The Folding Star:
A comparative reframing and extension of validity concepts in system dynamics

David C. Lane - London School of Economics, Houghton Street, London, WC2A 2AE, U.K.

Abstract - The paper reviews ideas on validation in both mainstream OR/MS simulation and SD. A tetrahedron model derived from the former literature is adapted to SD and proposed in a new form; the Folding Star. This framework structures ideas on the elements of SD activities as well as the roles and validation measures required. Further structuring using a tabular and hierarchical format results in an interpretation of current SD validity tests and the proposal of two extensions, concerning cultural and operational issues. The framework is able to demonstrate the validity aspirations of different SD activities and to indicate areas for future development in validation tests. It also shows the respective strengths and weaknesses of different SD activities and leads to the proposal of a new form, Extended SD, an engagement between SD and 'soft' OR which aspires to a comprehensive notion of validity.

Introduction

The employment of computer simulation models in the SD field might seem to indicate some commonalities with broadly similar activities within operational research and management science. Care is needed in seeking any commonalities, however, since SD has knowingly distanced itself from many of the ideas of OR/MS, this being seen as necessary for the establishment of the new discipline (Forrester, 1961). Nevertheless, it has been argued that some cautious re-engagement might prove fruitful for SD (Lane, 1994a). This paper pursues this agenda by reframing and extending the concept of validation within SD in the light of ideas and developments from OR/MS.

The ideas underlying validity within OR/MS and SD are considered in §§1 and 2. In §§3 and 4 the 'modelling-validation' tetrahedron of Oral & Kettani (1993) is amended using a number of other ideas from OR/MS and thus adapted to the SD field, the result being the 'Folding Star' framework. In §5 existing validity tests in SD are interpreted using this framework and new ones are proposed. In §6 the framework is employed to show how various forms of SD activity may be interpreted as having different concepts of validity, these being tailored to their particular aspirations. The merits of these different validity definitions are considered in the closing section and it is suggested that a richer and more structured understanding of validity might prevent both an over-emphasis on technical quality to the exclusion of implementability and the reverse of this.

§1 Validation in Mainstream OR/MS Simulation

It seems simple to assert that a model is required to be 'valid' but as soon as definitions of that word are sought, a veritable explosion of terms appears. Models are variously required to be; 'verified', 'acceptable', 'plausible', 'implementable', 'representative', 'realistic', 'reliable', 'credible', 'convincing', 'legitimate', 'effective', 'useful', 'usable' or 'used'. Actual tests of validity - however defined - are also subject to this confusion. In considering the contributions to the literature it is useful to employ two structuring tools. Firstly, it is important to understand the level at which any contribution is operating. 'Macro' level statements about validity operate at a theoretical level and seek to establish the epistemological basis of knowledge claims. 'Meso' level statements introduce a practical note and involve an operational perspective that seeks to establish what one means in general terms by the concept of validity. Finally, 'micro' level debates concern the actual tests that one would perform to support a given concept of validity. The second approach that aids understanding is to be clear about the relationship between validation tests. Increasingly authors have broken down their concept of validity into a framework of definable, employable tests which apply to specific attributes of a model and which then combine to constitute validity. These two perspectives are employed in interpreting the contributions that are reviewed in this section.

From the very creation of OR the concept of validity was deemed to be important. Additionally, it was felt to be readily understood. From the founding work of Ackoff (1956) it is clear that it was vital that a model represent the system under study and that establishing that representativeness was the main thrust of validation. However Landry *et al.* (1983) observe that an additional measure of validity was usefulness but that that aspect was so taken for granted that it was barely mentioned.

The focus on the predictive powers of models led to one of the most influential contributions at the macro level, Naylor & Finger (1967). Their meso concept of validity, "means to prove the model to be true" (B-93) and they propose a 'multi stage approach' to validation which draws from three strands of theory. Rationalism can be used to select the variables and parameters of a model but these are treated as hypotheses only. Statistical techniques are the means of empirically 'verifying' these postulates and then, "the final decision concerning the validity of the model must be based on predictions" (B-97). However, the subsequent critiques of this paper reduce the austerity of this approach by weakening the concept of prediction and introducing the notion of usefulness.

Part of the evolution of macro ideas must be commented upon separately at this point because Naylor & Finger was a key statement of the predictive approach to validation. Landry *et al.* (1983) describe how this notion was subsequently softened when models began to be used for studying the consequences of different alternative actions. Such models cannot be validated under a predictive approach and so usefulness was brought into the validation debate as an additional, subjective measure. Déry *et al.* (1993) offer a broad view of macro ideas in OR and propose that this shift be viewed as a move from a critical rationalist, or falsificationist, approach to a utilitarian, instrumentalist philosophy. This emergence is considered further in the following selection.

Fishman & Kiviat (1968) take a strong representativeness line. Their meso level definition is that validation, "tests whether a simulation model reasonably approximates a real system" (186) and their micro level contribution is the specification of statistical tests to analyse model output with this in mind. Van Horn (1973) operates at all three levels and extends Naylor & Finger. At the macro level he comments on the rationalistic component that since models concern people, physical processes and organisational structures the representation of these, "will possess varying degrees of *a priori* confidence" (249), though, "good models for human behaviour are hard to find" (249). He supports empiricism but observes that sensitivity testing can substitute . His meso assumptions are noteworthy. Validation is "the process of building an acceptable level of confidence" (247-8). He therefore accepts the non-existence of validity proofs and accepts that, "There is no such thing as 'the' appropriate validation procedure. Validation is problem dependent" (248). This is born out in his micro level contribution which offers eight validation actions and the comment, "The real task of validation is finding an appropriate set of actions" (257). Shannon (1975) further adapts the macro contribution of Naylor & Finger to produce a 'utilitarian' approach: 'modified rationalism' (face validity of structure), empiricism (Fishman & Kiviat's tests) and 'absolute pragmatism' (usefulness in predicting behaviour). That no absolute proofs of validity exist but that validation is a process of accumulating evidence to ensure representativeness and credibility is Shannon's meso stance. These views are supported by Quade (1980) who also re-crafts this idea to describe the need to ensure that a problem is appropriately conceptualised prior to model construction. Sargent (1982) offers a framework for different sub-types of validity but these are not well defined.

Gass (1983) is concerned with models for policy-analysis on non-existent systems. His meso comments are helpful: validation concerns the confidence that those outside the building process have in a model, a judgement made by users with a purpose in mind. 'Model' validity concerns representativeness and involves 'structural' validity (does the model reasonably represent the functioning of the system?) and 'replicative' validity (does it match data already acquired about the system's behaviour?). 'Data' validity concerns the accuracy of data, whilst 'logical mathematical' validity is verification. These three, with 'predictive' validity (can the model predict data subsequently acquired about system behaviour?), form 'technical' validity. Gass adds 'sensitivity analysis' (are the recommendations insensitive to parameters?) and 'implementation' validity (does the system actually respond as indicated?) to form 'operational' validity. Finally, 'validity' is made up from 'operational' and 'dynamic' validity (can the model be updated and reviewed easily?).

Landry *et al.* (1983) offer another framework but this is superseded by Oral & Kettani (1993). Finlay (1985) draws on Gass. He also breaks down the conceptualisation process into three stages that a modeller should go through: 'the backwards look' (is the problem defined well?), 'the sideways look' (does the problem look like one tackled before?) and 'the forward look' (what is the model needed for, what are its data requirements and how will management interact with it?). Balci (1990) advances matters somewhat at the meso level, describing 'credibility assessment' but his prime contribution is a 'hierarchy of credibility assessment stages' which build up into 'acceptability of simulation results'. He lists numerous (micro) validation test that may be used. Pidd (1992) discerns two approaches. 'Black box' validation is a final stage with models where the workings remain unknown to the user and predictive power is the goal. The focus is on replicative and predictive validity. 'White/transparent box' validation concentrates on confirming a plausible structure and employs face and conceptual validity as tests. Oral & Kettani (1993) offer a tetrahedron model to structure the 'modeling-validation process' of different forms of OR. This acts as a framework for different 'types' of validation and the authors anticipate that these will lead to the more effective application of specific (micro) tests. This model is considered further in §3.

In closing we return to the macro level and the study of Déry *et al.* (1993). With the instrumentalist, utilitarian approach which emerged, the measure of validity was the usefulness of models as intellectual instruments. However, during the 1970s the notion of usefulness was increasingly seen as implying usefulness in a specific social context and this lead to the relativist, or paradigmitic stance that the knowledge claims derived from a model were determined also by the social relationships within which it was built. It is from ideas such as these that 'soft' OR emerged and interest shifted to these social relations. 'Coercing' a team's understanding using the 'truths' of OR was inappropriate. Instead a 'negotiative' approach was advanced in which OR ideas were used to elicit and structure a team's ideas and to formulate a course of action which both solved their agreed understanding of the problem and 'attended to social realities' (Eden & Sims, 1979). The latter involved proposing solutions that were implementable within the organisational culture and that were acceptable to the participants. Checkland & Scholes (1990) describe the need to perform a 'cultural analysis' which deals with 'norms, roles and values' and so create 'culturally feasible' changes. A key element is 'problem structuring' which attaches importance to the conceptualisation of situations, the creation of shared understanding of the different perspectives and how those worldviews both create responses to problems and allow different methods of resolution for them. The emergence of 'soft' OR and its methods are described elsewhere (Rosenhead, 1989 and Lane, 1994a). For our purposes two strands are important. Firstly, the attention given to the social context of, and responses to, modelling; problems are not 'solved' but 'finished' (Eden 1987). Secondly, the need to view humans as interpreting and creating their social realities and to supply tools that support this. Checkland (1995) returns to the key issue. In validating 'soft' OR models it may only be necessary to agree that they are 'relevant' as a means of illuminating a worldview. Although such models do not support experiments, their validation approach would seem to presage the triumph of usefulness over representativeness.

§2 Validation in System Dynamics

SD employs the ideas of information feedback and non-linear causal relationships and these lead to the use of computer simulation. Importance is attached to the need to have a sharp issue focus and on being able to access the mental database and to represent the mental models of problem owners. The process of model building is therefore a means of making a group's assumptions explicit in order to facilitate learning (Forrester, 1961, 1968a&b, 1969 & 1971a). The purpose is to impart, "a better intuitive feel [which] improves . . . judgement about the factors influencing company success" (Forrester, 1961, p.45). This explains the meso level view of validity in SD: validation is spoken of in terms of the 'confidence' that those using the model have in it, that confidence being created by various tests which add to the model's 'plausibility'. As a result, "In the [SD] approach validation is an on-going mix of activities embedded throughout the iterative model-building process" (Richardson & Pugh, 1981, p.311). These tests and activities are described later.

