The Case for Thoroughly Testing Complex System Dynamics Models

Wayne Wakeland[1] and Megan Hoarfrost[2]
[1] Systems Science Ph.D. Program, Portland State University
[2] Stanford University
[1] P.O. Box 751, Portland, OR 97207
(503) 725-4975(v) (503) 725-8489 (f)
wakeland@pdx.edu

Abstract
In order to determine whether model testing is as useful as suggested by modeling experts, the full battery of model tests recommended by Forrester, Senge, Sterman, and others was applied retrospectively to a complex previously-published system dynamics model. The time required to carry out each type of test was captured, and the benefits that resulted from applying each test was determined subjectively. The resulting benefit to cost ratios are reported. These ratios suggest that rather than focusing primarily on sensitivity testing, modelers should consider other types of model tests such as extreme condition tests and family member tests. The study also finds that all of the different kinds of tests were either moderately useful or very useful--fully supporting the recommendations of the experts.

Keywords
Model testing, model validation, model verification, sensitivity analysis

This research was sponsored in part by the Thrasher Research Fund.

Introduction

Is it really practical and worthwhile to run all of the model tests advocated by system dynamic modeling experts such as Jay Forrester and Peter Senge (1980), George Richardson (Richardson and Pugh 1981), John Sterman (2000), Yaman Barlas (1996), and others. Although some modelers meticulously apply the full battery of recommended tests to every model they develop, we suspect that most do not. Why? One reason might be lack of time. Another reason might that is seems unlikely that the benefits of running more than a few tests would offset the cost of performing the additional tests. But is this really true?

To answer this question, we torture-tested a previously published model according to the guidelines recommended by John Sterman (2000). This model had already been tested to a certain degree, and appeared to have sufficient "face validity" to warrant its being used to analyze real world situations.

The significance of this study relates to the fact that the *benefit* of a particular model test may or may not be strongly correlated with its *cost* (the amount of time it takes to conduct the test). Knowing which tests are efficient in terms of their benefit to cost ratio would help users to select the best tests to perform when time constraints preclude running all of the recommended tests. As Sterman points out in a recent article (2002 pg. 521), "Many important tests are simply never done."

Background

Although Sterman (2000) emphasizes that "the word validation should be struck from the vocabulary of modelers" he certainly does not mean to imply that models should not be thoroughly tested, but rather that the notion of model *validity* per se is problematic. Modelers must do everything possible to verify that their models have been correctly implemented per their intentions, and further, they must thoroughly test their models well beyond their design limits in order to determine the model's useful domain of applicability.

Forrester and Senge published a seminal paper on the testing of system dynamics models in 1980. This classic paper described a battery of specific tests for building confidence in system dynamics models. These 17 tests were grouped into tests of model structure, tests of model behavior, and tests of policy implications. At about this same time, Richardson and Pugh (1981) published their system dynamics modeling textbook. This widely used textbook featured a chapter on model testing that encouraged modelers to deactivate feedback loops, to use test functions to disturb models from equilibrium, to conduct hypothesis tests, and to do sensitivity analysis. The latter was divided into numerical, behavioral, and policy sensitivity. Richardson and Pugh provided a table that organized model tests according to model suitability, model consistency, and model utility on one hand, and model structure vs. model behavior on the other hand. This table contains many of the same tests prescribed by Forrester and Senge. Tank-Nielsen (1980) provides further discussion on sensitivity analysis, and Mass and Senge (1980) show how model behavior tests influence the selection of model variables.

Barlas (1996) divided model tests into structure validity and behavior validity, and emphasized that the purpose of the model strongly influences the notion of model validity—echoing Zeigler (1976). Barlas further divides structure tests into direct structure tests, that do not require simulation, and structure-oriented behavior tests that do involve simulation. Structure tests may be empirical, theoretical, or implementation-oriented. The latter include such tests as formal inspections, walkthroughs, and semantic analysis. Behavior tests include extreme conditions tests, behavior sensitivity tests, modified behavior prediction, boundary adequacy, phase relationship test, qualitative features analysis, and the Turing test.

