Balancing the Criticisms: Validating Multi-Agent Models of Social Systems

Marcus A. Louie
Department of Engineering and Public Policy
Center for the Computational Analysis of Social and Organizational Systems
Carnegie Mellon University
Pittsburgh, PA 15213

Kathleen M. Carley
Computation, Organizations and Society
Institute for Software Research International
Center for the Computational Analysis of Social and Organizational Systems
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Using multi-agent models to study social systems has attracted criticisms because of the challenges involved in their validation. Common criticisms that we have encountered are described, and for each one we attempt to give a balanced perspective of the criticism. A model of intra-state conflict is used to help demonstrate these points. We conclude that multi-agent models for social systems are most useful when 1) the connection between micro-behaviors and macro-behaviors are not well-understood and 2) when data collection from the real-world system is prohibitively expensive in terms of time or money or if it puts human lives at risk.

Keywords: Multi-agent model; simulation; validation; social systems;

1 Introduction

Multi-agent models are touted as a method for analyzing "complex social systems," particularly those characterized by multiple interacting parts and non-linear behavior [11, 13, 14, 29]. As a result, these models are being used to examine a variety of policy domains including civil violence [16], the spread of infectious disease [12], the effects of government policies on the transportation of goods [5], and the effects of mutual influence on domestic water demand [26].

Researchers and policy makers are turning to these models for reasons of ethics, cost, timeliness and appropriateness. In some systems, such as those modeling the spread of infectious disease, testing experimental conditions would put the safety of people at-risk, creating an ethical problem. In other cases, real-time evaluation of an existing system may be prohibitively long. Simulation allows for rapid assessment. Simulation is also used when the cost of collecting data on the dependent variable is prohibitively expensive, or there are a large number of experimental conditions to test. For example, in a disaster, simulation can be used to rapidly evaluate many previously unexamined alternatives [12, 27]. In all of these cases, since the real-world system under study is considered a complex, non-linear dynamic system, multi-agent simulations are often used as it is considered to have the appropriate level of complexity.

As the use of multi-agent models has become more prevalent, a growing concern has arisen with how to validate such models. From a history of science perspective it is important to note that the most advanced methods of validation were developed in engineering fields for assessing models of technical systems that followed fundamental physical laws. In contrast, these large-scale multi-agent systems are used for examining socio-political systems where the fundamental underlying laws are not known. Multi-agent models of social systems are difficult

to validate because these models represent a new approach to simulation for which traditional validation methods are not always applicable. Given these challenges, we need to first ask what an appropriate validation process is for such models. Second, we need to know what value these models have even despite the challenges in their validation.

This paper proceeds in two parts. First, we synthesize previous work in simulation model validation to construct a validation strategy for models of social-systems based on the purpose of a model. Second, we outline common criticisms related to validating multi-agent models of social-systems. For each criticism, we seek to demonstrate that even if the legitimacy of the claim is granted, that the models can still be useful as a means for developing theories about the target system. A dynamic-network multi-agent model of intra-state conflict called the Regional Threat Evaluator (RTE) is used as an example. Note, this paper is not a presentation and validation of the RTE model. Rather, this paper is an analysis of the appropriateness of standard validation procedures for large scale multi-agent models such as the RTE.

2 Designing a validation process

Previous work on validation processes for simulation models can be broken into two principal threads. One thread addresses what the *major steps* are in validation [4, 9, 30, 35]. For example, Thomsen et al. [35] propose a trajectory of major validation steps for simulation models that are based on real data and whose purpose is to be used prescriptively. The other thread has focused on the specific *techniques* that might be used during each of the major steps of validation [3, 20, 23]. A well-recognized example from this thread is Law and Kelton's [23] treatment of statistical validation techniques for simulation models.

The validation process should be tied to the purpose and the context for which the model is being developed [6, 8, 9, 30, 35]. We distinguish a validation process from a validation

technique. A process is a series of steps taken to validate different parts a model such as verifying that the model mechanisms are representative of the real-world or comparing model output to historical data. Techniques are the individual methods used to judge whether each part of the model is "valid." Statistical tests such as t-tests are examples of techniques.

Sargent [30] provides an overview of different validation techniques, each providing different types and levels of validity. He notes that the desired level of validity is determined on the purpose of the model, but does not attempt to describe in detail what different purposes are and how they relate to the validation process. Burton [8, 9] complements Sargent's work by describing types of questions that are asked of simulation models while recognizing that the level of validation is still dependent on the question, or purpose, of the model.

In this section, we synthesize this related work and organize the types of questions that are asked of models and associate them with the types of validation that are appropriate for each type of question. To assist in giving structure to the synthesis, we use a conceptual description of the simulation-model development and validation process given by Banks et. al [4]. Figure 1 is a representation of the conceptual components and steps of the process, recreated from Sargent [30].

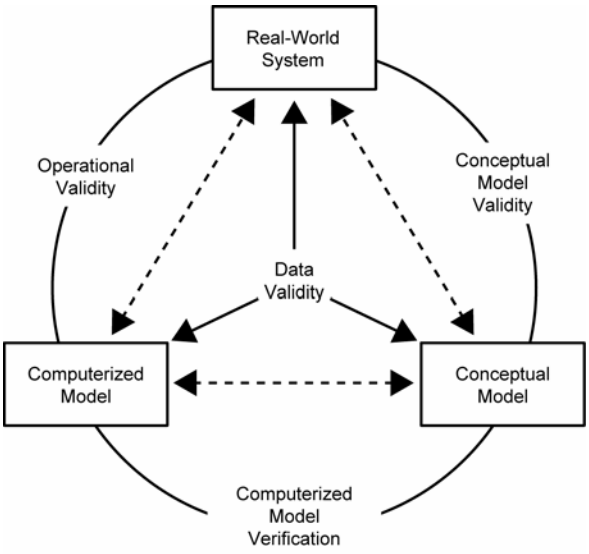

Figure 1. Conceptual depiction of the components of model development and validation. Figure is recreated from Sargent (1992).

2.1 Types of validation

The boxes in Figure 1 show three different parts of a simulation model that can be validated. This section identifies the different types of validation that can be performed for each of these model parts. Conceptual validity is determining the extent to which the model theories and the underlying assumptions are appropriate for the purpose of the model. Determining the validity of data involves making sure that the data are appropriate for the purpose of the model, that a sufficient amount of data exist to build and validate the model, and that the data are accurate with respect to the real system. The operational validity of the model is determining the extent to which the model produces output that matches the real system under investigation for the purpose in which it was developed.

Before a multi-agent model is written, a conceptual model is usually produced which identifies the assumptions, relevant entities, their actions, and relationships among entities. The next step is to formalize the concepts into mathematical relationships. Validating the algorithms

answers questions relating to whether the equations and computational procedure used represent the conceptual model. Subject matter experts usually validate the conceptual model.

Though this paper is focused on validating multi-agent models, we mention data validity because the quality of available data often constrains both model development and operational validation. In simulation models where data need to be translated from the real-world before going into the model, the data are subject to strong biases that reflect the background of the data collectors. Two data collectors may interpret qualitative data differently and so it becomes helpful to be aware of and to minimize the variation of model input produced by different biases of data collectors. Ideally, a formal data collection method would be established.