If we turn now to the macro level, we find that Bell & Bell (1980) considered refutationism as appropriate for SD since causal models offer clear test points. However, the strong practical thrust that SD shares with OR/MS lead Forrester & Senge (1980) to the conclusion that Naylor & Finger's multi-stage approach was appropriate. Radzicki (1990), in confirming the poor esteem in which economist hold SD, offers the diagnosis that a basic difference in research philosophy is located in the utilisation by the majority of economists of the logical empiricist approach, whilst SD can be seen as an example of pragmatic instrumentalism. Barlas & Carpenter (1990) similarly reject logical empiricism. However, they support the proposal that a Quinian, relativistic approach is appropriate. Validation is then, "inherently a social, judgmental, qualitative process" (p.148). Hence, "Validation is a matter of social conversation, because establishing model usefulness is a conversational matter" (p.157). This debate at the macro level is similar to that of OR/MS (as described in Déry *et al.*, 1993). So whilst we might locate SD within a functionalist social theory, there is an argument that an interpretative form of the approach is practised (Lane, 1994b). The use of group intervention methods based on the work of Schein (Lane, 1992) and a view of organisational culture from Argyris support this view (Senge, 1990a and Senge & Sterman, 1992).

Two works supply information on the micro level tests in the SD field. A series of tests are proposed in detail by Forrester & Senge (1980) and they have been organised into a table by Richardson & Pugh (1981) - see Table 1. The general nature of these tests merits comment. SD models are not appropriately judged using only standard statistical procedures. Firstly, SD models produce 'insight, not foresight'. Point prediction should therefore not be tested. Secondly, an SD model constitutes an assembly of causal hypotheses about relationships between variables which then support time-evolutionary behaviour. The shorthand for this idea is: 'the right behaviour for the right reason'. Judging the validity of a feedback model by fitting its output to behaviour data can be unrevealing. For example, it is possible to produce a correlation model giving good fit but implausible causal relationships. Or such an analysis might lose in the background error the results of a reinforcing loop which at that time was not dominant even though it might become of crucial importance in alternative runs. Finally, Mass & Senge (1978) show that regression can fail to infer from a data set the existence of a feedback link present in the model that generated the set. It is therefore vital that behaviour tests are done in association with tests of structure; building confidence in a model is a process of considering both. Naturally parameter values must be acquired and judged as being within acceptable ranges. Nevertheless, the distinctive part of the model testing process may be considered as two related activities. The variables are validated by judging whether an effective choice of variables has been made in order to express the desired activities and whether they have been connected well. Each detail of the structure must be examined, equation by equation, policy by policy. The confirmation of structure is tied in with the behaviour of the model, this being required to have characteristics close to that of observed data. Sterman (1984) suggests that system dynamicists have withdrawn too far from statistical tests of behaviour. He advocates adapted tests since this adds to credibility in a form widely used in mainstream simulation. The work of Barlas (1986) develops these tests of behaviour.

An additional contribution is that of Randers (1980), who proposes eight characteristics that an SD model should fulfil reasonably or very well. These will be considered in a later section.

§3 The 'Folding Star' Framework for the SD Modelling Process

Oral & Kettani (1993) seek to explore the modelling and validation processes of OR. Their framework is a tetrahedron in which the vertices, facets and edges all contribute to the explanatory power. The following two sections present an adaptation of this framework. This adaptation addresses some imperfections of the original by improving definitions and terminology in line with the OR literature, it introduces many ideas from Balci (1990) and is tailored for use with SD. The new framework is called the 'Folding Star' and its explanatory power is the subject of §§5 & 6.

3.1 Vertices - The four elements of SD modelling studies

Although Figure 1 shows the framework unfolded we conceptualise the SD modelling process as having four elements which are then the vertices of the three dimensional Folding Star.

The first element is 'Appreciation of the Situation', or AoS. This arises as a relevant group of individuals collects data from, and formulates views about, the world which, drawing on Checkland (1981), we may think of as consisting of 'natural systems', 'designed physical systems' and the cultural artefacts resulting from 'designed abstract systems'. Since the AoS holds the appreciation of these systems, it consists of the knowledge, understanding and interpretation of phenomena made by decision makers and policy formulators as well as aspects of the social context in which they exist. It is the source of all empirical data on real world systems as well as the interpretations of norms, values, perceptions and roles (Checkland & Scholes, 1990). These mental models are assumed to motivate some form of analysis for prediction, for the alleviation of an existing problematic behaviour, or for the assessment of a desired behaviour.

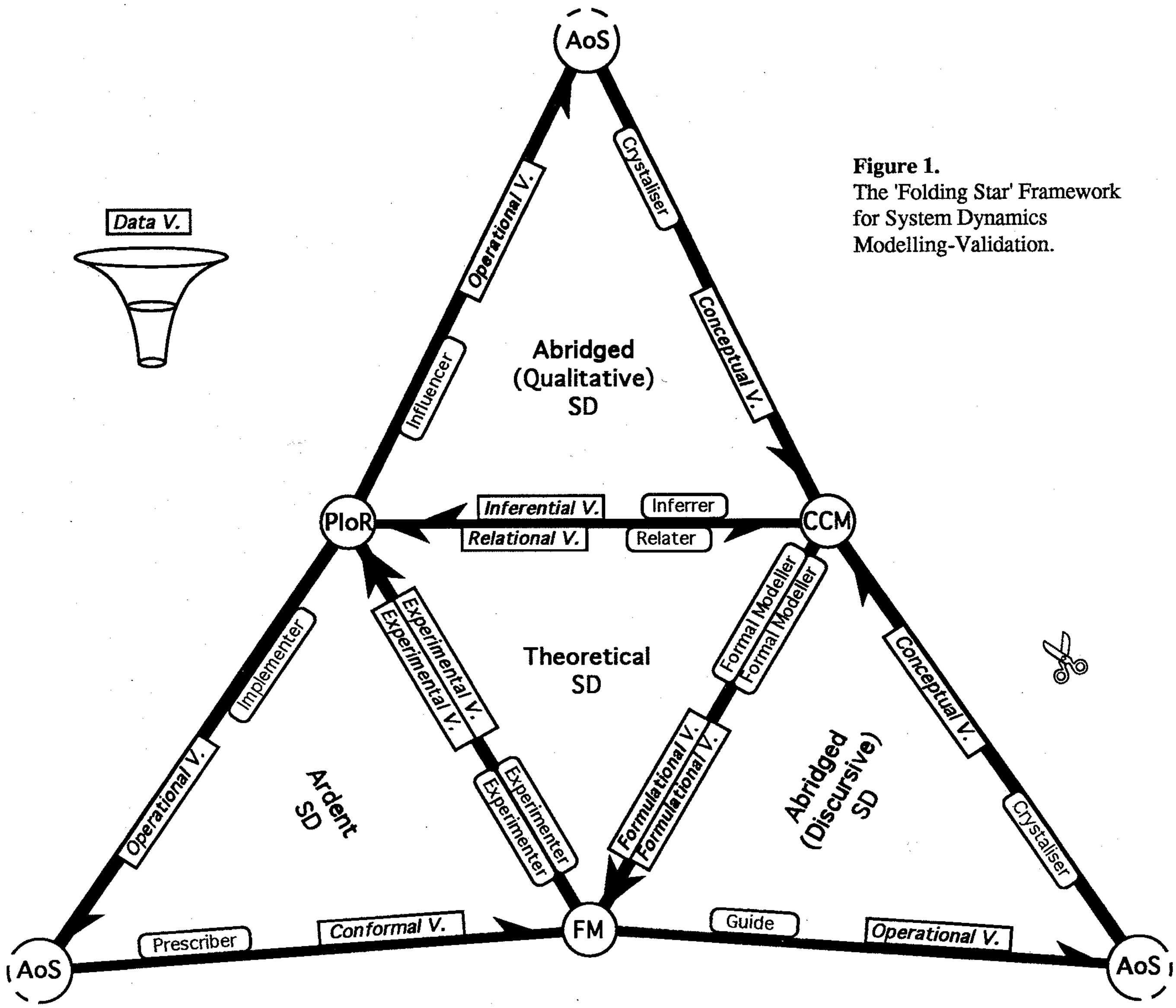

Figure 1.
The 'Folding Star' Framework for System Dynamics Modelling-Validation.

The 'Communicated Conceptual Model', or CCM, is a qualitative representation of the essential features of the AoS using an orderly framework which then allows the contained ideas to be communicated to other humans and so compared by them with the AoS in order to build understanding. Various diagramming tools would be employed for this (Morecroft, 1982) as well as the pictorial interface of some computer packages. It will also express the issue that is the focus of the study and contain the data that creates the 'dynamic hypothesis' required for an SD model, illuminated by an appreciation of different perceptions amongst the stakeholders of the issue and also located within an awareness of their norms, roles and values. The creation of a CCM implies an acceptance that SD is an appropriate tool and that the potential benefit of continuing with the approach justifies further effort. A crude distinction between the AoS and the CCM is that whilst the former deals with 'data', the latter structures, represents and shares this to form 'information'.

The 'Formal Model', or FM, is the Oral & Kettani namesake. It is a representation of the CCM in logico-mathematical terms. We may think of it as a computer simulation model in which all of the equations have been constructed so that the model is able to run a set of simulations, though board game equivalents fit this description too. The FM facilitates the conduct of experiments of ideas regarding the AoS without attempting to modify it or the real world (see Sterman, 1994). In Balci's terms this FM embraces both 'programmed model' and the 'experimental model'.

Whilst Oral & Kettani have a 'decision' vertex, we employ here the label 'Policy Insights or Recommendations', or PIoR. In order to represent different forms of SD, PIoR has both qualitative and quantitative elements. It may therefore simply involve insights, the understanding of the resistive properties to policy changes of the AoS (as represented in the CCM and/or FM) and of the points where policy interventions might alleviate current problems or generate desired modes of behaviour (Forrester, 1961). However, in addition to this the PIoR may concern an optimised solution of a model and a specific, numerical recommendation of the policies to install in the system (see Winch, 1977). This treatment then embraces a span of model deliverables from 'general understanding', through 'policy formulation' to 'detailed implementation' (Scholl, 1994).

