corrections are distributed among the parameters being used in the calibration problem, as the number of parameters being estimated increases, the 'correction' to each parameter becomes smaller. Small deviations from 'reasonable' values and wider confidence intervals make it more difficult to detect fundamental formulation errors, especially when a 'good fit' to historical behavior has been achieved.

(iii) *Nature of implied testing process*. The main critique that can be raised against AC is that it constitutes a confirmation test—rather than a falsification test—of the structure's ability to replicate historical behavior. By using AC to generate the best possible match to the historical behavior (taking the proposed structure as given), a modeler is attempting to obtain an affirmative response to the question: "Is the proposed structure capable of replicating the observed behavior?" While not a trivial test, this is a confirmation test of the DH, i.e., a test seeking evidence consistent with the DH. As discussed above, the ethos in testing a hypothesis should be to attempt to falsify it. Tests that focus on confirmation of current hypotheses have a diminished ability to identify and recognize anomalies that might lead to improvements (see Sterman, 1994). Unless there are obvious mismatches, the AC process is likely to confirm our

current beliefs and does not support a process aiming to reject the DH.

From the three types of issues raised against the AC process, it is clear that its outcome—the estimated parameter values and the judgment of adequacy of the model structure—should not be accepted at face value. AC's effectiveness to estimate the value of model parameters is dubious because of the "correct structure" assumption and the multiple degrees of freedom inherent to the process. The assessment of adequacy of model structure is also questionable since its based on a *confirmation* test of the DH. Thus, we are faced with a paradox: while model calibration, by requiring simultaneous adherence to observable behavior and structure, constitutes a stringent test of the DH, AC, the most powerful tool available for calibration, strips the process of its power to perform the test.

### 3.2. Proposed heuristics

Working around this paradox requires leveraging the strengths of AC without simultaneously overriding the process of testing the DH. AC is a reliable and efficient way to fit a structure to the observed behavior. However, to maintain the power of calibration as a test of the DH, AC has to be used judiciously. Three simple heuristics to frame the use of AC are enough to address the criticisms raised against it. The first two heuristics aim to increase the process's capability of identifying and diagnosing sources of differences. The third heuristic addresses the nature of the test being performed when using AC. Together these heuristics reorient the calibration process such that it becomes a testing strategy for dynamic hypotheses.

I. *Include in calibration problem all knowledge available about system parameters*. If parameters can be directly observed—or estimated from data below the level of aggregation—they should be treated as part of the known structure (Graham, 1980). Leaving a known parameter as a calibration lever increases the risk of an error being masked by small adjustments to that parameter. If knowledge about a parameter is not precise, but it can be limited to a feasible range, such information should be introduced into the calibration problem by restricting the search range for that parameter. Limiting the flexibility of a parameter to its feasible range, or altogether treating it as part of the known structure, forces the AC process to correct for deviations by 'fixing' other pieces of the structure where the uncertainty lies.

II. *Apply AC to the smallest possible calibration problems*. By "small," I mean a reduced number of equations and, consequently, parameters, i.e., partial-model testing (Homer, 1983). Working with small calibration problems reduces the risk of the structure being forced into fitting the data, increases the efficiency of the estimation (estimators with smaller variances), and concentrates the differences between observed and simulated behavior in the piece of structure responsible for that behavior.

III. *Use AC to test the hypothesis "The estimated parameter matches the observable structure of the system."* Calibration, as discussed above, constitutes a confirmation test of the question: "Is the inferred structure capable of generating the historical behavior?" This important question needs to be addressed. However, since AC yields the best possible set of parameters to match the observed behavior, setting the null hypothesis to test the *a priori* parameter estimates is a much more powerful test. The question we should be asking after AC is not whether the behavior was matched, but whether the estimated parameters are consistent with what we know about the system. Of course, testing the relevance of a parameter estimate presupposes success in matching the observable behavior. That is, it is not possible to evaluate the appropriateness of a parameter value if that parameter is embedded in a structure incapable of generating the observed behavior.

The first heuristic can be easily handled by omitting the directly observable parameters from the calibration problem, or by limiting its search space. This measure has the additional benefit of reducing the size of the calibration problem (heuristic II). Heuristic II requires knowledge of model structure, the role of individual parameters in determining the model's behavior, and the sources of data available. The goal is to partition the model as finely as possible, in order to focus the analysis

and generate more efficient estimators. Finally, heuristic III requires a way of thinking about calibration output beyond the fit to historical data. The next section explores the set of tests available to test the adequacy of fitness to historical behavior and structural evidence.

Before proceeding, it should be noted that it is not new to claim that calibration should be an *iterative* process of *controlled* experiments (partial-model structure) to *reject* the hypothesis linking structure to observed behavior. As can be inferred from the titles of his papers, Homer (1983, 1996, 1997) has been arguing for a similar process, and the approach is also consistent with Forrester's view of the modeling process and model testing (Forrester, 1971; Forrester and Senge, 1980). My proposal, however, goes further by tapping into the power of the AC procedures in a manner consistent with the ethos of model testing.

