Contributions to model validation: hierarchy, process, and cessation

Stefan N. Groessera,b,c* and Markus Schwaningerb

Abstract
In the domain of dynamic modeling and simulation, the assurance of model validity is a prominent challenge. An extensive number of contributions concerning model tests, terminology, and the epistemological foundations of validation have been elaborated. These contributions, however, do not fully answer the questions for novice modelers, namely, which validation tests to choose, when and how to apply them, and at what point to cease their formal validation efforts. Our intention here is to help close this gap by introducing a complexity hierarchy of validation tests, an integrative validation process, and a decision heuristic about when to stop formal validation efforts. The paper concludes by providing directions for future research. Copyright © 2012 System Dynamics Society.

Introduction
Validity is a critically important issue in every discipline of science. Model validity—the property a model has of adequately reflecting the system being modeled, contingent on the model's purpose (Kleindorfer and Geneshan, 1993)—is a primary measure of model utility and effectiveness. Validation efforts gradually establish, on the one hand, trust and confidence in a qualitative or quantitative model (Forrester and Senge, 1980; Balci, 1994; Barlas, 1996; Sargent, 2008); on the other hand, one must accept the impossibility of absolute validity or confidence (Zeigler et al., 2000; Sterman, 2002). Given that the quality of simulation models currently in use is in most cases unsatisfactory (Forrester, 2007), there exists a pressing need for improvement.

From the beginning of system dynamics modeling until today, the discussions about validation in the literature have been extensive and have revolved mainly around the following topics: validation tests, their description, and categorization (see Forrester, 1961; Forrester and Senge, 1980; Richardson and Pugh, 1981; Eberlein and Wang, 1983; Barlas, 1989, 1990, 1996; Balci, 1994; Coyle and Exelby, 2000; Sterman, 2000; Saysel and Barlas, 2006; Schwaninger and Groesser, 2009), the philosophical roots and epistemological underpinnings of model validation (Barlas and Carpenter, 1990; Déry et al., 1993; Schwaninger and Groesser, 2008, 2009), terminological clarifications (Gass, 1983; Finlay and Wilson, 1997; Zeigler et al., 2000; Sargent, 2008), validation processes (Oral and Kettani, 1993; Kleijnen, 1995; Barlas, 1996; Coyle and Exelby, 2000), and the objectives of validation (Weil, 1983;...
Introductory books about simulation modeling treat validation mostly only superficially relative to the accumulated knowledge on the topic (Law and Kelton, 1982; Bratley et al., 1987; Zeigler et al., 2000; Bossel, 2004); Sterman (2000, Ch. 21) and Richardson and Pugh (1981, Ch. 5) are notable exceptions.

Given the remarkable body of knowledge available, one might ask: what is still missing? In general, the current knowledge provides modelers with essential as well as sophisticated approaches for the validation of models. However, we experience the fact that novices and often even professionals with moderate modeling experience have difficulties in accessing and applying these validation approaches. This is because the existing categorization of the validation tests as well as the validation processes proposed in the literature are often perceived as too abstract and unspecific to be readily applied. Moreover, important decisions, such as the decision to cease formal validation, are currently not addressed sufficiently. These issues are often puzzling, especially for novice modelers, which might result in an actually low utilization of even basic validation techniques (Peterson and Eberlein, 1994; Sterman, 2000). Expert modelers have learned about these aspects by trial and error and by gradual familiarization with the validation knowledge distributed among many publications. For them, the paper offers a possibility for reflecting their often intuitive competences and approaches to validation.

First, we start out by stratifying the existing validation tests according to their capability to address the complexity embodied in a model. For this, we develop a heuristic in the next section—called validation hierarchy—which supports the modeler in selecting validation tests according to the criterion of matching model complexity. Thereby, we can answer questions concerning the kinds of tests which should be selected and which properties of a model they can test. This aspect is new, in that prior research has not treated validation tests from the perspective of varying degrees of complexity. Our second contribution is in providing a more comprehensive understanding of the validation process. For this purpose, in the third section we offer an alternative to detailed schemes, in an idealized outline of what we call an integrative validation process. By “integrative” we denote that the process considers both the granularity as well as the complexity of the validation objects. In the fourth section, we discuss the decision to cease formal validation efforts for a modeling project. The factors that constitute such a cessation threshold have not been addressed explicitly in the system dynamics community. In view of this gap, the paper develops a heuristic for ascertaining the saturation threshold of a validation procedure. The new concepts are then applied in the fifth section by means of a case study. The sixth section concludes and proposes future research in this area.

We intend to contribute to research in three ways. First, the validation hierarchy should substitute a systematic approach for a random or competence-based selection of validation tests. Second, the integrative process should help to operationalize the idea of modularized modeling and validating. Finally, the cessation heuristic should provide modelers with a guiding principle about when to cease formal validation efforts. This principle is especially helpful for the design phase of modeling endeavors and for communication with clients. Our paper can help to increase the appropriate application of validation techniques, which is a burning need for the development of the field of system dynamics (Peterson and Eberlein, 1994; Forrester, 2007).
Validation methods and tests

Method derives from the Greek preposition metá (towards) and the noun hodós (pursuit, path). Consequently, a method is defined as a coherent procedure or set of rules, directed toward the attainment of a goal. Essentially, there are two types of methods: algorithms and heuristics. An algorithm is a fixed procedure with a high level of robustness by which a well-defined objective is achieved. A heuristic is a procedure with a lower level of robustness for solving ill-defined problems. Heuristics are, for instance, rules of thumb, search criteria, and procedural principles. They support learning and discovery by experimentation. From the perspective of system dynamics, most available validation methods have the characteristics of heuristics. A system dynamicist does not have a fixed toolbox of validation methods at hand, whose application would automatically guarantee models with high validity. Instead, the available methods are heuristic devices which, if properly selected and correctly applied, enhance model validity. This explains the difficulty in selecting the most adequate tests—especially for novice modelers. Such an audience can benefit most from this paper; expensive trials and errors can be reduced. In the next section, we address the traditional categorization of validation methods in the literature. Our new way of organizing them according to a complexity hierarchy provides a new logic for selecting methods. This is explained below.

