Comput Econ (2007) 30:265–290
DOI 10.1007/s10614-007-9101-7

Validating Simulation Models: A General Framework and Four Applied Examples

Robert Ernest Marks

Received: 31 July 2006 / Accepted: 5 July 2007 / Published online: 3 August 2007

Abstract This paper provides a framework for discussing the empirical validation of simulation models of market phenomena, in particular of agent-based computational economics models. Such validation is difficult, perhaps because of their complexity; moreover, simulations can prove existence, but not in general necessity. The paper highlights the Energy Modeling Forum's benchmarking studies as an exemplar for simulators. A market of competing coffee brands is used to discuss the purposes and practices of simulation, including explanation. The paper discusses measures of complexity, and derives the functional complexity of an implementation of Schelling's segregation model. Finally, the paper discusses how courts might be convinced to trust simulation results, especially in merger policy.

Keywords Simulation · Validation · Agent-based computational economics · Antitrust · Functional complexity · Schelling · Sufficiency · Daubert criteria

JEL Classifications C63 · C15

1 Introduction

The apparent past reluctance of some in the discipline to accept computer simulation models of economic phenomena might stem from their lack of confidence in the behaviour and results exhibited by such models. Even if there are other reasons, better validation of such models would reduce any skepticism about their results and their

R. E. Marks (✉)
Australian Graduate School of Management,
Australian School of Business, The University of New South Wales,
Sydney, NSW 2052, Australia
e-mail: bobm@agsm.edu.au

usefulness. Leombruni et al. (2006) go further, arguing that a common protocol for conducting, validating, and reporting agent-based (AB) social simulations is necessary, including five cases of validation (see Sect. 6 below). Fagiolo et al. (2006) make a similar proposal, and Midgley et al. (2007) also argue for a common approach for establishing the "assurance" (programming verification and empirical validation) of AB models, introducing a five-step process for modellers.

This paper discusses the general issue of validation (for whom? with regard to what?) and its relationship to the use of computer models for explanation, and agent-based simulation models in particular. Section 2 delineates AB computer models from other simulation techniques, and discusses emergent behaviour in AB models. Section 3 discusses possible reasons for the apparent lack of enthusiasm in the profession for AB computer models, and discusses the simulation benchmarking that the Energy Modeling Forum has been undertaking at Stanford for thirty years. Section 4 contrasts sufficiency from computer models with sufficiency and necessity from closed-form solutions. Section 5, using data from a real-world market for branded coffee, discusses the general issues of validation. Section 6 provides a formalisation of validity, and a definition of its measurement. Section 7 discusses notions of complexity, and argues that functional complexity is the appropriate measure for the complexity of computer simulation models. We calculate the functional complexity of a well known simulation model, and ask whether such models might be too complex to be properly tested. As we discuss below, this will depend on the goals of the simulation. Although such complexity might not matter for qualitative AB models, the issue of "over-parameterisation" (Fagiolo et al. 2006) that we highlight here is pervasive with all kinds of AB models. Section 8 discusses the use of simulation models in competition (antitrust) regulation, and the use in the US of the *Daubert* discipline to satisfy the courts that the behaviour exhibited by merger simulations is sufficiently close to what would happen that policy can rely on it. Section 9 concludes.

How does this paper contribute to the literature? Two possible reasons for the reluctance of economics to use AB models to any great extent are, first, model sufficiency but not necessity, and, second, difficulties in verifying and validating such models, perhaps because of their high levels of complexity. Simulations are similar to existence proofs, but what of necessity? A discussion of sufficient and necessary conditions and simulations attempts to answer this question. The paper undertakes an extensive review of the debate on methodological issues concerning the use of simulation as a tool for scientific research, specifically, the validation of simulation models in general and of AB models in particular, stressing the trade-off between realism and parsimony. A review of the literature is included, including long-standing research at Stanford which uses simulations of disparate computer models of the same phenomena for policy purposes. A formal framework for choosing among different methods of validation is presented. Despite the functional complexities of agents, systems of agents need not be more complex than single agents; indeed, at a higher level, AB systems might be less complex. The paper concludes by bringing together a discussion of empirical validation, the measure of model complexity, and legal requirements for predictions from simulation models to be accepted as evidence in antitrust cases.

2 Agent-Based (AB) Computational Economics

AB models are being used more often in the social sciences in general (Gilbert and Troitzsch 2005) and economics in particular (Rennard 2006; Tesfatsion and Judd 2006). They are also being adopted in marketing research (Midgley et al. 1997), in political science, in finance (Sharpe 2007) and in the multidisciplinary world of electricity markets (Bunn and Oliveira 2003 et seq.). AB simulations are bottom-up: they have more than one level. At the lowest level the agents interact, and as a result behaviour of the system might emerge at a higher level. By emergence, we mean macro behaviour not from superposition, but from interaction at the micro level. It is this property of emergence that sets AB models apart from single-level simulation models, such as systems dynamics models (Forrester 1961), as discussed in Sect. 8 below.

In economics, the first AB models[1] used Genetic Algorithms as a single population of agents (Marks 1992; Miller 1996). A single population was acceptable when the players were not differentiated and when the flow of information from parents to offspring at the genotype level was not an issue (Vriend 2000), but when the players are modelling heterogeneous actors—in realistic coevolution, for instance—each player requires a separate population, not least to prevent the modelling of illegally collusive, extra-market transfers of information.

Moss and Edmonds (2005) argue that for AB models there are at least two stages of empirical validation, corresponding to the (at least) two levels at which AB models exhibit behaviour: the micro and the macro. The first stage is the micro-validation of the behaviour of the individual agents in the model, which they suggest might be done by reference to data on individual behaviour. The second stage is macrovalidation of the model's aggregate or emergent behaviour when individual agents interact, which Moss and Edmonds suggest can be done by reference to aggregate time series. Moreover, since the interactions at the micro level may result in the emergence of novel behaviour at the macro level, there is an element of surprise in this behaviour, as well as the possibility of highly non-standard behaviour,[2] which can be difficult to verify using standard statistical methods. As Moss and Edmonds note, at the macro level only qualitative validation judgments (about Kaldor's stylised facts, for instance) might be possible as a consequence. But, as we discuss in Sect. 7 below, at the macro level the complexity of an AB model might be less than its complexity at the micro level.

There has however been a reluctance in economics to embrace simulation in general or AB modelling in particular. This apparent aversion—or disdain—is mirrored in the discipline's approach to viewing the economic system as a whole—or parts of it such as markets—as complex adaptive systems, despite the recent publication of four papers in the June 2005 issue of the *Economic Journal* and the Tesfatsion and Judd *Handbook* (2006).[3]

[1] Arthur (2006) recalls the first attempts to use AB models in economics.

[2] For instance, leptokurtosis and clustered volatility might be observed inter alia; they can be highly non-Gaussian.

[3] Even *The Economist* (Economics focus, July 22, 2006, p. 82) has remarked on the "niche pursuit" that evolutionary economics in general—and AB economics in particular—has remained.

3 Why the Reluctance?

Two papers (Leombruni and Richiardi 2005; Leombruni et al. 2006) address the absence of AB simulations in the economics literature.[4] They reveal that only eight of 43,000 papers in the top economics journals[5] used AB modelling, and further argue that this was owing to two reasons. First, difficulties with interpreting the simulation dynamics and with generalising the results when the model parameters are changed. Second, no clear equations with which to use statistical methods to estimate the simulation model.

Fagiolo et al. (2006) argue that, as well as lacking a developed set of theoretical models applied to a range of research areas, AB simulators are handicapped by lack of comparability among the models that have been developed.

Nonetheless for almost thirty years, the Energy Modeling Forum at Stanford has been overseeing comparisons among simulation models, although not specifically AB models. The EMF was born out of the energy crises of the 1970s, when energy modellers turned to computer simulations to try to get answers to energy-related issues. Different researchers' models of the same phenomena are benchmarked and compared. The EMF did not build these models itself, rather it provided a forum for comparison of the models' behaviours, as Hill Huntington explains:

> EMF desperately tries to avoid "forecasts" and should not be considered a Delphi technique. Instead, we have focused on scenario comparisons where modelers use common assumptions to provide some badly needed consistency across estimates. Comparing how different models respond to a common change in assumptions hopefully provides some useful information about how these models behave. The technique is closer to the ceteris paribus approach of the relative importance of different assumptions or assumed conditions than being the modeler's best guess of what will happen using their own-best set of assumptions.
>
> As we proceed with issues like climate change, baseline conditions (without a policy) have become increasingly important, since they are then compared to a constrained policy case that fixes the emissions levels to some targeted level. In these cases, we have often asked the modelers to work with their best guess of a reference case, using their own assumptions. This approach sometimes creates problems because it makes it more difficult to compare results consistently, but it does provide other benefits.
>
> We often remind participants that the underlying uncertainty reflected in EMF studies often understates the true uncertainty, largely because we try to use

[4] Axelrod (2006) argues that this scarcity occurs across the social sciences, with no obvious focus for publishing studies that use AB methods. He identified 77 social-science articles published in 2002 with "simulation" in the title. These appeared in 55 different journals, only two of which published three or more articles. Similar dispersion is revealed by other methods of analysis.