## 4. Analysis of both historical and structural fit

Although it is much more difficult to hide formulation flaws when calibrating small model partitions, given the effectiveness of AC in matching historical behavior, it is still possible to overlook the signals that there is a problem with model formulation (see Example A in Section 5). In this section, I summarize the use of some tools to diagnose the adequacy of historical fit, and test the hypothesis of structural fit. A module to facilitate the analysis described in this section has been developed (Oliva, 1995), and is available online.

### 4.1. Does the model match the historical behavior?

There are multiple measures of fit of simulation output to historical data, and the selection of a measure should be based on the purpose of error analysis (Kleijnen and Sargent, 2000; Reichelt et al., 1996; Sterman, 1984). Beyond assessing the magnitude of the error, the traditional test to identify the sources of errors for historical fit of SD models is based on the Theil inequality statistics (Theil, 1966). The Theil inequality statistics decompose the mean-square-error ($\mathrm{MSE} = 1/n \sum_{t=1}^{n} (y_t - d_t)^2$) between simulated and actual series into three components: bias, unequal variation, and unequal covariation. Dividing each component by the MSE gives the fraction of the error that is due to unequal means, unequal variances, or imperfect correlation. For a full description of how to interpret these statistics for goodness of fit of systems dynamics models and the identification of systematic errors, see Sterman (1984).

Although Theil's statistics are good at flagging systematic errors in model formulation, they are not very effective in helping diagnose the output of AC. Two strengths of AC combine to reduce the power of these statistics as a diagnosis aid. First, AC normally generates very good fits between simulated and historical behavior. The AC process is geared to optimizing the fit between the simulated and historical behavior. If there is a mismatch between observed and simulated structure and behavior, AC will minimize the behavior mismatch and "hide" the source of error in the parameter estimates. Second, error functions based on the squared differences between simulated and historical data (the most common error functions utilized in MRO) minimize large errors. This results in a tendency for AC processes to replicate the historical mean and place most of the error in the unequal variance and unequal covariance components of Theil's statistics. The desired result of a calibration process is, of course, small residuals with zero-mean (unbiased). However, by performing well in these metrics and masking the sources of errors somewhere else, AC diminishes the diagnosis power of Theil's statistics. Typically, systematic errors in a formulation calibrated through AC will be signaled by a large value in the unequal variance component of the MSE. The signal, however, is not symmetric, as a small unequal variance does not mean that the model is good. Furthermore, clear interpretation of these numbers tends to be difficult because of the relatively small errors (mean average percent error $< 3\%$) that AC generates. To diagnose the source of error in a calibration problem, it is necessary to explore the residuals of the match (Sterman, 1984).

A simple inspection of the plot of residuals over time is helpful in detecting biases, trends, and cyclical components. The auto-correlation spectrum

of the residuals is helpful in identifying cyclical components of the time series and auto-correlated errors. Finally, a scatter plot of residuals against independent variables is useful to check for violations of the equality of variance assumptions (heteroskedasticity), and diagnose the structural sources of those deviations (see Oliva, 1995 for a description of these plots and their use in identifying behavioral and structural mismatches).

### 4.2. Does the model match the structure?

Once adequacy of fitness to historical behavior has been asserted, it is possible to move to the actual testing of the hypothesis whether the estimated parameters match what is known about the system structure. The test of estimated parameters can be broken down into three increasingly demanding tests: feasibility, consistency, and confidence.

Feasibility is the first test for the estimated values for parameters. Estimated time constants should be positive, initial conditions for physical stocks should obey the laws of conservation of matter, "fractions" should have values between 0 and 1, and the resulting formulation should be robust to extreme condition testing. The evaluation of the feasibility of estimated parameters is context-sensitive, and should be performed with full understanding of model formulation. A way to facilitate this test is, as suggested by heuristic I, to limit a parameter's search range to its feasibility area. However, if the result of the estimation is at one of the limits specified for the search space, we should question the adequacy of the DH or the model formulation. Normally, when AC results in parameter estimates at one of the limits in the search space, the fit to historical behavior is not very good. Estimates at a limit indicate that the model structure is being "bent to fit" the data up to a feasible limit (see Example B in Section 5). If the structure cannot be accommodated within the feasible solution space, the difference is reflected in the residuals. Removing the limiting constraint, and performing the calibration again, will allow the model to be freely adjusted to the data; yielding a more precise diagnosis of the formulation flaw.

A final assessment of parameter feasibility is to check if estimates are consistent with other parameters in the model. For example, we should reject a structure whose calibration yields delays shorter than the simulation computation interval, DT. Non-feasible parameter estimates are a clear sign of a formulation error, indicating that the proposed structure should be revised.

The next test is to determine if the parameter is consistent with what is known about the system structure. The estimated parameter needs to match other sources such as interviews and direct observations. See Graham (1980) and Forrester (1994) for a detailed discussion of other data sources for parameter estimates. Formally, the test of whether the estimated parameter ($\hat{\beta}$) is consistent with what is known about the system structure—the *a priori* estimate ($\beta_0$)—should be stated as the null hypothesis ($H_0 : \hat{\beta} = \beta_0$). If the *a priori* estimate differs from the computed estimate value, we need to reject the null hypothesis and reconsider the proposed formulation, the prior parameter estimate, or the DH (see Senge, 1978 for an example of formal tests of estimates).