3.2 Facets and Edges - Types of SD activity and associated roles

We would hope that any SD activity contributes to the 'appreciation' in the AoS or to the theory of the field. The different activities can be associated with different facets of the Folding Star. These are then bounded and defined by the respective edges which indicate the tasks, or roles, to be addressed during the process. The activities and roles are shown in Figure 1, which indicates by arrows the progression of tasks that constitute an activity. However, the boundaries retain some fluidity; SD interventions can involve the return to a previous task. There are four facets to the Folding Star and we associate with each a type of SD activity and its constitutive tasks or roles.

The 'Ardent SD' facet is concerned with those activities in which AoS, FM and PIoR are the main elements. On this facet we have studies in which the AoS is perceived in a direct and unproblematic way to be best addressed using the formalisms of an SD simulation model. Examples of this activity clearly include studies from the early years of SD before graphical interfaces were developed or ideas had emerged on the use of causal loop diagrams (CLDs) as a conceptualising tool (Goodman, 1974). However, this form also endures widely today, see Richmond (1987), and is significantly facilitated by the development of new software tools which ease the process of moving from AoS to FM (Richmond, 1985). Also, the use of generic structures implies a rapid move from AoS into FM (Forrester, 1969). Finally, we observe 'Strategic Management Simulation' (Lane, 1994b): the application of SD as a traditional simulation modelling approach by expert consultants as part of a planning process. Although some conceptual descriptions of the model will be used, such processes are best described as passing straight from AoS to FM so the modelling is left to experts and the model is a black-box. A consequence of this lack of involvement by management is that models tend to be rendered credible by the inclusion of large amounts of 'objective' data rather than by the use of participants' mental databases. Pugh-Roberts Associates provide an example of this activity (Lyneis, 1980). A result of the opaque process is that insights grounded in the experience of running a model are of little value; specific recommendations are required and these are frequently derived using optimisation.

The roles making up Ardent SD follow (see Figure 1). The 'Prescriber' either identifies the AoS as being of a known type and supplies a generic structure or is able rapidly to capture the relevant features whilst constructing the FM. The problems here are that that measure of relevance may be biased by the requirements of an SD model (Lane, 1993) and that a strong element of belief may be required of any clients. The move from FM to PIoR is achieved by an 'Experimenter'. He is responsible for designing and undertaking simulations which test and support the FM and which then generate improved and even optimal policies. Finally, The 'Implementer' is tasked with persuading the client group to operationalise any model recommendations. The nature of this activity means that the type of 'confidence' that will be the currency for this will involve an emphasis on the 'correctness of the model, its correct calibration and its representativeness.

On the 'Abridged (Qualitative) SD' facet both the FM and the experimenter role are removed as the process moves from AoS to CCM to PIoR and back. There may be time pressure or the view that an FM would not be useful, either because it would not be acceptable in the organisation or because it would be incapable of expressing the true richness of the AoS. This leads to the use of

CLDs, archetypes (Meadows, 1982) and 'qualitative SD' (Wolstemholme & Coyle, 1983) as means of getting insights into a system. These usages are attempts to create and employ CCMs which carry more meaning for untrained users. The emphasis in Abridged (Qualitative) SD is on tools which facilitate groups to take a systemic view of their environment and of the goals, actions and policies of the actors within it. The tools provide a language and a process through which opinions can be articulated clearly and discussed so that individuals learn about that environment and decide on action which will achieve agreed aims and which they therefore support. The richer and more participative nature of the tools used should mean this is a more negotiative, less coercive, approach. This facet ties in with what Senge (1990a) has called 'systems thinking'.

The roles in this activity commence with the 'Crystalliser', whose work embraces those described by Balci (1990) as leading to the creation of a 'communicative model'. The task here is to elicit information about the AoS and to perform problem structuring with the group. This should lead to a view of the issue and a choice of SD as the appropriate method. The Crystalliser then helps to shape the issue into an SD form, ensuring that the necessary data is brought out and that the benefits of the approach and the costs have been clarified. The 'Inferrer' works with the resulting CCM to derive insights and understanding. It is crucial to note that although the explanatory power of qualitative SD tools is supported by the underlying theory of SD they offer quite limited power in inferring insights. Abridged (Qualitative) SD closes with the role of the 'Influencer' whose job is to generate understanding and commitment to action amongst the participants. This is a more negotiative approach to moving from PIoR back to AoS.

The term 'Abridged (Discursive) SD' is applied to activities moving from AoS, through CCM to FM and back to AoS. This facet concerns the creation and provision of SD model-based management games or simulations which are used in training situations to enable experimentation and hence to support learning. This may involve direct interaction with a computer acting as a 'practice field' for managers (Sterman, 1988b and Senge, 1990a&b). Alternatively, the FM may have been transferred to a board game (see Jarmain, 1963 and Meadows, 1989). Whatever form the 'microworld' (the term widely used for such FMs) takes, there will be a protocol, or guidelines for running it. It is important to note that the issue addressed by such a microworld is specified by the group steering its creation and that it will have been strongly informed by elements of Theoretical SD, in that there will be a clear view as to what insights can, in principle, be gained from playing with the FM. The guidelines will advise on the best way to structure the users' experience so that they are introduced to these insights (see Lane, 1995). Nevertheless, no user interaction can be completely structured, it will inevitably be discursive to some extent.

The tasks needed to build and use such microworlds fall into three. The Crystalliser works with the client to specify the issue focus, or educational purpose of the microworld (Lane, 1995). This is a crucial phase, so it is assumed that the Crystalliser shapes ideas on this subject into an appropriate CCM. The 'Formal Modeller' then specifies and constructs the FM based on this CCM. In Balci's (1990) terms this will involve elements of model 'formulation' and 'representation' as well as 'programming' and some 'experimentation'. Additional emphasis will be placed on the interface or form of the microworld. The 'Guide' then introduces groups of users to the microworld and guides them through interaction with it. Such groups will start with a general interest in thinking about an issue and the Guide must shape this into specific support for the idea of addressing that issue using the available microworld. The Guide must also explain to a greater or lesser degree the underlying structure of the model - this will probably involve a conceptual description, perhaps using parts of the CCM - and help to de-brief users so that their experiences give rise to meaningful and relevant learning (Sterman, 1988b and Sterman & Senge, 1992). This SD activity is also part of Senge's (1990a) 'systems thinking'.

The central facet of the Folding Star, 'Theoretical SD', is consistent with the tetrahedron equivalent. On this facet are located two types of work. Firstly, there is modelling work done on situations that can be directly grasped by SD researchers so that a CCM can be produced in a straightforward way, or situations that are so divorced from a specific user that a CCM can only be produced by researchers. A fine example of the former is the work on Kuhn's theory of scientific development in which the original model (Sterman, 1985) and subsequent developments had CCMs drawn from Kuhn's writing. Similarly, the various global modelling studies fit the latter description (Forrester, 1971a and Meadows *et al.* 1972 and 1992). This work involves, "models . . . where it is difficult to identify the potential users . . . Such models are usually developed in the hope of raising awareness and thus influencing and shaping the perception of policy makers" (Oral & Kettani, 1993, 229). The second type of work on this facet involves the development of ideas, theories and approaches for the discipline of SD. So here we locate the ideas on the 'principles of systems' (Forrester, 1968b) and the applicability of tools (Richardson, 1985), work on chaos (Mosekilde & Larsen, 1988), accounts of 'generic structures' (Forrester, 1969 and Lane & Smart, 1994), and work on the 'validation' of microworlds (Bakken *et al.*, 1992).

When simulation is involved this facet deals in three roles, Formal Modeller, Experimenter and 'Relater'. The Relater moves between the PIoR and the researcher-created CCM and checks that the insights that have arisen from model experimentation are consistent and meaningful when related to the conceptualisation of the research issue as expressed in the CCM. Some iteration around the edges of the facet will result, in a style similar to Balci's (1990) 'redefinition' step which moves back from the stage 'simulation result' to 'system and objectives definition'.

One might wish to propose a fifth form of SD in which the CCM, FM and PIoR are all used in a practical way to inform the AoS. We call this 'Extended SD' and will deal with it later in this paper. For now we propose that the points of the Folding Star - Ardent, Abridged (Qualitative) and Abridged (Discursive) - capture the forms of SD-based interventions.

§4 The 'Folding Star' Framework and SD Modelling-validation Measures

In this section we begin to show how the Folding Star can add to the understanding of validation in SD at the meso and micro levels. We now introduce sub-types of validity. Some are specific to a facet whilst others are common to more than one. These two types are dealt with below in a facet-by-facet account and are related to roles and activities. However, we must first deal with a validity sub-type (or measure) which, following Oral & Kettani (1993), we treat as relevant to all facets.

'Data Validity', or DV, concerns the reliability and accessibility of the data that will be used at various stages of the modelling-validation process. Oral & Kettani refer to 'sufficiency', 'accuracy' and 'reliability' of data as factors increasing this measure and we can add that the concept is similar to that of Gass (1983). Oral & Kettani appeal to a broad definition of data and this is relevant for SD in which importance is attached to drawing information from 'mental databases' as well as textual and numerical ones (Forrester, 1961). The idea of using a model to manifest a worldview, or mental model, was present at the creation of the field (Forrester, 1961) and has been restated since (Forrester, 1980b & 1992). This is a distinctive feature of SD, though it has been ignored by some commentators (Flood & Jackson, 1991). The invocation in SD is therefore to use a wide range of data sources including externally stored data concerning tangible objects and knowledge about systems held only in the minds of system actors, e.g., values and goals. In consequence we can think of DV as having two features, 'facts' and 'interpretations'. Dealing now with the location of actions adding to DV, access to, and use of data are essential ingredients of all elements and roles of the modelling process. Therefore, although we treat the measure separately here, it is best seen as interpenetrating all parts of the Folding Star and as a feature of all other validity measures.

