Verification and Validation through Replication: A Case Study Using Axelrod and Hammond's Ethnocentrism Model

William Rand
Northwestern University
wrand@northwestern.edu

Uri Wilensky
Northwestern University
uri@northwestern.edu

Abstract

Recent years have seen a proliferation of agent-based models (ABMs), but with the exception of a few "classic" models, most of these models have never been replicated. We argue that replication has even greater benefits when applied to computational models than when applied to physical experiments. Replication affects model verification, in that it aids in determining if the implemented model reflects the conceptual model. It affects model validation, since a replication of a conceptual model may change the output from an implemented model and thus alter the correspondence between the model and the real world. Replication also affects validation by forcing the model developer and replicator to re-examine assumptions made in the original model. In addition replication fosters shared understanding of the details of modeling decisions within the research community. To facilitate the practice of replication, we argue for the creation of standards for both how to replicate models and how to evaluate the replication. In this paper, we present a case study of our attempt to replicate an ABM developed by Axelrod and Hammond. We detail our effort to replicate that model and the challenges that arose in recreating the model and in determining if the replication was successful.

Contact:
Dr. William Rand
Northwestern Institute on Complex Systems
Northwestern University
600 Foster Street
Evanston, IL 60208

Tel: 1-847-491-5734
Fax: 1-847-467-6938
Email: wrand@northwestern.edu

Key Words: verification, validation, replication, methods, ethnocentrism, social science

Acknowledgement: We would like to thank the Northwestern Institute on Complex Systems for their support during this work.

Verification and Validation through Replication

William Rand and Uri Wilensky

One of the foundational components of the scientific method is the idea of reproducibility (Popper 1959). In order for an experiment to be considered valid it must be replicated. This process begins with the scientists who originally performed the experiment publishing the details of the experiment. This description of the experiment is then read by another group of scientists who carry out the experiment, and ascertain whether the results of the new experiment are similar to the original experiment. If the results are similar enough then the experiment has been replicated. This process validates the fact that the experiment was not dependent on local conditions, and that the written description of the experiment satisfactorily records the knowledge gained through the experiment.

Agent-based modeling (ABM) is a new form of scientific experimentation. If ABM is to become an integral part of the scientific practice then it must develop standards of practice similar to those that have been developed in other disciplines. Thousands of agent-based models (ABMs) have been published in the last few decades (Epstein and Axtell 1996; Axelrod 1997), but with some notable exceptions very few of these models have been reproduced.

In fact, replication is more important within the realm of computational models than it is within the realm of physical experiments. Replicating a physical experiment proves that the original experiment was not a contingent event, and makes the experiment useable as a tool for the model replicator in their own research. Replicating a computation model has these benefits, but in addition replication of a computational model improves model verification, re-examines model validation, and facilitates a common understanding between modelers.

Despite these benefits the replication of computational models occurs infrequently. Part of the reason for this is that there is not a lot written about the replication attempts. Knowledge of how to replicate, and how to validate the results of a replication are not part of the common knowledge and education in the field of ABM. This in turn impedes others from carrying out replications. Building up a body of cases of replication would enable researchers to take a step back and extract general principles regarding the replication process. This paper embraces this agenda. We begin by describing the benefits and terminology of scientific replication. From there we move on to describe a particular case study that we undertook to replicate the Ethnocentrism model created by Axelrod and Hammond (Axelrod and Hammond 2003). We conclude by summarizing our experience and noting future directions of research for the methodology of model replication.

The Benefits of Replication

A successful replication of a physical experiment advances scientific knowledge because it demonstrates that the model outputs can be repeatedly generated and thus the original results were not an exceptional case. In addition, after a replication, the knowledge and data embodied by that experiment can be utilized by the model replicator as a tool to advance their own research agenda. Replication of a computational model gains similar benefits. By successfully replicating a computational model the replicator shows that the results of the original model were not a rare occurrence. Also, by replicating a computational model, the replicator has now built a new tool (i.e. the model) that can be used for the replicator's research.