Sterman relied heavily on these prior works when he wrote the chapter on model testing in his now classic textbook on business dynamics (Sterman 2000). The model tests recommended by Sterman are summarized in the next section.

No discussion of model testing would be complete without recognizing that testing occurs in various forms throughout the entire modeling process, as emphasized by Randers (1980) in his influential paper on model conceptualization that clarifies the highly iterative even recursive nature of the model construction process. He also discusses the role of the reference [behavior] mode as a guide to model structure. Luna-Reyes and Anderson (2003) review the literature regarding the system dynamics modeling process and discuss how qualitative methods may be used to strengthen each aspect of the process, including model testing. Relevant methods for model testing include interviews, focus groups, Delphi groups, and experimental approaches. These methods facilitate model assessment by domain experts.

Method

We chose to apply each of the model tests summarized in Table 21-4 in Sterman (2000, 859-61). Sterman indicates that this material is adapted and extended from Forrester and Senge (1980). The tests are as follows:

1. Boundary Adequacy: Does the selection of what is endogenous, exogenous, and excluded make sense?
2. Structure Assessment: Is the level of aggregation correct, and does the structure conform to reality?
3. Dimensional Consistency: Do the units of the model make sense, and does the model avoid the use of arbitrary scaling factors?
4. Parameter Assessment: Do parameter have real life meanings, and are their values properly estimated?
5. Extreme Conditions: Do extreme parameter values lead to irrational behavior?
6. Integration Error: Does the behavior change when the integration method or time step are changed?
7. Behavioral Reproduction: How well does the model behavior match the behavior of the real system?
8. Behavior Anomaly: Does changing the loop structure lead to anomalous behavior consistent with the changes?
9. Family Member: How well does the model "scale" to other members within the same class of systems?
10. Surprise Behavior: What is revealed when model behavior does not match expectations?
11. Sensitivity Analysis: Do conclusions change in important ways when assumptions are varied over their plausible range? Changes in conclusions could be numerical changes, behavior mode changes, or policy changes.
12. System Improvement: Does the model generate insights that actually lead to the hoped for improvements?

Our approach was to:

A. Perform each of the recommended tests as fully as possible, without regard for the time required,
B. Carefully document the results, including the time required for each test (its cost),
C. Subjectively assess the benefits of each test, and, as appropriate, capture the number of insights gained from the test and the number of model changes inspired by the test, and
D. Contrast the tests in terms of their benefit/cost.

This process is admittedly highly subjective, and therefore the results must be considered exploratory at best. We share them to obtain critical feedback and because so few comparisons of different validation methods are available in the literature.

The target model analyzes intracranial fluid volumes and flows in order to study intracranial pressure (ICP) dynamics for patients with traumatic brain injury. The model is designed to simulate multiple pathophysiologies and treatment options (Wakeland and Goldstein, 2005). To

give a sense for the complexity of the model, the model diagram is provided in Figure 1 (Note: the details of the model are not essential to the present paper).