[5] *American Economic Review, Econometrica, Journal of Political Economy, Journal of Economic Theory, Quarterly Journal of Economics, Journal of Econometrics, Econometric Theory, Review of Economic Studies, Journal of Business and Economic Statistics, Journal of Monetary Economics, Games and Economic Behavior, Journal of Economic Perspectives, Review of Economics and Statistics, European Economic Review, International Economic Review, Economic Theory, Journal of Human Resources, Economic Journal, Journal of Public Economics, Journal of Economic Literature.*

> common assumptions rather than to ask participants to use their own assumptions (which presumably could vary from each other by quite a bit). That is, there is both a model structure uncertainty and an input uncertainty. [Hill Huntington, pers. comm., 2006]

Although the EMF has been too busy comparing models and developing policy recommendations (see Weyant and Hill (1999), for example) to have published much on its methodology of model comparisons, Huntington et al. (1982) provides some insights. Its projects have provided policy-makers with robust understanding of complex interactions in the physico-social-economic systems modelled.

Axtell et al. (1996) introduced the term "docking" when a second team attempts to replicate another team's simulation model. They clarified three decreasing levels of replication: "numerical identity," "distributional equivalence" (the results cannot be distinguished statistically), and "relational equivalence" (the same qualitative relationships). Although the EMF did not explicitly attempt to dock simulation models, in the studies Huntington describes, they were almost always able to achieve relational equivalence, often distributional equivalence, but rarely numerical identity. One outcome for the EMF studies, however, came from asking how models based on different assumptions were able to achieve these levels of replication. The closer the replication from different models, the greater the confidence in the simulations' results.

We believe that a fourth reason[6] for the lack of interest from the profession at large in AB modelling is that simulation can, in general, only demonstrate sufficiency, not necessity. Since necessity is, in general, unattainable for simulators, proofs are also unattainable: simulation can disprove a proposition (by finding conditions under which it does not hold) but cannot prove it, absent necessity. But if there are few degrees of freedom, so that the space of feasible (initial) conditions is small, then it might be possible to explore exhaustively that space, and hence derive necessity.[7]

4 Sufficiency and Necessity

With some formality, it is possible to show how difficult it is to derive necessity using simulation. A mathematical "model A" is the conjunction of a large number of separate assumptions embodied in a specific implementation, with several equations that constitute a conglomeration of hypotheses and generalisations, as well as parameters and initial conditions that must be specified. So model A comprises the conjunction $(a_1 \wedge a_2 \wedge a_3 \ldots \wedge a_n)$, where $\wedge$ means "AND", and the a_i denote the elements (equations, parameters, initial conditions, etc) that constitute the model.

Sufficiency: If model A exhibits the desired target behaviour B, then model A is sufficient to obtain exhibited behaviour B. That is, $A \Rightarrow B$. Thus, any model that exhibits

[6] Epstein (2006) argues that, whatever their success in prediction, explanation, and exploration in the past, simulations are believed to lack the brilliant, austere beauty of an elegant theorem, a belief he argues is wrong. Beauty is, after all, in the eye of the beholder.

[7] Watson and Crick (1953) did that with their "stereo-chemical experiments"—simulations—for the structure of DNA. Note that the title of their 1953 paper included the phrase "*a* structure", not *the* structure, flagging sufficiency, not necessity (our emphasis).

the desired behaviour is sufficient, and demonstrates one conjunction of conditions (or model) under which the behaviour can be simulated. But if there are several such models, how can we choose among them? And what is the set of all such conjunctions (models)?

Necessity: Only those models A belonging to the set of necessary models $\mathcal{N}$ exhibit target behaviour B. That is, $(A \in \mathcal{N}) \Rightarrow B$, and $(D \notin \mathcal{N}) \not\Rightarrow B$. A difficult challenge for the simulator is not to find specific models A that exhibit the desired behaviour B, but to determine the set of necessary models, $\mathcal{N}$. Since each model $A = (a_1 \wedge a_2 \wedge a_3 \ldots \wedge a_n)$, searching for the set $\mathcal{N}$ of necessary models means searching in a high-dimensional space, with no guarantee of continuity, and a possible large number of non-linear interactions among elements.[8]

For instance, if $D \not\Rightarrow B$, it does not mean that all elements a_i of model D are invalid or wrong, only their conjunction, that is, model D. It might be only a single element that precludes model D exhibiting behaviour B. But determining whether this is so and which is the offending element is a costly exercise, in general, for the simulator. With closed-form solutions, this might not be trivial, but it might seem easier than determining the necessary set using simulation. Bar-Yam (2006, pers. comm.) argues, however, that necessity is only well defined in a class of models, and that "if anything, the necessary conditions are better defined in a discrete model [such as a simulation model] and are more difficult to establish in differential-equation models, hence the emphasis on their proofs."

Without clear knowledge of the boundaries of the set of necessary models, it is difficult to generalise from simulations. Only when the set $\mathcal{N}$ of necessary models is known to be small (such as in the case of DNA structure by the time Watson and Crick were searching for it) is it relatively easy to use simulation to derive necessity.[9]

We return to this issue of the degree of complexity of AB simulation models in Sect. 7 below.

5 Validation

What is a good simulation? The answer to this question must be: a good simulation is one that achieves its aim. But just what the aim or goal of a simulation might be is not obvious. There are several broad possibilities.[10] A simulation might attempt to *explain*

[8] Fagiolo et al. (2006, p. 29) speak of "backwards induction" of the "correct" model; in our terms, this is identification of the necessary set $\mathcal{N}$. Burton (2003) speaks of solving the "backward" problem to obtain "a few alternative plausible" (sufficient) explanations; his "forward" problem (Gutowitz 1990) is prediction, with the observation of emergence (Burton, pers. comm., 2006).

[9] Klein and Herskovitz (2005) provide an overview of the philosophical foundations of the validation of simulation models. They do not expand on our treatment here; there is, however, a relationship between the the hypothetico-deductive approach of Popper and the discussion in this section.

[10] Haefner (2005) lists seven possible goals: usefulness for system control or management, understanding or insights provided, accuracy of predictions, simplicity or elegance, generality (number of systems subsumed by the model), robustness (insensitivity to assumptions), and low cost of building or maintaining the model. Axelrod (2006) also lists seven: prediction, performing tasks, training, entertaining (see those ubiquitous games consoles), educating, existence proofs, and discovery; prediction, existence proofs, and discovery are the main scientific contributions.

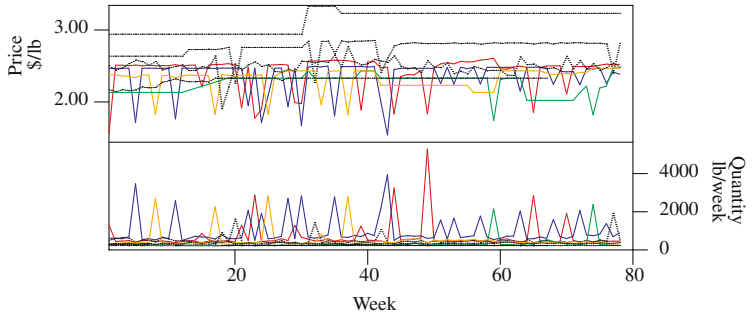

Fig. 1 Weekly prices and sales (*Source*: Midgley et al. 1997)

a phenomenon; it might attempt to *predict* the outcome of a phenomenon; or it might be used to *explore* a phenomenon, to play, in order to understand the interactions of elements of the structure that produces the phenomenon.[11]

Axelrod (2006) argues that simulation can be thought of a third way of undertaking scientific research: first, induction is the discovery of patterns in empirical data (not to be confused with mathematical induction); second, deduction involves specifying a set of axioms and proving consequences that can be derived from them; and, third, simulation, described by Axelrod as a "third way" of doing science—starting with a set of explicit assumptions, simulation does not prove theorems but instead generates data that can be analysed inductively, as a way of conducting thought experiments.

