Extending model validity concepts and measurements in system dynamics

DJW Arthur and GW Winch
Plymouth Business School, UK
Tel: +44 1752 233522 / Fax: +44 1752 233505
D.Arthur@plymouth.ac.uk, Graham.Winch@pbs.plym.ac.uk

System dynamics models are employed for a variety of purposes in socio-economic systems, including behavioural prediction, policy analysis, consensus building and for hypothesis testing about complex system behaviour. Models can be valid or effective in three ways: by precisely representing reality, through their potential to stimulate learning or by demonstrating utility by instigating organisational change. This paper proposes an overall framework for setting validity tests in the context of different modelling purposes, emphasising three main types of validity. A sequence is also proposed through which a validation process should move to be effective. Quantitative measures of validity and identification of different approaches to constructing prior consensus on validity are proposed. Validity profiles can characterise how model utility varies throughout a project. The validation framework affords explanatory power for the efficacy of different modes of modelling and can help to clarify model purpose.

Keywords: Validation; validation framework, profiles, process; consensus schemas; project impact evaluation

Introduction

Validation continues to be a challenging issue for both the study and the practice of model building in the management and social sciences. The challenge stems from critiques of published models in academic settings and the requirements to demonstrate quality assured products and processes in commercial and practitioner modelling projects. In the system dynamics field, Richardson's (1995, 1996) stocktaking reviews indicated that model validation was one of eight distinct areas needing development. He suggested that there had been little progress since Forrester and Senge's (1980) detailed description of a suite of 17 validation tests and Sterman's (1984) proposal for historical goodness of fit summary statistics. The President's address at the System Dynamics conference in Istanbul (Richardson, 1997) reported on advances under most of the other seven headings but at that time commented on the slow progress in validation and confidence.

Barlas (1996) also lamented the lack of published articles over the preceding decade – three papers on model validity were published in the System Dynamics Review from 1985 to 1995. However, Lane (1995, 1998) has offered broader views of validation and different types of modelling practice. Additionally, Coyle and Exelby (1999) discuss how commercial system dynamics models built in consulting projects are subject to different constraints and emphases compared to academic work.

There seems to be good reason for this lack of development: at the outset of the field, *Industrial Dynamics* (Forrester, 1961) provided a rather complete statement of validation philosophy. The approach departed radically from the dominant logical empiricist view of science used in management science and economics at the time. Instead, it was representative of a relativist and holistic philosophy of scientific knowledge, which despite criticism is no less a scientific basis for modelling (Barlas and Carpenter, 1990). Vázquez et al (1996) subsequently positioned system dynamics to encompass both objective and subjective philosophies of knowledge. Forrester proposed that models should be based on a view of validation that emphasises significance or relevance of the model relative to a particular purpose. A purpose must include the definition of a problem to be solved or opportunity to be seized. Model purpose is ultimately related to the impact achieved:

> The purpose of industrial dynamics models is to aid in the design of improved industrial and economic systems. How are we to judge if the models are suitable for this purpose? The ultimate test is whether or not better systems result from investigations based on model experimentation. By this criterion, the validity of industrial dynamics is not separable from the effectiveness of industrial dynamics as a total viewpoint and discipline. The pertinent test is that of improving management practice. (Forrester, 1961, §13.1).

Forrester restates (1976, 1994a) the aspirations of relevance of the system dynamics field to both real social and business problems and the domain of education, describing the field as an activist profession. In contrast, econometric and other social science modelling is oriented around observation and explanation, employing an absolutist view of model validation that seeks to calculate summary correlation statistics based on objective significance levels. By contrasting these two philosophies, Barlas and Carpenter (1990) and Barlas (1996) indicate why system dynamics models have been criticised for relying too much on subjective assessments of validity. However, these papers also demonstrate the ultimate reliance of the absolutist philosophy on subjective justification of the significance levels.

Nevertheless, the value and correctness of the system dynamics philosophy of validation is now being mirrored in mainstream Operational Research practice in two ways. First, in simulation modelling there is growing recognition of the importance of validation relative to model purpose (Brooks, 1999). Second, in soft OR methodologies, models built for interpretive purposes, contrasted with predictive ‘hard’ OR models, need approaches that more directly reflect the subjective nature of utility of these models.

The impetus for this paper arose out research[1] using a system dynamics model where the ultimate success of the modelling project was not derived from the technical aspects of validity. It stemmed instead from other dimensions of the process and outcomes of the modelling. Accordingly, we propose an assessment framework reflecting a growing interest in performance measurement for projects and organisational practice. The framework permits the evaluation of model validity from the customer’s perspective. Scholl’s (1992) benchmarking study of the system dynamics community reported that the tests described by Forrester and Senge (1980) are not much used. A validation framework and set of procedures should facilitate good modelling practice, although that ultimately remains a matter for professional motivation.

In this paper we do not consider the advances in formal validity tests (Barlas, 1996) and tools that have been implemented in system dynamics software, such as the Reality Check™ tool (Eberlein, 1994). This paper is instead linked to the study by Cavaleri and Sterman (1997). They recommend that systems thinking/system dynamics interventions should be designed in advance to undertake impact evaluation and they call for further research to this end. Confounding factors and environmental change impede retrospective analyses of longitudinal studies so that the reasons for the efficacy of interventions are ambiguous. The framework presented in this paper is intended as a contribution to the goal of structuring intervention assessments in modelling projects.

Historical review of validation

It is instructive to consider well-known system dynamics studies from the viewpoint of validation. In this paper, we do not present a detailed analysis of published models and their critiques (Ansoff and Slevin, 1968; Averch and Levine, 1971; Nordhaus, 1973; Zellner, 1980 etc). However, it is apparent that a significant fraction of the controversy arose because critics possessed narrower expectations and definitions of validity. The project to apply the generic *Urban Dynamics* (Forrester, 1969) to specific cities became embroiled in discussion with the advisory committee – comprising mainly experts in econometrics – hampered by concerns about data gathering and validity (Alfeld, 1995):

> Although we attempted to engage the committee in a constructive dialogue concerning the model’s utility, the members had neither the time nor the inclination to study system dynamics. Criticisms concerning data and validation techniques displaced issues of system structure and user needs.