Typically, a simulation model is compared to a real system in a sequence of steps. Often the first step is to calibrate the model to show the extent to which it can reproduce some target system. To determine whether the model has been overfit to that target system, one or more target systems are used to determine how well the model generalizes to other systems. Historical data from the target systems are usually used to calibrate the model and to see how well it generalizes. Prospective data from the system can also be used, and using this type of data from the system is referred to as forecasting. Thus, the goal is see how well the model can predict future conditions of the target system. A related form of forecasting attempts to not just forecast the future, but to attempt to change the future based on results from the model. In forecasting with an intervention, the intervention is used in the real system. The level of validity that has been achieved in previous tests will determine whether real resources and people will be committed to implementing the intervention.

2.2 Types of questions

We commonly think of two broad categories of questions in science and policy: the positive and the normative. In the positive, we observe the world, create a description and then seek an explanation. Theoretical models are proposed and then tested by creating hypotheses to answer questions to which we do not yet know the answers. In the normative, we seek to describe what is good, what should be done. This requires a confirmation that not only does the model make sense in the realm of the positive, but that its boundaries of explanation overlap to the subset of the normative for which the model is being applied. Somewhere between these questions is a third category of questions, questions of what is plausible. The plausible are explorations of *what might be* [9].

Simulation models are particularly useful in this context. We include system dynamics, celluar automata, and multi-agent models as types of simulation models. In the real-world we can devise controlled laboratory experiments if we would like experimental control over explanatory variables, but in much larger contexts such as processes of state-failure or disease epidemics, we cannot easily manipulate the environment to conduct a controlled experiment and if we could, it is not always appropriate or ethical for reasons mentioned previously. Further, we might think about data coming from the real-world as a single data stream. With the exception of comparative case studies and extrapolation from data, we are not able to explore the plausible beyond what has already happened. Using theory and simulation models allow questions to be posed that look beyond what has already happened to what could happen.

2.3 Synthesizing validation types and questions

In line with others who have stated that the validation process should be tied to the purpose of the model, we structure a framework for validating simulation models. Table 1

synthesizes our discussion on major validation steps and the purpose of a model. Our hope is that it provides added structure for others who are interested in validating their simulation models, moving validation efforts away from the diverse set of procedures cited by Kleindorfer et al. [21].

The columns of the table list the three broad types of questions that can be posed by simulation models. Under each are a series of cells that correspond to different validation steps (conceptual, operational, and data) and sub-steps (e.g., calibration, replication). In each cell is a value that indicates whether the sub-step should be included in a given procedure to establish the extent to which a model is valid. Computational model verification was purposefully left out of the table as we believe model verification has more to do with debugging and software engineering than it does with a comparison of the model to the real-world.

			Type of Question		
			Positive	Plausible	Normative
Type of Validation	Conceptual	Representative	Yes	Yes	Yes
		Proper Algorithms	Yes	Yes	Yes
	Operational	Calibration	Yes	Yes	Yes
		Replication	Yes	Yes	Yes
		Forecasts	Yes	Yes	Yes
		Hypothetical	No	Yes	Yes
		Intervention	No	No	Yes
	Data	Sensitivity to Parameter Bias	Yes	Yes	Yes

Table 1. The types of questions appear across the columns and examples of validation steps that are appropriate for each question appear down the rows. Table cells marked "Yes" indicate that the validation sub-step is appropriate for increasing the degree of validity for the corresponding type of question.

The framework offered in Table 1 builds on past efforts to give structure to the validation process of simulation models. Sargent [30] provides a conceptual picture of the varied ways in which a simulation model can be validated. This conceptual picture is combined with Burton's

[8, 9] enumeration of different questions that can be posed to and answered by simulation models. The result is Table 1, organized by the purpose of the model; it describes the different ways in which a model can be validated.

3 The RTE: A case example

The Regional Threat Evaluator (RTE) model is based on the integration of multiple theories of social, psychological and economic behavior that collectively account for why an agent takes action and what action gets taken. It is currently being developed from a similar model based on urban threat environments. Both models were developed as special-purpose models to answer specific types of questions.

The basic idea is that inter-group conflict is due to a combination of tension [19] and social comparison [17], the effects of which can be modulated by social pressure [18]. Agents who are more tense and who see themselves at more of a disadvantage relative to others are more likely to engage in hostile actions; whereas, lower tension and higher advantage lead to non-hostile actions. Agents who have influence over the agent in question can use that influence to escalate or de-escalate the impact of tension and social comparison. Specifically, an agent who is influenced by others who themselves are tense or feel deprived will feel more tense and deprived than will an agent surrounded by others who are less tense or less deprived. Social influence derives from shared attributes such as culture, knowledge, borders, and goals and co-evolves with those attributes [10]. It follows that the more heterogeneous a population and the more the lines of differentiation line up the greater the potential for hostility [7]. When agents decide to take action, the action and target are selected using a bounded-rationality cost-benefit analysis [1, 24, 33] subject to resource constraints [28]. The costs and benefits of taking a particular action against a particular target are also modulated by social influence. Thus, agents

are more likely to take the kinds of actions against the kinds of targets that social pressure suggests are appropriate and will be sanctioned by other agents for inappropriate action or target choice.

The RTE Model is a multi-agent network model of state failure (see Figure 2). In this model, boundedly-rational agents interact and take actions to achieve goals. When agents act they take into account what resources they have available, the cost and benefits of the action, and the opinions of others by whom they are influenced. These actions influence the likelihood of state failure. State failure is measured using nine factors and a composite indicator. These factors are lack of state legitimacy, potential for province secession, hostility, tension, level of corruption, level of terrorist activity, level of criminal activity, level of foreign military aid, and lack of essential services. State failure is also measured at the province level using similar indicators.