Explanation should result in arriving at sufficient conditions for an observed phenomenon to occur. To take an example, Fig. 1 presents prices and quantities of branded coffee sales by week in a single supermarket chain in the US Midwest in the 1980s, from supermarket scanner data.[12]

Several characteristics (or stylised facts, Kaldor 1961) are obvious from the graph: first, there is a great deal of movement in the market: prices and quantities of at least four brands are not at all stable, and are experiencing great week-to-week changes in their quantities sold. (These have been plotted with solid lines.) Second, some brands are not altering their prices much at all. (These four have been plotted with dotted lines.) Third, for the first group the norm seems to be high prices (and low sales), punctuated every so often by a much lower price and much higher weekly sales. If we tabulated each brand's price by its frequency at different price bands, other patterns might become clear; the first moments of price would reflect price dynamics over the period.

We could ask several questions about the historical data and the underlying data generator: What is causing these fluctuations? Is it shifts in demand, whether in aggre-

[11] Rubinstein (1998) lists four purposes: predicting behaviour, as a normative guide for agents or policy-makers, sharpening economists' intuitions when dealing with complex situations, and establishing "linkages" between economic concepts and everyday thinking. Burton (2003) lists three questions: asking *what is, what could be,* and *what should be.*

[12] The upper grouping depicts brands' prices by week (LHS); the lower grouping depicts brands' weekly sales (RHS); the four "strategic" brands are coloured, solid lines, the rest dotted. The brands' price are fixed for seven days by the supermarkets.

gate or for specific brands (perhaps in response to non-observed actions such as advertising)? Is it driven by the individual brand managers themselves? If so, are they in turn responding to the previous week's prices? Or it might be that the supermarket chain is moderating the behaviour of the individual brands. A further cause of the behaviour might be the non-price, non-advertising actions of the brands: aisle display and coupons.

If the profit-maximising behaviour of the simulated brand managers, together with some external factors or internal factors, led to behaviour qualitatively or quantitatively similar to the behaviour of the brands' prices and quantities sold seen in the figure, then we would have obtained one possible explanation. The issue of degrees of similarity is one of closeness of fit, and could be handled using statistical measures. Note, following Durlauf (2005), that by making the assumption of profit maximizing, we are going beyond merely seeking a set of equations exhibiting periodicity similar to the "rivalrous dance" of the brands in the figure.

Having sought patterns in the past, and calibrated a model,[13] it becomes possible to conduct a sensitivity analysis to answer the question: what will happen if such and such are the prevailing exogenous conditions? Here, the endogenous variables might include the prices and other marketing actions (i.e., aisle displays, coupons, and advertising) of one brand manager, or of a set of them in this market. This leads to the second broad goal, prediction.

For prediction, sufficiency suffices: there is no need to know which if any alternate conditions will also lead to the observed endogenous behaviours. That is, prediction does not require an understanding of necessity of the underlying exogenous variables. This might explain, as Friedman (1953) argued, that economic actors can behave as though they have solved complex optimisation problems, even though they remain ignorant of any formal representation of the problem or its solution.

Exploration is perhaps the most interesting example of what can be done with simulation models. What are the limits of this behaviour? Under what conditions does it change to another general form of behaviour? Just what ranges of behaviour can the system generate? How sensitive is the model behaviour (and hopefully the real-world behaviour) to changes in the behaviour of a single actor, or of all actors, or of the limits of interactions between players? Indeed, Miller's (1998) Automated Non-Linear Testing System technique deliberately "breaks" the model by searching for sets of parameter values that produce extreme deviations from the model's normal behaviour, as part of the validation exercise. Exploring the model in this way may, we hope, shed light on the underlying real-world data generated.

The aim of the simulation depends partly on who is simulating (or who is the client), and who will be presenting the results. The first step to convince others that your simulation is appropriate is to convince yourself. With friendly tools, even the naïve user can use simulation models to explore.[14] In Sect. 8 we discuss the criteria that the US courts use to determine the admissibility of expert testimony, and the extent to which

[13] Implicitly, this approach is what Brenner and Werker (2006) call "the history-friendly" approach to empirical validation, with its strengths and weaknesses. (See also Malerba et al. 1999).

[14] Resnick (1994) provides a clear example of a student using a NetLogo simulation to explore emergent behaviour.

AB simulation models might satisfy them, as a guide to validation. Bar-Yam (2003) discuss the value of including the stakeholders in the model's use in the validation of the model.

Leombruni et al. (2006) list five types of validity that theory- and data-based economic simulation studies must consider: theory (the validity of the theory relative to the simuland), model (the validity of the model relative to the theory), program (the validity of the simulating program relative to the model), operational (the validity of the theoretical concept to its indicator or measurement), and empirical (the validity of the empirically occurring true value relative to its indicator).

Throughout the paper, we focus on Leombruni et al.'s empirical validity rather than their program validity. The former (Manson's (2002) "output validation") asks how successfully the simulation model's output exhibits the historical behaviours of the real-world target system. The latter (Manson's "structural validation") asks how well the simulation model represents the (prior) conceptual model of the real-world system. Of course, if the work lacks theory or model or program validity, then it will in general be very difficult to obtain empirical validity.[15]

Following Rosen (1985), it is useful to think of two parallel unfoldings: the evolution of the real economy (or market or whatever) and the evolution of the model of this real-world phenomenon. If the model is properly specified and calibrated, then its evolution should mirror the historical evolution of the real-world phenomenon: we could observe the evolution of the model or the real-world evolution—both should reveal similar behaviour of the variables of interest.

6 Formalisation of Validation

Let set **P** be the possible range of observed outputs of the real-world system, here the prices, quantities, and profits[16] of the coffee brands of Fig. 1 each week. Let set **M** be the exhibited outputs of the model in any week. Let set **S** be the specific, historical output of the real-world system in any week. Let set **Q** be the intersection, if any, between the set **M** and the set **S**, $\mathbf{Q} \equiv \mathbf{M} \cap \mathbf{S}$. We can characterise the model output in five cases.[17]

(a) If there is no intersection between **M** and **S** ($\mathbf{Q} = \emptyset$), then the model is *useless.*
(b) If the intersection **Q** is not null, then the model is *useful,* to some degree. In general, the model will correctly exhibit some real-world system behaviours, will not exhibit other behaviours, and will exhibit some behaviours that have not historically occurred. That is, the model is both incomplete and inaccurate.
(c) If **M** is a proper subset of **S** ($\mathbf{M} \subset \mathbf{S}$), then all the model's behaviours are correct (match historical behaviours), but the model doesn't exhibit all behaviour that has historically occurred. The model is accurate but *incomplete.*

[15] Midgley et al. (2007) include program verification as one of their five steps in what they dub "assurance"—verification and validation—of the model.

[16] Midgley et al. (1997) and Marks (2006) describe how they calculate each brand's weekly profits, given the combination of marketing actions of all brands that week, and with prior knowledge of the brands' costs.

[17] This conceptual framework was introduced by Mankin et al. (1977).

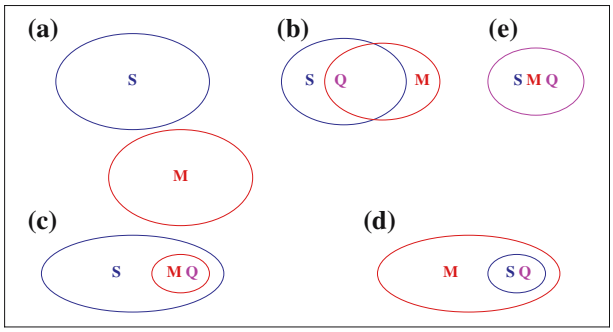

Fig. 2 Validity relationships (after Haefner (2005))

(d) If **S** is a proper subset of **M** (**S** ⊂ **M**), then all historical behaviour is exhibited, but the model will exhibit some behaviours that have not historically occurred. The model is complete but *inaccurate.*
(e) If the set **M** is equivalent to the set **S** (**M** ⇔ **S**), then (in your dreams!) the model is complete and accurate.

By *incomplete*, we mean that **S****Q** is non-null, so that the model does not exhibit all observed historical behaviours. By *inaccurate*, we mean that **M****Q** is non-null, so that the model exhibits behaviours that are not observed historically.[18] Haefner (2005) notes that the set boundaries might be fuzzy: not "in" or "out," but contours of the probability of belonging to the set. Figure 2 illustrates these relationships.

One goal of the modeller might be to attempt to construct and calibrate the model so that **M** ≈ **Q** ≈ **S** (case (e)): there are very few historically observed behaviours that the model does not exhibit, and there are very few exhibited behaviours that do not occur historically. The model is close to being both complete and accurate, for explanation. But this might be overfitting for prediction. In practice, a modeller examining sufficient conditions (existence proofs) for previously unobserved (counterfactual) behaviour might be happier to achieve case (d), where the model is complete (and hence provides sufficiency for all observed historical phenomena), but not accurate.[19] Of course, changing the model's parameters will in general change the model's exhibited behaviour (set **M**). In the calibration stage, we might well be happier if we could adjust the model parameters so that **M** ≈ **S**, in the belief that the changed set **M**′ with different parameters might well model a variant of historical reality.