§4.1 Validation for the Ardent SD Facet

Since the tasks of the Prescriber can be seen as a sub-set of Crystalliser and Formal Modeller, it is not surprising that the validity measure, 'Conformal Validity' has features in common with the measures on those roles' edges (see Figure 1). For now it is sufficient to say that Conformal Validity concerns the extent to which the FM correctly describes and represents the well conceived AoS. There are two points here and both derive from the fact that Ardent SD can drift into 'black box' modelling. Firstly, any checking and confirmation of the actual model structure is restricted if the model users are eased out of the building process by the rapid application of a generic structure or by the rapid ('ardent') use of software to build an FM. This point is described in §6. Secondly, the use of an FM early in a process can limit the richness of the cultural description of the AoS; this point is described in §6. Hence conformance will be judged more by the modellers, the users having to use an element of trust. Similarly, representativeness will be judged by the replicative validity of the model, perhaps more than by face validation of the structure (c.f. Gass, 1983). However, trust can be won and SD models can reproduce past behaviours if judged appropriately so a medium to good target for Conformal Validity might be appropriate for this SD activity.

The tasks of the Experimenter should add to 'Experimental Validity', EV. This concerns the design and the results of the experiments that are performed using the FM and breaks into two ideas which we call EV1 and EV2. Experiments are used to challenge or to support the structural assumptions that are made in the model and some return to the tasks of the Prescriber may be necessary before behavioural tests of the FM confirm its structure and increase EV1. Secondly, the experiments form the bridge from FM to PIoR and EV2 concerns the analytical quality of the insights. A modelling-validation process achieves high Experimental Validity if the FM does generate useful insights and if those insights are rigorously supported by runs with the model and have been demonstrated to be robust by sensitivity analysis. Depending on the situation, high precision may be required of the insights rather than a more qualitative result and so optimisation procedures might also be needed to increase EV2. This measure of validity can therefore only be high if an FM has been created and if careful analysis has been performed on it.

'Operational Validity' is a multi-faceted measure which will appear often and which is the subject of structuring in §5.3. Oral & Kettani comment that it concerns the influence that the modelling has had on the AoS, the degree to which the modelling process gives rise to changes in the AoS - changed actions or appreciations, either of which constitute improvements. Operational Validity is often the deciding factor in interventions and includes factors such as breadth, depth and rigour of insights, time and cost of the work, usability, usefulness, availability, transparency and enjoyability of the FM, synergy of any policy recommendations with values and congruence with roles, and the ability of the intervention to inspire action. These and later ideas draw also on both Gass (1983) and Balci (1990). Here we suggest appropriate Operational Validity aspirations if it is created by the Implementer. For our purposes, three features are relevant; the realism of the FM, the analytical quality of the PIoR and the satisfaction felt with the process. On this facet we should recall that there are forces pulling the intervention away from 'glass box' and into black box' modelling. However, this does not prevent a model being accepted as realistic if the Prescriber and the Experimenter have done their jobs and data has been collected carefully. On the second feature, experimentation should ensure a very high quality of PIoR. Finally, it may not be that the process has strongly connected with cultural elements of the problem, or that there has been consideration of the social implementability of the recommendations but such lapses may be overcome by the trust that the system dynamicists are able to inspire and by the limited time and cost of the work that results from a rapid move to an FM. These comments begin to define Operational Validity as it applies to Ardent SD and indicate that this SD activity might have a high goal for this measure.

§4.2 Validation for the Abridged (Qualitative) SD Facet

'Conceptual Validity', CptV, concerns the relationship between the CCM and the AoS and is a particularly important feature of the SD activity located on this facet. It therefore deals with the extent to which the CCM draws on the mental models of the group so as to express and begin to make sense of the AoS in the most appropriate and beneficial way and is accepted by the relevant actors as such. Whether SD is felt to be an appropriate approach, whether the information necessary for the application of SD is coming together and whether any qualitative models are acceptable devices for addressing the issue are also elements. This is a situation in which it is correct to say that, "Clients' ideas must not just be in a model, they must be seen to be in a model" (Lane, 1992, 68). If we call these the 'model' elements then we need also introduce 'cultural' aspects. By this is meant the extent to which the CCM expresses the social and cultural aspects of the situation, meaning the feelings of the individuals and the ideas and values established by the group as acceptable and useful, the goals they wish to reach, the policies that they find acceptable to reach them and the role restrictions that apply to group members and that might become relevant if policy changes are suggested. It is the particular feature of the CCM that it is an open representation of all of these ideas so Conceptual Validity is judged predominantly by the users.

'Inferential Validity' measures the extent to which the PIoR can be inferred from the qualitative models (and other information) in the CCM. It contains weak echoes of Formulational Validity and

Experimental Validity, indeed, the tasks carried out by the Inferrer are informed by a flow of ideas from the theory base of the Folding Star since it is by this route that qualitative approaches were developed (see Forrester, 1968c & 1969 and Meadows, 1982). The ability to deal with less functional aspects of a situation, e.g. 'personal mastery' (Senge, 1990a), might be expected to broaden the Conceptual Validity goal appropriate in this SD activity. Indeed, the use of qualitative tools to give less rigorous but faster insights into dynamic systems has come to prominence with the publication of Senge's book. However, there has been considerable discussion in the field about the effectiveness of such studies. CLDs are known to be problematic in revealing the behaviour of systems (Richardson, 1985) and archetypes have their own difficulties (Lane & Smart, 1994). Hence, Richmond's (1994) observation, "using [CLDs] to make inferences about behaviour is a treacherous business" (144). Similarly Forrester (1994) holds that, "[CLDs] do not provide the discipline to thinking imposed by level and rate diagrams" (252). Work has been done to make robust the inferences from qualitative CCMs but the complexity thus introduced is open to the criticism that it approaches the use of an FM (see Dolado, 1992).

Operational Validity is shared with this facet but it must be interpreted in the context of the Influencer's role and the nature of this SD activity. However, we can use the same three features as earlier. Although no Formal Modeller or Experimenter roles are played, the cultural strength of Conceptual Validity may make up for this so that the qualitative models used might aspire to being highly representative of the problems at hand. Conversely, it is hard to see that PIoR with high analytical quality is an appropriate target, as evidenced by Forrester's comment above. On the third feature, concerning the effectiveness of the intervention as a process, Abridged (Qualitative) SD could aspire to performing well since the richness of the Crystalliser's tasks and the resulting CCM might lead one to expect a strong appreciation of cultural and social aspect of the intervention, since awareness on this front might be expected to yield more meaningful recommendations which are more congruent with the roles of the group and the values of the organisation.

§4.3 Validation for the Abridged (Discursive) SD Facet

This facet shares two validity measure so we first deal with 'Formulational Validity', or FV. This measures the extent to which the Formal Modeller correctly carries out his programming tasks so that an FM consistent with the CCM results. We extend Oral & Kettani's original, employing concepts used by Balci. Three questions (measures FV1, FV2 and FV3) will prove useful: Is the extraction of the FM from the CCM constrained by language? Is the FM representative of the CCM? Is the FM programmed correctly? FV1 is then concerned with linguistic difficulties; the possibility that the discipline of the programming language of the FM has meant that elements of the CCM have been left out or distorted. FV2 concerns representativeness; the extent to which the FM can be shown to be consistent with the CCM regarding its structure and behaviour. Finally, FV3 deals with technical validity or verification; the bug-free nature of the FM and its conformance with SD model construction guidelines, e.g. having loops which only connect stocks to flows.

Conceptual Validity on this facet measures the features described earlier but requires the addition of a goal for what Balci (1990) calls 'feasibility assessment of simulation', since in this SD activity a computer model will be created and so consideration must be given by the Crystalliser to the educational requirements of the FM. Similarly, Operational Validity can be read as having largely the same features discussed previously. However, difficulties arise in setting appropriate aspirations for any of the validity measures on this facet since it is necessary to deal with the fact that the same users are not present throughout the process. Although, as was described previously, the Crystalliser works with those who commissioned the microworld, and these same people may work with the Formal Modeller, by the time the Guide's tasks are enacted, the initiators of the work will either be few among many users or may not be on the scene (see Senge, 1990b). Hence Conceptual and Formulational Validity may have high values for the SD practitioners but without restorative action the users would be moving from AoS to FM and straight back again! It is therefore necessary for Operational Validity on this facet to embrace the factors that are included in the definitions of these previous two validities since any credibility to be derived from them may be assumed to have been lost from the point of view of the microworld users.

§4.4 Validation for the Theoretical Facet

This SD activity shares two validity measures with other facets. If we accept that researchers are the source of the CCM used in any such work then establishing Formulational Validity involves the same judgements as were described earlier. The same comments applies to Experimental Validity.

From the perspective of validation, the distinctive feature of this facet is the need to create 'Relational Validity'. This is done by moving between PIoR and CCM in both directions. What is required from such a process is the confirmation that the PIoR derived from the FM correspond with, are meaningful in relation to, the CCM. This is quite similar to Oral & Kettani's original and Balci's idea of 'credibility assessment of simulation results' (Oral & Kettani, 1993 and Balci, 1990). For Theoretical SD this process acts as an appropriate alternative to Operational Validity. So, in establishing this measure it is necessary to ensure that the PIoR are not merely model artefacts with no relevance to the CCM. One must ask; Do the PIoR make sense? Are they realistic? Are they informative? Finally, are the PIoR sufficiently innovative in comparison with the original CCM that new understandings have been created. In this later case, the new knowledge is accumulated into the theory of SD, the firm base of the Folding Star, and attempts will be made to communicate the insights to people outside the SD field, Forrester (1971a), Meadows *et al.* (1972 & 1992) and Meadows (1991) being fine examples of this intention.

§5 System Dynamics Validation Tests and the 'Folding Star'

In this section we use the Folding Star to interpret and to extend SD validity tests. We start by reframing current tests in §5.1 and then in §§5.2&3 extend the suite of tests. However, some of the existing tests, particularly those of Randers (1980), are dealt with in the later sub-sections

5.1 Interpreting Current System Dynamics Validation Tests

The aim of this sub-section is to work between meso and micro levels by relating existing tests to the Folding Star. The findings are expressed in Table 1. Both rows, but only the first two columns, contain the tests proposed by Forrester & Senge (1980), the structure being drawn from Richardson & Pugh (1981). Some of the names have been altered for consistency and clarity but the links back to the original sources are indicated.

In establishing how these test contribute to the validity measures of the Folding Star we see that CptV, FV and EV are the concerns of these tests. Particularly strong connections to DV are also indicated. The tests in cell (R&C,S) contribute to the checking that the contents of models are relevant. Such checking can be done in relation to qualitative models drawn from the AoS and to FMs derived from CCMs. Therefore these tests add to both CptV and FV2. Similar comments apply to cell (S,S), though the variable appropriateness and the verificational elements justify the inclusion of FV1&3. The behaviour focus of the second column is seen to contribute to FV and EV. In cell (R&C,B) we see relevance and consistency - representativeness - tested using data on observed behaviours. This connects with FV2 and relates to the concept of 'replicative' validity, though this must be adapted (see Sterman, 1984). Similarly Mass & Senge (1978) show that care must be taken in applying any 'Statistical Tests'. Here we also label as EV2 those tests which begin to generate insights. In cell (S,B) representativeness is further checked but now using model runs - EV1 - and more tests contribute to the generation and sensitivity analysis of insights - EV2.