The Urban Dynamics team eventually solved the ensuing communication difficulties by avoiding methodological discussion and concentrating on concerns about immediate urban policy plans and outcomes for subsequent model sponsors. Alfeld (1974) concluded that the major failing lay in not pointing out that *Urban Dynamics* was an example to be followed and not a set of final conclusions to be implemented.

The *World Dynamics* (Forrester, 1973) and *The Limits to Growth* (Meadows et al., 1972) studies offer perhaps the best examples. The strongest criticisms centred mainly on the technical content of the models or omission of model structure and parameters. But critics did not consider to the same extent the value of the process of constructing the model, the pertinence to the model’s sponsors and the broader impacts of the messages presented. Clearly, the Club of Rome wanted the model to raise the profile of the impact on the global environment of economic development without having responsibility to operationalise remedies. Ormerod (1994) contrasts the model content and impact of the world modelling studies:

> Many criticisms can be – and indeed have been – made of the Limits to Growth exercises. Indeed, all of them, fair and unfair, weak and strong, have been made. One way in particular in which the studies were open to criticism from economists was their neglect of the role of prices………But whatever, the particular criticisms, the true and lasting significance of the Limits to Growth work was the development of a fundamentally different approach to understanding the workings of the economy to that of orthodox economics.

Forrester (1971) comments that it is the modelling process that conveys the greatest value of models – critics often believe that the author of a model has a greater commitment to the published structure than is actually the case. In practice, however, time and cost constraints limit the extent of detail or scope of a model. Enough resource must be devoted to documenting and communicating the model and the conclusions from it. Meadows argues (Meadows et al, 1982) that even if an unlimited modelling budget were available, she would ‘strive mightily’ to limit it. The model purpose provides a rationale for limiting the model content – it should include enough to solve the problem encompassed by the purpose.

Robinson and Pidd (1998) have commented on the lack of agreement amongst authors in the simulation modelling literature about the meaning of terms such as validity, credibility, acceptability and confidence. Nonetheless, their comprehensive interview research of simulation model providers and customers shows the perceived importance of validity. Of the most commonly cited factors leading to delivery of a simulation project and its success, model validity occurred the most frequently.

Validation framework

Richardson (1996) calls for a return to the central issue of model validation and that it should be unpacked into its several constituent parts. He proposes a taxonomy of decision environments since different modelling situations require different tests of validity and confidence. He points to the need to accumulate wisdom about the conditions under which various types of tests and procedures appear to be most appropriate.

One can imagine several different candidate taxonomies to classify decisions. One way would be to position different types of decision or model purpose on a two-dimensional matrix. For example, the axes of a matrix could be type of change or decision required versus nature of audience or client. Although such a taxonomy might be interesting to formulate, the choice of matrix dimensions is debatable. A taxonomy is also likely to become prescriptive, rather than aiding the development of judgment for assessing model validity. However, it is still useful to consider broad groupings of different model purpose types. Figure 1 suggests five groups of purposes although the list may be incomplete and there could be different nuances within each group. In conjunction with a validity scoring system explained below, a clearly defined model purpose in any project can be assessed individually and compared to another purpose if required. Hence, for the purposes of this paper, it is not necessary to discuss further how to classify decision situations.

Figure 1. Possible groupings of model purposes

A	Behavioural forecasting/scenario development
B	Insight generation – current/general or historical
C	Achievement of a specific organisational change, or consensus generation
D	Decision support/judgment/clarification of issues/remedial policy analysis
E	Hypothesis formulation about complex system behaviour

As the purpose of a model has such importance in any study, a validation framework should aid in measuring the salience of the purpose. *Industrial Dynamics* provides a reminder that there is no formal test to show how promising or feasible the goals of a modelling study are:

> The ultimate significance of our effort to design a better industrial system rests on the investigator's judgment of a specific objective" (Forrester 1961, § 13.2).

Several conclusions emerge from both the of system dynamics literature, systems methodologies and OR (mainly discrete event) simulation modelling (Lane 1995; Coyle and Exelby, 1999; Pidd, 1996; Robinson and Pidd, 1998; Robinson, 1999):

- there are three main dimensions of validity or confidence; individual tests can be grouped under these headings
  a) substantive (technical content or perceived representativeness of models)
  b) constructive (process effectiveness or learning by participants)
  c) instrumental (outcome or final utility of the project or intervention validity)
- each validity test can only be assessed to a degree of acceptability, given time and cost constraints
- perception of the validity of a model changes through the course of a project
- the significance attached to individual tests of validity can differ between model builder and customer
- some tests are essential, others are only desirable

Given that the total validity of a model is not a simple binary variable, one should not expect each validity test to be a simple accept/reject decision. A list of model tests can give an impression that each one should be passed completely. This must be so for only some of these tests, such as the dimensional consistency test. The portion of validation tests described by Forrester and Senge (1980) that are actually verification tests require complete satisfaction. For example, part of the structural validity test and the dimensional consistency test imply verification. Verification is defined as internal logical consistency of the equations – that is, bug-free coding – whereas validation assesses the relationship of the model to the real world.

Forrester and Senge indicate that some tests are more essential than others. Moreover, the tests requiring a more subjective judgment can only be passed to a degree of satisfaction, given constraints in available time and cost:

> In the early phases we can expect to establish only a degree of confidence that is less than complete. (Forrester, 1961, § 13.1)

System dynamics practitioners should be comfortable with the notion of quantifying soft or subjective variables and numerical quality ratings are commonplace. Lane (1995) suggests a validity scoring in different modes of modelling and applies this in a simplified way to the confidence that can be expected of generic structures (Lane, 1998). Huz et al (1997) provide a framework for evaluating systems thinking interventions using subjective (Likert scale) scoring.