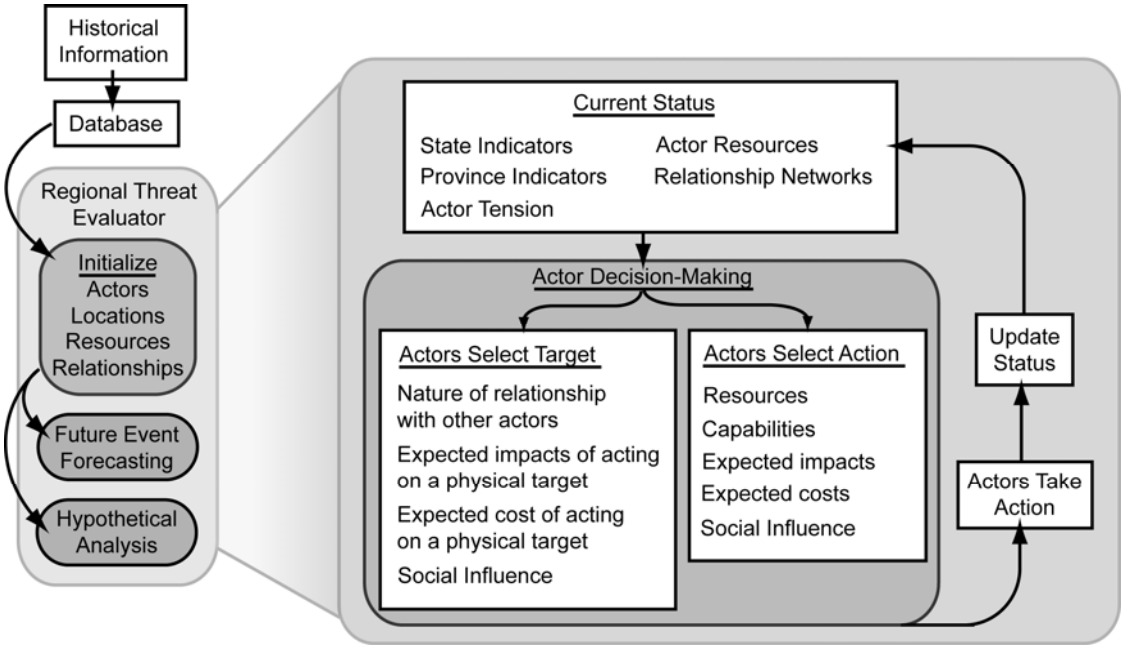

Figure 2. This top-level view of the Regional Threat Evaluator (RTE) describes the decision-making process of the actors and how the model characterizes the environment.

The model uses real-world data to ground the initial model parameters and then the actors, or agents, proceed to interact and take actions which consume or generate resources. Activity at the agent level then leads to changes in these agents, their resources, the non-agent targets, and these indicators. For example, forced migration of a population from one province to another is likely to decrease tension in the province left, increase tension and hostility and decrease essential services in the province migrated to, increase tension in the population that migrated and decrease their resources.

Data that are used to parameterize the initial conditions of the model, inform the scenario, and do limited validation came from a 32 different sources. These sources included: 6 national agencies (e.g., Indonesia-Tourism.com), 6 NGO's (e.g., United Nations and World Bank), 4 US Agencies (e.g., CIA and Dept. of Energy), 6 News Services (e.g., Bangkok Post and BBC), 10 Research and Academic Institutions (e.g., Terrorism Knowledge Base and Institute of Southeast Asian Studies), and 2 Corporate/Labor groups (Netcraft and International Telecommunication Union). In addition, specific information on the relevant entities and provinces came from various on-line news (e.g., Washington Post) and web-services (e.g., WikiPedia) from both US and foreign media. Illustrative websites used include http://www.uis.unesco.org, http://www.tkb.org, http://www.undp.or.id/pubs/ihdr2001/ihdr2001_full.pdf, and http://www.unescap.org/esid/psis/population/database/thailanddata/thailandfacts.htm. These data were used as a basis for 150 state indicators, 60 province/region indicators, and 30 entity indicators used to initialize the simulation model. In addition to the real data that were collected, regional experts on Indonesia and Thailand were also consulted. Experts were put together by the Defense Advanced Research Projects Agency. Representative affiliations include the Defense

Intelligence Agency, Office of Naval Research, and the U.S. Pacific Command. Face validation by the regional experts focused on whether the conditions represented in the model produced the expected outcomes in terms of the indicator variables overtime.

4 Balancing the criticisms

While reviews, editorial decisions, and discussions at professional conferences often contain critiques, it is rare in the literature to find systematic critiques of multi-agent models to study social systems. In fact, the authors are not aware of any articles that critique the method. Nonetheless, evidence suggests that the use of multi-agent models to study social systems has not yet been fully embraced. While there are a growing number of papers that use using multi-agent models published in top-tier social science journals, the number of such papers is still small. For example, Axelrod [2] points out that a simple search of the Social Science Citation Index in 2002 for the word "simulation" in the title turns up 77 articles spread among 55 different journals.

Based on observations from the literature and our own editorial and conference experience, we have compiled a list of what we perceive to be the most significant complaints against using multi-agent models to study social systems. Full weight is attempted to be given to each criticism before one or more counterpoints are offered to give a more balanced perspective on the role multi-agent models have for studying social systems. Throughout the paper, the RTE will be used to illustrate certain points.

4.1 Limited real-world data

Models, as representations of real-world systems, are typically evaluated by how well they can match the behavior of the target system. Evaluations are most credible when the

behavior of the target system can be captured by measurable data and when model output can be compared to the system using quantitative and objective methods. One of the criticisms of using multi-agent models to study social systems is that there is often a paucity of data from relevant real-world social-systems making it difficult to both ground the model in reality and to validate its output. When the purpose of a model does not require data, then this criticism is least relevant. For example, intellective models, which might be used to demonstrate the theoretical adequacy of an assumption or to help prove a concept, may not need data from any specific real-world system. It may be sufficient to use stylized facts to parameterize the model. However, when the purpose of the model is intended to describe the behavior of a specific real-world system, then data are necessary to achieve a credible degree of veridicality. In essence, the purpose of the model dictates how relevant this criticism is. This section is principally concerned with models whose purpose requires greater degrees of veridicality.

Data from the real-world are usually sought for two reasons, to ground the model parameters in the real world and to compare the model output with measurements from the real world. In other words, data are used to make sure that what goes in and what comes out of the model are related to the real-world.

Collecting data about a social-system may be difficult for several reasons. Suppose a multi-agent model of intra-state conflict is selected to study how networks of insurgent groups evolve and dissolve overtime and what effect this has on the stability of the state. Collecting data on the characteristics of the groups, their behaviors, and their relationships with one another is difficult for reasons of safety, access, and credibility. In a multi-agent model of the weaponized-spread of biological agents, where the demographics, social relationships, and behaviors of individual people are believed to influence the spread of diseases as well as what

effective policies may be, there are very few real-world incidents where biological agents have been released. In both of these cases, the lack of data does not mean that we cannot try to understand how a system may behave in a systematic and formal way. The consequences to human lives is obviously too great to perform real-world experiments to collect data, but multi-agent models could be used to represent what are believed to be the important structures and variables and then used to fill in the data that are too costly to collect from the real-world.

We outline these difficulties not to excuse model developers from grounding their models in real-world data or from performing quantitative model validation, but rather, to highlight these difficulties as *opportunities* for which multi-agent models can be employed to help learn about social systems that are otherwise outside of the scope of other formal methods. If data are missing, how, then, can multi-agent models be used to still learn something about the system under study?

Consider again the example of a multi-agent model of intra-state conflict. Corruption is known to be an indicator of state-failure, but real-world quantitative data to compare model outputs to are limited. The portion of aid that is skimmed from aid for personal gain is not reported, and there is good reason for this. It is difficult to collect this type of data because corruption can happen at many different levels and by many different actors. Maintaining transparency is time consuming and expensive. Further, aid agencies are reluctant to report known cases of corruption for fear that it reflects poorly on their ability to fund projects, potentially reducing the amount of donations they receive. The data on corruption that do exist are the opinions of subject matter experts, aggregated into a perception index. The data, however, are collected annually, a period of time that is larger than what the RTE produces [36]. Thus, any shifts in the level of corruption that are predicted by the RTE within time periods less

than one year cannot be compared to the real data. For example, when foreign aid is received by a country we might expect corruption levels to shift in the months after reception, but this dynamic is not possible to capture with data currently available.