5.2 Proposed Tabular Extension: Cultural Tests

In this sub-section we propose three areas where there is a need to develop tests that the Folding Star indicates are relevant. These are presented in Table 1 as a column on the right and are labelled, for reasons given below, 'Focus on Culture'. We may think of the tests as concerning the social elements of an SD activity. The tests address two areas of concern: the existence in a group of different perceptions of a problem situation and the need to attend to the 'social realities' of the group. We each in turn, describing them in more detail and proposing an extension to SD tests based on these ideas. The measures of validity from the Folding Star are then related to the tests.

Problems may be perceived in different ways. Not only is it important to establish a clear understanding of the problem at hand - this is readily accepted in the SD field - but it is important to expend sufficient time on the process. Quade (1980) cites as a potential pitfall of analysis 'insufficient attention to formulation'. It is therefore necessary to consider the diversity of opinions amongst a group before converging on a statement of the problem. The first new test, is therefore called 'Perspectives Boundary-Adequacy'. Processes contributing to such a test would be concerned with whether the CCMs support debate concerning different perspectives on the AoS, whether different ideas are considered and illuminated by this initial form of analysis. We would want to address ideas amongst the group on different policies that might be available, different possible ways of achieving goals. We would wish to consider broadly the range of goals that the group actually wishes to achieve. Testing whether a rich perspective on goals and policies has been debated is important since, "It will be possible to build many models of 'a system to launch a new product', each embodying a different world view" (Checkland, 1995, 50). The same applies to the need to address perspectives on the issue that will be the focus of the modelling. The final feature of the test is whether an SD activity has actually employed the best approach. Lane (1993) describes the difficulties of acquiring an appreciation of a problem situation without biasing one's perspective towards SD. And one might resist the representation of a social situation using an SD model. Even if a model is used in an organisational learning process in the style of Argyris (Senge, 1990a), there remains the matter of how SD models represent humans. The goal-seeking behaviour at the heart of SD probably cannot encompass Schein's 'complex model' of human motivation (Schein, 1980) but it is explicitly rejected by Vickers (and hence Checkland). He asserts that humans are best understood as individuals who 'appreciate' situations and bring their own interpretations to bear (Checkland, 1985) and this view is supported by Sagasti & Mitroff (1973). Can SD practitioners find a test of whether their approach adopts the appropriate perspective on a problem? Perhaps we must view this matter more broadly since the whole issue of methodology choice may be the most problematic part of the research agenda for OR/MS in the next decade. Finally, this test adds, as shown in Table 1, to Conceptual Validity, the 'culture' element.

The other two tests are also located in the second row of Table 1. They are more related to the social context of the group involved in modelling. The remarks concerning 'soft' OR in §1 are relevant here since the tests concern the diversity of roles, norms and values within the group and the ability of CCMs and FMs to express matters of this nature. In the terms of Checkland & Scholes (1990), modelling must be done in parallel with a 'cultural stream' which illuminates such concerns. Similarly Eden separates the 'process' from the 'content' part of a project. He criticises the urge to deal just with 'content' issues - the defined problem, the structure of a model - but advocates the need to attend to the 'social realities' that shape participants' responses, what he calls the 'process' (see Rosenhead, 1989). In Eden (1995) the reasoning is that methods have a social process dimension grounded in a social theory not measurable by strictly functionalist approaches. We propose two tests that measure an activity's performance with respect to these concerns.

Although the 'Norms/Values Boundary Adequacy' test is in the second row, there is some connection with the 'relevance and suitability' concerns of the first row. The focus is on the behaviour of participants as debated by and represented in any models. Are the goals considered or modelled consistent with the system states desired by the group? In discussing policies, are the actions based on goal and actual conditions acceptable to the group members? Testing an intervention using these questions would contribute both to CptV (the culture element) and to FV2.

The 'Roles Boundary Adequacy' test is addressed to the structure of a model, be it CCM or FM. The concern is the representation of the feedback links; are they consistent with the abilities of current actors to access, interpret and employ information in making decisions? This is an issue that is partially addressed by the work on bounded rationality by Morecroft (1983) in which he describes the need to consider and install in models the restrictions on actors' abilities to acquire information and the organisational and cognitive barriers that hinder the use of information that is available. Such testing relates both to CptV (the culture element) and to FV2.

Table 1 (Overleaf). Existing and Proposed Extension of Validity Tests for System Dynamics-based Studies. The first two columns are from Richardson & Pugh (1981), though their third row has been deleted. The right hand column and the tests in it are additions. The elements of the original columns are from Forrester & Senge (1980), the bracketed numbers revealing the relevant section. The names used for the tests are based on these sources and also Randers (1980). The ability of the tests to contribute to the Folding Star sub-types of validity are shown (bold)

Table 1	Focusing on STRUCTURE	Focusing on BEHAVIOUR	Focusing on CULTURE
Testing RELEVANCE TO AND CONSISTENCY WITH AoS Tests comparing the model representations with information, views & opinions about the system derived from the relevant actors.	• PARAMETER VERIFICATION [3.2] - Conceptual correspondence **CptV & FV2** - Numerical correspondence **DV** • STRUCTURE VERIFICATION/ FACE VALIDITY [3.1] **CptV & FV2** Are the structures in the CCM or FM right or convincing or plausible?	• BEHAVIOUR REPRODUCTION [4.1]**FV2&DV** Does the FM's behaviour match any historical data and/or the reference mode? • OTHER STATISTICAL TEST [3.6] & [4.9] • EXTREME POLICY [4.6] **FV2** When policies are pushed to extremes are the FM's behaviours reasonable? • MODE REPRODUCTION ABILITY [5.2] **FV2** With different past policies, does the FM yield behaviours consistent with other e.gs. of the system? • BEHAVIOUR PREDICTION [4.2] **EV2** Does FM reproduce the anticipated behaviour in future/hypothetical situations? • ANOMALOUS/SURPRISE BEHAVIOUR **EV2** [4.3] & [4.5] Have odd behaviours been studied to show that either: - They are anomalous, needing FM corrections to remove them? - The FM yields insights into a previously unrecognised mode?	• PERSPECTIVES BOUNDARY ADEQUACY **CptV** Do the models support debate on different perspectives in the AoS concerning: - Choice of modelling approach used? - SD issue addressed? - Goals to be achieved? - Policies for doing so?
Testing SUITABILITY FOR PURPOSE Tests focusing inward on the models; their construction & ability to yield useful results.	• STRUCTURE BOUNDARY ADEQUACY [3.4] **CptV & FV** Do the models contain sufficient and appropriate variables, policies and feedback loops to address the issue that they are being built to study? • DIMENSIONAL CONSISTENCY [3.5] **FV3** • EXTREME CONDITIONS IN EQUATIONS [3.3] **FV3** Are the outputs of policies reasonable if the inputs take extreme values?	• BEHAVIOUR SENSITIVITY [4.8] **EV1&2** Are the previous behaviour tests compromised by plausible changes in parameter values? • BEHAVIOUR BOUNDARY ADEQUACY [4.7] **CptV & EV1** Does the FM contain sufficient and appropriate variables, policies and feedback loops to address the issue when this is tested by adding new pieces of relevant structure and examining the resulting behaviour? • POLICY SENSITIVITY [5.4] **EV2** Are the suggested PIoR robust to plausible parameter changes? • POLICY BOUNDARY ADEQUACY [5.3] **EV2** Does the addition of more possibly relevant structure change the PIoR?	• NORMS/VALUES BOUNDARY ADEQUACY **CptV & FV2** Do the models support debate concerning, and represent the behaviour of the relevant actors': - Goals (are the desired states acceptable?) - Policies (are the actions based on discrepancies between goal and actual conditions acceptable within the culture?) • ROLES BOUNDARY ADEQUACY **CptV&FV2** Are the feedback links in the models consistent with the abilities of current actors in the system to access, interpret and employ information?

In closing, we should make two comments. Firstly, it is clear that judgements similar to those the proposed in the Cultural Tests are being used at the present but the situation is as described by Lane (1994a), "[SD] is no stranger to diverse viewpoints and copes with them on most occasions . . . The point is that whilst [SD] can certainly be applied to non-consensual and even divergent groups, this is less a function of the formal methodology than it is an outcome of the skills of a specific practitioner." (120). The tests given here are therefore a structuring of present activities as well as a proposal for future ones, the joint goal being the improvement of validation criteria. The second point is that there is a philosophical stance behind the Culture Tests, the interpretivism that is embraced by some 'soft' OR practitioners. The key idea is that there is no privileged mirror that one may hold up to the world. This is clear in the definition of the AoS given in §3.1. It follows then that there will be many different 'appreciations' in the AoS and that these will be constantly evolving, that this cultural diversity will need to be addressed in an intervention if a 'coercive' approach is to be avoided (Eden & Sims, 1979) and that performance on this front must be an element of any sensible framework for judging validity.

5.3 Proposed Hierarchical Extension: Operational and 'Combined' Validity Tests

In this sub-section we return to the concept of Operational Validity and propose an extension to the tests in the SD field. We have seen in §4 that this can be considered from three role perspectives: Implementer, Influencer and Guide. We propose an approach for structuring these interpretations of Operational Validity into a single hierarchy suited for SD. We go on to connect this idea with a 'Combined Validity' measure which can aid in forming a final judgement on any SD activity.

Figure 2 shows the SD validation hierarchy. It contains validity measures from the Folding Star and these therefore relate the hierarchy to specific, micro tests in the way described in §§5.1&2. The new measures are based on Oral & Kettani (1993), Gass (1983) and Balci (1990). They are proposed here as suitable elements of the confidence that users have in a modelling process. Although the figure suggests influences from Folding Star validities, the predominant role played by the new measures is agenda setting; they are intended to suggest areas in which other, specific tests should be applied or developed. However, there are connections to the tests in Richardson & Pugh (1981), to Rander's (1980) eight characteristics and to Forrester & Senge (1980).