In ordinary least squares estimation, the $t$-statistic for an estimate—the ratio of the distance between the computed estimate and the test value to the estimated standard deviation of the estimate ($(\hat{\beta} - \beta_0)/\hat{\sigma}_\beta$)—gives a sense of the confidence that can be placed in the estimate being equal to the test value. This statistical test, however, belongs to the single-equation class, and does not yield information on the significance of the parameter for model specification (Mass and Senge, 1980, p. 222).

Under MRO, for the case when the error function $f$ is defined as the square of the predicted error, and $w_i$ is set as the reciprocal of the variance of the predicted error ($1/\hat{\sigma}^2$), the objective function corresponds to the likelihood equation, and the parameter estimates are the maximum-likelihood estimates (Greene, 1997). Using the response surface of the likelihood equation, it is possible to determine confidence intervals for the parameter estimates by performing sensitivity analysis inside the hyper-volume defined by the parameters being estimated. The confidence intervals for the parameters are calculated from the curvature of the response surface by varying each parameter and

measuring the change in response (Peterson, 1980; see also Long, 1997, p. 87). The intervals are determined based on a percentage reduction of the objective function equal to the desired significance of the test. If the response surface around the optimal point is steep, small variations of the parameter will yield a significant drop in the objective function. The tightness of the reported confidence interval measures how useful the data are for estimating a parameter. Although not explicitly computing the variance of the estimate, this computation of the confidence intervals takes into consideration the full structure specified in the calibration problem (as opposed to a single equation). When confidence intervals of the parameters are reported, the test for $H_0$ is simple, as it is only necessary to test whether the *a priori* estimate falls within the reported confidence interval—testing against *a priori* feasibility interval $(\beta_{ll}, \beta_{ul})$ can be done through a couple of one-tailed tests ($\hat{\beta}_{ll} > \beta_{ll}$ and $\hat{\beta}_{ul} < \beta_{ul}$). The test is similar to a Wald test based on estimates of the unrestricted model (Greene, 1997; Long, 1997), but using the curvature of the response surface to infer the variance of the estimate.

The final test for the parameter estimates is to assess its confidence interval. Although there is no fixed rule to determine when a confidence interval is tight enough, the confidence test is here outlined as a necessary step to assess the quality of AC outcomes, and the power of the consistency test described above. Ultimately, the appropriateness of the interval, and its associate risk to incur in a type II error, is determined by the purpose of the model and the potential benefits of increasing its accuracy (Raiffa and Schlaifer, 1961).

It should be noted that the reported confidence intervals could also be interpreted as a sensitivity analysis. A tight confidence interval means that data-structure combination was effective in discriminating among parameter values. A wide confidence interval, on the other hand, implies that, at least for the variables in the error function, the model is not sensitive to variations in that parameter. This suggests that additional efforts to increase the precision of estimates of parameters that are not critical to the model's performance might be unwarranted. Additionally, because multiple compensating feedback loops, SD models are characterized by not being very sensitive to changes in most model parameter (Forrester, 1961), further making the case for not spending resources increasing the efficiency of the estimates. Nevertheless, sensitivity to changes in parameter values depends on the variable being used to measure the behavioral change. In some instances variables might be susceptible to small changes in parameters that were thought not critical during modular optimizations and greater precision is required.

If model calibrations are performed through the smallest model partitions possible, parameter changes should have direct impact on the outcome variables, and the confidence intervals for the estimated parameters will be small. Wide confidence intervals under these conditions generally mean that either the parameter has small impact in the selected error function, or there is significant multicollinearity among independent data series and changes in parameters do not affect the dependent variable. In either case, if more precise estimates are required, more data should be incorporated into the calibration problem to increase the efficiency of the estimates (reduce their variance). If a parameter has small impact in the selected error function, it might be necessary to obtain a data series for a variable more sensitive to its variations. Multicollinearity, on the other hand, results in estimates with large variance because only unique variations between the dependent and independent data series are used to estimate the parameters that govern their relationship. Data series in SD models normally share a time trend and tend to show strong multicollinearity, thus additional data needs to be collected to increase the power of the test. Additional data might be brought to bear upon the estimation problem by removing one parameter from the search space, formalizing the relationship between two parameters through model equations, or integrating more the data points in order to increase the variance among independent variables (Kennedy, 1992).

Given the characteristics of AC, the proposed tests (behavior match, feasibility of parameters, consistency of parameters, and confidence of tests) are incrementally demanding. However, in practice,

the process is iterative. Modifications to the estimation problem aimed at improving performance in one test frequently undermine the results of other tests that had previously been passed. The process—formulation, parameter estimation, analysis of fit, and model re-formulation—should be iterated until a specification of model structure that simultaneously matches observed structure and behavior is achieved. The following section illustrates the suggested iterative process for the calibration of a model segment used in a study for a retail lending operation of a bank. The process resulted in an improved formulation with direct consequences for the policy recommendations.