Figure 2 combines all of these ideas into a validity scorecard. To each test a qualitative score is assigned, measuring degree of confidence, and a weight of importance attached. The significance of the weighting is to allow modeller and client to emphasise in which areas they require a model to achieve its purpose. A total score can be summed for each dimension of validity and the contribution of each one to the total validity. Further comments in the example of the construction of a qualitative model are given below.

Model sponsors often seem to view the substantive validity dimension as a hurdle to be passed, typically through the evidence of historical fits against data for quantified models. The other two dimensions of validity are contributory, rather than conflicting, aspects of an overall confidence that the model is significant and measure the real value added in the modelling process. The ‘validity quotient’ is a fraction representing the actual validity against the total possible score. For the present, we do not propose an absolute hurdle figure to indicate a benchmark of acceptable validity.

The validity quotient should be seen as an explicit, comparative indicator – rather than a normative index – to assist the modeller in achieving a balanced effort, especially in group modelling environments.

Figure 2. Validity scorecard

			A	B	A×B
	DIMENSION	**TEST** [a]Tests from (Forrester and Senge, 1980)	Score (0-5)	Weight (1-3)	Product
FORMAL or SEMI-FORMAL	SUBSTANTIVE or content/ representativeness	Tests from Forrester and Senge (1980): Structure Behaviour Policy [must include essential tests and weight appropriately eg. dimensional consistency test] For qualitative models: Internal (logical) consistency [structure verification[a]] Comprehensiveness of overview [structural boundary adequacy[a]]			
		∑ Percentage contribution to total validity			
INFORMAL	CONSTRUCTIVE or process/learning	Surprise behaviour test[a] Process is stimulating/engaging Understanding of causal mechanisms Specific actions are indicated to management Appreciation of relation of system to the environment Insights/perspectives gained that have general applicability/transferability ……….			
		∑ Percentage contribution to total validity			
	INSTRUMENTAL or outcome/impact	System improvement test[a] Model facilitates development of later versions/variants in modelling process Is implementation likely? Model instigated a management process that was previously absent Mental models surfaced/modified? Timeliness Extension to the project/application/repeat work likely Acceptable cost Consensus generated Relevance (either predefined or emergent) ……….			
		∑ Percentage contribution to total validity			
	Validity Quotient				

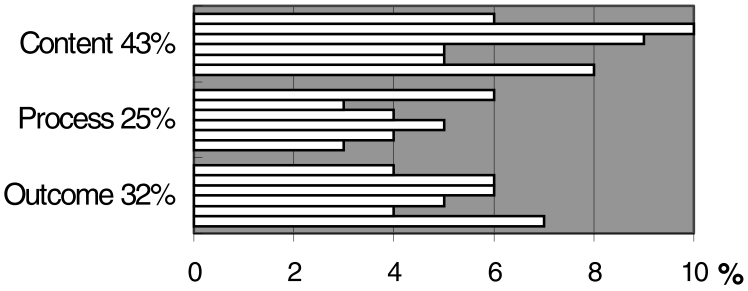
That interesting and useful models can be built should not however be seen as an excuse for carelessness in the technical aspects of model construction. Yet there is the opposite danger of building a technically correct model that is irrelevant. Forrester

(1961, § 13.2) hints at a model that has a high validity quotient but a deficient substantive validity:

> A simple and even inaccurate model may still be tremendously valuable if it yields only a little better understanding of the reasons for major success and failure.

The contributions to total validity, calculated as a percentage for each test, can be used to build up a profile, which is illustrated in Figure 3. Developing a clear purpose for a model is an oft-repeated injunction in the system dynamics literature (eg. Richardson and Pugh, 1981) but it is a task for which there are few guidelines. The responses to *World Dynamics* indicate the ease to which model purpose can be misunderstood (Andersen and Richardson, 1980). The profile serves to characterise the nature of the modelling effort and may provide a key for structuring model purpose, if an expected profile can be drawn up at the outset of modelling.

Figure 3. Generalised example validity profile



This suggests that scorecard in Figure 2 should list a broad range of tests, so that the areas in which a particular model demonstrates weaker validity are illuminated, even if they are given a lower weighting. Ideally, the number of tests in each of the three dimensions should be equal to allow for an even potential contribution across the three dimensions.

The process of model validation

Forrester (1961, 1994b), Randers (1980) and Richardson and Pugh (1981) all provide guidelines on the modelling process, noting its iterative nature. Roberts (1978), Weil (1980) and Coyle and Exelby (1999) identified the importance of considering implementation issues throughout a consulting assignment, not just as an afterthought. Forrester and Senge (1980) and Randers (1980) also mention the cycling, multistage nature of validation.