Based off input from the regional experts, a tsunami followed by an infusion of foreign aid is expected to have two main effects: 1) foreign aid has a marginal initial impact on corruption levels and 2) lower initial levels of corruption should lead to lower increases in corruption in response to foreign aid. Table 2 summarizes the effects, the data used to evaluate the RTE, and the method used to compare the simulation output to what is expected from the real world.

Expected Effect	Simulation Data	Real Data	Method of Evaluation
Foreign aid has a marginal initial impact on corruption levels	Overtime level of corruption	Subject matter experts	Inspection and face validation
Lower initial levels of corruption should lead to lower increases in corruption in response to foreign aid.	Overtime level of corruption	Subject matter experts	Inspection and face validation

Table 2. Listed for each expected effect is the type of data coming from the RTE, data about the real world that is used for comparison, and the method of comparison.

When constructing rules that affect corruption levels, we turned to regional experts and published reports on Indonesia to hypothesize about the relevant factors driving corruption levels. While a lack of quantitative data on corruption levels prevents us from comparing model output to data from the real world using statistical analyses, sensitivity analysis can support validation by showing whether factors have effects that they are expected to have. Based on input from the regional experts and reports, the following variables and equation was used as a first-order approximation to update levels of corruption (Equation 1):

- Initial corruption level, C_0
- Adjustment to corruption due to aid type and level, A
- Minimum level of acceptable service in the region, S_T
- Current level of service in the region, S
- Region volatility, V

$$C' = C_0 + (1 - C_0) \cdot (S_T - S) \cdot A \cdot V \cdot C_0$$

Equation 1. A modified version of this weight and adjustment formula is used to update most other variables in the model. This equation considers the difference between current levels of essential services versus what the region's perceived minimum level of essential services is. All variables are assumed to have values between 0 and 1 inclusive.

Updates to corruption use a weight and adjustment formula that is used to update most other variables in the model. The formula includes the $(S_T - S)$ term, which weights the adjustment by the difference in the minimum level of acceptable service in a region and the current level of services being provided. The modification is motivated by the idea that less aid is grafted when there is more of a perceived need for aid.

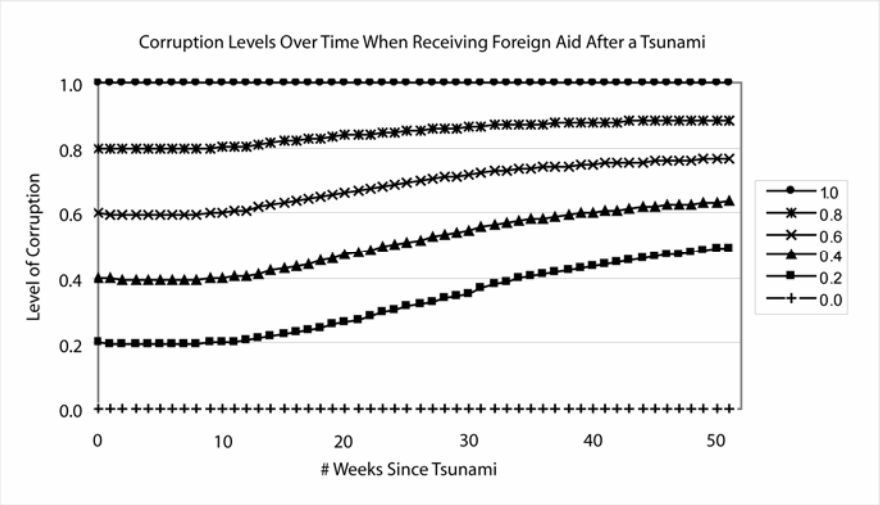

Figure 3. Overtime levels of corruption for a range of initial corruption levels.

Figure 3 shows how the initial corruption level before the tsunami affects the future level of corruption once aid starts to flow into the state. The model was parameterized and run using data for Indonesia. All tested variations in initial corruption levels show that future increases in levels of corruption start to happen around week 16. The stability of this feature in model output gives us confidence that initial corruption levels will not significantly alter when increased levels of corruption begin to occur. Regional experts agreed that changes in levels of corruption were low during the first 5 months as most people were in need of food, water, and medical care. Once reconstruction of housing and infrastructure begins, corruption becomes most prevalent as contractors are hired [34]. The government of Indonesia focused on immediate relief (food, water, medical aid, and temporary shelters) until June 2005, thereafter the focus shifted to rehabilitation and reconstruction. Thus, there is agreement in the literature on corruption, what is known about Indonesia's recovery plan, and the regional experts' beliefs about corruption levels in Indonesia. The model's results approximated the expected rise in corruption levels, but predicted the rise about one month early.

Initial corruption levels did affect how quickly future corruption levels increased. It is expected that less corrupt countries would lead to less of an increase in future levels of corruption, but Figure 3 shows that lower initial levels of corruption lead to greater increases in corruption than greater initial levels of corruption do. The discrepancy indicates that the underlying conceptual model may be flawed. As a result the operating conditions should be constrained to countries with relatively high corruption levels (as in Indonesia and Thailand) or the theory should be amended to include countries with lower levels of corruption.

Even without data, we were able to develop and use the RTE to test a conceptual model of corruption in a systematic and formal way. It allowed us to identify potential shortcomings in

the conceptual model, shortcomings that might not have been identified without undertaking a modeling effort. The value of multi-agent models lie in their ability to explore and inform us about how a system might operate under different conditions. The lack of data is not a fault of the method, but the model can be developed and used to exploit what is already known about the system to determine what the data might look like. Many systems that we would like to study do not have the appropriate data to answer the types of questions we might desire to ask.

4.2 Subject matter experts can disagree

If model outputs are not compared to real data, then by what other means do we have? Subject matter experts may be used in cases when the data are known not to exist, are expensive to collect, or the time frame for collecting the data surpasses the funding time frame of model development. Using subject matter experts for model validation and calibration can be more troublesome than using quantitative data because experts can disagree with each other. Disagreement opens up a series of questions: what are the origins of the disagreement, does the range of expert opinions significantly impact the model results, and how should different expert opinions be used [25]. Even when there is agreement, experts can still be wrong.

Disagreement between experts does not necessarily mean that the value of the model is indeterminable. Experts vary in their experience, foundations of knowledge, and assumptions. These differences among experts can lead them to make conclusions that conflict with one another about how a real-world system ought to behave. In general, model construction and simulation is useful as a means of testing assumptions, not necessarily to determine whether one assumption is correct or not, but also to determine what effects an assumption has if incorporated into a model. This latter style of assumption testing may be particularly useful when experts do not agree.