However, in addition to these benefits of physical experiment replication, the process of replicating a computational model can aid the scientific community in other ways. Replication of computational models aids in the processes of model verification, model validation, and in developing a shared understanding of modeling. *Model verification* is the determination of whether the implemented model corresponds to the conceptual model. *Model validation* is the determination of whether the implemented model corresponds to and explains some phenomenon in the real world. The shared understanding that is gained through replication is the creation of a set of terms and best practices that can be utilized by model developers to communicate about their models.

Replication aids in the model verification process because if two distinct implementations of a conceptual model are able to produce the same results then that supports the hypothesis that the original model correctly implements the conceptual model. During the model replication process if differences between the original model and the replicated model are discovered, it is not guaranteed that the replicated model needs to be fixed. It may also be the case that the original model is not a verified, correct implementation of the conceptual model.

Replication aids in model validation because validation is a process that determines a correspondence between the outputs from an implemented model and measures of the real world. As a result, if the replicated model produces different outputs than the original model then that raises questions as to which outputs correspond more to real world data. If the replicated model's outputs are closer to the real world data that suggests that the replicated model is more valid than the original model. Moreover model replication may call into question the details of the

original model and how they correspond to the real world. If this is the case then it may be determined that the original model is not valid because there is not sufficient correspondence between the original model and the real world. In addition, replication forces the model replicator to examine the face validity of the original model by re-evaluating the original mapping between the real world and the conceptual model, since the replicator must eventually re-implement those same concepts.

Finally replication creates a language of modeling. By developing a suite of "best practices" (Jones 2000) with respect to replication, the field of ABM as a whole will advance. In much the same way that statistics has a shared understanding of what is meant by "mean" and "standard deviation" and when to apply statistical tests, replication of experiments over time might define terms of art like "shuffled list" and "time-step" and may develop tests for comparing ABMs like "distributional equivalence" and "relational alignment."

Since ABM is a relatively new methodology most research has aimed at showing off the power of ABM and not at developing core guidelines for the utilization of ABM. However there are exceptions. In fact we view this paper as contributing to the recent conversation about how to replicate agent-based models and why it is important. For instance, Hales, Rouchier and Edmonds held a workshop on model-to-model analysis (Hales, Rouchier et al. 2003). In addition, though it is still not widespread there have been other attempts to replicate ABMs (Axtell, Axelrod et al. 1996; Cohen, Axelrod et al. 1998; Fogel, Chellapilla et al. 1999; North and Macal 2002; Edmonds and Hales 2003; Rouchier 2003; Galan and Iquierdo 2005).

Terminology within Scientific Replication

Before we embark on the main thrust of the paper it will be useful to establish some definitions. Terms such as *model, conceptual model,* and *replication* are often used loosely, and different papers use the terms in different connotations. For our purposes in this paper, we will establish the following definitions. By the term *model* we mean a simplified representation of a real-world process or object. A *conceptual model* is a textual description of a real-world process or object. An *implementation* of a conceptual model (also known as an *implemented model*) is a formalization of that model into a computational format such that the model can be given input and generates output. In this paper we are primarily concerned with agent-based implementations of models. *Agent-based models* (ABMs) utilize numerous autonomous, heterogeneous agents that follow simple rules as their basic ontological units. However, many of the issues that we will discuss apply equally well to other forms of implemented models, such as systems dynamics models or cellular automata models.

Though many conceptions of replication may exist, for the purposes of this paper we will define *replication* as the implementation (*replicated model*) by one scientist or group of scientists (*model replicators* or *replicators*) of a conceptual model described and already implemented (*original model*) by a scientist or group of scientists at a previous time (*model builders* or *builders*). The process of the implementation of the replicated model must differ in some way from the original model building process. For instance, it could be by a different group of individuals or on a different software platform. Since replication refers to the creation of a new implementation of a conceptual model based on the previous results of an implementation, *original model* and *replicated model* always refer to implemented models. Moreover since this paper is concerned with replicated models any reference to a researcher's model is a reference to their model implementation (e.g. Axelrod-Hammond model, Rand-Wilensky model).