## 5. Example

During the 1990s, a major retail bank in the UK sought to cut costs by moving back-office operations from branches to centralized processing centres in more affordable locations. While the strategy to centralize and standardize the back-office operations had improved the productivity of the lending officers and the quality of the lending book, employees with direct customer contact had some concerns about the level of service provided. Our hypothesis, developed in previous work in the service industry (Senge and Sterman, 1992; Senge and Oliva, 1993), suggested that efforts to maximize throughput drove employees to work harder and, eventually, to reduce the attention given to customers. In the absence of accurate assessments of service quality and customer satisfaction, managers construe the reduction of attention given to customers as productivity gains, and, consistent with their objective of minimizing cost, reduce their estimates of required service capacity. Eroding standards mask this underinvestment in service capacity as servers, their managers, *and* customers come to expect mediocre service and justify current performance based on past performance. To explore this proposition we developed a formal model that integrates the structural elements of service delivery, as well as the goals, expectations, and choices of the actors in the situation (see Oliva and Sterman, 2001 for a full description of the model).

Table 1 lists the calibration problem for a set of equations representing the 'eroding goal' structure in which the employees' standard for time allocated per customer order (*Desired TO*) is adjusted to past performance (*Time per order*) (see Fig. 1). Consistent with previous work on adjustment of expectations (Cyert and March, 1963; Lant, 1992; Levinthal and March, 1981), the original formulation for this structure included a time constant to modify the speed of the adjustment process (*Time to adjust DTO*). The employees' standard of service is in turn used in conjunction with the *Desired order fulfillment rate* to determine the *Desired service capacity*. The relative difference between

Table 1
Calibration problem

**Minimize:**
$\sum$(Time per order $(t)$ – **Time per order(t)**)$^2$ for $\{t|$**Time per order**$(t) = \text{value}\}$
**Over:**
Initial DTO $> 0$; $\alpha < 0$; t to adjust up $> 0$; t to adjust down $> 0$
**Subject to:**
Time per order = max(Desired TO $*$ Effect of WP on TPO, Min processing TPO)
Desired TO = INTEG (DTO chg, Initial DTO)
DTO chg = (Time per order – Desired TO)/Time to adjust DTO
Time to adjust DTO = IF THEN ELSE (Time per order > Desired TO, t to adjust up, t to adjust down)
Desired service capacity = **Desired order fulfillment rate** $*$ Desired TO
Work pressure = (Desired service capacity – **Service capacity**)/**Service capacity**
Effect of WP on TPO = EXP (Work pressure $*$ $\alpha$)
Min processing TPO = 0.6

**Bold** variable names represent the historical time series for the variable.
For clarity the time subindex has been eliminated from the constraint equations.

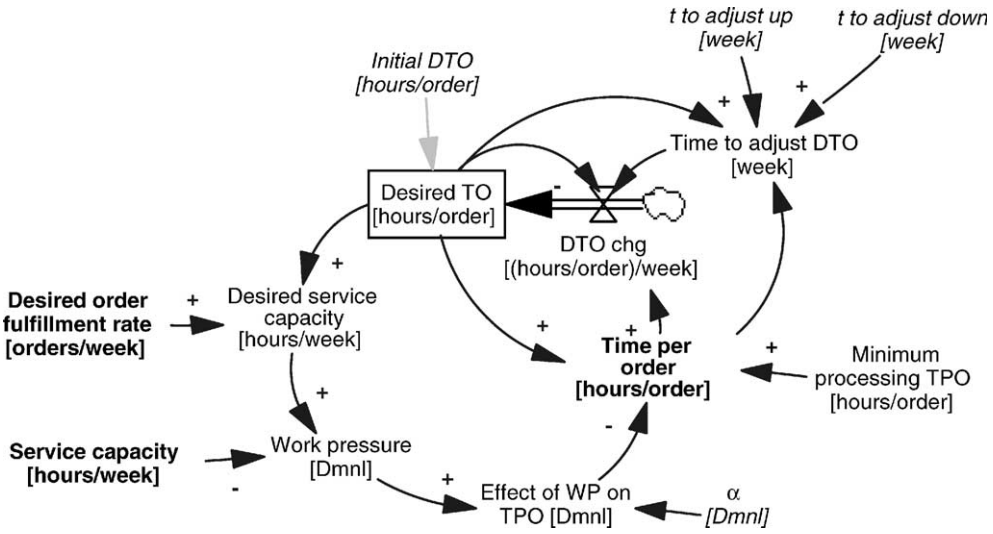


Fig. 1. Allocation of time per order.

*Desired service capacity* and *Service capacity* is defined as *Work pressure*. Finally, the actual time allocated to each order (*Time per order*) is adjusted to eliminate *Work pressure*—the higher work pressure leads to lower time per order in order to reduce the backlog of customer orders and the *Desired order fulfillment rate* (not shown in equations). The effect of *Work pressure* on *Time per order* was specified as an exponential function ($\textit{Time per order} = \mathrm{EXP}(\textit{Work pressure} \times \alpha)$), where the parameter $\alpha$ determines how aggressively employees cut *Time per order* in response to *Work pressure*.