Assumptions come in a variety of forms, but we broadly categorize these as *assumptions of representation*, *assumptions of process,* and *assumptions of value*. Assumptions of representation are those that specify what the important agents of the model are, at what level the agents are modeled, what variables are used to describe the agents, and what relationships exist between the agents. Assumptions of process specify how the agents make decisions, how they respond to events, and how their relationships affect their behavior. The order of the procedural steps is also a process assumption as ordering can have non-trivial effects on the results. Assumptions of value have to do with how the model is grounded in the real-world. If real data exist to parameterize the model sometimes a translation needs to occur that converts the real data into model-specific values. For example, in the RTE agents have influence relationships with each other. The real-world data on how much one agent influences another has to be inferred from publicly available reports, news stories, and from subject matter experts. This qualitative data has to then be translated into a value the model can manipulate, in this case a normalized value between 0.0 and 1.0 inclusive. Whatever reasoning is used to translate the qualitative data into a model readable data are assumptions of value. The reasoning makes two assumptions: 1) what the real-world is actually like and 2) what values used in the model represent the real world.

Once it has been decided that a multi-agent model is the appropriate method for studying a system, there is some limitation to how widely experts may disagree about its representation. Nonetheless, experts may disagree about the level of analysis; should the agents represent individuals, groups, or both, what types of relationships are important to model? Once a representation has been set, experts may still disagree about the processes by which agents take actions, or the rules that their behavior is governed. Even supposing representation and processes are agreed upon, how the model should be parameterized can also be in contention.

The contention here is not over the precise value of a parameter in the model, but over a conceptual value. For example, is the relationship between two agents neutral or extremely hostile, not whether the value representing the relationship is assigned a value of -1.0 or -0.75. The latter is an issue of model calibration, not an issue over the actual state of affairs in the real world.

The usefulness of multi-agent models is that they allow assumptions about agents and their relationships to be tested, which is naturally important when it is already believed that a multi-agent model is an appropriate representation of a real-world social system. Unlike technical systems that often follow fundamental physical laws, explanations for behavior in social-systems is more nuanced, sometimes with multiple theories being offered to explain the same phenomenon. In these cases, multi-agent models can be used to systematically test different theories to see what the consequences are if a given theory is assumed. Figure 4 gives a graphical depiction of how multi-agent models can be used in conjunction with the predictions and assumptions of subject matter experts to iterate and improve the conceptual model of how a social system behaves. As an example, relationships among insurgents and insurgent groups have been cited as significant in influencing how effective these groups are. One position believes that strong alliances among the insurgent groups are detrimental to the state. Groups are able to coordinate with each other and share needed resources. Another position believes that conflict among insurgent groups is worse than if they were strongly aligned. Though the groups spend less time fighting the state, their attacks against each other demoralize citizens of the country, fueling sentiment against the state. The RTE, as a dynamic-network multi-agent model, is capable of representing these two different assumptions.

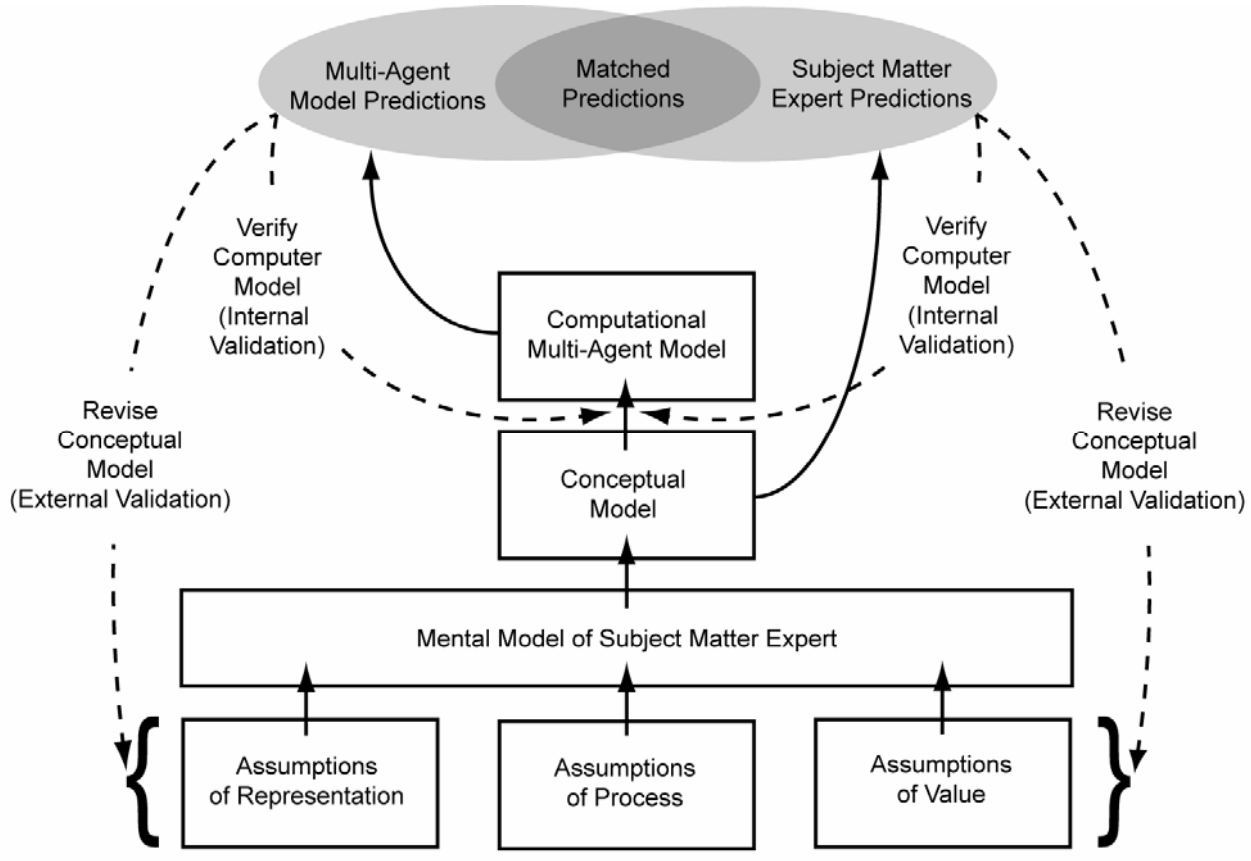

Figure 4. Multi-agent models can be used to refine a conceptual model by comparing the output from the multi-agent model to predictions from subject matter experts. Model predictions that do not match expert predictions may indicate that either the computer model or conceptual model needs to be revised.

4.3 Defining the operating domain

Large parameter spaces, characteristic of multi-agent models and especially of dynamic-network multi-agent models, permit a potentially large response surface. The challenge then becomes determining over which ranges and sets of parameters the model is capable of producing valid results. Defining the operation domain of multi-agent models is difficult for at least two reasons: many assumptions are implicit and the parameter space is very large. Most multi-agent models are programmed using a special modeling language or a more general language such as C++. Regardless of how it is implemented, the assumptions of the model are

embedded in the code. In contrast to a regression model where many of the assumptions of a regression model are known when the model is defined, the construction of a multi-agent model requires that assumptions be clearly stated outside of the model or that those interested in assumptions read code. Stating the assumptions outside of the code requires basically writing another document that is as long as or longer than the code itself. In writing up such assumptions, as in writing up a verbal theory, it is easy to inadvertently omit assumptions. Checking assumptions is difficult because it requires that the code or a suitable representation of the code be made available. In practice, however, making the code available may invalidate intellectual property claims.