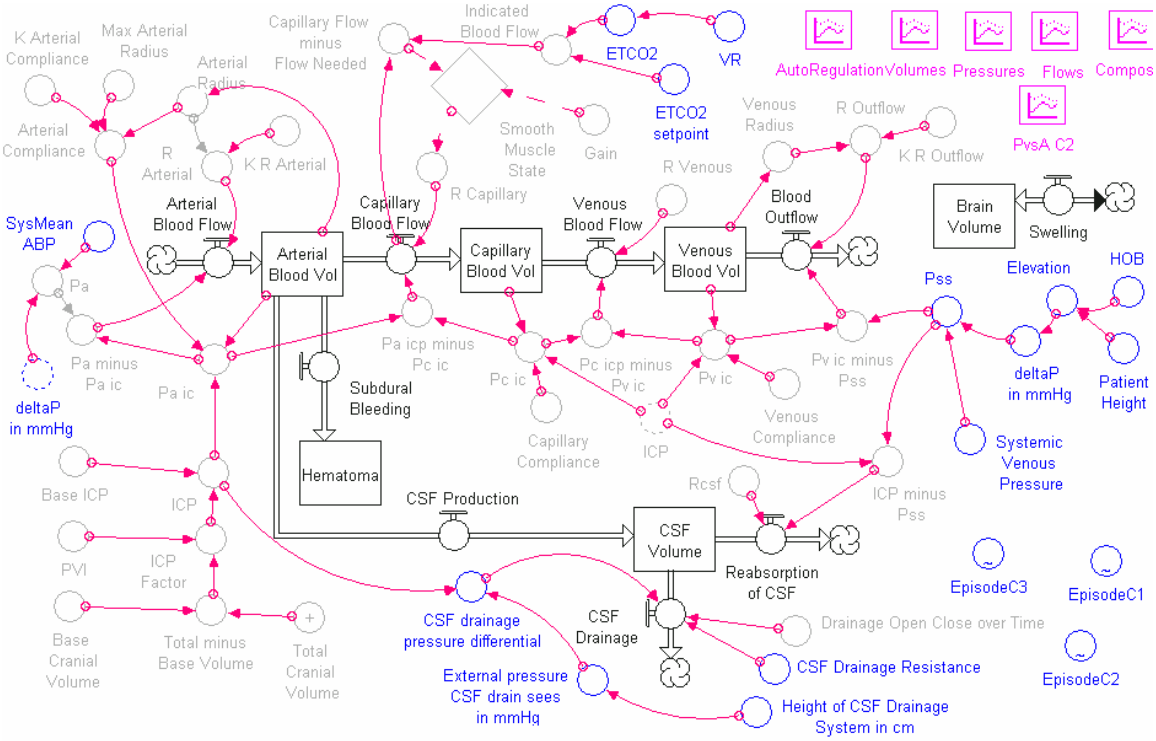

Figure 1. Diagram of subject model of ICP dynamics

We felt that this type of model was a good candidate for studying the model testing process for two main reasons. First, the physiology represented in the ICP model is complex enough to require very accurate integration. This is because the flow rates are very high relative to the size of the storages. A small error in the calculated flow could cause the stock to empty in a single time step if the time step is too large. Second, the parameters in the model must be carefully balanced in order initialize the model in steady state. Consequently the model seemed to be fragile rather than robust.

Figure 2 shows the feedback structure, with self loops excluded. The structure is relatively simple, but tightly coupled, with many parameters influencing the gain of multiple loops. There are thirteen feedback loops: three self loops, where state variables directly influence their own rate of change, six simple loops involving two state variables, three loops involving three state variables, and one loop that includes four of the state variables. We will refer to the loop analysis later in the results section.

Results

The results are presented in the order that the tests were performed. The time required to perform each test is listed in parentheses after the heading for each test.

Boundary Adequacy Tests (~2 hours, estimated)

During the initial model design decisions were made regarding what would be excluded