The error function for the calibration problem was specified with *Time per order* as the only data-variable, and historical data on *Service capacity* and *Desired order fulfillment rate* (bold in Table 1 and Fig. 1) were used as known inputs. The *Desired order fulfillment rate* (incoming orders plus desired backlog adjustment) was standardized in a way that one-hour-per-standard-order represented a neutral standard of service. With this data series and the model equations it was possible to specify a calibration problem to estimate the parameters *Initial DTO*, *Time to adjust DTO*, and $\alpha$. Note that the *Minimum processing time per order* was not included in the calibration problem since this parameter was directly observable from other sources, thus reducing the size of the calibration problem. The output of the calibration problem (with 95% confidence intervals) and some tools for the analysis of fit are reported in Fig. 2A. The fit to historical behavior is quite good. 95% of the MSE is concentrated in the imperfect covariation component of the Theil inequality statistics, mean-absolute-percent-error (MAPE) is below 2%, and $R^2$ is almost 0.8. Furthermore, there is no systematic bias, and the residuals do not show a significant trend. The estimated parameters all have the expected sign and tight confidence intervals—a feasible structure. However, the value for the estimated *Initial DTO*, 10% below the neutral standard of service (0.9 hour/std. order), implied an evolution of the *Desired TO* not consistent with what was known about the system. Interviews with employees suggested a high initial standard of service, and that they had felt pressure to reduce it over the last year. When asked how they felt the lending centres were working for the bank, here is what two lending officers had to say

> ... the feedback you get back [from the customer] is "I'm dehumanized, I just became a number. I can no longer talk to you as a person, you just treat me as a number." ... we have lost the customers along the way.

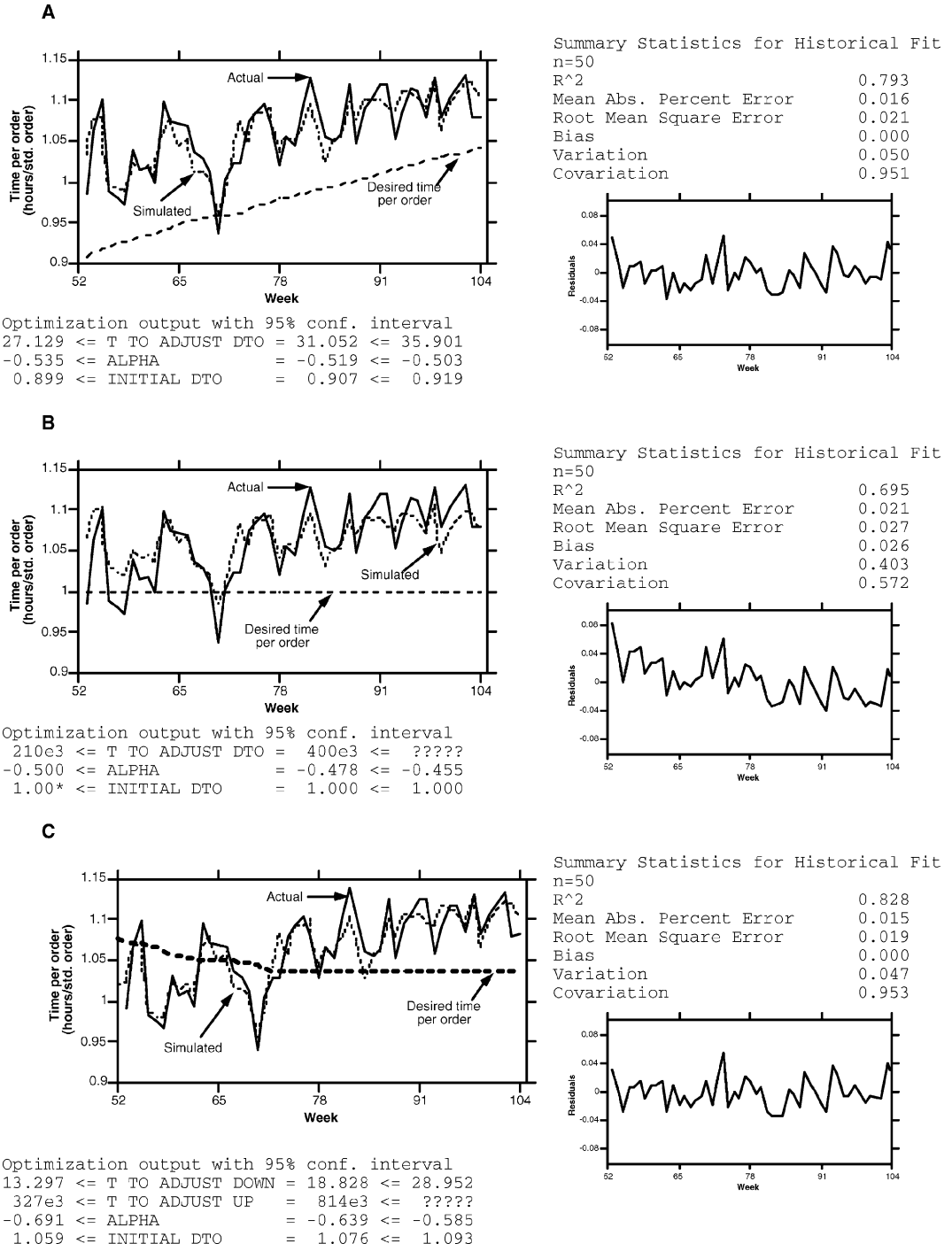


Fig. 2. Example of iterative formulation, calibration, and analysis.