This suggests a measure of validity which balances what we might call (from statistics) the Type I error of inaccuracy with the Type II error of incompleteness. In order

[18] One referee would prefer the term *redundant* here, arguing that such a model might tell the modeller something about what could yet happen in the world, with larger sets **S** and **P**.

[19] As Fagiolo et al. (2006) remind us, "in-sample" data are relevant when the goal is description or replication of the phenomenon; "out-of-sample" data can be used to test the model's efficacy as a prediction model. But, given the scarcity of good time-series against which to both calibrate models (using "in-sample" data) and then predict (against "out-of-sample" data), there have been few predictions using AB models. Froeb, in Woodbury (2004), supports this claim for antitrust simulations (see Sect. 8 below).

to define these measures, we need a metric (a ratio scale) defined on the sets. Call it $m()$.

We can define[20] *inaccuracy* α as

$$\alpha \equiv 1 - \frac{m(\mathbf{Q})}{m(\mathbf{M})}, \tag{1}$$

and *incompleteness* γ as

$$\gamma \equiv 1 - \frac{m(\mathbf{Q})}{m(\mathbf{S})}. \tag{2}$$

A measure of degree of validation V could be a weighted average of inaccuracy α and incompleteness γ:

$$V \equiv v(1-\alpha) + (1-v)(1-\gamma) = m(\mathbf{Q})\left(\frac{v}{m(\mathbf{M})} + \frac{1-v}{m(\mathbf{S})}\right) \tag{3}$$

The value of the weight v, $0 \leq v \leq 1$, reflects the tradeoff between accuracy and completeness.

For well-behaved measures $m()$ of set size, the first partial derivatives of validity V with respect to both the size of the model set **M** and the size of the historical set **S** are negative, so the smaller each of these two sets, the higher the measure of validity, cet. par. The partial derivative of V with respect to the size of the intersection set **Q** is positive, so the larger the intersection, the higher the measure of validity, cet. par.

It might be possible to reduce incompleteness by generalising the model and so expanding the domain of set **M** until **S** is a proper subset of **M**, as in case (d). Or by narrowing the scope of the historical behaviour to be modelled, so reducing the domain of **S**. It might also be possible to reduce inaccuracy by restricting the model through use of narrower assumptions and so contracting the domain of **M**. If **M** is sufficiently small to be a proper subset of **S**, as in case (c), then the model will never exhibit anhistorical behaviour.

This process of constricting the model by narrowing the assumptions it builds on is not guaranteed to maintain a non-null intersection **Q**, and it is possible that the process results in case (a), with no intersection. This is reminiscent of the economist looking for his lost car keys under the street light (**M**), instead of near the car where he dropped them in the dark (**S**). Advocates of simulated solutions, such as Judd (2006), have argued that it is better to "have an approximate answer to the right question, than an exact answer to the wrong question," to quote Tukey (1962).

Haefner (2005) notes that most published validation exercises (at least in biology) focus on the size of **Q** or, at best, on model accuracy. In economics we are not only interested in understanding the world, but also in changing it, by designing new systems. Design often requires prediction of counterfactuals, situations that have never been observed, and simulations—with their demonstrations of sufficiency—are one way of grounding the design process, as Marks (2006) discusses. Designing

[20] Mankin et al. (1977) introduce the concepts *model reliability* and *model adequacy* effectively defined as $m(\mathbf{Q})/m(\mathbf{M})$ and $m(\mathbf{Q})/m(\mathbf{S})$.

counterfactuals is rare or non-existent in the natural sciences. (We discuss counterfactual simulations further in Sect. 8.)

Model completeness or adequacy is difficult to evaluate when most historical datasets are small and sparse. For any given level of fit between a model and the historical, better models are those with lower ratios of the degrees of freedom to the data points fitted (Simon and Wallach 1999). This is because, with sufficient degrees of freedom, any pattern in the historical data could be tracked, but with such overfitting, it would be difficult to predict outside the historical data set.

How appropriate are the relationships of Fig. 2 to our three broad goals of simulation in the social sciences in general and in economics in particular? As with the natural sciences, when seeking explanation, the closer the model's behaviour to the real-world behaviour the better, as in case (e). As discussed above, prediction in the social sciences is often handicapped by small numbers of prior observations **S** and predictions can be counterfactual; this implies that case (d), the inaccurate case, might be more appropriate. The opposite case, the incomplete case (c), might be more appropriate for exploration: as we change the underlying assumptions, the model set **M** will appear to move across the real-world set **S**, as a searchlight might pick out objects of interest flying across the night sky.

7 Complexity of Agent-Based Simulations

By simulating bottom-up behaviour, AB models of such social interactions as market exchanges more closely represent the way phenomena at the macro level, such as prices and aggregate quantities, emerge from behaviour at the micro level than do reduced-form simulations using systems of equations. But there is a trade-off: complexity.

Some, such as Van Damme and Pinkse (2005), argue that since the world, or at least those phenomena in the world we wish to model, is complex, the models must be complex too.[21] But with complexity comes a challenge: to validate.

Can we put a number to the complexity of such models? Yes, in fact several. Consider four measures of complexity (Bar-Yam (1997)): algorithmic, behavioral, descriptive, and functional. Algorithmic (or Kolmogorov) complexity of a system is a measure of the complexity of the procedures used in transforming inputs to outputs, where inputs are included in the conjunction of Sect. 4 above ($a_1 \wedge a_2 \wedge a_3 \ldots \wedge a_n$). It is a generalisation of the Shannon (1948) information content of the program algorithm expressed as a minimal-length string. This measure, however, requires reducing the algorithm to a minimum length, which is not easy. Moreover, any estimate of algorithmic complexity of a system must also include the complexity of the compiler, the operating system, and the hardware.

Behavioral complexity is a measure of the probabilities of the possible outputs exhibited by the model. Strictly, it is the amount of information necessary to describe the probabilities of these behaviours. Descriptive complexity of a system is measured by the amount of information necessary to describe it, including its behaviour.

[21] To paraphrase Einstein, as complex as necessary, but not more.

As such, descriptive complexity is a superset of behavioral complexity. In general, when building simulation models of economic phenomena, we are more interested in how the inputs might map to outputs, than in the means of modelling.[22]

Functional complexity (Bar-Yam 2003) is a measure of the complexity of possible mappings from inputs to outputs of the model. When focussing on validation—to what extent combinations of inputs result in the model **M** exhibiting "correct" outputs **S** (where there are levels of correctness, and, possibly, contours of the probability that a specific conjunction of inputs will exhibit a target set of outputs)—the appropriate measure of complexity is that of functional complexity. Moreover, this does not require us to determine the minimum description of an algorithm or a model state, or indeed to ask much at all about the programs, compilers, or operating systems.

Bar-Yam (2003) defines functional complexity as the relationship of the number of possible inputs and the number of possible outputs:

$$C(f) = C(a)2^{C(e)} \tag{4}$$

The complexity of the function of the system, $C(f)$, equals the complexity of the actions of the system $C(a)$ times two raised to the complexity of the inputs variables of the system, $C(e)$. The three complexities, input $C(e)$ and output $C(a)$ and functional complexity $C(f)$, are defined by the logarithm (base 2) of the number of possibilities, or equivalently, the length of its description in bits. As motivation for this equation, Bar-Yam argues that this follows from recognizing that a complete specification of the function is given by a look-up table whose rows are the actions ($C(a)$ bits) for each possible input, of which there are $2^{C(e)}$.

Bar-Yam further adds that this definition applies to the complexity of description as defined by the observer—apart from knowing possible inputs and outputs, the function could be a black box—so that each of the quantities can be defined by the desires of the observer for descriptive accuracy. This dovetails well with this paper's purpose.

Equation (4) is similar to a derivation of ours in 1989 (see Marks 1992) when modelling a p-player repeated game with players as stimulus-response automata, responding to the state of play as they know it: with p players, each choosing from a actions and remembering m rounds of play, the number of possible states each must be able to respond to is a^{mp}. When modelling players as bit strings with unique mappings from state to action, what is the minimum length of each string? The string would require Ceiling[$\log_2(a)$] bits per state, or a total of Ceiling[$\log_2(a)$]a^{mp} bits per player.

For example, a player in an iterated Prisoner's Dilemma, a is 2, m is 1, and p is 2, resulting in 4 possible states. With only two possible actions, $\log_2(a) = 1$, and the minimum-length bit string is 4. Using Eq. 4, $C(e) = \log_2(2^2) = 2$, $C(a) = \log_2(2) = 1$, and so $C(f) = 1 \times 2^2 = 4$, nothing more than the minimum-length bit string. More generally, $C(e) = \log_2(a^{mp})$, $C(a) = \log_2(a)$, and so log functional complexity is given by $C(f) = \log_2(a)2^{\log_2(a^{mp})} = a^{mp}\log_2(a)$ per player.