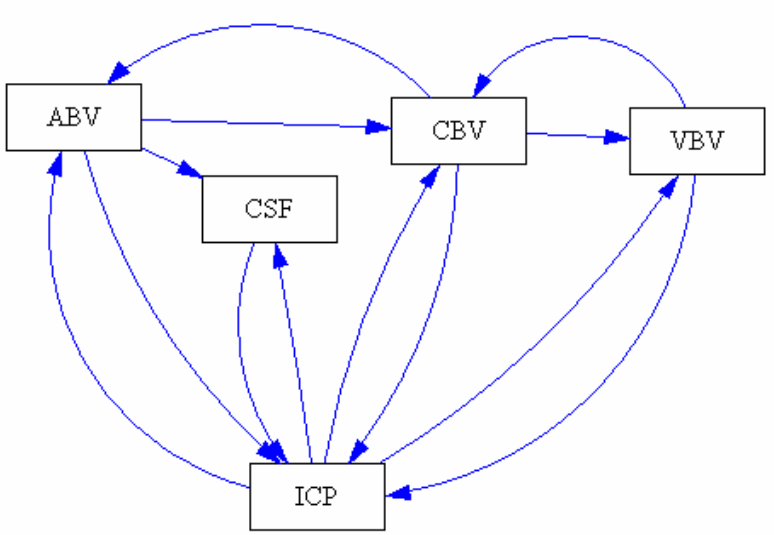

Figure 2. Feedback structure of subject model, with self loops excluded.

entirely, what would be an exogenous input, and what would be calculated endogenously? These decisions were reviewed and deemed to be appropriate. The benefits from these tests were moderate.

Integration Error Tests (<1 hour, estimated)

If model behavior changes when DT is reduced, then the original DT was too large and cannot be used. Through initial integration error tests, we determined that the ICP model is very sensitive to DT an integration method. Consequently we chose to use 4^{th} order Runge Kutta and the smallest possible DT value allowed by the simulation package, 0.000125. These tests were not time-consuming and were essential for assuring accurate integration of the equations. The benefits from these tests were judged to be high.

Structure Assessment Tests (2 hours, estimated)

Most of the structure assessment tests were done during the modeling process, which is consistent with Sterman's recommendation that modelers test their model as they build it. Structure assessment tests are vital to creating an insightful model. These tests involved acquiring information from outside experts on intracranial physiology. One example of a structure assessment test that was done later involved testing several representations for arterial pressure: 1) a constant, 2) a repeating graphical function, and 3) a sine wave. These tests showed that although the model was initially designed to use a constant mean arterial pressure (MAP), the model behavior remained credible when the MAP constant was replaced by periodic signals similar to the pulsatile arterial blood pressure signal. The benefits from these tests were moderate.

Parameter Assessment (5 hours, estimated)

As soon as we had a model structure that resembled the target system in pertinent ways, exogenous parameters were adjusted to create model behavior that matched reference data. This process was essential, and the benefits from performing these tests were high.

Behavioral Reproduction Tests (12 hours)

Early testing also included behavioral reproduction tests. These tests were mostly utilized to test whether the model is able to reproduce the range of behaviors embodied in reference data recorded for real patients. One of the first trials was to reproduce impact of draining cerebrospinal fluid (CSF), a standard treatment for lowering ICP during brain trauma. Estimates for the amount of CSF drainage that corresponded to a given degree of ICP reduction were reviewed by clinicians for reasonableness, since the actual amount of fluid drained is not routinely captured in the Intensive Care Unit. The benefits from performing these tests were high, and resulted in several insights and model changes.

Behavior Anomaly Tests (5 hours)

Behavior anomaly tests were done at the same time as the behavioral reproduction tests. Behavioral anomaly tests are beneficial because they show that the model is able to correctly simulate abnormal (pathophysiological) behavior as well as normal behavior. In this case, the elevation of ICP during traumatic brain injury is often caused by subdural or epidural bleeding. Model refinements were required to obtain correct behavior in the model when these different pathophysiologies were incorporated. For example the logic for the flow of blood from the venous cavity during elevated ICP was initially too simple and had to be modified. This led to additional structure assessment tests, parameter tests, etc. The benefits from performing these tests were high, and led to multiple insights and model changes.

Sensitivity Analysis (37 hours)

The majority of testing time was spent on sensitivity analysis. We chose to do this testing manually for two reasons. First, we gained greater knowledge about the fundamentals of the model by carrying out each test manually. Second, there were so many outcome variables to monitor in response to changes in input parameters that it was easier to do the analysis by making the parameter changes individually and manually. We observed the impact by watching five different variables: ICP, venous blood volume, CSF volume, capillary blood flow and arterial blood flow. Note that some authors recommend varying multiple parameters simultaneously, we chose to vary parameters one at a time.