Traditional categorizations of validation tests

Validation tests have been categorized in several different ways. Forrester and Senge (1980) have grouped validation tests into structure tests, behavior tests, and tests of policy implications. Barlas’s (1996) categorization is: direct structure tests, structure-oriented behavior tests (also indirect structure tests), and behavior pattern tests. Schwaninger and Groesser (2009) have integrated and elaborated these categories of validation tests into a multidimensional framework termed a “Validation Cube”. The cube organizes the most relevant issues for validation among three dimensions: domains of validation, layers of reflection, and levels of resolution. In the first dimension, the cube distinguishes between three validation domains: context, structure, and behavior. The context domain, for instance, is about averting the danger of a higher-order mistake, e.g. either by defining the problem erroneously or choosing the wrong methodology. The latter two domains, by and large, follow the established areas of validation (see Forrester and Senge, 1980; Barlas, 1996). The second dimension—“layers of reflection”—refers to epistemology and methodology. In the third dimension, the cube differentiates three levels of resolution—micro, meso, and macro—which refer to a model’s degree of granularity. The categorizations of the Validation Cube are of the descriptive type. In order to contribute to an efficient model development, we propose an insightful new way of organizing the available validation tests according to their potential for addressing complexity.

A hierarchical approach to validation tests

By complexity in this context we denote the variety of potential behaviors a concrete dynamic system or a respective model can exhibit (Schwaninger, 2009). In the pursuit of an efficient development of models, we organize the available validation tests with regard
to their adequacy of coping with such complexity. This is a way of fleshing out the areas of application for the tests, by stratifying them according to their capacity for dealing with complexity—an undertaking not yet ventured in the field. To elicit the tests’ capacity for complexity absorption, we start by differentiating levels of complexity inherent in a system dynamics model. Thereafter, we relate layers of model complexity on the one hand and levels of test capability for absorbing this complexity on the other.

We differentiate between five layers of complexity in a simulation model, represented in the ‘model hierarchy’ (Figure 1, left). The levels in the hierarchy are (I) elements, such as stocks, flows, auxiliary variables, causal relationships, causal polarities, and combinations thereof, (II) single feedback loops, (III) combinations of feedback loops, and (IV) the complete model. We demarcate the context of the model (V) with components such as the boundaries of the model, the type of complexity modeled, i.e. detail complexity or dynamic complexity, and the purpose of the model (Sterman, 2000). Put differently, the model-related context accounts for all that is left outside the immediate modeling endeavor. The degree of complexity increases from bottom to top of the scheme, in that a model as a whole (macro structure) in principle can generate a higher variety of behaviors than a single feedback loop (meso structure), which in turn has a higher potential variety than any single element of a model (micro structure). A combination of elements (flow rate and stock; level I) is already capable of producing dynamic behavior. However, the interaction of a single feedback loop (level II)—or a combination of loops (level III)—can produce a much richer range of behaviors. It might be feasible that for some systems a single feedback loop is able to generate the system’s most important dynamics. However, in most cases more complex combinations of feedback loops are required to describe complex system behavior adequately. The level of the model as a whole (level IV) is supposed to generate the full dynamics of the model. Finally, since the context of the model (level V) contains all the issues that are not addressed by the model itself, it consequently exhibits the highest level of complexity.

The battery of validation tests developed in the field of system dynamics (see, for example, Forrester and Senge, 1980; Richardson and Pugh, 1981; Lane, 1995; Barlas, 1996; Coyle and Exelby, 2000; Sterman, 2000; Schwaninger and Groesser, 2009) shows a cogent
correspondence to the complexity levels of the model hierarchy depicted here. A parallel scheme can be constructed for the different types of validation tests (Figure 1, right), visualizing the fact that a match occurs between the different levels in the hierarchy of model complexity and the distinct types of tests. Table 1 shows the validation tests that belong in each level of the validation hierarchy, the assignments of the tests being indicative, not exclusive. On the first level, the **elementary level**, model validity is examined through direct structure tests. Examples include the test for assessing the existence of variables, the test of parameter value (Oliva, 2003), the test for direct extreme conditions, and the test for internal dimensional consistency. On the second level, **simple dynamics** of individual feedback structures are tested. The respective tests are for behavior sensitivity, integration error, and indirect extreme condition. These first two validation levels address the micro-structure level of a model.

On the third level, model structures with **multiple feedback loops** are tested. These exhibit a higher level of dynamic complexity. Correspondingly, feedback loop knockout,

<table>
<thead>
<tr>
<th>Level</th>
<th>Validation hierarchy</th>
<th>Existing tests</th>
<th>References</th>
</tr>
</thead>
</table>
loop dominance, and behavior sensitivity tests are indicated. In these cases, the analyst
tries to systematically modify the model structure by selectively switching on or off or
changing the gains of individual feedback relations in the model, to investigate and under-
stand the structure and interactions of the feedback loops (Taylor, 1980; Ford, 1999;
Güneralp, 2006; Kampmann and Oliva, 2008). This level of validation takes place on the
meso-level of model structure. The fourth level is that of the whole model (macro-level),
which potentially exhibits the full dynamics. Here, model validity is assessed by means
of the tests of outcome behavior and, inter alia, the test for behavior sensitivity (e.g. Yücel
and Barlas, 2011), the behavioral test for adequacy of the model boundary (e.g. the possi-
bility of simplifying model structure, as in Saysel and Barlas, 2006), the test for behavior
reproduction (Sterman, 1984), tests of oscillatory behavior (Eberlein and Wang, 1983),
the test of autocorrelation (Barlas, 1990), and the Turing test (Balci, 1994).1 On the fifth
level, the meta-level, tests relating to the modeling context and purpose are assembled.
The test of purpose is, in principle, the most important test to execute; it inquires whether
the model purpose as given by the client has been achieved. In the same direction, the
test of appropriateness of structure evaluates whether the current model structure is appro-
priate for the client group, or whether the structure has to be changed to relate more to
their needs (Richardson and Pugh, 1981). A third important evaluation is to question
whether the methodology is adequate to address the issue at hand (Schwaninger and
Groesser, 2009).