As described here, '<u>Operational Validity</u>' concerns the question; Did, or will, the model get tried or paid attention to? Following the ideas from the OR/MS literature this is divided into two different measures, representativeness and usefulness, with names echoing this debt.

The measure '<u>Perceived Representativeness of Models', or PRoM</u>, is a judgement made on CCMs, FMs or both. It is a more broadly drawn form of Rander's second characteristic, 'descriptive realism' (or 'R2'). The concern now is whether a model's structure, data and (if relevant) behaviour represent the system that the users wanted to consider and it is therefore built up from Folding Star validities. Conceptual Validity (both elements) and Data Validity contribute to PRoM, the latter connection relating to R8, 'formal correspondence with data'. It is similarly influenced by Formulational Validity, specifically FV1 and FV2, the definition of the second indicating a linking with R3, 'mode reproduction ability'. Behaviour tests indicate a contribution from Experimental Validity and there can be a trade-off between the contributions to PRoM of Data Validity and Experimental Validity. As van Horn states, "empirical testing [c.f. Data Validity] . . . often has a lower cost substitute - sensitivity testing" (van Horn, 1971, p.251). .

The second sub-measure proposed for Operational Validity is called 'Usefulness of Intervention' and this is judged in terms of the analytical part of the SD activity and of the needs and responses of those participating in the process. It is itself immediately divided into two.

The '<u>Analytical Quality of PIoR', or AQ</u>, clearly relates to technical content. It has four contributing sub-measures, derived particularly from Balci (1990). 'Insight Generating Capacity' concerns the basic issue of whether a model does lead to any PIoR. This is a concept is drawn from both R1, 'insight generating capacity', and Richardson & Pugh's 'generation of insights' test. This test also relates to 'Relevance & Fertility of PIoR', however, this measure is predominantly inspired by Randers (1980) via R5, 'relevance', and R7, 'fertility' and indicates the need to consider whether the PIoR are innovative and important. With 'Rigour & Robustness of PIoR' the focus turns to the extent to which any insights are supported by simulations with an FM and shown by sensitivity analysis to be sturdy. The issue here is how an activity scores against an appropriate target. Finally, 'Precision of PIoR' concerns the nature of the PIoR, from qualitative insights to quantitatively precise recommendations.

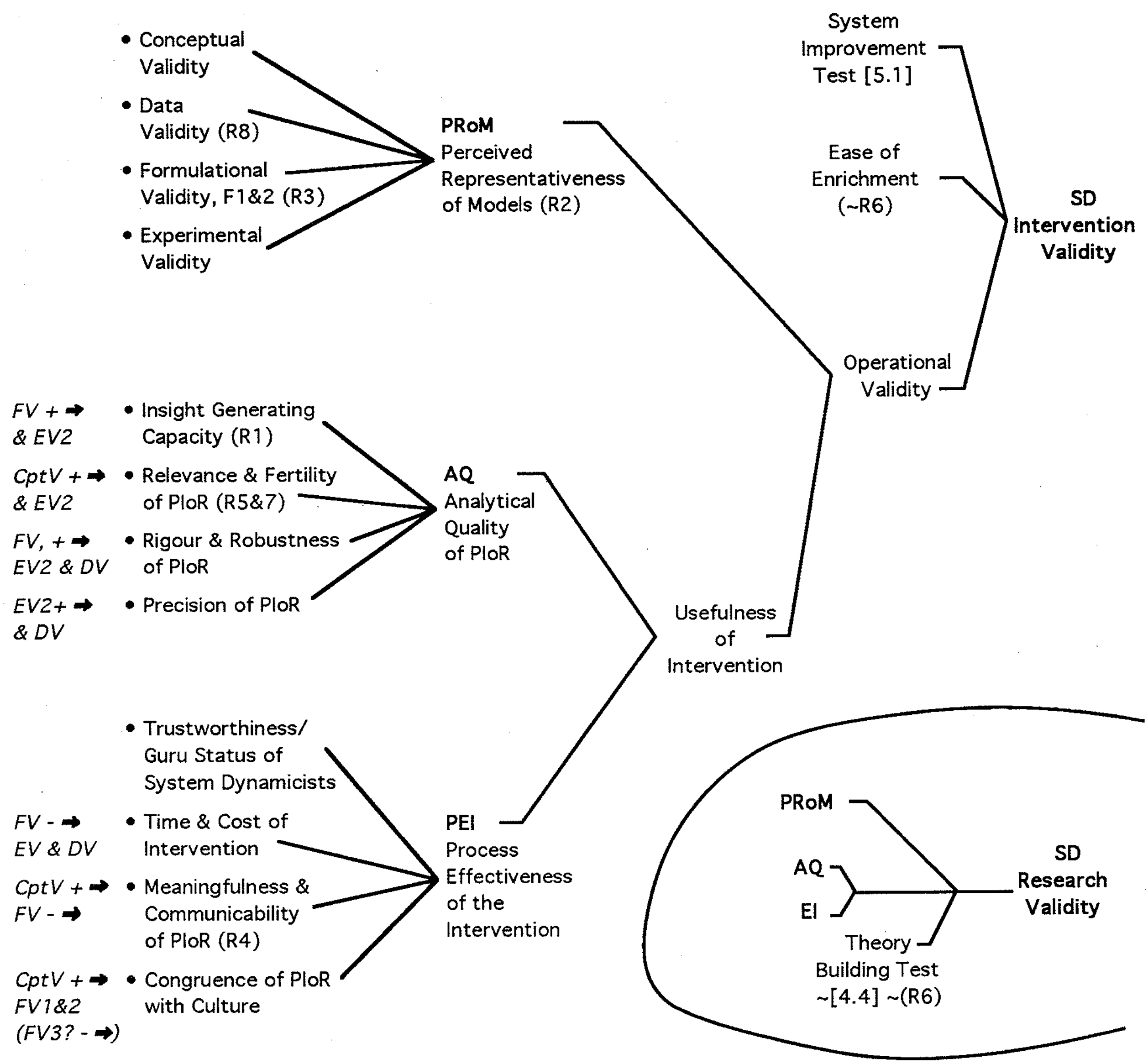

Figure 2. Proposed Hierarchical Extension of Measures Constituting 'Operational' and 'Combined' Validity. The main figure shows the breakdown of measures for different types of interventions. Theoretical activities require amended measures, shown in the inset. The (R) numbers indicate links to Randers' (1980) model characteristics.

The 'Process Effectiveness of the Intervention', or PEI, relates to the participants' responses to the social process of modelling rather than to a model. The measure, 'Trustworthiness/Guru Status of System Dynamicists', derives from Finlay (1985) since there will be a better reaction to the intervention if users respond well to those supporting the activity. 'Time & Cost of Intervention' must be included, though again this is measured against a target. With 'Meaningfulness &

Communicability of the PIoR' we are addressing many questions: How easily available are the models? How transparent are the assumptions in them? (this is related to R4, 'transparency') Is it easy to explore the model and any runs? Is it fun to do so? How much did the relevant actors participate in building the models and uncovering the PIoR? Finally, 'Congruence of PIoR with Culture' encourages us to ask about the social implementability of any proposal arising from the modelling. This is a recognised issue in the OR/MS world; "It is of very little use for analysts to compare alternatives that the policy makers cannot adopt because they involve action[s] . . . [that] have features that any potential observer could see to be unacceptable" (Quade, 1980, 24-25).

Let us now consider in broad terms the hierarchy that generates Operational Validity in Figure 1. On the right of the figure is a 'combined' validity measure, called 'SD Intervention Validity'. This is an attempt to add structure to the SD concept of 'confidence' and the measure is a surrogate for this. In addition to Operational Validity, we have two further areas of importance constructing this measure. 'Ease of Enrichment' involves concerns about the ability of any models to be updated with new data or used to test the effect of new policies. This is similar to a test from Gass (1983) though his term 'dynamic validity' has been dropped for obvious reasons and replaced with one from Randers (1980) since it involves most of his R6. The final measure expresses the prime aspiration of SD. We therefore call this the 'System Improvement Test' drawing on Forrester & Senge (1980), to ask: Did it work? Did the SD intervention improve the system?

The terms used above indicate an authorial bias in this 'combined' measure towards organisational interventions. This hierarchy might therefore seem to be relevant only to activities in the points of the Folding Star, perhaps excluding Theoretical SD. The contribution that simulation can make to theory building is considerable so we would obviously wish to validate the models used for such purposes. This is an area of some debate in SD, as evidenced by the interplay in *System Dynamics Review* 8(1) from 1992 concerning Sterman's (1985) model of Kuhn's theory and the validation of theoretical models. However, we can develop the validation hierarchy to address this issue to some extent, as shown in the insert to Figure 2. We use PRoM, AQ and parts of PEI, removing the measures 'Trustworthiness/Guru Status of System Dynamicists' and 'Congruence of PIoR with Culture' to get just 'PE'. We introduce a fourth measure, 'Theory Building Test'. The precise nature of this test may still constitute a research area for the field but the sense of it is the question; Does the modelling work add to the theory base of SD and/or will it be paid attention to? We would relate this test to elements of R6 in Randers (1980) and would certainly include a known test, the 'Family-member test', from Forrester & Senge (1980). These four elements then form 'SD Research Validity'. We accept that this additional hierarchy probably requires more fleshing out than the main one but this formulation relates back to the Folding Star and therefore ensures that that framework can contribute to the understanding of Theoretical SD.

§6 Validity Aspirations in the System Dynamics Activities

In this section we use all of the previous work to study the different SD activities. As in Landry *et al.* (1983), we are dealing not with achieved but rather the desired levels of validity. We shall find that the activities may be interpreted as having different concepts of validity, tailored to their aspirations. We perform this study using the five types of validity employed in Figure 2. However, we need to interpret the remaining validities in terms of these. The results are in Table 2, the 'transformations' in this table following from the definitions of the validities given previously.

We are now in a position to use the hierarchy of Figure 2 to find the Operational validities for the SD activities in the Folding Star. We break that validity measure down into three; PRoM, AQ and PEI, as shown in the figure. The results are summarised in Table 3, the entries of which follow from previous descriptions. The key point to notice is that all five activities aspire to good scores for Operational Validity but that those scores are seen as arising for different reasons.