[22] A caveat is that we'd like most of our agents to engage in purposeful behaviour, up to some bound of rationality.

Because of the exponential nature of the definition of functional complexity, if the complexity of the inputs or environmental variables to the model is larger than, say, 100 bits, the functional complexity will be huge, and cannot reasonably be specified.[23]

From this discussion, one could conclude that the system complexity is simply the per-player complexity times the number of players or agents. In the iterated Prisoner's Dilemma, the simulation model would have log functional complexity of 8, since there are two players. More generally, the log functional complexity of models of p agents choosing from a actions and remembering m previous rounds of play of a repeated interaction would be $pa^{mp}\log_2(a)$, for state-based stimulus-response games. But for models with more than one level, such as AB models, this would be misleading.

Functional complexity per player or agent (measured by the minimum bit-string length for each) is not the system's or model's functional complexity. The interactions of the agents may lead to the emergence of higher-level patterns, which is often the object of interest of such models. Although the log functional complexity of a model with p players might appear to be simply p times the agent's log functional complexity, this ignores the appropriate scale of observation and the emergence of higher-level patterns. Bar-Yam (1997) argues that a system of N agents need not be N times as complex as a single agent, since agents, however heterogeneous, may have much in common; indeed, the system might be less complex than the complexity of a single agent as we shift our observation from the scale of the individual agent to the system scale. That is, at the macro level of the system, the observer is blind to the complexities at the micro (individual agent) level, which fall below his threshold of observation, whether spatial or temporal; instead, the observer looks for patterns at the macro level.[24]

For a whole class of micro models the same macro behaviour will emerge, which means that the complexity at the micro scale, although real enough, need not preclude validation of models to exhibit target macro behaviour: with regard to the macro behaviour, many variables are irrelevant. This is known as universality (Bar-Yam 2003). Another way of putting this is that the emerging behaviour at the macro scale is insensitive to fluctuations in the values at the micro scale. This is reflected in the non-increasing complexity profile. It follows that, before the measure of complexity can have any practical meaning in validation, we must specify the scale of discussion (is it a bitwise comparison or is it at a more macro level?) or the order of magnitude of the accuracy of validation.

The realisation that the complexity of the AB system at the macro level might well be less than the complexities of the individual agents at the micro level means that Moss and Edmonds' (2005) recipe for validation at the two levels might be difficult to achieve: although the macro level validation might be attainable, at least the qualitative validation of the stylised facts, attempts to validate the agents might founder, because of the agents' complexity. Does this matter? Not in cases where Bar-Yam's

[23] In Marks (2006), $a = 8$, $m = 1$, and $p = 4$, resulting in bit strings of length $\log_2(8) \times 8^4 = 12,288$, the log functional complexity of this model's bit-string agents, resulting in log functional complexity for the AB model of 98,304.

[24] The complexity of a system observed with a certain precision in space and time is characterised in a "complexity profile" (Bar-Yam 1997). In general, such profiles are non-increasing in scale.

universality holds. Indeed, the demonstration that emergent behaviour at the macro level is robust to different micro agents strengthens the power of the simulation.

Two observations provide support for a lower level of complexity at the macro level of the system than at the micro level. First, a complex system is frequently comprised of interrelated subsystems that have in turn their own subsystems, and so on until some non-decomposable level of elementary components is reached. Second, interactions inside subsystems are in general stronger and/or more frequent than interactions among subsystems (Courtois 1985).[25] These two observations support the notion that, whatever the micro behaviour (within limits), the macro behaviour that emerges is invariant, so that the complexity of the system at the macro scale is less than the complexity of the aggregation of micro subsystems.

But some complex systems are not decomposable, or at least not decomposable under certain external conditions. The challenge is to identify the class of models that are less complex at the macro scale than at the micro, and that capture the properties of a real-world system, that is, to identify the class of models that exhibit universality. Related to this issue is the problem of testability of representations through the validation of the mapping of the system to the representation (Bar-Yam 1997).

How is a model's functional complexity at any scale related to the degree of validation, V, of Eq. 3? The validity relationships of Fig. 2 have been plotted assuming a common scale of measurement between the real world and the model. This is important: inadvertent measurement at, say, a smaller scale for the model and a larger scale for the real world would result in different precisions for the sets **M** and **S**, which could mean measuring micro behaviour of the model and attempting to compare it with real-world macro behaviour. Depending on the behaviour of the two, V would be estimated as higher or lower than its true value.

7.1 The Functional Complexity of Schelling's Segregation Model

In order to put numbers on the measure of functional complexity (Eq. 4), we have chosen a specific implementation of Schelling's Segregation model (Schelling 1971, 1978): the version freely available for use in NetLogo (Wilensky 1998). Although not strictly an AB model (Gilbert and Troitzsch 2005), it is a model in which interactions at the micro scale (between each cell and its eight neighbours) lead at a higher scale to the emergence of segregated regions, under some conditions. Schelling (2006) explains he used the model to demonstrate that segregated neighbourhoods could arise even when households possessed a degree of tolerance for other kinds of people living next door. It is true that this model is essentially qualitative, and so does not require calibration to historical parameters in order to be useful, but it is a very well known model and serves as an appropriate model for this discussion of complexity.[26]

[25] This is also the basis of the definition of "nearly decomposable systems" (Simon and Ando 1961).

[26] Moreover, another version (Cook 2006) of this model has been used by Leigh Tesfatsion to conduct learning exercises in developing measures of segregation (see Frankel and Volij 2005), and in testing hypotheses about segregation with different levels of the model 's parameters (see Tesfatsion 2006). Cook's model is MS Windows-specific.

Appendix A displays the NetLogo code for Segregation, after the comments have been stripped. The ASCII code includes 1687 7-bit characters; but after compression with the Unix program gzip, which reduces the size of the file using Lempel-Ziv coding (LZ77), the compressed size is 609 8-bit characters. In bits, the size of the code fell from 11,809 bits to 4,872 bits, after compression. The compressed size is a measure of algorithmic complexity, although it ignores the bulk of the NetLogo program, to which the code in the Appendix is only an input, or environmental variable. To obtain the full measure of algorithmic complexity of Segregation, we need the underlying NetLogo code and more as well.[27]

With the definition of functional complexity, this need is obviated: we only need to have the measures of inputs (environmental variables) and outputs, and then use Eq. 4. For the NetLogo implementation of Segregation, these are the inputs (environmental variables):

1. The number of "turtles," or inhabitants. This number (call it N) ranges between 500 and 2500 in Segregation, which requires 11 bits to specify, strictly $\log_2(N - 500)$.
2. The tolerance, or percentage of a similar colour desired as neighbours in the eight adjoining "patches". This number ranges from 0 to 8, which requires 4 bits to specify.
3. The initial randomisation. Each of the turtles is either red or green. Leaving aside how this random pattern is generated, each of the up to 2500 turtles requires a colour bit. Colour increases by 1 the number of bits ($\log_2(N - 500)$) required to specify the initially occupied patches.

The NetLogo implementation of Segregation has as outputs two measures and the pattern:

1. A measure of the degree of segregation, such as the percentage of similar patches.[28] This number ranges from 0.0 to 99.9, which requires 10 bits to specify.
2. The percentage of unhappy turtles. Ranging from 0.0 to 99.9, this too requires 10 bits to specify. (In the long run, this almost always converges to 0, so long as there are enough empty patches.)
3. The pattern. The pattern appears on a square of 51×51 patches, where each patch can be red, green, or black (unoccupied). This requires $51 \times 51 \times 3 = 7{,}803$ possibilities, which requires 13 bits to specify.

If we ignore the first two measures of the output, summaries which anyway could be calculated from the final pattern, Eq. 4 indicates that the functional complexity of this implementation of Segregation is:

$$C(f) = C(a)2^{C(e)} = 13 \times 2^{(\log_2(N-500)+4+1)} \tag{5}$$

The power of 2 in this equation is bounded above by 16, so the maximum log functional complexity of this implementation of Segregation is $13 \times 2^{16} = 851{,}968$, measured

[27] For this reason, we do not attempt to measure the descriptive complexity (or its subset, the behavioural complexity) of Segregation.

[28] As Frankel and Volij (2005) discuss, there is no consensus on how to measure segregation. This is at least a simple measure.

in bits. That is, there are at most $2^{851,968}$ possible mappings from sets of inputs to unique outputs for this model.