> ... because of the various pressures on us, we are going to be asked to be more proactive in selling ... We just don't have the relationship basis to sell effectively. The customers have said that they become a number; and in a way they have... It is difficult to sell that way.

According to these interviews, and other indicators of service quality for the same period, the employee's standard of service should start high and erode over time (compare this to the raising value of *Desired TO* in Fig. 2A). To incorporate this additional information into the calibration problem, the search space for the *Initial DTO* parameter was limited to be greater or equal to one—a neutral standard of service. Results are reported in Fig. 2B, and, at a first glance, they are not too bad: MAPE is 2%, the bias component of the MSE is small, and $R^2$ is almost 0.7. The large estimated value of the *Time to adjust DTO* suggests that *Desired TO* does not adjust to past performance. $\alpha$ has the right sign and tight confidence intervals; again, a plausible structure. However, one parameter (*Initial DTO*) is at the limit of its feasibility range, and the residuals show a significant downward trend (also reflected in the high value for the unequal variation component of the MSE), indicating that a part of the behavior still has not been fully captured by the structure.

Further exploration of the situation, in part guided by the downward trend of the residuals, revealed that employees had experienced a significant learning curve during the period when data were collected. The effect of the learning curve was incorporated into the historical data on *Service capacity*, and multiple alternative formulations were tested through iterations of this process. A match of the historical behavior and what was known about the system's structure was finally achieved when an asymmetric adjustment process for the *Desired TO* was considered. Asymmetry of adjustments means that employees adjust their internal standard at different rates depending on whether the actual performance is above or below the internal standard. This is reflected in the formulation by allowing different time constants (*t to adjust up* and *t to adjust down*) to govern the adjustment of *Desired TO*. The results of the calibration (see Fig. 2C) show that when work pressure forces actual *Time per order* to fall below the desired level, the desired level erodes quickly—with a time constant of 19 weeks. There is, however, no evidence of any upward revision in *Desired TO* when work pressure is low (*t to adjust up* $\approx\infty$), despite the fact that actual *Time per order* exceeded *Desired TO* in more than half the data set. This formulation change had a direct impact on the policy recommendations emerging from the study once the asymmetric adjustment was identified as an erosion of quality and not a productivity gain; the new understanding on the formation of expectations created the need for aggressive management driven quality initiatives (Oliva, 2001; Oliva and Sterman, 2001).

The example was selected to show how good fit of behavior is not enough when calibrating a model (Example A), and how—despite constraints—the AC process is capable of generating fairly good matches to the observed behavior (Example B). Assessing the goodness of fit of an AC outcome requires a full consideration of the estimated parameters as they shape the model structure. Although the initial formulation provided a good behavioral match to the observed variables, the iterative testing process described above yielded an improved formulation with important implications for the behavior of the system.

## 6. Conclusions

The argument for calibration as a testing strategy for a DH has been articulated as follows. SD interventions can only be as good as the DH that is used for policy design, thus careful consideration must go into building confidence in a DH. A DH explicitly posits a causal link between structure and behavior. Although it is impossible to verify a hypothesis, science has refined a systematic approach for increasing the confidence in a stated hypothesis and ruling out alternative explanations, namely, experiments designed to falsify the hypothesis. SD models are well suited for this experimental approach since they are logically sound and need to be relevant to the problem situation, i.e., they are empirically testable.

Calibration explicitly attempts to link structure and behavior, thus making it a more stringent test than matching either structure or behavior alone.

AC techniques are capable of generating an optimal fit to historical data from a given structure and set of parameters. However, because of the assumption of correct structure and the effort to match the historical behavior, AC techniques are typically used to confirm the DH—can the structure match the observed behavior? Three heuristics are suggested to increase the power of AC as a testing tool: do not override known (observable) structure, tackle small calibration problems (modularize), and use AC to test the hypothesis: "the estimated parameter matches the observable structure of the system." Finally, a set of increasingly demanding tests were outlined to guide assessment of the AC output in the context of hypothesis testing. Ultimately, the best calibration is not the calibration that generates the best fit to historical behavior, but the calibration that best reflects the observed structural characteristics of the system and *simultaneously* captures the observed historical behavior.

Placing this paper in a broader context, it should be noted that calibration is only the first step for testing a DH and should really be viewed as part of the model building process. Forrester (1961, p. 133) argues that "confidence in a model arises from a twofold test—the defense of the components and the acceptability of over-all system behavior." The proposed testing strategy aims only to increase the defensiveness of the model components. Full testing of a DH also requires tests at the system level (Oliva, 1996), and assessment of the model's application domain (Graham et al., 2002; Oliva, 1996). Further research should look into formal ways of partitioning a model for estimation purposes, and integrating these strategies into a rigorous battery of tests for dynamic hypotheses.

## Acknowledgements

The author would like to thank David Lane, Jack Homer, and John Sterman for their helpful comments on an earlier version of this paper. I am also grateful to the anonymous referees who offered valued comments on the originally submitted form of this paper.