In terms of the validation hierarchy discussed here, an effective validation process requires
a balance between the complexity of the model level under study and the complexity absorp-
tion potential of the respective tests chosen to validate it. The gain from this heuristic is that it
gives the modeler a rule for choosing the appropriate types of tests depending on the location
of the validation object in the model hierarchy. A question not yet answered is how
validation tests should be executed over time and relative to one another. This is the topic
of the next section, on the validation process.

**Validation process**

Conceptually, the validation process is a learning process, in which understanding is
enhanced through the interaction of a formal model with a mental model (Morecroft,
2007, p. 375). As this process evolves, both the formal model and the mental model of
the modelers change, leading to a successive approximation of the formal model to real-
ity: as models are used, they are adapted as a function of feedback from the real world
(Sterman, 2000, p. 88), and as they are tested, modelers fit them to the properties of
the real world.

In a more operational language, the validation process is a procedural sequence of
activities in time by which the modeler applies different validation methods. In this sec-
tion, we develop the concept of an integrative validation process in a different dimension,
which will be specified shortly. The term integrative validation process highlights the
nature of validation as being enmeshed in the modeling process. Richardson and Pugh
put it as follows: “Validation is an on-going mix of activities embedded throughout the
iterative model-building process” (Richardson and Pugh, 1981, p. 311). First, we turn to
the existing conceptualizations of validation processes.
Conceptualizations of validation process in use

It was Barlas (1996) who first suggested a validation process with three stages: empirical and theoretical direct structure tests, then structure-oriented behavior tests, and finally behavior pattern tests. Most other contributions concerning the validation of simulation models in the field of management science (Law and Kelton, 1982; Bratley et al., 1987; Zeigler et al., 2000) have failed to conceptualize the validation process on the level of detail delivered by Barlas. The existing research has treated validation as a uniform process while neither fully accounting for the domains of validation, nor the levels of resolution, nor the various levels of complexity. Hence it is not yet clear what the different processes of validation are and how they might follow each other, or how they are intertwined.

An integrative validation process

The integrative validation process provides guidelines about which steps to choose during a modeling endeavor. It accounts for three dimensions which are important for model validation: the domains of validation, the levels of resolution, and the levels of complexity in the validation hierarchy. For the first, we have already shown that it is useful to differentiate between tests for the relevant context of the modeling project, for the model structure, and for the model behavior (Schwaninger and Groesser, 2009). The second dimension is the level of resolution which distinguishes micro, meso, and macro structures. Finally, the third dimension uses the level of complexity, which the validation tests are capable of addressing best (see above). In Figure 2, we explicate the integrative validation process. The process is idealized and clarifies under which conditions the validation tests of the different categories should be applied. In this paper, it is not feasible to provide explicit examples for each test; this would be a task for future research. Here, we intend to provide an overall framing.
System dynamicists have argued that modeling can start either with conceptual modeling or with quantitative modeling (Forrester, 1985; Wolstenholme, 1999; Sterman, 2000). For us, the approach taken depends on the specific circumstances, but in principle formal validation efforts should begin as early as possible, as an integral part of the modeling process. Already while formulating a preliminary model purpose during the development of a conceptual model, mainly the tests of the context dimension can be used to check the framing of the model. Also structure and parameter verification tests (see below) are already useful at this point. We are not dealing with this phase in detail here.

As the simulation model is being built, the integrative validation process starts at the lowest level of complexity. This is adequate in the case of a novice modeler; experts might start at a higher level since they might use readily available model components or modules when building the new model. The question how the use of existing model structures would influence the validation process is one for future research. For now, we assume the case of a novice modeler. Let us first consider the levels of resolution (micro, meso, macro) and the domains of validation (model structure, model behavior, and model context). To reiterate, the available validation tests are stratified in the categories (I) to (V) of the validation hierarchy (see above and also Table 1). Now, these categories of the validation hierarchy are capable of associating the domains of validation and the levels of resolution with the adequate validation tests. In Figure 2 one can see that the tests in category (I) are useful for testing the model structures on the micro level of resolution. The categories (II), (III) and (IV) utilize model behavior and are used to test small meso structures (II), large meso structures (III), or macro structures (IV) of a model. The tests in category (V), i.e. tests concerning the model context, are used mainly on the macro level. Figure 2 shows in addition that tests on the meta level (V) and the elementary level (I) are distributed over the complete horizon of the modeling and validation process. This distribution captures the fact that some tests on the complexity levels (V) and (I) are omnipresent during the modeling process, e.g. parameter and structure verification test, purpose test, and adequacy of methodology test. One reason for this omnipresence is that these tests can be executed mentally; no computational support is required. However, it is necessary that the modeler embrace the attitude of applying these tests in a continuous manner throughout the modeling process. What we have here is a disciplined thinking process, as is common in theorizing in general (Weick, 1989; Schwaninger and Groesser, 2008). The continuous use of the mentioned tests is indicated by the bright-grey color (instead of dark-grey).

With this information, we have established a relation between the three dimensions of “domains of validation”, “levels of resolution”, and “validation hierarchy”. What is missing is the procedural aspect of validation, i.e. where to start and how to proceed. This procedural aspect is represented by the stages from (a) to (e), which denote distinct sets of validation activities carried out over time (Figure 2).

Again, we assume a novice modeler. In that case, the first stage is to concentrate on the micro foundations of a quantitative model, where the tests for the elementary model structure are undertaken (a). The available tests for the lowest level of model complexity (I) are listed in Table 1. For instance, the test to apply early is the parameter verification test to ensure that each element which has been included in the model actually represents an element of the concrete system to be modeled. Next, it has to be verified that interwoven individual elements, e.g. a stock-and-flow relationship, in fact exist in this combination in the real system (structure verification test). After checking their existence, one must ensure that the set of elements is dimensionally consistent (dimensional consistency test). Having
assured the validity of the micro structure, then tests which capture the *simple dynamics* (II), i.e. the interactions of structural components, are applied on the micro level (b). With these tests, the feasibility of an individual feedback loop is tested, e.g. if a stock value evolves within reasonable bounds even if parameters are set at extreme levels (indirect extreme condition test), and if the computational results are independent of the integration time constant (integration error test). When errors are encountered, the modeler has to iterate to the lower level of complexity, i.e. the level of individual elements (I). This is indicated by the arrows between (a) and (b) in Figure 2.