In the case of the RTE, sensitivity analysis and replication helped to define how far out the current structure of the model can be pushed and still get answers that approximate the real-world. For the RTE, sensitivity analysis showed that the level of corruption before foreign aid starts to flow affects how quickly corruption increases when aid is received. However, it also showed that lower initial levels of corruption resulted in steeper increases in corruption over time, which is not what we might expect especially when corruption levels are very low. In these cases corruption should not increase or increase only marginally since we would expect uncorrupt governments not to skim from received aid. Sensitivity analysis of the RTE helped to set a boundary by demonstrating that the model may not operate correctly when the government is not corrupt.

Doing this analysis while informative, only covered a small portion of the response surface. This is a typical problem for multi-agent models. In general one needs to use data-farming techniques to fully evaluate the response surface. Even then, given the size of typical parameter spaces there may not be sufficient computer storage space for the results from a

comprehensive analysis. In addition, data-farming environments, such as that at the Mauii high performance computer center are not easily nor routinely available to most researchers and require that the model be written with certain web enabling features.

Using the model on a different set of data provided an additional method for defining the operating domain. The dynamics for corruption were developed using what was known about Indonesia. Though the RTE matched regional experts' beliefs about corruption in Indonesia and what is known through published reports, it may not appropriately replicate what is occurring in Thailand. Corruption in Indonesia is assumed to stem from aid flowing into the country whereas corruption in Thailand might also occur over land disputes between resort owners and the local population. The latter cause of corruption is not represented in the RTE, clearly marking a boundary for where the model can accurately represent corruption dynamics. In general, replication is a powerful and effective technique from a time and space constraint perspective. However, it requires both multiple subject matter experts and multiple modeling teams. Given tight resources, having such multiplicity is often viewed as redundant and a waste of resources by funding agents rather than as a necessary component of validation.

4.4 Applicable techniques of validation

Whether a true preference for statistical and quantitative techniques of validation exists, most of the literature on validation techniques covers these types. Kleijnen [20], in a discussion of validation and verification for simulation models, describes statistical techniques that can be used to validate models depending on the types of real data that are available. He notes three scenarios: 1) no data, 2) real data that can be compared to the model output, and 3) real data that can be used to ground the model in reality in addition to real data that can be compared to the model output. In all three cases, he places emphasis on the statistical and quantitative techniques

that can be used. We point this out as not a fault with the paper. Understanding what statistical techniques are available in a given situation of available data is important, but there is, nonetheless, an assumption that statistical techniques are the preferred method of establishing a case of model validity. In another well-known reference of simulation modeling, Law and Kelton [23] reserve a chapter to the validation of models. Brief mention is given to subject matter experts. Most time and explanation of method is given to statistical techniques of comparing distributions and time-series data.

This preference for statistical and quantitative techniques can put validating multi-agent models at a disadvantage, especially those whose purpose requires veridicality. As mentioned previously, multi-agent models can suffer from a lack of available data, both for grounding and to compare model output to what has happened in the real world. Consequently, the types of statistical validation techniques available are those with the lowest power, those that carry the least weight in convincing others that the model is "valid."

Practically speaking, most stakeholders are more comfortable with establishing a quantitative and objective case of model validity. However, it would be imprudent to dismiss the multi-agent modeling as not valuable for this reason alone. First, progress is continually being made in developing validation techniques and validation processes for multi-agent models. Second, multi-agent models allow theory to be developed about social processes that other methods of study are not amenable to doing. This is a theme that is discussed in several previous papers including this one. Thomsen et al. [35] point out that most research in the social sciences occurs at a single level of analysis, whereas research using multi-agent modeling nearly always happens at two levels. The bottom-up approach to using multi-agent models to study social systems permits researchers to link theory at the micro-level with theory at the macro-level [15].

Multi-agent models should not be considered an inferior method for studying social systems because familiar statistical techniques cannot be applied to validate all parts of the model. It does indicate that more work should be focused on techniques for validating multi-agent models. Some progress has been made in this area. Examples in the literature include Schreiber and Carley's method of calibrated grounding. Their validation approach combines two basic approaches: initializing a model with data that is grounded using a real-world system under study and to then to internally calibrate the model [31, 32]. In the initialization stage, the model uses empirical data that are grounded in the real-world system to set the model parameters. Internal calibration confirms that the model's internal processes and parameters are related to those of the real-world system.

The challenge for multi-agent modelers may first be to recognize the domains in which these types of models are most useful and then to develop techniques for establishing the validity of their models, while also noting that traditional techniques of simulation validation may not be an available method.

4.5 Theory integration and validation by parts

Most scientific theories have a *ceteris paribus* assumption, that the theory is valid while other factors are constant. It is a necessary assumption to have in order to rule out possible other relevant factors and determine the relative effect of a few factors of interest. This assumption, however, is at odds with multi-agent systems where part of the expressiveness of the model is derived from the interaction of multiple theories. The *ceteris paribus* assumption of a theory defines the parameter space over which the theory is thought to hold. While these conditions are constant the theory is believed to be true. Changing the conditions no longer guarantees that the theory is valid. The dynamics of simulation models may be drawn from multiple existing

theories, for example, on how agents are thought to make decisions about what actions to take each time step and how actions affect an agent's characteristics. If the dynamics of one theory violates the *ceteris paribus* assumption of another theory, then it is uncertain whether the dynamics of that particular theory is valid and consequently whether the model as a whole is valid. Determining whether the model is valid is an open question and answering that question is one we will begin to address shortly. First, however, we will compare the simulation method with comparable methods in the social sciences to see how the integration of multiple theories into a single model can complement other methods.

One of the hallmarks of the scientific method is that across experimental trials, all conditions are controlled for. In the social sciences this desire for controlled experimentation is a liability to how well the results generalize to other domains. In a highly controlled environment, as one might find in a laboratory setting, the results of the study may have a high degree of internal validity, but may lack sufficient external validity. Experimental results may lack external validity for a number of reasons. In this case we are concerned about the possibility that the conditions so carefully controlled for in the laboratory setting will, in the real world, vary naturally and may alter how people behave. Similarly, theories which rely on a *ceteris paribus* assumption may be invalid once they are coupled or integrated into a simulation model, as the model, just like in the real world, permit a violation of the assumption.

Not all research of social systems is so carefully controlled. Theory development through observation of a natural social system bypasses control over the conditions, but in doing so gains confidence that experimental trials are in the natural conditions. Methodologically, the method is challenged by having to construct a case for why all relevant variables were measured and

controlled and that no exogenous variables were driving any correlations observed in the data generated by the system.