In most of the sensitivity tests, it was observed that with an increase in ICP, venous blood volume, CSF volume, and capillary and arterial blood flows decreased by roughly the same intensity. ICP is an important factor of the real system and is related to most of the parameters in the model, so it is a good general indication of how the model reacts to changes. Figure 3 shows the impact on peak ICP when several different model parameters were increased and decreased by 20%. Two parameters could not be decreased by 20% without making other changes in the model. This result made sense physiologically.

One might expect that the influence of a parameter might be related to the number of feedback loops that the parameter influences. Table I contains a row for each parameter shown in Figure 3, ordered by their influence on model behavior. There are seven columns, one for each group of parameters, where a group is defined as a set of parameters that influence the same feedback loops. The columns are ordered from left to

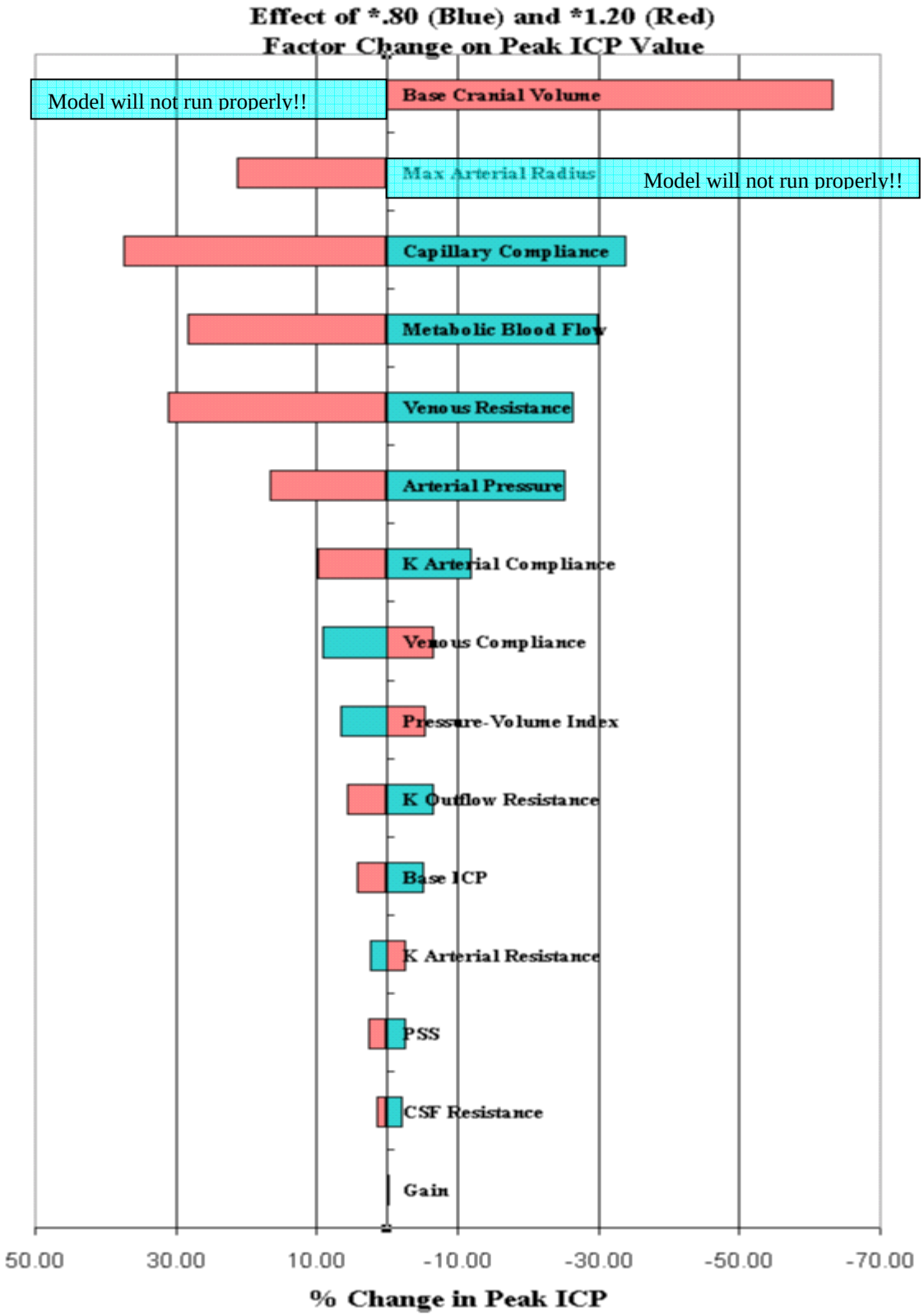

Figure 3. A "tornado" diagram that shows the impact of on Peak ICP value when selected parameters are increased and decreased 20%. Note that for the two parameters at the top of the diagram, the model will not tolerate a 20% decrease.

Table I. Correlation between parameter influence and number of loops. Columns are ordered by the number of loops impacted by each group of parameters. X's indicate group membership. A diagonal pattern would indicate high correlation.

Parameter group:	G1	G2	G3	G4	G5	G6	G7
Number of loops influenced:	12	10	8	6	4	2	1
Base Cranial Volume			X				
Max Arterial Radius				X			
Capillary Compliance	X						
Metabolic Blood Flow		X					
Venous Resistance					X		
Arterial Pressure				X			
K Arterial Compliance		X					
Venous Compliance					X		
Pressure-Volume Index			X				
K Outflow Resistance						X	
Base ICP			X				
K Arterial Resistance				X			
PSS						X	
CSF Resistance							X
Gain		X					

right based on of the number loops its members influence, 12 for G1 (Group 1), 10 for G2, etc. The "X" in each row indicates the group to which that parameter belongs. If the influence of a parameter were highly correlated with the number of loops that it impacts, the X's would be positioned near the diagonal, and the X's in a given column would be clustered together. Table I indicates that the correlation is weak, a result that was unexpected.