After testing the micro structures, the validation proceeds to the meso structures (c), shifting to the level of *multiple dynamics* (III). The model which is tested at that point is composed of several already validated micro structures. The most practical approach would be first to apply the loop knock-out test that activates additional micro structures or feedback relationships by means of individual switches. That way, the impact of the activated structure on the overall model behavior can easily be traced and evaluated. Once a meso structure with several feedback loops has been created, tests have to focus on the sensitivity of that structure’s behavior (behavior sensitivity test) and on its close match with the expected behavior of the real system. If this test is passed, it is reasonable to apply a loop dominance analysis to determine what feedback structure is dominant at what time instance and, again, to compare this to the dominant structure expected in reality. Finally, the application of the boundary adequacy test can help to determine whether the existing structure is able to generate realistic dynamics. In case the meso structures fail to pass one of these validation tests, iterations between (c) and (b) are required (Figure 2). In order to track, localize, and correct an error, it may be necessary to follow the chain down to the level of individual elements, iterating from (b) to (a).

Once the meso structures have been built, either additional micro structures or other meso structures will be linked together to form macro structures (d). These macro structures are tested using tests that can account for the *full dynamics* of a model (Figure 2, IV). These are, for example, the test for behavior pattern reproduction (Sterman, 1984) and the test for autocorrelation (Barlas, 1990). Again, it might be necessary to iterate back to meso structures (from (d) to (c)), or to micro structures (from (c) to (b) as well as to (a)), to discover reasons for erroneous or surprising model behavior. It is these iterative modeling and validation steps which can contribute significantly to substantive learning and model understanding (Mass, 1991; Homer, 1996). This procedure is an embodiment of the principles of modular modeling and validation (Balci, 1994) as well as partial model testing (Homer, 1983).

Having validated the full model, the same is evaluated with respect to its context (e). For that purpose, the validation tests on the *meta level* (V) are applied. One of the first steps is to check whether the model fulfills its purpose, e.g. whether it supports a decision about an investment policy (test of purpose). A further important test is to analyze the sensitivity of policy implications given different model structures (policy sensitivity test; Forrester and Senge, 1980; Kirchner, 1984). A test at a later stage is for ascertaining whether a modeling project is capable of improving the system performance (system improvement test). In case a test on the meta level has failed, a model reformulation and subsequent validation at a lower level of complexity are required.

As has been stated, the respective next stage should be tested only after the test of the previous stage has been passed. If a test on a higher level fails, the model ought to be revised on a lower level. In principle, it might be that such model revisions change other features of the model on the level of multiple and full dynamics. Consequently, it might
be required to spiral down to the elementary level to ascertain validity on the micro-structural level. For the novice modeler, this is expected to be labor intensive, but insightful nevertheless; expert modelers tend to know more intuitively which aspects of the model are affected and to react accordingly. However, even they might be surprised by counterintuitive results of simulations, and reap new learning opportunities. The validation process may require a heavy investment of time, but what we are laying out here is an ideal process; limited resources may entail tradeoffs (see next section concerning the cessation threshold).

As mentioned earlier, the modeler learns during the validation process. The interaction of a modeler’s mental model and a formal simulation model constitutes a continuous process of formulating hypotheses, their testing by means of simulation, outcome evaluation, and a subsequent reformulation of the initial hypotheses (Morecroft, 2007; Schwaninger and Groeßer, 2008). This interaction enhances both the accuracy of the modeler’s mental model as well as the formal simulation model. In addition, the increasing level of accuracy of a modeler’s mental model also influences the validation process in at least three ways: first, a modeler focuses his data-sampling efforts on the aspects of the model which are most vulnerable to validation tests, because a more accurate mental model helps to prioritize the efforts of searching and sampling data. Thereby, the validation process of testing all structures with the same attention is adjusted to emphasize particularly critical structures. Second, besides focusing on critical structures, a modeler with a more detailed mental model will have a higher iteration frequency and shorter validation processes. From his cumulative simulation experience, the modeler knows which tests to apply to yield maximal insights into improving the validity of the model. In addition, the experience develops a more intuitive selection of the validation tests for the specific modeling situation. One could frame this intuition as the learned capability to recognize behavior patterns, diagnose the likely cause of erroneous behaviors more correctly, and then select a test that is best suited to address the situation. A third possible result might also be that a more accurate mental model traps a modeler in his mental and formal model and reduces his ability to perceive disconfirming evidence. To conclude, a modeler with a more accurate mental model of the situation to be modeled gains flexibility and effectiveness in executing the validation process. However, at the same time, he risks being locked into the existing formal model and thereby discarding evidence that could help in advancing the modeling project. One way to estimate model advancement is to track the changes of a formal model and interpret successive changes in the formal model as learning (Schaaffernicht, 2006).

To conclude this discussion, we summarize the benefits of the integrative validation process. It integrates the domain of validation, level of resolution, and degree of complexity of the validation tests (Figure 2). Moreover, we have laid out the possible impacts that changes of the modeler’s mental model can have on the validation process. In addition, it has become obvious that the characteristics of validation, in particular the length and accentuation of validation activities, vary among the five levels of complexity. This variance is symbolized by the number of blocks on each level of the figure. Each rectangle abstractly represents a different validation test on a given complexity level. The question that now remains concerns how long the validation process should go on. When should a modeler cease formal validation? This is the topic of the next section.
Cessation of formal validity testing

Rationale of the cessation decision

We assume that most often the decision to cease formal validation efforts is not approached consciously (Simon, 1964); rather, the modeler continues modeling until the model is just “good enough” for the target group—going by an intuitive criterion for when to cease modeling (Kleindorfer and Geneshan, 1993)—or until budgetary restrictions force the project to be finished. Neither approach is satisfactory. A more effective use of available resources could result if the issue of ceasing formal validation were treated explicitly and normatively. Consequently, the validity of models—understood as “usefulness with respect to some purpose” (Barlas, 1996, p. 84)—could be improved significantly and at the same time could support discussions with the involved stakeholders. This leads us to the question of when to end the validation process. Other methodologies face the same challenge of finding the appropriate degree of validity at which to stop both sampling and modeling. To illustrate this common ground, we use two widely accepted research methodologies as examples: The grounded theory approach, representing a qualitative approach, and multivariate regression modeling, representing a quantitative approach.

