APPENDIX B

Table 1: Criteria for quality of SD models and distribution in the literature

Type	N	Criteria	Description	Forrester & Senge, 1980	Meadows, 1980	Richardson and Pugh, 1981 (apud Lane, 1985)	Barlas, 1989	Balci, 1994	Lane, 1995	Barlas, 1996
			Tests							Distrib
Test of model structure: Before testing the behavior and outputs of the model, this step of testing will focus on the structure of the model and the system it is simulating.	1	Structure-verification	By comparing the structure of the model with the structure of the real system the model represents. To pass, the model should not contradict the knowledge about the real system.	x		x	x	x	x	x
	2	Parameter-verification	Model constants should correspond conceptually and numerically to the real system. Conceptually means the variable modelled and the real-world unit must be congruent, and match, meaning the same entity. Numerically states for a plausible range between the numerical real data and the value presented in the model. Parameter estimation discussion escapes the scope of the present paper.	x	x	x		x	x	x
	3	Extreme-conditions	Despite this being relative to the numerical model, it is brought here because the structure of the model must allow for the extreme combinations of state variables. To make the extreme conditions test, the auxiliary equations on which rates depend must be able to assume an imaginary maximum and minimum value and still unfold into plausible model behavior, or in the present case, consistent structural design.	x	x	x	x		x	x
	4	Boundary-adequacy (as structure)	Looks for the adequacy of the aggregation level and at the same time tries to understand if the model is capturing the relevant structures of the system. Despite discussing the aggregation level and boundaries of the model, the basic logic of the test would be to include a desired structure from the real system not yet represented in the simulation and compare the results towards a significative improvement that justifies the marginal effort.	x	x	x		x	x	
	5	Dimensional-consistency	By checking the coherence of the rates and stock dimensions, a structural problem might be revealed. Generally, the presence of a “scaling parameter” (Sterman’s fudge factor) with no match in reality is a symptom of a model with structural problems.	x		x		x	x	x
These tests are focused on the outputs of the model, usually numeric, except when very specific behavior pattern can be inferred from the relations embedded in the qualitative model.	6	Behavior-reproduction	Is the test that compares what the model is producing with the real data. This kind of test includes specific tests such as symptom generation (model can reproduce the problem at stake), frequency generation (model reproduces the periodicity of data), relative phasing (model reproduces the fluctuations and phasing of the data), multiple modes (model can reproduce different behaviors from the system) and behavior characteristics (other behavior reproduction tests)	x	x	x		x	x	
	7	Behavior-prediction	It is similar to a behavior reproduction test but not focused on reproducing historical data but on future behavior. Here two main categories arise: pattern-prediction (the model generates qualitatively correct patterns of behavior) and event-prediction (the model should reproduce a sharp drop/rise in a behavior). System dynamics models do not usually focus on good point prediction of future events, and this must be taken into account when analysts are adopting this sort of model.	x		x	x	x	x	x
	8	Behavior-anomaly	It reveals anomalies in the model structure by comparing the model behavior with the behavior of the real system.	x		x		x	x	x
	9	Family-member	It's relative to the degree of generalization the model might have. The recommendation is, that by adjusting specific few parameters, the model can reproduce a family-level behavior, instead of a case-specific behavior. One example would be in reproducing the behavior of any company of technology instead of just one specific.	x						
	10	Surprise-behavior	The model can generate a behavior present in the system but not yet perceived. When this kind of surprise appears, the analyst can search for the causes and compare data, gaining trust in model robustness and confidence.	x		x			x	x