Developing theory through simulation allows researchers to retain control over all of the variables in the system while allowing individual theories developed to interact with one another in the model of the larger, more complete system. Theory development of social systems need not be isolated to very controlled situations, but can instead be simultaneously be developed across multiple levels of analysis using multiple different theories to drive the dynamics. While an initial conceptual model may be based on individual theories that rely on *ceteris paribus* assumptions and can consequently be violated in the computational model, simulation modeling allows theories to be revised and developed across the whole system under study.

Simulation offers an alternative to social system analysis when natural experimentation and laboratory experiments cannot be used for practical or ethical reasons.

5 Discussion

Validation of multi-agent models used to study social systems face a number of challenges. Determining whether these models are useful despite these challenges is much less a question of whether they are useful, but more a question of *how* they can be useful. The criticisms together with their counterpoints help describe the role multi-agent models are likely to best serve in studying social systems. The criticisms serve to constrain how the models should be used while the counterpoints push against those constraints and expand the method's potential usefulness. The reader can use Table 3 as a quick summary of the discussion in section 4. In this section, we go beyond the list of criticisms and counterpoints to give a more holistic description of how multi-agent models are used to study social systems despite their validation challenges.

Consider once again the different types of simulation model validation discussed earlier: conceptual model validation, operational validation, and data validation. The criticisms outlined in this paper affect the extent to which each of these components can be validated. In turn, each of the different types of model validation gives different types of credibility to a model.

Unobservable data are easily the most significant factor affecting a multi-agent model's ability to be operationally validated. Typically, simulation models are operationally validated by measuring how accurate a model's predictions are with respect to the real-world system given the model's stated purpose. If data from the real-world system are not available, then it is impossible to make an objective comparison between the model output and real-world system. Consequently, among stakeholders, multi-agent models will often lack a sufficient degree of credibility to ask questions about future conditions or favorable policy changes for specific social systems.

Using multi-agent models to study theoretical aspects of a social system still makes sense, even if data from the real world are unobservable. They are of particular use when the theory connecting agent-level behavior to system-wide, emergent behavior is not well-understood [35]. Multi-agent models allow researchers to implement conceptual models of agents, their behavior, and their relationships with others into a simulation model. Running the simulation produces system-wide behavior that is derived from these agent-level specifications. When theory about certain types of behavior is not well-understood or when there are competing explanations for the same phenomena, then constructing multiple versions of the conceptual and computerized models, each version representing competing theories, allow consequences of competing explanations to be investigated at the system and agent levels.

Operational validation is also affected by the lack of quantitative techniques to compare model predictions to the real world. Often this is because the data that are sought from the real-world are not observable. Other times the nature of the data coming from the real-world social-system violates assumptions of normality or other assumptions of statistical tests that are used to compare distributions of data [26]. Alternate techniques are occasionally developed to handle some of the questions one might use a multi-agent model to help answer. For example, multi-agent models of social systems are sometimes used to investigate the relationships that exist between the agents as a function of the characteristics of the agents. Ordinary least squares regression cannot be used for the analysis because the observations are not independent. Instead, the multiple regression quadratic assignment procedure was developed to handle this type of relational data [22]. Though procedures may be developed, missing data will still prevent models from being quantitatively compared to the real world.

When objective techniques of model validation are not possible, subject matter experts are often utilized to provide face validation. Subject matter experts are known to disagree with each other and thus their use can pose an additional challenge to both operational and conceptual validation. If, however, theory is not yet well-developed then simulation can provide a platform by which different subject matter experts or competing theories can be implemented into a computerized model. Using simulation models to test a variety competing theories is consistent with Davis et al.'s [15] proposition that simulation models are particularly useful in developing theory that is "rough" and "not yet logically precise and comprehensive."

While missing empirical data can mar a simulation model's validation process, it is a simulation model that can help shed light on a system's properties and dynamics in the face of missing data. In social systems, interactions between entities can be unobservable, especially

interactions that are purposefully obscured from observation such as terrorist activities or sub-national groups fighting the state. The constructs of a multi-agent model allow interactions among system entities that are normally unobservable in the real world to be simulated. Simulation allows synthetic data to be created under a variety of conditions that are otherwise difficult to experiment with or observe in the real world. This sort of systematic experimentation, or sensitivity analysis, can be used to understand how a system might behave given a specific formalized model of how the agents behave. Again, in cases where it is not well-understood how the agents behave competing models can be used to investigate how various assumptions affect a system's behavior.

Criticism	Counterpoint
A limited amount of data are available to ground and validate model.	Simulation models can be leveraged to explore what relations are likely to exist if more comprehensive real-world data were available. The models do this by filling in information in a way that is consistent with known data and processes. Example real-world systems include military conflict and disaster environments.
Subject matter experts can disagree.	Simulation permits the expert mental models to be formalized, tested, and revised as needed. Further, they can be used to ask how different the outcomes would be if an expert *A* rather than expert *B* were right. Hence, the formal computer model acts as a tool to guide and mature mental models. This can be particularly useful when our understanding of how a social system behaves is not well-understood.
Defining the operating domain is difficult because of the large parameter spaces.	A descriptive model, one with a large parameter space, allows one to explore factors that are suspected, but not yet known to have an influence on the target system. Mental models can be refined, including simplification, as the model is used to develop a better understanding of the system.
Large input parameter spaces are capable of explaining nearly everything.	More parameters let you effectively capture more of a domain with higher fidelity. Extraneous parameters can be removed as knowledge of the target system matures.
Traditional quantitative techniques for validation are not applicable.	Validation techniques need to match the purpose for which the model is being used. As multi-agent models are good at connecting theories of individual behavior to emergent behavior of the system, methods like sensitvity analysis, which need less data, are useful in determining how important factors at the individual level influence the emergent, system behavior.
Ceteris paribus assumption of individual theories may be violated when they are integrated.	Theories of individual behavior can be integrated in multi-agent models while experimental control is maintained.

Table 3. Each criticism is accompanied by its counterpoint to show how multi-agent models are still useful to study social systems.

6 Conclusion

Use of multi-agent models to study social systems has garnered criticism because of the challenges that exist in validating the models. In this paper we described the space in which multi-agent models can be useful to study social systems. We started by synthesizing the literature on simulation and multi-agent model validation, creating a framework for how models ought to be validated based on a model's purpose. Next, we presented important challenges to validating multi-agent models of social systems and for each challenge we presented reasons why these types of models can still advance our understanding of a social system. We used a

dynamic-network multi-agent model called the Regional Threat Evaluator to help us demonstrate these points.

By presenting the challenges and their corresponding counterpoints together, we were able to present a balanced picture for what role these types of models have in advancing our understanding of social systems. We believe that the models may be most useful when 1) the connection between micro-behaviors and macro-behaviors are not well-understood and 2) when data collection from the real-world system is prohibitively expensive in terms of time or money or if it puts human lives at risk.