In grounded theory (Glaser and Strauss, 1967; Glaser, 1978; Strauss and Corbin, 1997, 1998), “the general rule when building theory is to gather data until each category is saturated” (Glaser and Strauss, 1967, pp. 61–62). Here, “category” stands for a theme or variable that covers and abstracts information in the empirical data. The pertinent term “theoretical saturation” signifies that data have to be sampled until (a) no new or relevant data regarding a category seem to emerge, (b) the category is well developed in terms of its properties and dimensions demonstrating variation, and (c) the relationships among categories are well established and validated. If the sampling process ceases too soon, the theory will be unevenly developed and lack density as well as precision (Glaser, 1992; Strauss and Corbin, 1998). In grounded theory, the rules that govern cessation in theory building efforts are qualitative. Thus, with criteria such as “no new or relevant data seem to emerge”, “well established and validated” and “well developed”, in principle it is up to the perception of the researcher to decide when the criteria are fulfilled and thus when to stop validation and theory-building efforts and proceed with model analysis. With grounded theory, one can say that considerations around cessation focus first on the output or product of the research—that is, a model with “well-developed” categories. Second, they concentrate on the input–output ratio of research efforts—namely, when further data do not seem to change the model. Third, they take into account multiple objects or categories—not only one main object.

In multivariate regression modeling, quantitative indices are used to define the model’s goodness of fit (e.g. Kennedy, 2003). Probably the most used diagnostic metric is adjusted $R^2$: the coefficient of determination. Adjusted $R^2$ ($R^2_{\text{adj}}$) is a modification of $R^2$ that adjusts for the number of predictors in a model. While $R^2$ will never increase when a predictor is dropped from a regression equation, $R^2_{\text{adj}}$ may increase. Specifically, if the $t$-ratio for a predictor is less than one, dropping that predictor from the model will increase the $R^2_{\text{adj}}$. It is the proportion of variability in a dataset that is accounted for by the specified statistical model. If the value of $R^2_{\text{adj}}$ exceeds thresholds that are defined by the scientific peer group (the range of 0.6–0.8 is commonly used; Kennedy, 2003), researchers are motivated to end their analyses. Multivariate regression modeling uses an output-oriented
measure for the definition of model validity; it does not account for any input factors (e.g. additional data). Furthermore, it concentrates on a numerical measure for a single (outcome) variable that is rigidly defined by a range of acceptable values.

Both of the examples show, first, that they possess qualitative or quantitative criteria; second, that they concentrate on input–output- or output-oriented measures; and third, that they consider single or multiple objectives when deciding at what point to stop efforts to improve a model or theory and continue with model analysis. We have also seen that to a certain extent it is at the researchers’ discretion to decide when the criteria are sufficiently fulfilled. The relevant difference between the mentioned methodologies and system dynamics research is that the former explicitly address “cessation” or “saturation” in model development in their respective literature. This is where system dynamics can learn and advance.

The discussion about cessation

Cessation is perhaps the most important, but least discussed, decision about validation. Following our discussion of the validation hierarchy and the integrative validation process, we proceed to analyze the decision to cease formal validation efforts. Most of the research that refers to the system dynamics methodology has hardly attended to this matter. For instance, Coyle and Exelby (2000) suggest that one ought to steadily improve the validity of a model until it reaches a satisfactory level. However, no further information is provided concerning when a satisfactory level might be achieved, and which factors help to attain that level.

Accounting for the current practice of simulation modeling, Finlay and Wilson (2000) develop a contingency framework containing 12 factors that determine when formal validation efforts are ended in practice. Olphert and Wilson (2004) use expert evaluations to validate that framework. Research by Anastasakis and colleagues (2008) supports the results obtained by Olphert and Wilson. These studies suggest that the most important factors that determine the length of formal validation efforts are the deadline of the decision for which the model is constructed, the level of risk associated with the decision, the complexity of the situation, and modelers’ expertise. In contrast to the works just quoted, we are taking a normative stance.

Instead of contemplating when modelers manage to stop themselves or when clients force them indirectly to stop due to budgetary restrictions, we want to discuss when modelers should cease and what factors influence this decision. The issue of budgetary restrictions is an important one, since it seems to determine how the resources available to a modeling project are allocated between model development, validation, analysis, interpretation, and the generation of implications. We will briefly take up this issue when we discuss costs of validation. However, we cannot address the issue comprehensively; this is also a subject of future research.

A heuristic for the decision to cease formal validation

We have seen that the concept of a threshold for ceasing validation and accepting a model is comparable to the examples of grounded theory and multivariate regression modeling. Also in simulation modeling, it is often “too costly or time consuming to determine that a model is absolutely valid over the complete domain of its intended applicability. Tests
and evaluations are conducted until sufficient confidence is obtained that a model can be considered valid for its intended application” (Sargent, 2008, p. 157). The validation cessation threshold (VCT) is defined by a value which, if reached, calls for ceasing formal validation activities. Consequently, a modeling effort to fulfill a low VCT is lower than that to fulfill a high VCT. The VCT is not a fixed value; it depends on a variety of contingency factors such as a target group’s experience with modeling, the relative importance/risk of decision, model size, costs of validation, the target group’s expectations, data availability, data intensity, potential degree of validity of the model, and a modeler’s level of expertise (Figure 3). The four direct determinants of the VCT will be discussed next.

Target group’s experience with modeling
A target group’s experience, i.e. the experience of final clients, with simulation modeling ranges on a continuum from nonexistent to extensive. It is assumed that a target group with more experience pays closer attention to the modeling process and its interaction with the modeler. Thereby, first, they can influence the modeling process and ensure that the model fulfills the intended purpose; and second, they can see the degree of validity of the model on a detailed level over an extended period of time. Hence an experienced target group does not need a high VCT that is inflated by exaggerated demands (Figure 4). They grasp the essential validity of a model during the modeling process and by means of basic but insightful tests.