Test of model behavior: T that can be forecasting or ba should be applied when the that allows testing. Structur most of the behavioral tests causal	11	Extreme-policy	Very similar to the structural case, but here the focus of the analysis is on the behavior of the model. Under extreme conditions of rate equations (imaginary maximum and minimum), the behavior of the model must be coherent, and should not present any anomaly behavior.	x		x		x	
	12	Boundary-adequacy (as behavior)	As a behavioral test, the inclusion of a desired system structure into the model should be reflected in a better behavior of the model.	x		x		x	
	13	Behavior-sensitivity	By changing the values of parameters to other plausible values, the model can reveal which parameters are more/less sensitive. More sensitive parameters are the focus of policy investigations. In general, system dynamics models are low sensitive to parameter change and more prone to reproduce structural behavior (Forrester & Senge, 1980)	x		x	x	x	x
plications: This kind of tests are basically a stem behavior after the application of policies with the outputs forecasted in the simulation.	14	System-improvement	Considers whether the behavior of a system improved after the implementation of the policies tested in silico. Complications here occur due to the time necessary to build confidence in the model, and the lag between the application of new policies and the results being observable, which might take years in some cases. In addition, it is very difficult to irrefutably connect the observed result with the policy change realized years before the result appears (many other elements might have influenced the system).	x					
	15	Changed-behavior-prediction	Goes for whether a specific behavior of the real system changes as predicted in the model after a determined policy is applied. This test can also be applied inversely when the model is trained to generate the change in system behavior, and with that, the analyst can understand the causes of the current behavior in the system.	x		x		x	
	16	Boundary-adequacy (as policy)	Concerns testing how the change in the boundaries of the model would affect the policy recommendations created by the simulation. In addition, the same policy can be tested for its adequacy if implemented outside the original boundaries set in the model.	x		x		x	
	17	Policy-sensitivity	Should reveal to which extent the policy might be affected by changes in specific parameters. By checking a parameter sensitivity, the model should be able to show the degree to which the policy might be influenced by these parameter changes.	x		x		x	
Tests of Policy Im comparison of the sy tested in the model,	18	Statistical	Statistical tests in causal models represent a contended terrain with discussions of its rationalities and usefulness going back even before the foundation of system dynamics as a field (e.g.,(Keynes, 1939). Despite that, many modelers use statistical tests and techniques to help validate models. The authors call attention tough, to the limitations of such tests when regarding causal models and system dynamics (Forrester & Senge, 1980; Mass & Senge, 1978)	x		x		x	
Analytical quality of PIoR	19	Insight generation capacity	Whether a model does lead to any Policy Insight or Recommendation					x	
	20	Relevance and Fertility of PIoR	Whether the Policy Insight or Recommendation is innovative and important					x	
	21	Rigor and Robustness of PIoR	Considers whether the insights are supported by a Formal Model and the sensitivity analysis showed it to be sturdy.					x	
	22	Precision of PIoR	Shows the nature of the Policy Insight or Recommendation, from more general qualitative insights to specific quantitative information					x	
Process Efectiveness of the Intervention: relates to the participants responses to the modeling process more than to the model	23	Trustwhortiness or Guru status of the system dynamicist	An affinity with the modeler can enhance positively the modeling process and the PIoR implementation					x	
	24	Time & Cost of the intervention	Should be measured against a target and inform the level of satisfaction with the results against the target investment					x	
	25	Meaningfullness and communicability of the PIoR	How available are the models? Is it easy and fun to explore the models and search for results? How much do the relevant actors participate in the model building?					x	
	26	Congruence of PioR with culture	This test verifies the social implementability of any Policy Insight or Recommendation. The point is that makes no sense to propose actions/policies that involve actions considered unacceptable or unbearable for a potential observer.					x	

Other dimension: Aditions from other authors	27 Perspectives in Boundary-adequacy	Do the models support debate on different perspectives in the AoS concerning: a) choice of model used; b) SD issue addressed; c) goals to be achieved; and d) Policies for doing so?		x	
	28 Norms/values in boundary adequacy	Do the models support debate concerning and represent the behavior of the relevant actor´s: a) goals (are the desired states acceptable?); b) Policies (are the actions based on discrepancies between goal and actual conditions acceptable within the culture?)		x	
	29 Roles in boundary Adequacy	Are the feedback links in the model consistent with the abilities of current actors in the system to access, interpret, and employ information?		x	
	30 Ease of Enrichment	Concerns about the ability of any model to be updated with new data, or used to test the effects of new policies		x	
	31 Audit or crossvalidation	Measure how adequately a model study is conducted concerning established standards, practices, and guidelines (or experience in the case of Cassidy et al., 2021). Done by someone not involved in the modeling process.	x		
	32 Higher-level Model review	A higher management level test of the model´s appropriateness to the systems definition and study objectives, adequacy of underlying assumptions, adherence to standards, modeling methodology used, model representation quality, structure, completeness, consistency, and documentation	x		x
	33 Turing Test	Experts are presented with two sets of data, one from the model and one from the system. Under the same input conditions, the experts are asked to differentiate these sets. If they can differentiate them, their comments are valuable to increase the model quality, if they cannot, the confidence in the model validity increases.	x		
	34 Walkthroughs	Are group exercises to test the overall documentation for any errors? Does not test performance.	x		x
	35 Presentation	Refers to the adequacy of the presentation of the model to the relevant audience, considering their level of scientific understanding or others	x		
	36 Documentation	Refers to the adequacy of the process of making every step in the modeling process replicable by taking a formal process or writing assumptions, discussions, updates, or a change in previous steps regarding the modeling process		x	
	37 Purpose	What is the purpose of this model? The idea is to understand the alignment of the technique, resources, personnel, and objectives with the purpose of the model.			
	38 Usefullness	Who will operate the model, the modelers or third parties?			
	39 Replicability	Are you sure that independent third parties can reproduce your model and all your results only using your written documentation?			
	40 Loop dominance	The loop dominance test compares the loops in the model with the modeler's or client's assumption about which are the dominant feedback loops in the real system.			

AoS: Appreciation of the Situation; PIoR: Policy insight or recomendations; CCM; conceptual causal model; FM: formal model.
Source of descriptions:
1-18: according to Forrester & Senge, 1979
19-29: according to Lane, 1995
31-36: according to Balci, 1994
37-39 according to Sterman, 2000
40 - according to Swchaninguer & Grosser, 2020

ution in the literature

Sterman, 2000	Quadrat-U, Seong, 2010	Quadrat-Ullah, 2012	Checkland, Poulter, 2020	Schwaninger et al. 2020	Cassidy et al., 2021	Criellard et al., 2022
x	x	x		x	x	x
x	x	x		x		x
x	x	x		x		
x	x	x	x	x		x
x	x	x		x		
x				x	x	
x				x		
x			x	x		
x				x		

x	x	x
x	x	x
x		
x		
x		
x	x	
x	x	

x x

x x

x

x

x

x

x x

x

x

x

x