A *successful or valid replication* is one in which the replicators are able to establish that the replicated model creates outputs similar to the outputs of the original model. Different targets exist for the level of similarity between model outputs. Axelrod (Axelrod 1997) examined exactly this question. He developed three criteria for a replication experiment. The first criterion, "numerical identity" is difficult since it entails showing that the original and replicated model produce numerically the exact same results. One of the reasons this is difficult is that it has been shown that running the same program on the same machine with the same parameters does not guarantee numerical identity (Belding 2000). The second criterion is "distributional equivalence." Here the goal is to show that the two implemented models are statistically indistinguishable from each other. It should be noted that it is impossible to prove that two models are distributionally equivalent due to the problem of induction and the stochastic nature of these models. However, one can show that given the current data there is no proof that the models are not distributionally equivalent (Axtell, Axelrod et al. 1996; Edmonds and Hales 2003). The final criterion is "relational alignment." Relational alignment exists if the results of the two implemented models show similar relationships between input and output variables, e.g. if input variable x is increased in both models then if output variable y increases in the original model it should also increase in the replicated model.

For the replication effort reported in this paper, we decided to attempt to achieve distributional equivalence between the replicated and original models. To determine distributional equivalence we must show that output data from the replicated model are statistically indistinguishable from the output data from the original model. However,

because ABMs can produce large amounts of data, it is necessary to establish focal measures for a replication effort. Instead of trying to show distributional equivalence for every output variable, we choose a few measures and demonstrate distributional equivalence for those metrics alone. For this project we identified three such measures, which are described later in this paper. To detail this effort we will begin by explaining the original conceptual model, then describe the original model, and finally narrate the replication of the original model.

The Original Model

At Northwestern University's 2003 conference on Complex Systems, Axelrod presented a conceptual description of an ABM that explored the evolution of ethnocentrism. The results of the Ethnocentrism model that were presented at the NICO conference were from an implementation by Hammond in coordination with Axelrod in the Ascape ABM toolkit (Parker 2000). This model has agents that immigrate into a world, interact with each other on the basis of their type and a strategy, give birth to new agents, and die. In the original implemented version of the Ethnocentrism model, "immigration", "interact", "birth" and "death" were described as rules in the Ascape environment. These rules correspond roughly to the descriptions of events in the conceptual model.

The most important general result of this model was to show that "ethnocentric" behavior arose under a wide variety of circumstances. In this case ethnocentric behavior was demonstrated by having a high percentage of ethnocentric genotypes and a large amount of behavior consistent with ethnocentrism. Moreover it was shown that cooperation levels remained high, indicating that many individuals were meeting and cooperating with individuals of their own color. These qualitative results were described numerically through three measures: percentage of cooperation (COOP), percentage of ethnocentric genotypes (CD_GENO), and percentage of behavior where individuals cooperated with someone of their same type or defected again some of a different type (CONSIS_E).

The particular results were important because we were aiming to establish distributional equivalence. All of the measures were made in the last 100 time-steps of a 2000 time-step run. The percentage of ethnocentric genotypes (CD_GENO) was 76%. In addition 74% of all interactions resulted in cooperation (COOP) and the percentage of interactions that were consistent with ethnocentrism was 88% (CONSIS_E). After developing the original model implementation, Axelrod and Hammond wrote several papers on the subject and published the results (Axelrod and Hammond 2003; Hammond and Axelrod 2005). They also made available on a website the original source code for the implemented model and the data that they had collected for their publications.

The Replication Experience and its Validation

Wilensky (Wilensky 1999) first replicated the Axelrod-Hammond model in the NetLogo language on the basis of the talk that Axelrod gave at Northwestern University in 2003. Due to the ease of use of NetLogo, Wilensky was able to replicate this model during Axelrod's talk and Wilensky showed this first version to Axelrod soon afterward.