Relative importance/risk of decision
Simulation modeling is used to understand dynamics and to inform decisions. Every decision has a level of importance and risk associated with it. Investment decisions about expensive, durable physical assets, for instance, are considered important decisions, since large financial obligations over a long term are linked to this decision. Decision theory suggests that the more important or risky a decision maker perceives a decision to be,
the higher will be the evidence required that the decision is truly beneficial (Kahneman and Tversky, 1979); hence a higher VCT will result (Figure 4).

Costs of validation
Validation is a labor-intensive process. The associated costs comprise costs for data and costs for executing the validation process. The first component depends on the data intensity, which improves the power to withstand falsification trials and hence builds up the level of confidence in a model. Also, data gathering and analysis are accounted for (e.g. access costs, transaction costs, and property rights costs; Balci and Sargent, 1981). It is difficult to quantify the costs of validation on empirical grounds. A publication by the U.S. Department of Defense about this topic estimates the costs for validation activities in the range of 5–17.5%, with an average of approximately 12% of the total modeling budget. These numbers have to be read with caution, since the research design is not reported (Defense Modeling and Simulation Office, 1996; Back et al., 2000).

Other factors being equal, we assume that the VCT is negatively correlated with the costs of validation (Figure 4). This is because only limited financial resources are available for projects, and the clients expect to receive useful results from a modeling study compared to not conducting the study at all. For high costs of validation, clients are willing to accept a lower VCT to obtain results they would otherwise not have.

Potential degree of validity of the model
This is the level of validity of the model that seems to be achievable given the means available, e.g. theoretical and empirical evidence, as well as the situational conditions. More
specifically, the potential validity of the model is enhanced by higher levels of data intensity and more extensive expertise in a modeler (Figures 3 and 4). The potential validity expresses the highest level of trust and confidence in a model, perceived as feasible by the client. The VCT will be lower, the higher the potential validity of the model is perceived.

Model size
The size of a model can be measured by the number of its stocks, flows, auxiliary variables, and feedback loops (Schaffernicht and Groesser, 2011; Groesser and Schaffernicht, 2012). With model size, the level of complexity of the model increases exponentially. The more complex a model is—irrespective of whether it is detail complex or dynamically complex—the more time it takes for a modeler to understand the model, to verify the syntax when adding variables, and to validate the model’s structure and behavior. As a consequence, the amount of time for formal validity testing for each new element increases exponentially, since all existing connections as well as the new ones have to be checked to ensure reasonable and accurate model behavior. Hence the costs of validation increase exponentially with the size of a model (Figure 5).

Target group’s expectations
The degree of a target group’s expectations defines when formal validation efforts can be stopped (Richardson and Pugh, 1981; Sargent, 1992). A model’s purpose is always related to a target group. Here, we distinguish two major target groups: practitioners and academics (Weil, 1983; Coyle and Exelby, 2000). For the first, a model might be valid, and hence formal validation is more likely to cease, once the model conveys important insights, and is reasonably grounded empirically, e.g. by a single case of high relevance. In general, the more accurately a model corresponds to the structure and dynamics of the situation under study,
the more that practitioners are willing to accept it as valid. The scientific audience, on the other hand, is interested in models and insights on a generic, conceptual level, and hence that target group calls for validation techniques based on larger sample sizes and statistically significant relationships (for an early system dynamics model that fulfills these scientific criteria see Hall, 1976). In other words, academics usually rank the goal of generality (also referred to as external validity) higher than practitioners. Depending on the target audience, validation methods specific to the relevant audience deserve special emphasis during the validation procedure (e.g. statistical estimation of parameter values for academic audiences), while other methods will be deemed less important (e.g. the family member test for a practitioner audience). Figure 5 shows the qualitative relationship between the expectations of the target audience and the impact on costs of validation.

Data intensity and data availability
A second determinant of the costs of validation is data intensity. The use of data is a fundamental aspect of validation (Balci and Sargent, 1981). Data intensity concerns the degree to which empirical data are used for the modeling project. When we refer to data, we assume that these can be used for system dynamics modeling (e.g. aggregated data, unbiased; Graham, 1980; Forrester, 1994). The resulting data intensity influences costs of validation and depends on the type of data (numerical data, verbal data, mental data; Forrester, 1994), their level of certainty, and their respective availability. These three aspects are captured by the variable data availability in Figure 3. Both data availability and data intensity correlate positively with each other (Figure 5). The question about whether additional data can be gathered to increase data quality and intensity is always one issue for consideration. It is assumed that the costs of validation increase exponentially with the level of data intensity (Figure 5).

A modeler’s level of expertise
An additional consideration concerns the contingency effect which the modeler’s level of expertise has on the data intensity and also on the potential degree of validity of the model. We argue that an expert can obtain higher levels of potential model validity given a certain data intensity than a novice modeler, because the expert not only uses all of the potentials inherent in a given dataset, but also procures additional data, e.g. by highly focused interactions with clients or institutions that possess the required data. A conceptualization of this is shown in Figures 6 and 7. The expert modeler uses quantitatively more data with

Figure 6. Assumed effects of modeler’s level of expertise and data intensity on potential degree of validity of the model
higher levels of reliability from more diverse types of sources (mental data, verbal data, and numerical data), which results in higher levels of data intensity.

To summarize, we have introduced several factors that help to determine the level of VCT. The interaction effects between the factors, which exist in reality, are shown in Figure 3, e.g. between high costs of validation (suggesting a low level of VCT) and high levels of importance of the decision (suggesting a high level of VCT). It is the interplay of these factors that determine the resulting VCT. For now, let us assume that we encounter an ideal modeling situation, i.e. high levels of data availability, the modeler’s expertise, the target group’s experience with modeling and low levels of model size, the target group’s expectations, and risk associated with the decision. In this best case, the VCT would be very low and hence the validation efforts for formal testing would cease relatively fast compared to a situation in which the ideal conditions did not exist. Conversely, the VCT would be very high in a case in which the previous assumptions would be the opposite. In such a worst-case situation, it would be difficult, perhaps even impossible, to reach the VCT. Then, modeling would be stopped when financial resources were insufficient.