In general, the parameters to which the model was highly sensitive were consistent with prior expectations in terms of the model equations, and the testing reinforced the need to accurately measure or estimate these particular parameters. Interestingly, although the change in peak ICP value changed significantly in response to the many of the parameter changes, the time that it took for ICP to reach a new steady state value changed very little. This may indicate that the time constants implicit in the model are a systemic property rather than a specific parameter value.

Despite the significant time investment, the benefits from conducting these tests were moderate, and led to few new insights and no model changes. However, the results, when presented in a format that is easily understood by potential model users, could help to inspire their confidence. In the future we would scale back, but not eliminate, sensitivity testing. We would also consider varying multiple parameters simultaneously, as has been recommended by other researchers.

Extreme Condition Tests (9 hours)

Extreme condition tests were mainly carried out in parallel with sensitivity analysis. Upper and lower limits for every exogenous factor were established, with all other factors remained constant. Table II shows the lower and upper limits for fourteen parameters. This information is highly correlated with the sensitivity analysis results, as

one might expect. The limiting values for parameters with high sensitivity tend to be much smaller than for parameters with low sensitivity. When the value for a parameter is set outside of its limits, the model behavior cannot be relied upon. The benefits from performing these tests were high, and lead to multiple insights, including clarification of the model's useful domain of applicability.

Table II. Range of acceptable values for fourteen model parameters. The table is sorted in descending order, with the range of parameter at the top being the most restrictive. Parameters 1 is restricted in both directions, whereas parameters 2-4 are primarily restricted in terms of how much they can be reduced. Parameters 5-8 are more restricted in terms of how much they can be increased.