However, the oral description of the conceptual model had some ambiguities in it. Looking at Axelrod and Hammond's description in the original paper, we see the following text: "2. Each agent receives a initial value of 12% as its Potential To Reproduce (PTR). " (Axelrod and Hammond 2003) This would seem to suggest the potential to reproduce is reset to its base level at each time period. The order of events was clear in the paper, but not in the original verbal description. Did it occur after a reproduction event? Did it occur at the beginning of a model step? Questions like these were clarified via communications between Wilensky and Axelrod. A "final" version of the replicated model was sent to Axelrod to examine. At this point Wilensky thought that the NetLogo code for the model captured the rules of Axelrod's conceptual model but he asked Axelrod to determine if the replication was indeed correct. Axelrod gave the Wilensky model to Rader who ran the same experiments that were originally run on the Axelrod-Hammond implemented model on the Wilensky model and compared the results. This was done in order to determine whether the replicated model created results "similar to" the original model.

Rader, in fact, discovered that there were differences between the two model implementations that indicated that statistical equivalence had not been achieved (Rader 2005). The Wilensky model resulted in less ethnocentrism and more cooperation than the Axelrod-Hammond model. Given the simplicity of the ethnocentrism conceptual model, this was surprising. Wilensky asked Rand to investigate the differences between the two model implementations with the goals of (a) understanding the mechanisms that led to the divergence in the results, (b) determining which (or whether both) of the two model implementations were externally valid, and (c) modifying the Wilensky model to achieve statistical equivalence with the Axelrod-Hammond model.

Rand noticed that the basic method of agent interaction appeared to be different in the Wilensky model than in the descriptions of the Axelrod-Hammond model. Wilensky had implemented the interaction as a two-way simultaneous prisoner's dilemma where each agent played against each neighbor and it was determined immediately whether both agents would cooperate or defect. This was similar to some previous models of the prisoner's

dilemma, but was different than the way the interaction was described in Axelrod and Hammond's ethnocentrism papers. This was a difference between the two model implementations, and during conversations between Rand and Hammond they both agreed that this aspect of the Wilensky model needed to be modified in order to have the Wilensky model replicate the Axelrod-Hammond model.

As a result, Rand modified the Wilensky model to utilize the method of interaction the Axelrod-Hammond model. However, after these changes the new implemented model still produced different results from the Axelrod-Hammond model. In fact, as Rand and Wilensky discussed these different results, they realized that the new Wilensky-Rand implementation of the model produced the same outputs as the original Wilensky version, despite the apparent difference in the method of interaction. It turned out whether agents acted independently and at different times from each other or concurrently and at the same time, the results were identical. As long as each agent interacted with each of its extant neighbors once, and the potential to reproduce was updated correctly then the net effect of all interactions was the same. This was in part due to the commutative nature of the interaction event.

Rand then went back to investigate Rader's results. In order to determine if the Wilensky model was a successful replication of the Axelrod-Hammond model, Rader had introduced new measures into the Wilensky model that corresponded with measures in the Axelrod-Hammond model. Rand corresponded with Rader and was able to obtain a copy of the Wilensky model with the new measures introduced by Rader that had been utilized in determining the success of the replication. While examining the Rader version of the Wilensky model, Rand realized that Rader had added several methods to the model implementation in order to produce exactly the same measures as the original Axelrod-Hammond model. Upon examination of some of these measures it became clear that Rader had misinterpreted, based upon the variable names, some of the results that were already being calculated by the Wilensky model. In particular, Rader had assumed that measures of genotypes were measures of interactions. For instance Rader interpreted "cc-count" as the number of cooperate-cooperate events that had occurred, not as the number of altruistic agents that existed, which was its actual definition. This confusion is understandable given the differences in the implementation of the interaction phase. Since these measures were the basis for the claim that the Wilensky model exhibited more cooperation and less ethnocentrism than the Axelrod-Hammond model, it seemed possible that this was the cause of those differences.

Thus Rand modified the model to output correctly these measures. This new implementation did generate different results than the Rader version of the Wilensky model. However the new Wilensky-Rand model was still statistically distinguishable from the Axelrod-Hammond model. Besides having different averages, the variance of the measures of interest was much higher in the Wilensky-Rand model than it was in the Axelrod-Hammond model.