## References

Barlas, Y., 1989. Multiple tests for validations of system dynamics type of simulation models. European Journal of Operational Research 42 (1), 59–87.

Barlas, Y., 1994. Model validation in system dynamics. In: Wolstenholme, E., Monaghan, C. (Eds.), Proceedings of the 1994 International System Dynamics Conference. System Dynamics Society, Stirling, Scotland, pp. 1–10.

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

Bell, J.A., Bell, J.F., 1980. System dynamics and scientific method. In: Randers, J. (Ed.), Elements of the System Dynamics Method. Productivity Press, Cambridge, MA, pp. 3–22.

Bell, J.A., Senge, P.M., 1980. Methods for enhancing refutability in system dynamics modeling. TIMS Studies in the Management Sciences 14 (1), 61–73.

Bunge, M., 1967. Scientific Research. Springer-Verlag, New York.

Checkland, P.B., 1981. Systems Thinking, Systems Practice. Wiley, New York.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Houghton Mifflin, Boston.

Cyert, R., March, J., 1963. A Behavioral Theory of the Firm. Prentice Hall, Englewood Cliffs, NJ.

Eden, C., 1990. Part III. Mixing methods—Introduction. In: Eden, C., Radford, J. (Eds.), Tackling Strategic Problems. Sage Publications, London, pp. 90–91.

Flew, A. (Ed.), 1984. A Dictionary of Philosophy, second ed. St. Martin's Press, New York.

Forrester, J.W., 1961. Industrial Dynamics. MIT Press, Cambridge, MA.

Forrester, J.W., 1971. The model versus a modeling process. System Dynamics Review 1 (2), 133–134.

Forrester, J.W., 1979. An alternative approach to economic policy: Macrobehavior from microstructure. In: Kamrany, N.M., Day, R.H. (Eds.), Economic Issues of the Eighties. John Hopkins University Press, Baltimore, pp. 80–108.

Forrester, J.W., 1994. Policies, decisions, information sources for modeling. In: Morecroft, J.D.W., Sterman, J.D. (Eds.), Modeling for Learning Organizations. Productivity Press, Cambridge, MA, pp. 51–84.

Forrester, J.W., Senge, P.M., 1980. Tests for building confidence in system dynamics models. TIMS Studies in the Management Sciences 14, 209–228.

Gass, S.I., 1983. Decision-aiding models: Validation, assessment and related issues for policy analysis. Operations Research 31 (4), 603–631.

Graham, A.K., 1980. Parameter estimation in system dynamics modeling. In: Randers, J. (Ed.), Elements of the System Dynamics Method. Productivity Press, Cambridge, MA, pp. 143–161.

Graham, A.K., 2002. On positioning system dynamics as an applied science of strategy. In: Davidsen, P.I. (Ed.), Proceedings of the 2002 International System Dynamics Conference. System Dynamics Society, Palermo, Italy.

Graham, A.K., Choi, C.Y., Mullen, T.W., 2002. Using fit-constrained Monte Carlo trials to quantify confidence in simulation model outcomes. In: Proceedings of the 35th Hawaii Conference on Systems Science. IEEE, Big Island, HI.

Greene, W.H., 1997. Econometric Analysis, third ed. Prentice Hall, Upper Saddle River, NJ.

Homer, J.B., 1983. Partial-model testing as a validation tool for system dynamics. In: Proceedings of the1983 International System Dynamics Conference. System Dynamics Society, Chestnut Hill, MA, pp. 919–932.

Homer, J.B., 1996. Why we iterate: Scientific modeling in theory and practice. System Dynamics Review 12 (1), 1–19.

Homer, J.B., 1997. Structure, data and compelling conclusions: Notes from the field. System Dynamics Review 13 (4), 293–309.

Kennedy, P., 1992. A Guide to Econometrics, third ed. MIT Press, Cambridge, MA.

Kleijnen, J.P.C., 1995. Sensitivity analysis and optimization of system dynamics models: Regression analysis and statistical design of experiments. System Dynamics Review 11 (4), 275–288.

Kleijnen, J.P.C., Sargent, R.G., 2000. A methodology for the fitting and validation of metamodels in simulation. European Journal of Operational Research 120 (1), 14–29.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.

Lakatos, I., 1974. Falsification and the methodology of scientific research programmes. In: Lakatos, I., Musgrave, A. (Eds.), Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge, pp. 91–196.

Lane, D.C., 1999. Social theory and system dynamics practice. European Journal of Operational Research 113 (3), 501–527.

Lant, T., 1992. Aspiration level adaptation: An empirical exploration. Management Science 38 (5), 623–644.

Levinthal, D., March, J., 1981. A model of adaptive organizational search. Journal of Economic Behavior and Organization 2 (4), 307–333.

Long, J.S., 1997. Regression Models for Categorical and Limited Dependent Variables. SAGE, Thousand Oaks, CA.

Lyneis, J.M., Pugh, A.L., 1996. Automated vs. 'hand' calibration of system dynamics models: An experiment with a simple project model. In: Richardson, G.P., Sterman, J.D. (Eds.), Proceedings of the 1996 International System Dynamics Conference. System Dynamics Society, Cambridge, MA, pp. 317–320.