With Ardent SD, PEI might be set at a low level of accomplishment. Moving rapidly to an FM can decrease time and cost and this is a significant factor but it is overcome by the lower scores in PEI which result from the low score for the culture element of Conceptual Validity. The potentials for Formulational Validity (converted from Conformal) and Experimental Validity give the highest score for AQ; in this SD activity we have a well-constructed model which is studied carefully to give well-grounded PIoR. These effects would be expected to overwhelm any reduction caused again by the low Conceptual Validity (culture) score via 'Relevance & Fertility of PIoR'. A similar argument applies to PRoM, though here the culture issues must yield a medium aspiration.

Table 2	Conformal V.	Inferential V.	Relational V. (in Theoretical SD)	Relational V. (in Extended SD)
⇒ Conceptual V. Culture Model	• ⬤	—	n/a ⬤	● ● (Adds to)
⇒ Formulational V. FV1 FV2 FV3	● ❍ ⬤	• ● n/a	—	—
⇒ Experimental V. EV1 EV2	—	• •	—	—

Table 2. (Above) Transformation of Four Validity Sub-types into Standard Three Types.

Key (both tables): • - low; ○ - low, possibly medium; ● - medium; ❍ - medium, possibly large and ⬤ - large.

Table 3. (Below) Validity Aspirations of Different System Dynamics Activities.
Maximum possible levels broken into four validity sub-types and summed into 'Operational Validity'.

Table 3	Ardent	Abridged (Qualitative)	Abridged (Discursive)	Theoretical	Extended
Conceptual V. Culture Model	• ⬤	⬤ ●	● ⬤	n/a ⬤	⬤ ⬤
Formulational V. FV1 FV2 FV3	● ❍ ⬤	• ● n/a	● ● ⬤	● ⬤ ⬤	⬤ ⬤ ⬤
Experimental V. EV1 EV2	⬤ ⬤	• •	(⬤ ○)	⬤ ⬤	⬤ ⬤
Data Validity Interpretations Facts	• ⬤	⬤ •	○ ⬤	n/a ⬤	⬤ ⬤
Operational V. PEI AQ PRoM	• ⬤ ●	⬤ • ❍	❍ ● ●	(⬤ ⬤⬤)	⬤ ⬤ ⬤

In the case of Abridged (Qualitative) SD, the desired Formulational and Experimental validities are limited because there is no FM; low levels only are provided via Inferential Validity. The emphasis here is on the cultural element of Conceptual Validity and the interpretations aspect of Data Validity. This bias is synthesised in the desired make-up for Operational Validity. The PRoM might be medium, perhaps high depending on the responses of the users to the qualitative approach and AQ could only ever seek a low level of validity. The target is clearly the PEI elements, wherein this activity seeks to perform very well. This activity emphasises a rich conceptualisation, like much of 'soft' OR, expending much effort on the 'backward look' (Finlay, 1985).

With Abridged (Discursive) SD two issues arise. The issue of ownership has been discussed in §4.3. The models may have Conceptual and Formulational (and even Experimental) validities for their commissioners but effort must be made by the Guide to establish these with a new audience, a task known to be difficult. So although the FM may be enjoyable to use, the Meaningfulness &

Communicability of the PIoR might be lower. Similar remarks relate to the AQ aspiration but this leads to the second issue: the users are acting as Experimenters. They may be guided but there will still be an element of serendipity in their ability to discover PIoR. They cannot substitute for the Experimenter. This would surely reduce EV2 and so a medium AQ and PRoM is set.

Theoretical SD can expect to be strong on Formulational and Experimental validities, though the handling of the work solely by researchers may lead to a rather technical specification, so reducing FV1. Here Conceptual Validity is mapped from Relational Validity; the culture element is not applicable without users but high achievement on the model element might be expected. As explained in §5, Operational Validity is not the measure here but the cell has been completed. This indicates high aspirations, though the truncated form of PEI should be remembered. The missing element is, of course, the Theory Building Test since this leads up to SD Research Validity.

Having seen the different ways that SD activities aspire to be valid, we can propose a hybrid, the 'Extended SD' mentioned at the end of §3. The first step of this proposed form moves from AoS to CCM, embracing the conceptual and cultural richness of Abridged (Qualitative) SD offered by the Crystalliser. However, the attachment to rigorous simulation formulation and experimentation is preserved; 'white/glass box' modelling is done in moving from CCM to FM and on to PIoR. These can then be related back to the AoS, adding to Cultural Validity (see Table 2, last column). Operationalisation is then achieved by a combination of Implementer and Influencer tasks. This is then a form of Ardent SD front-ended by careful problem structuring and interpenetrated by attendance to social realities. The aspirations are high - indicated by the desired scores in Table 3.

§7 Conclusions and Issues Raised

The case made in this paper is based on a comparison of SD with mainstream simulation in OR/MS. One conclusion is that the ideas of validity in the two are not so far apart. Different terms are used for similar ideas, though our distinctive vocabulary sometimes points up a real difference in the aspirations of our discipline. But there are many similarities. Improved understanding of the points of contact show that our field is much less of a loner than we tend to portray it. Sterman (1984) shows that SD models can stand up to many mainstream simulation credibility tests and this first conclusion supports his view; SD is already obeying many of the same rules as simulation and recognition of this can only add to the credibility of our field.

Secondly, the Folding Star, and the tabular and hierarchical extensions offer a structure which indicates that different SD activities take a different view on the types of validity relevant to them - those edging the appropriate facet. However, they can still be successful in those terms. Ardent SD aims to access the strong simulation theory of SD but cannot hope to perform too well on the cultural factors and so there is a reduction of process effectiveness (PEI), resulting from the low targets on Conceptual (cultural element) and Data validities. Abridged (Qualitative) SD may attempt a richer social intervention but at the expense of low analytical quality (AQ). Indeed, this type of 'Systems Thinking' is bordering on being merely an 'empathetic' approach, having little attachment to the benefits of formulation and experimentation. This approach has much in common with 'soft' OR processes but they lack the distinctive feature of SD: the provision of simulation models for the conduct of meaningful experiments. Abridged (Discursive) SD "attempt[s] to solve the representation of human behaviour by inserting a person directly into the simulation" (van Horn, 1971, 249), and thus reintroduces simulation and reaches for cultural richness but this is hampered by the change of participants (and consequent loss of cultural elements) and the partially-structured form of experimentation. A conclusion of this paper is therefore that the Folding Star acts as a map of the strengths and weaknesses of these three types of SD-based interventions. It is important to understand these different aspirations of different types of SD.

The Folding Star illustrates that the current validity tests are predominantly technical ones, gathered around the theory base of SD. Is it so easy to conceptualise a situation? Always interpreting situations solely in terms of our model of humans and organisations is a very specific way of viewing a problem, perhaps too influenced by ideas on physical systems (c.f. van Horn). Is it so straightforward to influence users' views? Not emphasising the social realities of groups is to border on a coercive consulting style. Standing between the base of the Folding Star - our ideas of SD - and the real world of natural and designed physical system is the AoS. The theory base is not a privileged mirror in which we may view reality directly, it is always and only from the mental models of relevant actors that we can acquire knowledge. It is only by influencing these elements of the AoS that we can change the world. SD requires more structured work on methods of appreciating situations and of influencing them. The third comment is therefore that the steps from and back to the AoS may be the hardest ones to understanding - that is why they are the longest edges of our framework, yielding not a tetrahedron but a star to be folded up. However, this paper attempts to set an agenda for this work and to propose a structure with which it might be begun.

The final conclusion is that an Extended SD is called for since it would overcome the limitations described above and would seem to offer comprehensive, 'broad-band', benefits (c.f. Eden). We might argue that the software available today blurs the distinction between a CCM and an FM, so that Ardent SD already offers these benefits. To some extent this is true but the definition of CCM used here is much more than a qualitative map, a user-friendly interface and some involvement of users in model building. Achieving Conceptual Validity requires a careful management of the whole social 'mess' of problem solving. At the macro level, this would be a new form of multi-stage validation but with the addition of interpretative considerations (c.f. Naylor & Finger and Checkland). Extended SD would therefore require problem structuring tools and a social awareness that nurtures the whole process. It must also involve the rigour of simulation models relevant to the participants, carefully constructed and producing useful analysis and commitment to action (see AQ and PEI). Some suitable approaches in 'soft' OR are already available. The integration of SD with Eden's SODA and Checkland's SSM have been proposed in Lane (1994a) and the last is detailed in Lane & Oliva (1994). This might be a way of moving towards Extended SD, the approach that this work recommends as offering improved SD Intervention Validity.

Hollinghurst's (1995) Folding Star calls shepherds back to the fold and it also promises embrace. The hope of this paper is that the Folding Star, combined with the tabular and hierarchical extensions, will help readers to embrace a better framed view of SD validation. In accepting the attractions of qualitative modelling whilst revealing its severe limitations, it may also advance a richer view of interventions but still return practitioners to the fold of rigorous simulation.