	Factor	Lower Boundary (As a fraction of the baseline value)	Upper Boundary (As a fraction of the baseline value)
1	Venous Resistance	0.80	1.75
2	Max Arterial Radius	0.82	3.67
3	K Arterial Resistance	0.77	5.77
4	Arterial Pressure	0.85	no limit
5	PSS	0.00	1.27
6	Pressure-Volume Index	0.07	1.45
7	Gain	0.00	1.59
8	K Arterial Compliance	0.04	1.94
9	Base Cranial Volume	0.00	6.40
10	Base ICP	0.01	8.41
11	K Outflow Resistance	0.00	15.00
12	Capillary Compliance	0.00	34.50
13	Venous Compliance	0.25	no limit
14	Metabolic Blood Flow	0.02	no limit

Surprise Behavior Tests (5 hours, estimated)

We encountered surprise behavior as we ran sensitivity analysis tests and extreme condition tests. Before making a run, we would hypothesize what should happen when we pushed the "run" button. When the behavior surprised us, we had to either alter the model or adjust our mental model after a careful comparison of the model behavior and the hypothesized behavior. For example, while increasing the outflow resistance constant during a sensitivity analysis, both ICP and venous blood volume increased in response. Until that point, ICP and venous blood volume had never increased or decreased in the same direction in response to a change. This result was inconsistent with our initial hypothesis--that a change in ICP would cause an opposite direction change in venous blood volume and vise-versa. Upon deeper reflection, we realized that it did indeed make sense that greater outflow resistance would cause less outflow and therefore greater venous blood volume; and that this would increase cranial volume and therefore increase ICP. So it was our mental model that needed to be altered. And we learned something about the robustness of the model from this "surprise" behavior. The benefits from performing these tests were high.

Family Member Tests (2 hours)

Like extreme condition tests, family member tests can identify problems in the model without being particularly time-consuming. When we tried to run the model with a smaller or larger patient in mind, it was not as simple to do as we had hoped. In addition to changing the blood volumes and base cranial volume, there were several other changes that had to be made in order for the model to run correctly when scaled. This particular test showed that many design changes would be needed to make the model easy to scale. The benefits resulting from these tests were high.

Dimensional Consistency (1 hour)

Sterman recommends testing dimensional consistency early on in the modeling process because it is generally straightforward, and can help to identify structural problems and logic flaws even before they are incorporated into the model. However, despite knowing this when we built the model, we elected not to initially utilize the dimensional analysis feature provided in the simulation package. This was an error in judgment, as we found that several conversion factors were missing, and that several equations included scaling factors with no real world meaning. This problem could have been avoided had we taken dimensional consistency more seriously during the model construction phase. The benefits from doing these tests [after the fact] were moderate.

System Improvement Tests (15 hours, estimated)

Later, logic was added to the model to reflect various physiologic challenges that were part of an approved research protocol. The challenges included changing the head of the bed and changing the respiration rate. Data for specific subjects that underwent the research protocol was used to estimate model parameters, by minimizing the mean absolute deviation between predicted ICP and actual ICP. During this process we learned that although the autoregulation logic in the model was able to replicate the response to changes in the head of the bed, it could not replicate the response to changes in the respiration rate. This will require model improvements. That such improvements will be needed was not a surprise. The benefits from these tests were high.

Table III summarizes the model testing results. The cost in terms of time has been categorized as low medium and high, with low being three or less hours, medium being four to ten hours, and high being more than ten hours.

Discussion

Although the result shown in Table I that parameter influence is only weakly correlated to the number of loops impacted by the parameter was initially surprising, the reason for this finding is quite simple. The influence of a given parameter depends almost entirely upon its specific mathematical relationship in the model. For example, the most highly influential parameter, base cranial volume, earned that distinction because in the model, a quantity with a similar value is subtracted from base cranial volume. Therefore small percentage changes in its value are amplified in the model.