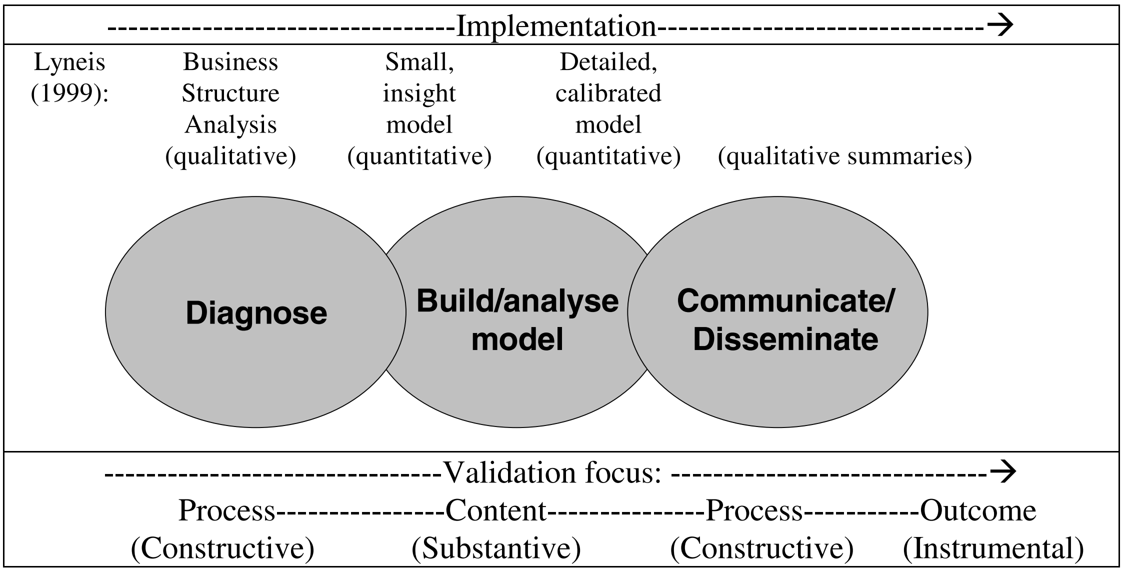
The three dimensions of validity identified above seem to need different emphases throughout the modelling process. Where there is a distinct client or problem owner, the modelling process relies on extensive communication at all stages. Figure 4 combines in a simplified way both product and process elements for a discrete modelling intervention. It shows how Lyneis' (1999) first three modelling stages (business structure analysis, small 'insight' models, detailed calibrated models) can be linked to the roles that a system dynamics consultant must perform (Els, 1997):

1. Diagnose (elicit/articulate problem)
2. Use expertise (model construction and analysis)
3. Influence (communicate results)

It is interesting to note that the detailed, calibrated model stage suggested by Lyneis affords a different kind of process facilitation: the detailed model allows selling of the results by addressing specific questions in a way earlier models cannot. The process stages for a qualitative-only study are broadly similar. The process-content-process stages in Figure 4 approximately mirror the diagram of the 'Intertwined Project Learning' Process suggested by Wolstenholme (1999) and are discussed below.

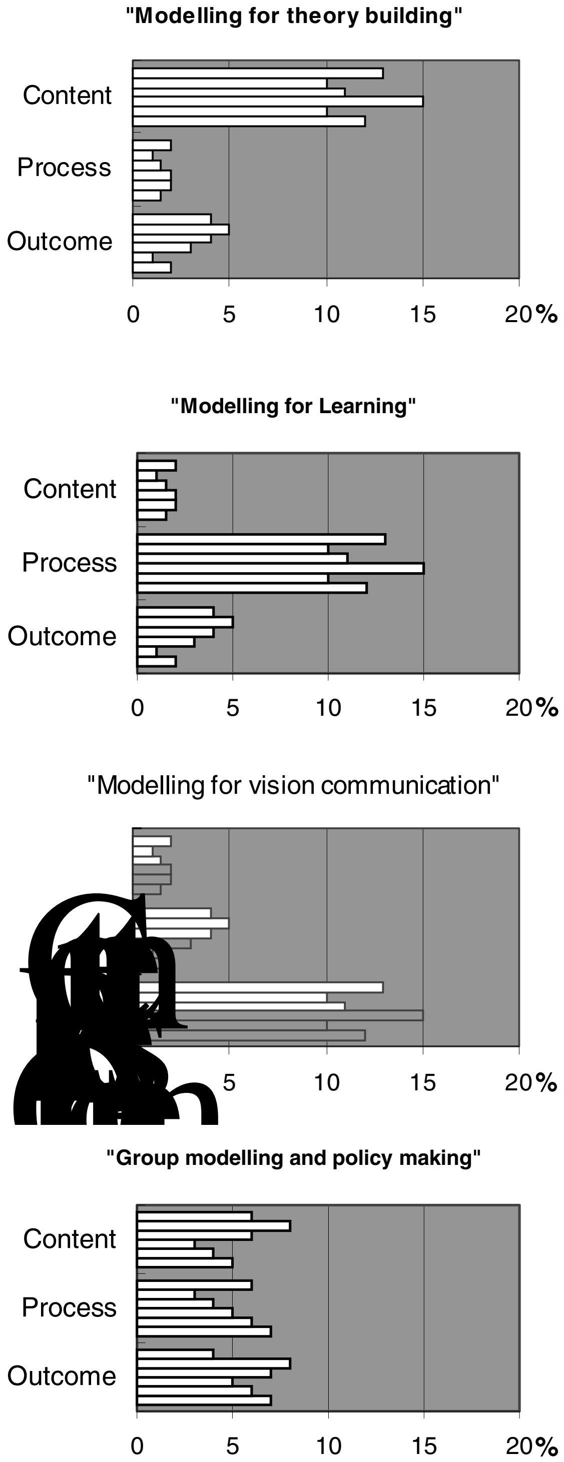
Although the three dimensions of validity offer contributions to a total validity, it may also be the case that where one dimension of validity is lacking, progress in a modelling project is effectively curtailed. For example, even where the process facilitation is effective, a project cannot proceed where model that is incapable of replicating historical behaviour to the satisfaction of the client – if that is what is expected. In the study reported by Akkermans and Bosker (1994), it was the process elements of the project design that caused the difficulties, not the model itself.

Figure 4. Schematic process of model construction, facilitation, implementation and validation



Successful modelling projects probably rely on a balanced set of substantive, constructive and instrumental elements of validity. Robinson and Pidd (1998) show that customer perception of model quality varies through a simulation model project. Consequently, it is useful to consider how the validity profile shown in Figure 3 can alter as a project progresses, with the emphasis shifting on to different dimensions. Figure 5 shows four types of validity profile, identifying each type with a putative label of different modelling modes. The labels are intended to be indicative rather than definitive.