But the large blobs of colour (of segregated neighbourhoods) or the dappled areas (well integrated) are what the eye is looking for, not micro patterns at the patch level: parsimony would suggest a much lower number than 2^{13} to specify the possible meaningful outputs of this qualitative model.[29]

Now, the Segregation model is a highly abstract model, and "it is not clear what data could be used to validate it directly" (Gilbert and Troitzsch 2005). But that's the point: if it were more specific, it would require even more inputs, which might double the functional complexity of the model for each additional bit required to specify the inputs, from Eq. 4. Although it might be difficult to know what data to use to validate a simpler, more abstract model, on the other hand a more realistic model might have a much greater complexity measure, as the possible mappings from inputs to possible outputs grow in number.

This suggests a trade-off: a less-complex model might in principle be easier to validate by virtue of its relative lack of complexity, while a more realistic model, which in theory could be fitted to historical data as part of its validation, might be enormously functionally complex, although the required scale of accuracy must be specified—at some more aggregate level usually.

Given the complexities of AB models, validating such models at both micro and macro levels may appear daunting, and given the difficulty of determining the class of models insensitive to many micro variables (or robust across Monte Carlo experiments with random micro variables), some might argue that such models must be taken on faith. We would argue that this pessimism is not warranted: complete validation is not necessary, and docking of different AB models of the same phenomenon will engender confidence in the robustness of their results. Indeed, simulations of economic phenomena are used in most countries to enable policy makers to have some idea of how the national economy is performing, and the EMF brings together simulators who model energy/environmental systems in order to illuminate policy issues, as discussed in Sect. 3 above.

Simulations are also used in a narrower domain by economists considering possible market consequences of proposed mergers between companies, so-called merger simulations, the results of which must be robust enough to convince the court, in many instances, that the competition regulator's decision is the right one. Convincing lawyers and judges of the validity of the predictions of one's economic simulation model is not easy. How might this be accomplished?

8 The *Daubert* Criteria and Simulations

When there is a need for a counterfactual, a prediction of what might happen to social welfare in a market in which two previously independent firms merge horizontally,

[29] One referee suggests that Segregation is, in practical terms, very much less complex than I have described it here: validity of the model at the macro level can occur for a large set of final patterns (output 3), so that a summary measure of segregation to perhaps only one significant place might be sufficient (output 1).

the economists of the US Federal Trade Commission and the European Competition Commission[30] have used so-called merger simulations to predict how market conditions would change and so whether economic welfare as measured by the change in social welfare would rise (perhaps because of economies of scale in production) or fall (perhaps because of reduced competition leading to higher prices than absent the merger).

Since merger decisions are often challenged in court, where costs and damages can run into millions of dollars, there is a high requirement for validity of the predictions, especially when approval might be withheld as a consequence of the simulation. Froeb (in Woodbury 2004) contrasts the academic referee spending "a little bit of time" on each review with the $100,000 referee reports in damages cases.

So merger simulations, to be taken seriously in court and in policy circles, need to be credible. How do they achieve this credibility, or how to they fail to do so?

Van Damme and Pinkse (2005) list the inputs to the model, when simulating market equilibria, in merger simulations: a description of the demand side (usually an estimated system of equations), a mode of competition (often Bertrand, although this is an unresolved issue in the literature, awaiting further empirical evidence, see Woodbury (2004)), a description of the supply side (firms' cost functions), and an assumption of firm behaviour (almost always profit-maximising).

Van Damme and Pinkse argue that two factors determine the precision of the results of any empirical study: *bias* (up or down in the results) and *variance* (from sensitivity of the results to changes in the data). Model complexity, they argue, plays a large role in the bias/variance tradeoff: very simple models which are a poor description of market reality may exhibit bias; but richer models, requiring more data, tend to exhibit more variance, unless more data is available. Richer datasets, and more complex models, can reduce both the bias and variance of the exhibited behaviour, but greater model complexity has its own traps, they argue. And yet, as we argued in Sect. 7 above, AB models at a macro level might be less complex than the aggregation of their agents' complexities and still exhibit macro behaviour qualitatively similar to the historical phenomena.

Van Damme and Pinkse (2005) present a hierarchy of modelling methods, as the available data and modelling resources grow: calibration, estimation, and modelling individuals' choices (with, for instance, supermarket scanner data).[31] As they put it, "Calibration entails the use of minute amounts of economic data to infer relevant quantities in a structured economic model." (p. 9) Absent exact models and sufficient data, the validity of the exhibited behaviour of the model is questionable, especially when making forecasts, such as presented in court testimony. Because of the implicit assumptions of lack of errors in both data and model, calibration is not able to provide confidence intervals for the exhibited behaviour, as Haefner (2005) argues a good

[30] In 2004 the European Commission amended the substantial test for mergers from "dominance" to the change in the level of effective competition in the market, thus approaching the 1982 US test, and requiring the competition watchdog to do more than predict the market share of the merged entity (Werden et al. 2004; Van Damme and Pinkse 2005).

[31] In Midgley et al. (1997) and Marks et al. (2006), we used scanner data when estimating the demand reactions to brand marketing actions: coupons, aisle displays, and promotional advertising, as well as price. See Fig. 1 above.

validation method should do. Estimation and modelling using individuals' market choice data can do so, but these econometric techniques are not AB simulation modelling.

In the US the courts have set a high standard for the merger simulators at the F.T.C. and the plaintiff companies, the so-called *Daubert* discipline:[32] As Stephens et al. (2005) recount, the 1993 *Daubert* decision made an explicit link between the reliability of an expert's testimony and the expert's use of scientific knowledge derived by use of the scientific method. The 1999 *Kumho Tire* decision extended the *Daubert* standards to testimony based on "technical" or "other specialised" knowledge.

Werden (Woodbury 2004) lists three conditions for admissibility of expert economic testimony, including the results of merger simulations: first, the witness must be an expert in the relevant field of economics;[33] second, the testimony must employ sound methods from the relevant field of economics; and, third, the testimony must apply those methods reliably to the facts of the case: simulation must be grounded on the facts. While lawyers have successfully argued with presumptions based on market shares and precedent, testifying economists must use the tools of economics, including simulations, properly.

In a paper rousing the system dynamics (Forrester 1961) simulation modelling community to meet the current standards for expert witness admissibility, Stephens et al. (2005) summarise the *Daubert* criteria and sub-criteria. Table 1 summarises this.

System dynamics simulations are quite different from AB simulations: possessing only a single level—they model the individual, or the firm, or the organisation, or the society—they cannot model interactions between scales or levels, and so do not exhibit emergent behaviour (Gilbert and Troitzsch 2005). But their relative simplicity means that they are more likely able to satisfy the criteria of Table 1.[34] It would take another paper to adequately explore the extent to which AB models might satisfy these criteria, that nonetheless present a standard for AB modellers to aspire to.

This non-lawyer will hazard a guess at how *Daubert* would apply to the Segregation model. The tolerance measure and the migration of unhappy households (the micro behaviours) are clearly testable hypotheses. That neighbourhood segregation occurs is also clear, although the mechanism for this must be related to the micro behaviours. Indeed, such tests at the micro and macro levels are required by *Daubert*, together with peer review. Moreover, general acceptance of segregation generated by such a

[32] *Daubert v. Merrell Dow Pharmaceuticals, Inc.,* 509 US 579 (1993); *Kumho Tire Co. v. Carmichael.,* 526 US 137 (1999).

[33] That the *Daubert* discipline places some emphasis on the people who oversee the simulation echoes the recent emergence of "companion modelling", in which the simulator is partnered by other stake-holders in the model-building and experimenting (simulating with different parameters) iterated stages (Barreteau et al. 2003). That is, developing confidence in a model is not simply a mechanical process, but involves the relationships between the simulation modellers and the people who use the simulation results or who are affected by others' use of them. This can also be seen in the EMF projects: simulation modelling, and its benchmarking, is almost always sponsored by policy-makers.

[34] Moreover, a list of 23 peer-reviewed journals in which system dynamics work has been published (Stephens et al. 2005, Table 2) reveals that this methodology is much more associated with managerial and engineering issues than with micro-economic and market-related issues: only two of the 23 journals could be regarded as economics journals.