Table III: Summary of Costs and Benefits for Different Model Tests

TEST	COST	BENEFIT	BENEFIT/COST
Integration Error	Low	High	Very High
Extreme Condition	Low	High	Very High
Family Member	Low	High	Very High
Structure Assessment	Low	Moderate	High
Dimensional Consistency	Low	Moderate	High
Boundary Adequacy	Low	Moderate	High
Parameter Assessment	Medium	High	Medium High
Behavioral Anomaly	Medium	High	Medium High
Surprise Behavior	Medium	High	Medium High
Behavioral Reproduction	High	High	Medium
System Improvement	High	High	Medium
Sensitivity	High	Moderate	Low

Table III indicates that every model test was found to be at least moderately useful. This, not surprisingly, fully supports what master modelers have been recommending for decades. Table II also indicates that despite the popularity of sensitivity analysis, there are a number of other model tests that merit attention. This is consistent with results found in another domain (software process modeling using hybrid simulation) that in order to reveal the nonlinearities in a model one must employ *broad range sensitivity analysis* (Wakeland et al 2004) rather than the more typical sensitivity analysis that focuses on relatively small perturbations near the nominal values of parameters. Broad range sensitivity analysis varies parameters over their full range of plausible values.

Although we do not advocate skipping any of the tests, modelers with a limited time budget would be well advised to take advantage, for example, of the extreme condition tests and the family member tests, since they are likely to yield insights without requiring a considerable amount of time. The answer to the question we set out to answer is a resounding, "Yes, it **does** make sense to thoroughly test complex system dynamics models!"

However, this is an extremely preliminary study, based one application of the method to a single target model. While the cost measure is reasonably objective, the benefit measure is highly subjective. The methodology must be sharpened to better measure the actual benefits from the tests, and must then be applied to many more target models. It is very likely that the results will vary considerably from model to model.

Acknowledgement

The authors very much appreciate the many helpful suggestions and additional references provided by the anonymous reviewers.

References

Barlas, Yaman. 1996. Formal Aspects of Model Validity and Validation in System Dynamics. *System Dynamics Review 13*(3):183-210.

Forrester, Jay, and Peter Senge. 1980. Tests for building confidence in S-D models. *TIMS Studies in the Management Sciences 14*:208-228.

Kambiz E. Maani, and Robert Y. Cavana. 2000. *Systems Thinking and Modelling: Understanding Change and Complexity.* Prentice Hall Pearson Education, New Zealand.

Luna-Reyes, Luis Phillip, and Deborah Lines Anderson. 2003. Collecting and analyzing qualitative data for system dynamics models: methods and models. *System Dynamics Review 19*(4):271-296.

Mass, Nathaniel, and Peter Senge. 1980. Alternative Tests for Selecting Variables, in Randers (ed.). *Elements of the System Dynamics Method.* Productivity Press, Cambridge, Mass.

Randers, Jorgen. 1980. Guidelines for Model Conceptualization, in *Elements of the System Dynamics Method.* Productivity Press, Cambridge, Mass.

Richardson, George, and Alexander Pugh. 1981. *Introduction to System Dynamics Modeling with DYNAMO.* MIT Press, Cambridge, MA.

Sterman, John. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World.* Irwin McGraw-Hill.

Sterman, John. 2002. Reflections on becoming a systems scientist. *System Dynamics Review 18*(4):501-530.

Tank-Nielson, Carsten. 1980. Sensitivity Analysis in System Dynamics, in Randers (ed.). *Elements of the System Dynamics Method.* Productivity Press, Cambridge, Mass.

Wakeland, Wayne, Robert Martin, and David Raffo. 2004. Using design of experiments, sensitivity analysis, and hybrid simulation to evaluate changes to a software development process: a case study. *Software Process Improvement and Practice 9*(2):107-119.

Wakeland, Wayne, and Brahm Goldstein. 2005. A computer model of intracranial pressure dynamics during traumatic brain injury that explicitly models fluid flows and volumes. *Acta Neurochirurgica* (in press).

Zeigler, Bernard P. 1976. *Theory of modelling and simulation*, Wiley, New York.