Figure 5. Validity profiles and different modelling modes



It is also possible to position the three dimensions of validity in a hierarchy of increasing breadth of judgment (Figure 6) when facing a critic's questioning about the impact of a modelling project.

The validity profiles are themselves dynamic throughout a study, or validity could be said to be emergent. Delauzun and Mollona (1999) report on a quantified model to study the future funding of the BBC World Service. It had a strong impact amongst a largely sceptical audience but the impact was different from that originally intended. Their article is evidence that when a model's purpose is defined in terms of effect on the audience, it often seems difficult to predict the nature of that effect. Validity profiles and emergent validity may be useful concepts to plan, or understand, the way in which modelling projects can be successful or unsuccessful.

Figure 6. The hierarchy of validation stages and sceptic's questions

Increasing breadth of judgment	Validation stage	Sceptic's questions
↓	Substantive	What's the point of a comprehensive and highly representative model if no-one learns anything from it?
	Constructive	What's the use of a model that aids in developing insight if the system actors aren't able to implement the insights or policy recommendations?
	Instrumental	Was it really the model (or the facilitation process) that made the difference?

In the consulting case studies described by Winch (1993), the incidental benefit of consensus building reported came from highly detailed models. These models were critically validated with very high emphasis on the substantive dimension – they were built to evaluate specific strategic decision options. Had they been built primarily to develop consensus or cohesiveness in the management teams then the highly detailed, expensive-to-construct models would not have been necessary. Indeed, these objectives may well have been much more easily and effectively achieved through simpler models, such as the 'insight model' phase as described by Lyneis (1999).

Qualitative vs quantitative modelling from a validation viewpoint

Hall et al (1994) remark that Forrester and Senge's (1980) validation procedures show a close likeness to those used for building confidence in qualitative research causal theory. Wolstenholme (1999) comments on an ongoing dialogue between qualitative and quantitative modelling practice in system dynamics, showing that both can be used in a complementary way. It is useful to consider this debate from the viewpoint of model validity.

By considering the three dimensions of validity, both qualitative and quantitative models can be valid whilst relying on differing emphases of content, process and outcome elements, assessed against a model's purpose. Lyneis (1999) shows how both qualitative and quantitative modelling must be used together. He indicates that the purposes of individual renderings of the model have different purposes at different stages of the modelling process. Elicitation and communication of the model are more important at the beginning and end of the modelling process respectively (see Fig. 4).

Coyle (1998) summarises arguments for and against the utility of modelling projects solely using causal loop (or influence) diagrams. The substantive validity of such diagrams is limited to the structure verification (relationship accuracy) and structure boundary adequacy (comprehensiveness) tests described by Forrester and Senge (1980). The extent to which it is possible to infer dynamic behaviour and policy conclusions reliably from a qualitative model depends on its size and is still a matter for debate in the system dynamics field. For instance, Risch et al (1995) discuss corporate strategy analysis and dynamic inferences based on causal loop polarities. However, causal loop diagrams are not amenable to the detailed scrutiny of Forrester and Senge's behaviour and policy tests. The main dimension of validity is whether the causal map provides an illuminating overview of a set of issues; the usefulness must then arise from the educational benefit of thinking in a systematic and holistic way. Coyle's comments suggest that a qualitative study can still demonstrate instrumental validity by indicating a programme of ameliorative action. Where recommendations are not made more explicit by quantification, the ease of implementation must nevertheless be reduced and hence the impact assessment of qualitative interventions remains an area for further research.

In summary, the breadth of current system dynamics practice has the advantage of requiring a comprehensive overview of the three dimensions of validity. In particular, if a modelling study moves through the qualitative and quantitative phases described above and in Figure 4, a validation framework can increase the ease by which the phases are traversed.

Validation of soft OR models, scenarios and management flight simulators

Lane (1995) identified the many points of contact of different modes of system dynamics modelling with validation methods used in soft operational research. Here, we analyse using validation profiles some of the different types of modelling.

Soft OR models are built for interpretive purposes, that is, they tend to focus on constructive validity and view a model as mainly a learning device (Pidd, 1996). In the Soft Systems Methodology, the requirement that the model should correspond with the real world is dropped. Substantive validity is then defined as whether the model has been competently built and the defensibility of its internal relationships (Checkland, 1995). Cognitive mapping (Eden, 1989) similarly relies on an axiomatic or reflexive definition of substantive validity. Lacking the rigour of a quantified and simulated model, instrumental validity must depend on later subjective reports about how effectively the model galvanised action. Although system dynamics studies also rely on subjective impact assessments, they can be further corroborated by performance metrics arising from revised policies and data from measured system level variables.

Scenarios, as narratives about hypothesised futures, share some similarities in validity profiles attributable to soft OR models. Their primary aim is to stimulate organisational learning. The necessary features of scenarios are:

- relevance (directed towards a significant purpose);
- vividness (challenging/stimulating new thinking, ie. constructive validity);
- plausibility (substantive validity);
- internal consistency (verification).

Yet the substantive validity of scenarios differs from soft OR models. The components of a scenario should relate directly to the real world. Furthermore, the second and third features listed above are distinctive since they represent a balance that the scenario should be sufficiently different from extant thinking whilst still being credible. The dynamic evolution of scenarios is normally omitted, and along with it Forrester and Senge's (1980) behavioural validity tests. Winch (1999), however, discusses how scenarios can be made more dynamic. They can possess instrumental validity by raising the organisation's perception of the external environment and lead to greater readiness to respond to change.