Table 1 Criteria for expert witness testimony under *Daubert* (After Stephens et al. 2005)

1.	The preferred testimony should be based upon a testable hypothesis or hypotheses:
1.1	The hypothesis must have explanatory power relative to the case—it must explain the how and why of the observed behaviour. More than merely descriptive, the hypothesis must also be predictive.
1.2	The hypothesis must be logically consistent—it must contain no internal inconsistencies.
1.3	The hypothesis must be falsifiable, through empirical testing.
2.	The hypothesis must have been tested to determine the known or potential error rate.
3.	Hypotheses must have been formed and tested and analyses conducted in accordance with standards appropriate to the techniques employed, including standards regarding:
3.1	Consistency with accepted theories.
3.2	The scope of testing—". . . the more severe and the more diverse the experiments that fail to falsify an explanation or hypothesis, the more corroborated, or reliable, it becomes . . ." (Black et al. 1994)
3.3	Precision—precision is easier to test than generalisation.
3.4	Post-hypothesis testing—can it explain more than existing data?
3.5	Standards at least as strict as those for non-forensic applications.
4.	The techniques used should have been peer-reviewed:
4.1	The initial peer review before publication.
4.2	The post-publication peer review.
5.	The techniques used should have been generally accepted in the scientific community.

mechanism is now well established. We discuss below how the validation framework of Sect. 6 might be applied to counterfactual simulation models.

Werden (Woodbury 2004) argues that *Daubert* requires that all assumptions and simplifications in the model must be justified—using economic theory, industry data, consistency with stylised facts, including the present—and that sensitivity analysis is necessary in order to demonstrate how predictions depend on assumptions.

Froeb et al. (2004) explain that the techniques and models used in merger simulations would be more reliable and hence acceptable if they had been validated, retrospectively, in their predictions of post-merger behaviour, of firms, and of markets, but that lack of data has meant very little out-of-sample predicting and validation. Absent such testing of AB simulation models, simple models fully justified, with clear sensitivity analysis of the exhibited behaviours, should be considered an absolute minimum for the courts and lawyers.

Which of the relationships of Fig. 2 would best satisfy the *Daubert* criteria? An incomplete model (c) might be useless if the model did not simulate historically observed data, especially if it's to be used to predict, to simulate counterfactuals. But an inaccurate (or, as one referee would have it, a redundant) model (d) might be of value to the court if the counterfactual scenario were believed to fall into the region $\mathbf{M}\backslash\mathbf{Q}$. In Eq. 3, set the weight v to unity, to value inaccuracy, not incompleteness, in the degree of validation measure V.

Whether a standard check-list for the validation of AB simulation models can be, first, developed by the AB simulation community, and, second, adopted by, say, journal editors as a necessary (but not, of course, sufficient) condition for consideration for publication of papers presenting AB simulation results remains to be seen. Not just Leombruni et al. (2006), and Midgley et al. (2007) but also Fagiolo et al. (2006)

propose agendas for concentrated methodological research by the AB modelling community, as a first step.

9 Conclusion

As Shannon (1948) taught us, the notion of description is linked to the idea of the space of possibilities. We are drawn to the use of computer models of economic phenomena for at least two reasons: first, to escape the restrictions on the model necessary to obtain closed-form solutions, and, second, to explore the space of possibilities. But the very attractiveness of exploring this vast space creates problems when we want others to have confidence in our results. Because of the many degrees of freedom inherent in computer models compared to closed-form solutions—even if the closed-form restrictions enable solutions to uninteresting problems only—the skeptic seeks validation of our results, a validation which is more difficult precisely because of the complexities implicit in the large space of possibilities.

This paper has attempted to provide some formalism to measure the extent to which a model is validated, and at the same time to demonstrate how difficult it is to validate these complex models, even if at a larger scale the model's macro behaviour might be insensitive to many micro variables' specific values. We have examined simulation models of coffee rivalry at the retail level, Schelling's classic simulation simulation, the EMF's comparisons of different simulation models of the same phenomena, and the US courts' requirements for the use of the output from simulation models as evidence in antitrust cases.

In almost all cases, there is a trade-off between realism and parsimony, with greater realism demanding more variables and greater degrees of freedom. The challenge is to identify the class of models whose macro behaviour is robust to changes in these variables' values. Nonetheless, the *Daubert* discipline, and other recent moves to include the stakeholders in the validation process, point the way forward for simulation modellers.

In this paper, we have argued that. although simulations can prove existence (of a model to exhibit a specific behaviour) and so enhance explanation (of the historical phenomenon resulting in that behaviour), it is difficult for simulation to derive the necessary conditions for models to exhibit the specific behaviour. This might, we argue, be one reason for economists' evident reluctance to use computer simulations.

We have formalised validation and derived five possibilities for the relationships between the historical observations and the model behaviour. We discussed the relevance of each case for the simulator, depending on which of the three simulation goals was uppermost—explanation, prediction, or exploration—arguing that different goals favour different validation relationships.

A discussion of complexity followed, in which we argued that functional complexity is the appropriate measure for simulation models. As we discussed, for AB simulation models the macro behaviour might be less complex than the micro behaviour of individual agents, or the aggregate of all agents, which turns the Moss and Edmonds (2005) approach to AB model validation on its head.

The counterfactual predictions of "merger models" must satisfy a high level of validity to be accepted as evidence by competition tribunals and antitrust authorities and courts. Our research might usefully provide a further basis for the discipline to strengthen acceptance of computer simulations of economic phenomena through adoption of new procedures of model development and assurance (Midgley et al. 2007).

Acknowledgements We thank Rich Burton, Yaneer Bar-Yam, Ian Wilkinson, Shayne Gary, Peter McBurney, Hill Huntington, Leigh Tesfatsion, the participants at the Fourth UCLA Lake Arrowhead Conference on Human Complex Systems, and two anonymous referees for their comments and assistance.

Appendix A: Listing of Segregation

```
globals [
 percent-similar
 percent-unhappy
]

turtles-own [
 happy?
 similar-nearby
color?
 other-nearby
 total-nearby
]


to setup
 ca
 if number > count patches
  [ user-message "This pond only has room for "
+ count patches + " turtles."
   stop ]


 ask random-n-of number patches
  [ sprout 1
   [ set color red ] ]

 ask random-n-of (number / 2) turtles
  [ set color green ]
 update-variables
 do-plots
end

to go
 if not any? turtles with [not happy?] [ stop ]
 move-unhappy-turtles
 update-variables
 do-plots
end

to move-unhappy-turtles
 ask turtles with [ not happy? ]
```

```
  [ find-new-spot ]
end

to find-new-spot
 rt random-float 360
 fd random-float 10
 if any? other-turtles-here
  [ find-new-spot ]
patch
 setxy pxcor pycor
end

to update-variables
 update-turtles
 update-globals
end

to update-turtles
 ask turtles [


  set similar-nearby count (turtles-on neighbors)
   with [color = color-of myself]
  set other-nearby count (turtles-on neighbors)
   with [color != color-of myself]
  set total-nearby similar-nearby + other-nearby
  set happy? similar-nearby >=
(% similar-wanted * total-nearby/100)
 ]
end

to update-globals
 let similar-neighbors sum
values-from turtles [similar-nearby]
 let total-neighbors sum
values-from turtles [total-nearby]
 set percent-similar
(similar-neighbors / total-neighbors) * 100
 set percent-unhappy
(count turtles with [not happy?]) / (count turtles) *
100
end

to do-plots
 set-current-plot "Percent Similar"
 plot percent-similar
 set-current-plot "Percent Unhappy"
 plot percent-unhappy
end
```

Source: Wilensky (1998).

References

Arthur, W. B. (2006). Out-of-equilibrium economics and agent-based modeling. In Tesfatsion, Judd, K.L. (2006), pp. 1551–1564.