Acknowledgements

This work was supported in part by the AFOSR under a MURI with GMU (600322) on adversarial modeling, the National Science Foundation (NSF DMS-0437239), the Office of Naval Research (ONR N00014-06-1-0104), and the National Science Foundation IGERT in Computational Analysis of Social and Organizational Systems (NSF DGE-9972762). At the conceptual level, RTE is based, on an earlier model called Acumen, developed under DARPA (FA8650-05-C-7222). Additional support was provided by CASOS – the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University (http://www.casos.cs.cmu.edu/). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Air Force Office of Sponsored Research, the National Science Foundation, DARPA, the Office of Naval Research, or the U.S. government. We would also like to thank Michael P. O'Connor for his assistance in developing the RTE as well as the anonymous reviewers for their insightful comments and suggestions.

References

[1] G. T. Allison and P. Zelikow, Essence of Decision: Explaining the Cuban Missile Crisis (Longman, Reading, 1999).
[2] R. Axelrod, Advancing the Art of Simulation in the Social Sciences, Japanese Journal for Management Information Systems, 12 (2003)
[3] O. Balci and R. G. Sargent, A Methodology for Cost-Risk Analysis in the Statistical Validation of Simulation Models, Communications of the ACM: Simulation Modeling and Statistical Computing, 24 (1981) 190-197.
[4] J. Banks, D. Gerstein, and S. P. Searles, "Modeling Processes, Validation, and Verification of Complex Simulations: A Survey," in: Proc. SCS Simulators Conference, Vol. 19 (The Society for Computer Simulation, Orlando, 1987).
[5] M. Bergkvist, P. Davidsson, J. A. Persson, and L. Ramstedt, "A Hybrid Micro-Simulator for Determining the Effects of Governmental Control Policies on Transport Chains," in: Proc. Joint Workshop on Multi-Agent and Multi-Agent Based Simulation, Vol. 3415 / 2005 (Springer Berlin / Heidelberg, New York, 2005).
[6] J. H. Bigelow and P. K. Davis, "Implications for Model Validation of Multiresolution, Multiperspective Modeling (MRMPM) and Exploratory Analysis," RAND, Santa Monica, CA MR-1750-AF, 2003.
[7] P. M. Blau, Inequality and Heterogeneity: A Primitive Theory of Social Structure (The Free Press of Macmillan Co., New York, 1977).
[8] R. M. Burton, Computational Laboratories for Organization Science: Questions, Validity and Docking, Computational & Mathematical Organization Theory, 9 (2003) 91-108.
[9] R. M. Burton and B. Obel, The validity of computational models in organization science: From model realism to purpose of the model, Computational & Mathematical Organization Theory, 1 (1995) 57-71.
[10] K. M. Carley, A Theory of Group Stability, American Sociological Review, 56 (1991) 331-354.
[11] K. M. Carley, Computational organizational science and organizational engineering, Simulation Modelling Practice and Theory, 10 (2002) 253-269.
[12] K. M. Carley, D. B. Fridsma, E. Casman, A. Yahja, N. Altman, L.-C. Chen, B. Kaminsky, and D. Nave, BioWar: Scalable Agent-Based Model of Bioattacks, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 36 (2006)
[13] P. Davidsson, "Multi Agent Based Simulation: Beyond Social Simulation," in: Proc. Multi-Agent-Based Simulation (MABS 2000), Lecture Notes in Computer Science, Vol. 1979 (Springer-Verlag, Boston, 2000).
[14] P. Davidsson, Agent Based Social Simulation: A Computer Science View, Journal of Artificial Societies and Social Simulation, 5 (2002)
[15] J. P. Davis, K. M. Eisenhardt, and C. B. Bingham, Developing Theory Through Simulation Methods, Academy of Management Review, 32 (2007) 480 - 499.
[16] J. Epstein, "Modeling Civil Violence: An Agent-Based Computational Approach," Proceedings of the National Academy of Sciences 99 suppl. 3:7243-7250, 2002.
[17] L. Festinger, A Theory of Social Comparison Processes, Human Relations, 7 (1954) 117-140.
[18] N. E. Friedkin, A Structural Theory of Social Influence (Cambridge University Press, New York, 1998).

[19] D. L. Horowitz, Ethnic Groups in Conflict (University of California Press, Berkeley, 1985).
[20] J. P. C. Kleijnen, "Validation of models: statistical techniques and data availability," in: Proc. 1999 Winter Simulation Conference, Vol., Phoenix, 1999).
[21] G. B. Kleindorfer, L. O'Neill, and R. Ganeshan, Validation in Simulation: Various Positions in the Philosophy of Science, Management Science, 44 (1998) 1087 - 1099.
[22] D. Krackhardt, Predicting with networks: Nonparametric multiple regression analysis of dyadic data, Social Networks, 10 (1988) 359-382.
[23] A. M. Law and D. W. Kelton, Simulation Modeling and Analysis (McGraw-Hill, New York, 1999).
[24] E. J. Mishan, Economics for Social Decisions: Elements of Cost-Benefit Analysis (Praeger, New York, 1973).
[25] M. G. Morgan and M. Henrion, Uncertainty: A Guide to Dealing with Uncertainty In Quantitative Risk and Policy Analysis (Cambridge University Press, Cambridge, 1990).
[26] S. Moss and B. Edmonds, Sociology and Simulation: Statistical and Qualitative Cross-Validation, American Journal of Sociology, 110 (2005) 1095-1131.
[27] Y. Murakami, K. Minami, T. Kawasoe, and T. Ishida, "Multi-agent simulation for crisis management," in: Proc. IEEE Workshop on Knowledge Media Networking (KMN'02), Vol. (IEEE, Kyoto, 2002).
[28] J. Pfeffer and G. R. Salancik, The external control of organizations: A resource dependence perspective (Harper & Row, New York, 1978).
[29] V. Roske, Opening up military analysis: Exploring beyond the boundaries, Phalanx, 35 (2002) 1-8.
[30] R. G. Sargent, "Validation and Verification of Simulation Models," in: Proc. 1992 Winter Simulation Conference, Vol. (IEEE, Piscataway, 1992).
[31] C. Schreiber and K. M. Carley, Going Beyond the Data: Empirical Validation Leading to Grounded Theory, Computational & Mathematical Organization Theory, 10 (2004) 155 - 164.
[32] C. Schreiber and K. M. Carley, "Agent Interactions in Construct: An Empirical Validation using Calibrated Grounding," in: Proc. 2007 BRIMS Conference, Vol., Norfolk, VA, 2007).
[33] H. A. Simon, Models of Bounded Rationality (MIT Press, Cambridge, MA, 1982).
[34] N. Stansbury, "Exposing the foundations of corruption in construction," in *Global Corruption Report 2005*. Berlin: Transparency International, 2005, 36-40.
[35] J. Thomsen, R. E. Levitt, J. C. Kunz, C. I. Nass, and D. B. Fridsma, A Trajectory for Validating Computational Emulation Models of Organizations, Computational & Mathematical Organization Theory, 5 (1999) 385 - 401.
[36] Transparency International, "Transparency International Corruption Perceptions Index 2005." Berlin, Germany, 2005.