References

Ackoff, R.L. 1956. The Development of Operations Research as a Science. *Opns. Res.* **4**:265-295.
Bakken, B., J.Gould & D.Kim. 1992. Experimentation in learning organizations. *EJOR* **59**(1): 167-182.
Balci, O. 1990. Guidelines for Successful Simulation Studies. In O.Balci, R.P.Sadowski & R.E.Nance (Eds.) *Procs. of the 1990 Winter Simulation Conf.* (pp.25-32). Piscataway, NJ: IEEE.
Barlas, Y. 1986. Multiple tests for validation of system dynamics type of simulation models. *EJOR* **42** (1): 59-87.
Barlas, Y. & S.Carpenter. 1990. Philosophical roots of model validation: two paradigms. *SDRev.* **6**(2):148-166.
Bell, J.A. & J.F.Bell. 1980. System Dynamics and Scientific Method. In J.Randers, (1980).
Bell, J.A. & P.M.Senge. 1980. Methods for enhancing refutability in system dynamics modeling. In Legasto *et al.*, 1980.
Checkland, P.B. 1981. *Systems Thinking, Systems Practice*. Chichester: Wiley.
__. 1985. From Optimization to Learning: A Development of Systems Thinking for the 1990s. *J. Opl. Res. Soc* **36**(9): 757-767.
__. 1995. Model Validation in Soft Systems Practice. *Systems Research* **12**(1): 47-54.
Checkland, P.B. & J.Scholes. 1990. *Soft Systems Methodology in Action*. Chichester: Wiley.
Clemson, B, Y.Tang, J.Pyne & R.Unal. 1995. Efficient Methods for Sensitivity Analysis. *SDRev.* **11**(1):31-49.
Déry, R., M.Landry & C.Banville. 1993. Revisiting the Issue of Model Validation in OR. *EJOR* **66**: 168-183.
Dolado, J.J.. 1992. Qualitative Simulation and System Dynamics. *SDRev.* **8**(1):55-81.
Eden, C. 1987. Problem Solving or Problem Finishing? In M.C.Jackson & P.Keys (Eds) *New Directions in Management Science* (pp.97-107). Aldershot, UK: Gower.
__. 1995. On Evaluating the Performance of 'Wide-band-GDSS's. *EJOR* **81**: 302-311.
Eden, C. & D. Sims. 1979. On the Nature of Problems in Consulting Practice *OMEGA* **7**(2): 119-127.
Finlay, P.N. 1985. *Mathematical Modelling in Business Decision-making.* London: Croom Helm.
Fishman, G.S. & P.J.Kiviat. 1968. The Statistics of Discrete-event Simulation. *Simulation.* **10**(4):185-195.
Flood, R.L. & M.C.Jackson. 1991. *Creative Problem Solving: Total Systems Intervention.* Chichester: Wiley.
Forrester, J.W. 1958. Industrial Dynamics: A Major Breakthrough for Decision Makers. *HBR* **36** (4): 37-66.
__. [1961] 1985. *Industrial Dynamics*. Cambridge, MA: MIT Press.
__. 1968a. Industrial Dynamics - A Response to Ansoff and Slevin. *Man. Sci.* **14**(9): 601-618.
__. 1968b. *Principles of Systems*. Cambridge, MA: MIT Press.
__. 1968c. Market Growth as Influenced by Capital Investment, *Industrial Management Review* **9**(2): 83-105.
__. 1969. *Urban Dynamics*. Cambridge, MA: MIT Press.
__. 1971a. *World dynamics*. Cambridge, MA: Wright-Allen Press.
__. [1971b] 1985. "The" model versus a modeling "process". *SDRev.***1**(2):133-134.
__. 1980a. System Dynamics - future opportunities. In Legasto *et al.*, 1980.
__. 1980b. Information Sources for Modelling the National Economy. *J. Am. Stat.Assoc.* **75**(371):555-579.
__. 1992. Policies, decisions and information Sources for Modeling. *EJOR* **59**(1): 42-63.
__. 1994. System Dynamics, Systems Thinking, and Soft OR. *SDRev.* **10**(2-3): 245-256.
Forrester, J.W. & P.M.Senge. 1980. Tests for building confidence in system dynamics models. In Legasto *et al.*, 1980.
Gass, S.I. 1983. Decision-aiding Models. *Opns. Res.* **31**:603-631.
Goodman, M.R. 1974. *Study Notes in System Dynamics*. Cambridge, MA: MIT Press.

Hollinghurst, A. 1995. *The Folding Star*. London: Vintage.
Jarmain, W.E. (Ed.) 1963. *Problems in industrial dynamics*. Cambridge, MA: MIT Press.
Landry, M., J.-L.Malouin & M.Oral. 1983. Model Validation in Operations Research. *EJOR* 14:207-220.
Lane, D.C. 1992. Modelling as Learning: A consultancy methodology for enhancing learning in management teams. *EJOR* **59**(1): 64-84.
__ 1993. The Road Not Taken: Observing a process of issue selection and model conceptualization. *SDRev*. **9**(3): 239-264 .
__. 1994a. With A Little Help From Our Friends: How system dynamics and 'soft' OR can learn from each other. *SDRev*. **10**(2-3): 101-134.
__. 1994b. Social theory and system dynamics practice. *Procs. of the 1994 Int. System Dynamics Conf. - System dynamics: methodological and technical issues* (E.Wolstenholme and C.Monaghan, Eds.), pp.53-66. Stirling, Scotland: Univ. of Stirling. Available from the author.
__. 1995. On a Resurgence of Management Games and Simulations. *J. Opl. Res. Soc.* **46**(5):604-625.
Lane, D.C. & R.Oliva. 1994. The Greater Whole: Towards a synthesis of SD and SSM. *Proceedings of the 1994 Int. System Dynamics Conf.: Problem-solving methodologies* (E.Wolstenholme and C.Monaghan, Eds.), pp.134-146. Stirling, Scotland: Univ. of Stirling and *EJOR*, to appear.
Lane, D.C. & C.Smart. 1994. Mad, Bad and Dangerous to Know? The evolution and limitations of the 'generic structure' concept in S.D. *Procs. of the 1994 Int. System Dynamics Conf. - System dynamics: methodological and technical issues* (E.Wolstenholme and C.Monaghan, Eds.), pp.67-77. Stirling, Scotland: Univ. of Stirling. Also *SDRev.*, to appear
Legasto, A.A., J.W.Forrester & J.M.Lyneis (Eds.). 1980. *System Dynamics*. TIMS Studies in the Management Sciences Vol. 14. Oxford: North-Holland.
Lyneis, J. M. 1980. *Corporate Planning and Policy Design*. Cambridge, MA: Pugh-Roberts Associates, Inc.
Mass, N.J. & P.M.Senge. 1978. Alternative tests for the selection of model variables. *IEEE transactions on systems, man and cybernetics* June, **8**(6): 450-459.
Meadows, D.H. 1980. The Unavoidable A Priori. In J.Randers, 1980.
__. 1982. Whole earth models and systems. *Coevolution Quarterly* Summer: 98-108.
__. 1991. *The Global Citizen.* Washington, DC: Island Press.
Meadows, D.H., D.L.Meadows, J.Randers and W.W.Behrens. 1972. *The Limits To Growth.* New York: Universe Books.
Meadows, D.H., D. Meadows and J. Randers. 1992. *Beyond The Limits*. London: Earthscan.
Meadows, D.L. 1989. Gaming to implement system dynamics models. In *Computer-based management of complex systems* (P.M.Milling and E.O.K.Zahn, Eds.), pp.635-640. Berlin: Springer.
Morecroft, J.D.W. 1982. A Critical Review of Diagramming Tools for Conceptualising Feedback System Models. *Dynamica*. **8**(I): 20-29.
__. 1983. System Dynamics: Portraying Bounded Rationality. *OMEGA* **11**(2):131-142.
Mosekilde, E. and E.R. Larsen. 1988. Deterministic chaos in the beer production-distribution model. *SDRev*. **4**(1-2): 131-147.
Naylor, T.H. & J.M.Finger. 1967. Verification of Computer Simulation Models. *Man. Sci.* **14**(2):B92-B101.
Oral, M. & O.Kettani. 1993. The Facets of the Modeling and Validation Process in Operations Research. *EJOR* **66**:216-234.
Quade, E. 1980. Pitfalls in Formulation and Modelling. In G.Majone & E.Quade (Eds.) *Pitfalls of Analysis* (pp.23-43), Chichester: Wiley.
Radzicki, M.J. 1990. Methodologia oeconomiae et systematis dynamis. *SDRev*. **6**(2):123-147.
Randers, J. (Ed.) 1980. *Elements of the System Dynamics Method.* Cambridge, MA: MIT Press.
Richardson, G.P. 1985. Problems with causal-loop diagrams. *SDRev.* **2**(2): 158-170.
Richardson, G.P. and A.L. Pugh III [1981] 19--. *Introduction to System Dynamics Modeling with DYNAMO*. Cambridge, MA: Productivity.
Richmond, B. 1985. STELLA: Software for bringing system dynamics to the other 98%. In D.F.Andersen, N.B.Forrester & M.E. Warkentin (Eds) *Procs. of the 1985 Int. Conf. of the System Dynamics Society: Volume II*. (pp.706-718). Boston, MA: System Dynamics Society.
__. 1987. The Strategic Forum: From vision to operating policies and back again. High Performance Systems Publications, 145 Lyme Road, Suite 300, Hanover, NH 03755, USA.
__. 1994. Systems Thinking/System Dynamics: Let's just get on with it. *SDRev*. **10**(2-3):135-157.
Rosenhead, J. (Ed.) 1989. *Rational Analysis for a Problematic World.* Chichester: Wiley.
Sagasti, F.R. & I.I.Mitroff. 1973. Operations Research from the Viewpoint of General Systems Theory. *OMEGA*. **1**(6): 695-709.
Sargent, R.G. 1982. Verification and validation of Simulation Models. In F.E.Cellier (Ed.) *Progress in Modelling and Simulation* (pp.159-169), London: Adademic Press.
Schein, E.H. 1980. Organizational Psychology (3rd Ed.). Englewood Cliffs, NJ: Prentice Hall
Scholl, G.J. 1994. Results of the 1993 System Dynamics Society Benchmarking Study. *Procs. of the 1994 Int. System Dynamics Conf. - Methodological and technical issues* (E.Wolstenholme and C.Monaghan, Eds.), pp.226-236. Stirling, Scotland: Univ. of Stirling.
Senge, P. 1990a. *The Fifth Discipline: The art and practice of the learning organization.* New York: Doubleday/Currency.
__. 1990b. Catalyzing Systems Thinking Within Organizations. In *Advances in Organization Development*. F.Massarik, (Ed.), Norwood, NJ: Ablex.
Senge, P & J.D.Sterman. 1992. Systems Thinking and Organizational Learning. *EJOR* **59**(1): 137-150.
Sterman, J. 1984. Appropriate summary statistics for evaluating the historical fit of system dynamics models. *Dynamica* **10**(winter): 51-66.
__. 1985. The Growth of Knowledge: Testing a Theory of Scientific Revolutions with a Formal Model. *Technological Forecasting and Social Change*. **28**(2): 93-122.
__. 1988a. A Skeptic's Guide To Computer Models. In *Foresight and National Decisions* (L.Grant, ed.), pp.133-169. Lanham, MD: Univ. Press of America.
__. 1988b. People express management flight simulator, software available from Sloan School of Management, Cambridge, MA 02139, USA.
__. 1994. Learning In And About Complex Systems. *SDRev*. **10**(2-3): 291-330.
van Horn, R.L. 1971. Validation of Simulation Results. *Man. Sci.* **17**(5):247-258.
Winch, G.W. 1977. Optimisation Experiments with Forecast Bias. *Dynamica* **2**(3).
Wolstenholme, E.F. & R.G.Coyle. 1983. The Development of System Dynamics as a Methodology for System Description and Qualitative Analysis. *J. Opl. Res. Soc.* **34**(7): 569-581.