Axelrod, R. (2006). Simulation in the social sciences, in Rennard (2006), pp. 90–100.
Axtell, R., Axelrod, R., Epstein, J., & Cohen, M. D. (1996). Aligning simulation models: A case study and results. *Computational and Mathematical Organization Theory, 1*, 123–141.
Barreteau, O., and others (2003). Our companion modelling approach. *Journal of Artificial Societies and Social Simulation, 6*(1), http://jasss.soc.surrey.ac.uk/6/2/1.html (accessed 2006/07/30).
Bar-Yam, Y. (1997). *Dynamics of complex systems*. New York: Perseus Press. http://www.necsi.org/publications/dcs.
Bar-Yam, Y. (2003). Unifying principles in complex systems. In M.C. Roco & W.S. Bainbridge (Eds.), *Converging technologies for improving human performance: Nanotechnology, biotechnology, information technology and cognitive science* (pp. 380–409). New York: Springer. http://www.necsi.org/projects/yaneer/ComplexSystems.pdf (accessed 2006/07/30)
Black, B., Ayala, F.J., & Saffran-Brinks, C. (1994). Science and the law in the wake of *Daubert* a new search for scientific knowledge. *Texas Law Review, 72*, 715–751.
Brenner, T., & Werker, C. (2006). A practical guide to inference in simulation models, *Max Planck Institute of Economics, Evolutionary Economics Group*, Jena, #0602. https://papers.econ.mpg.de/evo/discussionpapers/2006-02.pdf (accessed 2006/06/23).
Bunn, D. W., & Oliveira, F. S. (2003). Evaluating individual market power in electricity markets via agent-based simulation. *Annals of Operations Research, 121*, 57–77.
Burton, R. M. (2003). Computational laboratories for Organizational Science: Questions, validity and docking. *Computational & Mathematical Organizational Theory, 9*, 91–108.
Cook, C. (2006). Home page: The schelling segregation model demo http://www.econ.iastate.edu/tesfatsi/demos/schelling/schellhp.htm (accessed 2006/07/31).
Courtois, P.-J. (1985). On time and space decomposition of complex structures. *Communications of the ACM, 28*(6), 590–603.
Durlauf, S. (2005). Complexity and empirical economics. *The Economic Journal, 115*(June), F225–F243.
Epstein, J. M. (2006). *Generative social science: Studies in agent-based Computational modeling*. Princeton: P.U.P.
Fagiolo, G., Windrum, P., & Moneta, A. (2006). Empirical validation of agent-based models: A critical survey, LEM working paper 2006/14, Pisa, Italy: *Laboratory of Economics and Management, Sant'Anna School of Advanced Studies*, May. http://www.lem.sssup.it/WPLem/files/2006-14.pdf (accessed 2006/07/30).
Forrester, J. W. (1961). *Industrial dynamics*. Camb: MIT Press.
Frankel, D. M., & Volij, O. (2005). Measuring segregation, mimeo, Iowa State, http://www.econ.iastate.edu/faculty/frankel/segindex_all20.pdf (accessed 2006/07/31)
Friedman, M. (1953). *Essays in positive economics*. University of Chicago Press.
Froeb, L., Hosken, D., & Pappalardo, J. (2004), Economic research at the FTC: Information, retrospectives and retailing. *Review of Industrial Organization, 25*, 353–374.
Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the Social Scientist*, (2nd ed.). Buckingham: Open University Press.
Gutowitz, H. (1990). Cellular automata: Theory and experiment. In *Proceedings of a workshop sponsored by The Center for Nonlinear Studies. Los Alamos* (pp. 7–14). The Hague: North-Holland.
Haefner, J. W. (2005). *Modeling biological systems: Principles and applications* (2nd ed.). New York: Springer.
Huntington, H. G., Weyant, J. P., & Sweeney, J. L. (1982). Modeling for insights, not numbers: The experiences of the Energy Modeling Forum. *OMEGA: The International Journal of the Management Sciences, 10*(5), 449–462.
Judd, K. L. (2006). Computationally intensive analyses in economics, in Tesfatsion and Judd (2006), pp. 881–893.
Kaldor, N. (1961). Capital accumulation and economic growth. In F.A. Lutz, D.C. Hague (Eds.), *The theory of capital* (pp. 177–222). London: Macmillan.
Klein, E. E, & Herskovitz, P. (2005). Philosophical foundations of computer simulation validation. *Simulation & Gaming, 36*, 303–329.
Leombruni, R., & Richiardi, M. (2005). Why are economists sceptical about agent-based simulations? *Physica A, 355*, 103–109.
Leombruni, R., Richiardi, M., Saam, N. J., & Sonnessa, M. (2006). A common protocol for agent-based social simulation. *Journal of Artificial Societies and Social Simulation, 9*(1). http://jasss.soc.surrey.ac.uk/9/1/15.html (accessed 2006/06/22).

Malerba, F., Nelson, R., Orsenigo, L., & Winter S. (1999). History-friendly models of industry evolution: The computer industry. *Industrial and Corporate Change, 8*(1), 3–40.
Mankin, J. B., O'Neill, R. V., Shugart, H. H., & Rust, B. W. (1977). The importance of validation in ecosystem analysis. In G.S. Innis (Ed.), *New directions in the analysis of ecological systems, part 1 simulation council proceedings series* (Vol. 5: pp. 63–71). California: Simulation Councils, La Jolla. Reprinted in H.H. Shugart and R.V. O'Neill, eds. *Systems ecology* Dowden, Hutchinson and Ross, Stroudsburg, Pennsylvania, 1979, pp. 309–317.
Manson, S. M. (2002). Validation and verification of multi-agent systems. In M.A. Janssen (Ed.), *Complexity and ecosystem management*. Cheltenham: Edward Elgar.
Marks, R. E. (1992). Breeding optimal strategies: Optimal behaviour for oligopolists. *Journal of Evolutionary Economics, 2*, 17–38.
Marks, R. E. (2006). Market design using agent-based models, in Tesfatsion and Judd (2006), pp. 1339–1380.
Marks, R. E., Midgley, D. F., & Cooper, L. G. (2006). Co-evolving better strategies in oligopolistic price wars, in Rennard (2006), pp. 806–821.
Midgley, D. F., Marks, R. E., & Cooper, L. G. (1997). Breeding competitive strategies. *Management Science, 43*(3), 257–275.
Midgley, D. F., Marks, R. E., & Kunchamwar, D. (2007). The building and assurance of agent-based models: An example and challenge to the field. *Journal of Business Research, 60*, 884–893. http://www.agsm.edu.au/∼bobm/papers/Midgley-Marks-Kunchamwar.pdf.
Miller, J. H. (1996). The coevolution of automata in the repeated prisoner's dilemma. *Journal of Economic Behavior and Organizations, 29*, 87–113.
Miller, J. H. (1998). Active nonlinear tests (ANTs) of complex simulations models. *Management Science, 44*(6), 820–830.
Moss, S., & Edmonds, B. (2005). Sociology and simulation: Statistical and quantitative cross-validation. *American Journal of Sociology, 110*(4), 1095–1131.
Rennard, J.-P. (Ed.) (2006). *Handbook on research on nature-inspired computing for economics and management*. Hershey, PA: Idea Group.
Resnick, M. (1994). *Turtles, termites, and traffic jams: Explorations in massively parallel microworlds*. MIT Press.
Rosen, R. (1985). *Anticipatory systems: Philosophical, mathematical, and methodological foundations*. Oxford: Pergamon.
Rubinstein, A. (1998). *Modeling bounded rationality*. MIT Press.
Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Psychology, 1*, 143–186.
Schelling, T. C. (1978). *Micromotives and macrobehavior*. New York: Norton.
Schelling, T. C. (2006). Some fun, thirty-five years ago, in Tesfatsion & Judd (2006), pp. 1639–1644.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27*, 379–423 July, 623 August.
Sharpe, W. F. (2007). APSIM: An asset price and portfolio choice simulator. In *Investors and markets: Portfolio choices, asset prices and investment advice*. Princeton University Press. http://www.stanford.edu/∼wfsharpe/apsim/apsim.pdf (accessed 2006/07/30).
Simon, H. A., & Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica, 29*, 111–138.
Simon, H. A., & Wallach, D. (1999). Cognitive modeling in perspective. *Kognitionswissenschaft, 8*, 1–4.
Stephens, C. A., Graham, A. K., & Lyneis, J. M. (2005). System dynamics modeling in the legal arena: Meeting the challenges of expert witness admissibility. *System Dynamics Review, 27*(2), 95–122.
Tesfatsion, L. (2006). Conducting experiments with Chris Cook's Schelling Demo, mimeo, Iowa State, http://www.econ.iastate.edu/classes/econ308/tesfatsion/segex2.VIITrento.pdf (accessed 2007/05/31).
Tesfatsion, L., & Judd, K. L. (Eds.) (2006). In K. Arrow & M. D. Intriligator (Eds.), *Handbook of computational economics, volume 2: Agent-based computational economics* in the series *handbooks in economics*. Amsterdam: Elsevier Science.
Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33*, 13–14.
van Damme, E. E. C., & Pinkse, J. (2005). Merger simulation analysis: An academic perspective, TILEC Discussion paper No. 2005-013 http://ssrn.com/abstract=869737 (accessed 2006/07/30).
Vriend, N. (2000). An illustration of the essential difference between individual and social learning and its consequences for computational analyses. *Journal of Economic Dynamics and Control, 24*, 1–19.

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure of deoxyribose nucleic acid. *Nature, 4356*, 737–738, April 25.

Werden, G. J., Froeb, L. M., & Scheffman, D. T. (2004). A *Daubert* discipline for merger simulation, *Antitrust Magazine*, *18*(3), 89–95.

Weyant, J. P., & Hill, J. (1999). Introduction and overview, The costs of the Kyoto protocol: A multi-model evaluation, *The Energy Journal, 20* (Special Issue) vii–xliv, May.

Wilensky, U. (1998). NetLogo segregation model. Center for connected learning and computer-based modeling, Northwestern University, Evanston, IL. http://ccl.northwestern.edu/netlogo/models/Segregation.

Woodbury, J. R. (Ed.). (2004). Whither merger simulation? *The Antitrust Source,* May, http://www.abanet.org/antitrust/source/05-04/whither.pdf (accessed 2006/07/30).