Reflecting on the comprehensiveness of the factors currently considered, we can say that there are certainly additional factors that influence the cessation decision: for instance, budgetary restrictions, as noted at the beginning. In addition, we have not weighed the factors relative to each other since this depends strongly on the situational circumstances. These are topics for future research. However, we believe that we have made explicit the major factors that are constitutive for any decision about when to terminate a validation procedure.

**Example case study**

In the following, we describe a case study (Groesser and Ulli-Beer, 2006, 2008; Groesser and Bruppacher, 2007; Müller *et al.*, 2012) to explicate our discussion of the VCT and the decision to cease formal validation efforts. We do not address the validation process (see earlier section) owing to space limitations. The case study will be described in relation to the individual factors about the VCT as discussed above (summarized in Figure 3).

**Case study description**

The diffusion of energy-efficient building technologies is a focal point for policy makers in Switzerland, in order to reduce energy demand and the emission of greenhouse gases. A 3-year project (2006–2009) at the University of Bern analyzed managerial and organizational
adaptation processes that contribute to the diffusion of pioneering energy-efficient technologies in the Swiss building sector from 1970 until 2009. For this project, a system dynamics simulation model has been developed which sheds light on dynamic interactions between economic, social, behavioral, and contextual factors and thereby explains the diffusion of energy-efficient technologies. For the model development, qualitative and quantitative data have been gathered by means of more than 30 expert interviews, four workshops, and archival as well as online data.

Target group’s experience with modeling

The target group of this project is a general, multidisciplinary scientific audience plus policy makers and practitioners in the realms of the residential built environment. For the project, system experts from the different areas of the built environment have been involved in the development of the conceptual model and computer simulation; none of them had significant prior experience with simulation modeling of the type of system dynamics. This yielded a relatively high VCT, because the model had to be valid from the perspective of an inexperienced but interested and critical audience. Questions of the following kind had to be answered. What are the benefits for the individual participant? What can this method provide? Why are the details of each area not fully covered? Why should the participants care about a more comprehensive perspective?

Relative importance/risk of decision

The modeling endeavor was a fundamental research project funded by a major Swiss research institution. The purpose was to understand the interrelations in the system, their likely effects on diffusion of technology, and the resulting system’s dynamics—rather than having a practical model for solving a certain issue for a specific client. Lacking such a client, there was no final decision which this project had to support. Even though the topic of energy-efficient technologies and their dissemination is perceived as important by a general audience, the actual impact of our project is assumed to be small. Hence the resulting VCT is low.

Data availability and data intensity

For this project, all three databases (numerical, textual, and mental) have been utilized to the best extent possible. More than 30 expert interviews as well as four workshops with group model-building activities have been conducted to elicit relevant aspects of the system under study and to validate the model. By means of this approach, rich qualitative data have been generated that are reflected in the model. Textual databases (e.g. scientific literature) have been utilized to back up relationships in the model. These two types of data have been used to a large extent. Numerical data, however, were only sparsely available. Only three, partially biased, time series lent themselves to the quantification of the model. Consequently, the model has been parametrized according to the criteria of reasonability, consistency, and coherence. Sampling of additional numerical data, e.g. energy demand for houses from 1970 to 2009 according to construction permissions would possibly, in principle, have been retrievable from archival data: however, the costs for gathering these data rendered this
approach infeasible. In addition, no expert in the group could provide these data or refer us to an institution that could realistically make them available. Hence the actual provision of the type of data which the scientific community would consider as most reliable was very limited. From this standpoint, the data intensity of the model is considered low. The intensity is also influenced by the modeler’s experience, which will be detailed next.

Modeler’s level of experience

In retrospect, the experience of the modeler of the project with regard to empirical modeling was low to medium, and has increased to high levels over the duration of the project. However, since the data gathering was executed at the very beginning of the research project, the data intensity of the model could not be improved significantly at the end of the project: as explained earlier, obtaining (numerical) data was infeasible at later stages. However, the modeler has utilized the available data (textual and mental) to the best extent possible. In sum, the data intensity was still at a low level given the irreversibility just described.

Potential degree of validity of the model

For this, two aspects are taken into account. First, data intensity, as has been pointed out, was rather low, which resulted in a rather low potential level of validity; in other words, it was relatively easy to reject the model. The second aspect, a high level of experience in the modeler, which has increased during the course of the project, moderated this first development and led to a higher potential level of validity than otherwise would have occurred. However, the modeler could not reverse the level of data intensity, e.g. he could not provide concrete parameter values where none were at hand. Nevertheless, he used a set of other validation tests, not a behavior reproduction test, in this way ensuring that the model could not be refuted given a high level of internal validity. Hence the result was a medium level of potential validity of the model.

Model size

The simulation model endogenously embodies the multiple actors of the system, e.g. private and organizational building owners, tenants, architects, the physical building system, and standard-setting bodies. The target was to understand and represent the perspective of the different actors in detail. This resulted in a large simulation model with several thousand elements on the technical level (e.g. several multidimensional subscripts have been used). The time and hence costs of verifying and validating this model on the structural level became immense, leading to large costs of validating the model.

Target group’s expectations

Since there was no clearly identified target group, it was difficult to forecast expectations. Consequently, at the time of its development the model had been designed for a scientific audience as well as for policy makers and practitioners. In other words, the modeler had tried to satisfy the demands of several target groups, which increased the costs of validation enormously.
Costs of validation

The large model size, especially the very detailed model structure, and the wide-ranging expectations of the target group, led to high costs for validating the model. These large internal project costs could not be reduced by the relatively low level of data intensity. This resulted in a relatively low level of VCT.