Microworlds (management flight simulators) are also becoming commonly used as packaged management learning devices. Considering the three dimensions of validity may shed light on the continuing debate about their utility. There is accumulating evidence that the effectiveness of microworlds lies in the gain in understanding and involvement they create – their constructive validity. Milling and Maier (1997) show that microworlds seem to assist learning but are less effective as decision support tools, which was an expectation participants expressed in their trials. Hall et al. (1994) conclude that even with lengthy explanations and facilitation, users cannot determine underlying causal feedbacks embedded in a management game. This suggests that these models can have a low substantive validity, from the user's perspective even if not from the modeller's. Providing cognitive feedback via pop-up causal diagrams in a microworld leads to faster learning but does not improve ultimate performance against a benchmark (Langley, 1995). If microworlds do not confer a depth of insight of how model structure leads to behaviour, evidence of instrumental validity by improved performance in actual management tasks is unsurprisingly elusive, reminiscent of Forrester's (1961, §20.5) comments on the weaknesses of the black-box nature of management games.

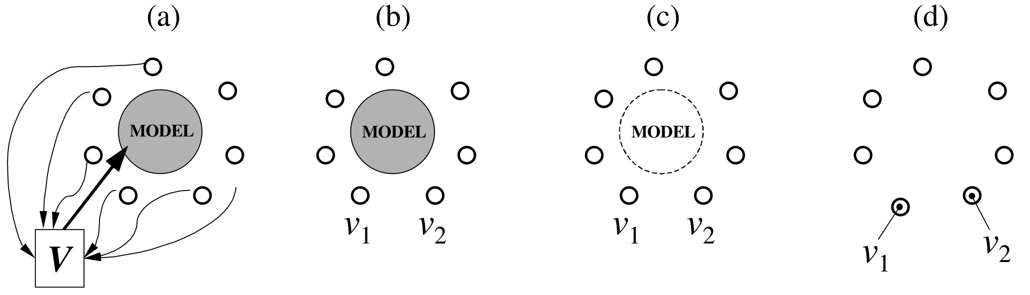
Constructing prior consensus on validity

Models can be built in a variety of group circumstances where viewpoints on the criteria for validity can develop in differing ways. Figure 7 suggests four schemas to conceptualise this process, where:

a) the validation framework, and moreover its content, can be agreed as a discrete and fixed set of criteria by all participants before the modelling starts, even if there are divergent views about the form of the model itself;
b) there is a plurality of viewpoints, such that agreement on the criteria for validity can only emerge over the course of the project;
c) the validity criteria cannot be agreed, because of methodological or other disputes;
d) there is no group modelling project at all; the model is constructed by individuals for their own personal purposes and the perception of validity criteria are entirely subjective and internal – the 'solipsism' described by Lane (1994; Burrell and Morgan 1979); however, there may be other individuals with similar perceptions or some published work to form a basis for the model.

The issue of tailoring the validation process to study objectives was raised by the authors' experiences with two applied research projects currently in progress. These projects involve the design and use of computer-aided visioning tools – simulators designed to help management teams envision their future organisations after major planned changes have been implemented. In thinking about the nature of the user groups for these particular models, the model-development process and the objectives, early experiences in the development phases led to critical thinking about the generally accepted models of validation and the process of validation as characterised in the schemas in Figure 7.

Figure 7. Validation schemas for group processes



Validation schema V	Individual schemas v_i $v_1 \neq v_2 \xrightarrow{\text{over time}} v_1 \approx v_2$	$(v_1 \neq v_2)$	$(v_1 \cap v_2 = ?)$
Group 'hard' model building eg. Winch (1993)	Group model building eg. Vennix (1996) in pluralist contexts	Non-agreed validation schemas lead to non-agreed model	Model and validation schema reside within each individual.
Validation schema ***V*** pre-agreed, where there is consensus about the model's purpose and content	Validation schema emerges,as individuals' schemas are elicited and the modelling proceeds.	eg. US Dept of Housing and Urban Development (HUD) disagreements re. *Urban Dynamics* (Alfeld, 1995)	No group process necessarily exists, so no group model is built, although the mental models of a collection of individuals might have a common theme.

The first case is a proof-of-concept research project, previously described by Winch et al (1997), aiming to apply system dynamics and management flight simulator technology to envision fundamental change in firms, including small-medium enterprises. The model is designed to serve as a communication device between a top strategy team or individual and senior managers. The change at the core of the model can derive from product or process technology innovations and/or other major structural changes in internal or external environments. This application's circumstances critically pointed to a model that is generic, but which can be parameterised to an individual firm. The approach to conceptualising such a model is outlined by Winch and Arthur (1998), and early trials with collaborating firms are also briefly documented in Winch et al. (1999).

A similar approach is being employed in the second, separate project concerning investments and benefits with IT-assisted teaching and learning (ITATL) in the higher education (HE) sector. ITATL implementation projects have largely tended to be isolated processes and are not being made by individual institutions within a strategic framework. Many projects appear to start from scratch and not build on previous experience and do not contribute to the organisation's learning curve let alone an education system wide one.

Considering the set of schemas above (Figure 7), for both projects we clearly expected a validation process related to schema (a). However, the set of model attributes that defines its validity (V) in such cases would clearly be very different to the standard 'hard' tests for models typical for those designed to give quantitative evaluations of specific options. This pointed to a need for a framework to define in developers' minds what would constitute a valid model in each individual case, and perhaps brokering an agreement between the model developers and users.

Referring to the validation framework described earlier, the validity of these visioning models is derived mainly from the instrumental dimension and secondarily from the constructive dimension. This is because the core model is generic and its degree of substantive (structural, behavioural and policy) validity is limited, although the parameterisation of the model leads to a face acceptability. Since the process facilitation did not include in-depth explanation of causal structures in the model, the 'modelling for learning' validity profile (Figure 5) typical for microworlds does not apply.