Martinez, I.J., Richardson, G.P., 2001. Best practices in system dynamics modeling. In: Hines, J.H., Diker, V.G. (Eds.), Proceedings of the 2001 International System Dynamics Conference. System Dynamics Society, Atlanta.

Mass, N.J., Senge, P.M., 1980. Alternative test for selecting model variables. In: Randers, J. (Ed.), Elements of the System Dynamics Method. Productivity Press, Cambridge, MA, pp. 203–223.

Miller, J.H., 1998. Active nonlinear tests (ANTs) of complex simulation models. Management Science 44 (6), 820–830.

Miser, H.J., 1993. A foundational concept of science appropriate for validation in operational research. European Journal of Operational Research 66 (2), 204–234.

Mitroff, I., 1972. The myth of objectivity or why science needs a new psychology of science? Management Science 18 (10), B613–B618.

Morecroft, J.D.W., 1983. System dynamics: Portraying bounded rationality. Omega 11 (2), 131–142.

Morecroft, J.D.W., 1985. Rationality in the analysis of behavioral simulation models. Management Science 31 (7), 900–916.

Morecroft, J.D.W., 1988. System dynamics and microworlds for policymakers. European Journal of Operational Research 59 (1), 9–27.

Oliva, R., 1995. A Vensim module to calculate summary statistics for historical fit. D-4584, System Dynamics Group, Massachusetts Institute of Technology, Cambridge, MA. Available from <http://www.people.hbs.edu/roliva/research/sd/>.

Oliva, R., 1996. Empirical validation of a dynamic hypothesis. In: Richardson, G.P., Sterman, J.D. (Eds.), Proceedings of the 1996 International System Dynamics Conference. System Dynamics Society, Cambridge, MA, pp. 405–408.

Oliva, R., 2001. Tradeoffs in responses to work pressure in the service industry. California Management Review 41 (4), 26–43.

Oliva, R., Sterman, J.D., 2001. Cutting corners and working overtime: Quality erosion in the service industry. Management Science 47 (7), 894–914.

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263, 641–646.

Peterson, D.W., 1980. Statistical tools for system dynamics. In: Randers, J. (Ed.), Elements of the System Dynamics Method. Productivity Press, Cambridge, MA, pp. 143–161.

Raiffa, H., Schlaifer, R., 1961. Applied Statistical Decision Theory. Division of Research, Harvard Business School, Boston.

Reichelt, K.S., Lyneis, J.M., Bespolka, C.G., 1996. Calibration statistics: Selecting a statistic and setting a standard. In: Richardson, G.P., Sterman, J.D. (Eds.), Proceedings of the 1996 International System Dynamics Conference. System Dynamics Society, Cambridge, MA, pp. 425–428.

Richardson, G.P., Pugh, A.L., 1981. Introduction to System Dynamics Modeling with DYNAMO. MIT Press, Cambridge, MA.

Roy, B., 1993. Decision science or decision-aid science? European Journal of Operational Research 66 (2), 184–203.

Schweppe, F.C., 1973. Uncertain Dynamic Systems. Prentice Hall, Englewood Cliffs, NJ.

Senge, P.M., 1977. Statistical estimation of feedback models. Simulation 28 (June), 177–184.

Senge, P.M., 1978. The system dynamics investment function: A comparison of the neoclassical investment function. PhD Thesis, Sloan School of Management, MIT, Cambridge, MA.

Senge, P.M., 1990. The Fifth Discipline: The Art and Practice of the Learning Organization. Doubleday Currency, New York.

Senge, P.M., Oliva, R., 1993. Developing a theory of service quality/service capacity interaction. In: Zepeda, E., Machuca, J.A.D. (Eds.), Proceedings of the 1993 International System Dynamics Conference. System Dynamics Society, Cancun, Mexico, pp. 476–485.

Senge, P.M., Sterman, J.D., 1992. Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. European Journal of Operational Research 59 (1), 137–150.

Smith, J.H., 1993. Modeling muddles: Validation beyond the numbers. European Journal of Operational Research 66 (2), 235–249.

Sterman, J.D., 1984. Appropriate summary statistics for evaluating the historical fit of system dynamics models. Dynamica 10 (Winter), 51–66.

Sterman, J.D., 1989. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Management Science 35 (3), 321–339.

Sterman, J.D., 1994. Learning in and about complex systems. System Dynamics Review 10 (2–3), 291–330.

Sterman, J.D., 2000. Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin McGraw-Hill, Boston.

Theil, H., 1966. Applied Economic Forecasting. North-Holland Publishing, Amsterdam.

van Horn, R.L., 1971. Validation of simulation results. Management Science 17 (5), 247–258.

Wilson, J.R., 1997. Doctoral colloquium keynote address: Conduct, misconduct, cargo cult science. In: Andradóttir, S., Healy, K.J., et al. (Eds.), Proceedings of the 1997 Winter Simulation Conference. IEEE, Piscataway, NJ, pp. 1405–1413.

Wolstenholme, E.F., 1990. System Enquiry: A System Dynamics Approach. Wiley, New York.