Validation cessation threshold

In light of this explicit discussion of each factor that influences the VCT, we can conclude that the resulting VCT was relatively low for this project. This was especially the case because the importance of the model’s purpose for most of the participating experts was not relevant to their daily professional or personal life; not all the details that the experts perceived as being relevant could be included in the model, resulting in a lack of implications for daily decision making, which is characterized by detail complexity. In addition, the costs of validation have been large and, given the limited budget, consequently the VCT had to be reduced. The low level of the target group’s experience with modeling, and the medium degree of potential validity of the model, did not contribute to improve the resulting VCT. The natural path to have taken would have been to reduce the model to the crucial aspects of importance from the perspective of dynamic complexity. However, budgetary restrictions did not allow for this.

By means of this descriptive case study, we have demonstrated how our framework for the cessation of formal validation can be applied. The “proof” of the utility of both the framework (Figure 3) and the discussion about cessation is revealed when the framework is applied during the definition of a modeling project and also reflected upon while modeling. The systematic consideration of the elements in the framework can guide modelers in their efforts to develop models of high validity, because it clarifies the relative importance of the different determinants of the modeling process, which enables higher validity.

Conclusion and future research

We have conceived of validation as a rich and well-defined effort by means of which the confidence in a model is gradually enhanced. Validity, then, is always a matter of degree, never an absolute property. “Well defined” here is not meant in the sense of a rigid algorithm, but as the rigorous application of a validation process and validation tests. In this paper, we have developed, first, a hierarchy of validation tests in accordance with their capability for attending to model complexity. The different levels of complexity absorption in these tests are matched with the corresponding levels of complexity present in the model. It is thereby ensured that the validation tests are applied in their ideal area of application, that they are utilized most beneficially, and that all aspects of the model’s complexity are taken care of. Thereby, modelers are provided with a new access to the toolbox of validation tests—they will now be able to select the appropriate validation test according to the level of complexity at which they are working. Second, we have introduced an integrative validation process that pervades all phases of modeling. More specifically, we have shown ways of incorporating three dimensions into the integrative validation process: the domains of validation—context, behavior, structure—the levels of resolution—micro, meso, macro—and
the levels of complexity—elementary level, simple dynamics, multiple dynamics, full dynamics, and meta level. Finally, we have discussed the question of validation cessation, concerning the point in time at which formal validity testing should be terminated. Our analysis resulted in a set of decision determinants by which a VCT can be ascertained. One implication for a system dynamics project is, for instance, that one ought to concentrate on those variables to which the model’s behavior is highly sensitive, thereby saving resources from use on variables to which the model’s behavior is robust. We have tried to explicate all three contributions by means of examples.

More research about validation is necessary to increase its uptake, especially by novice modelers. Future research should account for the importance or priority of the different tests as perceived by practitioners and academics. This, of course, is relevant since resources for validation—i.e. time, monetary resources, expert knowledge, and appropriate data—are scarce. Furthermore, we have addressed neither the issue of accuracy in the interpretation of model results, nor the questions of if and how validation of a model should be conducted via a target group’s interpretations. The validation of the implications stemming from a model output differs from expert validation of the model itself (e.g. done by walkthroughs; Balci, 1994; Barlas, 1996), since it relates to the practical implications based on the behavioral results of the model, not on the model’s structure. The risk associated with misleading interpretations is a demanding subject, which further research needs to investigate. In principle, the question must then be addressed: “how can we transfer to the client the knowledge about the relationship between model structure and the behavior it produces?”

This paper has contributed in three aspects to validation. What is ultimately required will be even more—a comprehensive, dedicated validation methodology. We believe that a continuous dialogue and reflection about the state of validation in the field of system dynamics is a necessary step for further scientific progress. Even though Finlay and Wilson (1997) and others have emphasized the crucial impact that a validation methodology has on the quality of operational research models, a comprehensive validation methodology for simulation models is still a project for the future.

Acknowledgements

We thank Henry Weil as well as participants in the methodology track at the International System Dynamics Conference 2009 in Albuquerque, and two anonymous reviewers of System Dynamics Review for their helpful comments. In addition, we thank Silvia Ulli-Beer and Ruth Kaufmann-Hayoz for their support throughout the Swiss National Science Project (405440-107211/1), which has been used to apply the concepts developed here. Moreover, this publication was supported by the Swiss National Science Foundation (PB/SGP1_133613). In addition, we are grateful to the System Dynamics Group at the Massachusetts Institute of Technology for their support.

Notes

1. We intentionally excluded commonly known procedures, e.g. evaluation of the model’s face validity, walkthroughs, or group model building, from Table 1. These are procedures that provide the environment in which validation tests are executed. For instance, a structured walkthrough meeting assembles a group of experts who together inspect the structure
and behavior of a simulation model and discuss its validity, and detect and document faults (Balci, 1994). To achieve this purpose, the tests as outlined in Table 1 are applied.

2. It is also possible that the relationships follow other patterns, e.g. the pattern of S-shaped growth. Since the shapes depend on the project specifications, we can only argue that the patterns as hypothesized in Figures 4–6 are reasonable; we do not argue that they are generic.

Biographies

Stefan N. Groesser is Professor of Strategic Management at the Bern University of Applied Sciences, Switzerland. In addition, he is a Senior Researcher for strategic management and system dynamics at the University of St Gallen, Switzerland, and a visiting scholar at the System Dynamics Group at MIT Sloan. He has received degrees from the University of Stuttgart, Germany, in business administration and economics, from the University of Bergen, Norway, in system dynamics, and from the University of St Gallen, Switzerland, in Management. Stefan’s research interests include strategic management, business models, mental models, and simulation methodology.

Markus Schwaninger is Professor of Management at the University of St Gallen, Switzerland. His research is focused on the study of complex socio-technical systems, with a methodological focus on system dynamics and organizational cybernetics. His research projects to date have been related to organizational intelligence, the design, transformation and learning of organizations, and to systemic issues of sustainability. Schwaninger is author of over 200 publications in six languages, including Organizational Transformation and Learning (Wiley, 1996; with Espejo and Schuhmann), and Intelligent Organizations (Springer, 2006/2009). He has lectured widely, on four continents, and is involved in international, transdisciplinary research projects.

References


Copyright © 2012 System Dynamics Society

DOI: 10.1002/sdr