In the first case, depiction of business futures to senior management was engaging, partly by dint of the simulator interface, leading to validity on the constructive dimension. However, the main impact of the model seemed to be more in terms of its instrumental role in stimulating a new type of communication process, a willingness to carry out further model-facilitated work and its relevance in eliciting issues related to substantial changes in business structures and markets. In the second case, a major objective is to persuade HE managers to think about the factors in their own organisation that would lead to maximum benefit from ITATL. This will include the placing of their thinking about these developments within their institution's overall strategic thinking and planning – even though this broader perspective is not to be included in the model. This process will be effective – and the model therefore valid – if there is any movement in these directions in their thinking habits.

Conclusions

Considering validation from the broader perspective suggested in this paper can lead to a balanced quality assessment of a model on three dimensions. The first assesses the reliability of the content of the model – relative to a clear purpose. The second considers the value of the modelling process. The third evaluates the model's ultimate impact. Assessing validity on these dimensions can in turn aid the formulation of the purpose of a modelling project.

System dynamics can benefit from a validation framework that facilitates the development of modelling judgment and procedures for good practice. The framework proposed in this paper can explain how differing modes of modelling are effective. Validation schemas can help modellers to identify the social processes needed to negotiate the significance of a model and modelling process.

That a model can be useful should not be seen as an excuse for poor technical content. But the system dynamics professional must be willing to face the critical questions that go beyond building good models. Has anyone gained understanding or enhanced intuition about the feedbacks and dynamics of systems they are involved with? In what ways or to what degree? Still further, does the improved insight allow them to manage those systems to a greater performance level? Have new policies been created that allow people to achieve more of their potential?

Validation frameworks and profiles such as proposed in this paper can provide a basis for eliciting these decisive "so what?" questions. Providing a validation framework at the outset of a modelling project would be a way of gaining prior consensus on how model sponsors will benefit and in what areas. Validation profiles indicate from where confidence can be expected to originate and how it might change during the assignment, similar to the aims of Quality Assurance plans. Together, they also provide tools for greater clarity of thinking to distinguish means from ends.

Further accumulated evidence is needed to demonstrate the effectiveness of the proposed framework in closing the validity perception gap between model builders and customers. Nonetheless, the validity of a model is secondary to the relevance and significance of the model purpose and assumptions. Broad but explicit measures of the components of model validity are worthwhile if they contribute to clarity about goals and assumptions.

References

Akkermans H and Bosker J (1994). Design guidelines for participative business modelling projects: lessons from an unsuccessful case study. *Proceedings of the 1994 International System Dynamics conference, Stirling*. Problem Solving Methodologies, 15-24.

Alfeld LE (1974). Urban Dynamics and its critics. In: *Readings in Urban Dynamics*, vol 1. Mass NJ (ed). Wright Allen Press, Inc. Cambridge, MA.

Alfeld LE (1995). Urban dynamics – the first fifty years. *System Dynamics Review* 11(3) 199-217.

Andersen DF and Richardson GP (1980). Towards a pedagogy of system dynamics. In: *TIMS Studies in the Management Sciences, vol 14 – System Dynamics,* Legasto AA, Forrester JW and Lyneis JM (eds). North-Holland.

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

Arthur DJW and Winch GW (1998). Strategic Model Conceptualisation using Resource and Stakeholder Service Fulfilment Concepts. *Proceedings of the 10th European Simulation Symposium*, Nottingham, UK, October.

Averch H and Levine RA (1971). Two models of the urban crisis: an analytical essay on Banfield and Forrester. *Policy Sciences* 2(2) 143-158.

Brooks RJ (1999). Simulation validation – mission impossible? *Newsletter of the Operational Research Society*. April, no. 340, reporting on: Simulation Study Group: The Practice of Simulation Verification and Validation, 17 February 1999, University of Warwick, UK.

Burrell G and Morgan G (1979). *Sociological paradigms and organisational analysis*. Heinemann, London.

Cavaleri S and Sterman JD (1997). Towards evaluation of systems thinking interventions: a case study. *System Dynamics Review* 13(2) 171-186.

Checkland P (1995). Model validation in soft systems practice. *Systems Research* 12(1) 47-54.

Coyle RG (1998). The practice of system dynamics: milestones, lessons and ideas from 30 years' experience. *System Dynamics Review*. 14(4) 343-365.

Coyle RG and Exelby D (1999). The validation of commercial system dynamics models. *System Dynamics Review* (forthcoming).

Delauzun F and Mollona E (1999). Introducing system dynamics to the BBC World Service: an insider perspective. *Journal of the Operational Research Society. Special Issue: System Dynamics for Policy, Strategy and Management Education.* RG Coyle and JDW Morecroft (eds). 50(4) 364-371.

Eden C (1989). Using cognitive mapping for strategic options development and analysis (SODA). In: *Rational Analysis for a Problematic World.* J Rosenhead (ed). John Wiley, Chichester.

Els SA (1997). Better results through better communication techniques in consulting. In: *Systems approach to learning and education into the 21st Century – Proceedings of the15th International System Dynamics Conference, Istanbul.*

Forrester JW (1961). *Industrial Dynamics*. MIT Press, Cambridge, MA. Republished by Productivity Press, 1994.

Forrester JW (1969). *Urban Dynamics*. MIT Press, Cambridge, MA. Reissued by Productivity Press, 1994.

Forrester JW (1971). "The" model versus a modeling "process". *System Dynamics Review* [1985] 1(1) 133-134.

Forrester JW (1973). *World Dynamics*. Wright-Allen Press (2nd ed). Reissued by Productivity Press 1994.

Forrester JW (1976). Educational implications of responses to system dynamics models. In: *TIMS Studies in the Management Sciences, vol 2: World Modelling: A dialogue*. Churchman CW and Mason RO (eds). North-Holland.

Forrester JW (1994a). Policies, decisions and information sources for modeling. In: *Modelling for learning organisations* Morecroft JDW and Sterman JD (eds). Productivity Press, Oregon.

Forrester JW (1994b). System dynamics, systems thinking and soft OR. *System Dynamics Review* 10(2-3) 245-256.

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

Hall R, Aitchson PW and Kocay WL (1994), Causal policy maps of managers: formal methods for elicitation and analysis. *System Dynamics Review* 10(4) 337-360.

Huz S, Andersen DF, Richardson GP and Boothroyd R (1997). A framework for evaluating systems thinking interventions: an experimental approach to mental health system change. *System Dynamics Review* 13(2) 149-170.

Lane DC (1994). With a little help from our friends: how system dynamics and soft OR can learn from each other. *System Dynamics Review* 10 (2-3) 101-134.

Lane DC (1995). The folding star: a comparative reframing and extension of validity concepts in system dynamics. In: *System Dynamics 1995 – Proceedings of the International System Dynamics Conference, Tokyo.* Shimada T and Saeed K (eds).

Lane DC (1998). Can we have confidence in generic structures? *Journal of the Operational Research Society*. 49(9) 936-947.

Langley P (1995). *An experimental study of the impact of on-line cognitive feedback on performance and learning in an oil producers microworld*. Unpublished PhD thesis, University of London.

Lyneis JM (1999). System dynamics for business strategy: a phased approach. *System Dynamics Review*. 15(1) 37-70.

Meadows DH, Richardson J and Bruckmann G (eds 1982). *Groping in the dark: the first decade of global modelling*. John Wiley.

Meadows DH, Meadows DL, Randers J and Behrens WW (1972). *The limits to growth*. Universe Books, A Potomac Associates Book, New York

Milling PM and Maier FH (1997). On the effectiveness of corporate gaming environments and management simulators. In: *Systems approach to learning and education into the 21st Century – Proceedings of the15th International System Dynamics Conference, Istanbul.*

Nordhaus WD (1973). World dynamics: measurement without data. *Economic Journal*. 83: 1156-1183.

Ormerod P (1994). *The death of economics*. Faber and Faber, London.

Pidd M (1996). *Tools for thinking – modelling in management science*. John Wiley and Sons, Chichester.

Randers J (1980). Guidelines for model conceptualization. In: *Elements of the system dynamics method*. Randers J (ed.). Productivity Press, Cambridge, MA.

Richardson GP (1995). Problems for the future of system dynamics. In: *System Dynamics 1995 – Proceedings of the International System Dynamics Conference, Tokyo.* Shimada T and Saeed K (eds).

Richardson GP (1996). Problems for the future of system dynamics. *System Dynamics Review.* 12(2) 141-157.

Richardson (1997). President's address 15th International System Dynamics Conference, Istanbul. Website: http://www.albany.edu/cpr/sds/gprspeech97.htm.

Richardson GP (1999). Reflections for the future of system dynamics. *Journal of the Operational Research Society. Special Issue: System Dynamics for Policy, Strategy and Management Education.* RG Coyle and JDW Morecroft (eds). 50(4) 440-449.

Richardson GP and Pugh AL (1980). *Introduction to system dynamics modeling with DYNAMO.* MIT Press, Cambridge.

Risch JD, Troyano-Bermúdez L and Sterman JD (1995). Designing corporate strategy with system dynamics: a case study in the pulp and paper industry. *System Dynamics Review* 11(4) 249-274.

Roberts EB (1978). Strategies for effective implementation of complex corporate models. In: *Managerial Applications of System Dynamics.* Roberts EB (ed).

Robinson S and Pidd M (1998). Provider and customer expectations of successful simulation projects. *Journal of the Operational Research Society*. 49(3), 200-209.

Robinson S (1999). General concepts of simulation quality. Unpublished PowerPoint™ presentation, Simulation Study Group: The Practice of Simulation Verification and Validation, 17 February 1999, University of Warwick, UK.

Scholl G (1992). Benchmarking the system dynamics community. *System Dynamics Review* 8(3) 263-266.

Sterman JD (1984). Appropriate summary statistics for evaluating the historical fit of system dynamics models. *Dynamica* (10) II, Winter 51-66.

Vázquez M, Liz M and Aracil J (1996). Knowledge and reality: some conceptual issues in system dynamics modelling. *System Dynamics Review* 12(1) 21-37.

Vennix JAM (1996). *Group model building*. John Wiley & Sons, Chichester.

Weil H (1980). The evolution of an approach for achieving implemented results from system dynamics projects. In: *Elements of the system dynamics method.* Randers J (ed.). Productivity Press, Cambridge, MA.

Winch GW (1993). Consensus building in the planning process: benefits from a 'hard' modelling approach. *System Dynamics Review* 9(3) 317-328.

Winch GW (1999). Dynamic visioning for dynamic environments. *Journal of the Operational Research Society. Special Issue: System Dynamics for Policy, Strategy and Management Education.* RG Coyle and JDW Morecroft (eds). 50(4) 354-361.

Winch GW, MacDonald J, Sturgess S (1997). Frameworks and Tools for Computer-Aided Visioning. In: *Systems approach to learning and education into the 21st Century – Proceedings of the15th International System Dynamics Conference, Istanbul.*

Winch GW, Arthur DJW and Grey C (1999). Building and Conceptualising Generic Model Structures to address Fundamental Business Change. *International System Dynamics Conference*, Wellington, N.Z., July

Wolstenholme EF (1999). Qualitative vs. quantitative modelling: the evolving balance. *Journal of the Operational Research Society. Special Issue: System Dynamics for Policy, Strategy and Management Education.* RG Coyle and JDW Morecroft (eds). 50(4) 354-361.

Zellner A (1980). Comment on: Information Sources for Modeling the National Economy, JW Forrester. *Journal of the American Statistical Association*. 75(371) 567-569.

---

[1] Research funded by UK Economic and Social Research Council Innovation Programme: "Computer Aided Visioning for Fundamental Industry Change"