To determine why the results from the model implementations were still different, Wilensky and Rand again examined the NetLogo source code for the model. Wilensky and Rand realized that the Axelrod-Hammond model utilized a different order of events. In the Axelrod-Hammond model the order was immigrate, interact, birth, death. In the Wilensky-Rand model the order was immigrate, birth, death, interact. This appeared, at first, to be an unimportant difference between the two implemented models, but as Wilensky and Rand discussed this difference in ordering, they realized that there might be an effect on the reproduction rate of the population as a whole. With the birth event happening after immigration but before the interaction event, new immigrants would not have a chance to raise their reproduction rate before the birth event resulting in fewer births and thus a smaller population. In fact, through experimental analysis, Rand showed that the net result of this ordering of events was fewer individuals in the population. Since there were fewer individuals in the population, the variance of the measures was higher due to the effect of small numbers. Aligning the order of events lowered the variance of the Wilensky-Rand model and the replicated model might now be in statistical agreement with the Axelrod-Hammond model.

Thus the Wilensky-Rand model was modified once again. The result of this modification was that the variance of the measures of interest in the Wilensky-Rand model decreased, but it was still higher than it was in the Axelrod-Hammond model. Upon reexamination, another difference between the two implementations was discovered. In the Axelrod-Hammond model the list of agents was shuffled before the reproduction event took place. In the Wilensky-Rand model the list was in an unshuffled but arbitrary order. When the list was unshuffled some agents would be repeatedly preferentially selected. Since the space for reproduction is a scarce resource, the preferentially select agents would have a greater chance of reproducing. This would cause a bias throughout the Wilensky-Rand model toward these preferentially selected agents. If these agents were, for example, a particular color then that could systematically bias the results. This would in turn increase the variance in the measures of interest. Rand modified the Wilensky-Rand model to account for the differences in agent ordering. When the results were generated, this version of the Wilensky-Rand model produced results that were statistically indistinguishable from the Axelrod-Hammond model on the three main measures. Distributional equivalence had finally been achieved.

Conclusion

As this experience illustrates, model replication is not as straightforward a process as it seems. There are many careful considerations that must be addressed by both model replicaters and model developers. Model replication is a critical component of the scientific process. In addition computational model replication is even more beneficial to the scientific community than the physical experiment replication. Since computational model replication can affect the model verification process, it can alter our view of conceptual models. The effect of model replication on the validation process directly increases our knowledge about the real world. In the replication experiment detailed here, differences were discovered between the two models, and the replicated model had to be modified to produce the original results. As a result of this successful replication, the original Axelrod-Hammond model now has additional evidence that it is a correctly verified model. Moreover, the model now has a higher level of validity since the results have been borne out by two separate implementations. However, the process that was required to determine that the replication was successful was complicated and involved some unforeseen problems. By accumulating best practices of replication, the ABM community can start to place ABM on a firmer footing. This paper contributes to the recent conversation on model replication; specifically this paper articulates the benefits of replication with respect to verification and validation, presents a terminology for discussing replication, and illustrates these results with a case study. However, it is only careful consideration of these issues by both model replicators and developers that will lead to the widespread replication of ABMs as standard methodology.

References

Axelrod, R. (1997). Advancing the Art of Simulation in the Social Sciences. Simulating Social Phenomena. R. Conte, R. Hegelsmann and P. Terna. Berlin, Springer-Verlag**:** 21-40.

Axelrod, R. and R. A. Hammond (2003). The Evolution of Ethnocentric Behavior. Midwest Political Science Convention, Chicago, IL.

Axtell, R., R. Axelrod, et al. (1996). "Aligning Simulation Models: A Case Study and Results." Computational and Mathematical Organization Theory **1**: 123-141.

Belding, T. C. (2000). Numerical Replication of Computer Simulations: Some Pitfalls and How To Avoid them, University of Michigan's Center for the Study of Complex Systems.

Cohen, M., R. Axelrod, et al. (1998). "CAR Project: Replication of Eight "Social Science" Simulation Models." from http://www.cscs.umich.edu/Software/CAR-replications.html.

Edmonds, B. and D. Hales (2003). "Replication, Replication and Replication: Some Hard Lessons from Model Alignment." Journal of Artificial Societies and Social Simulation **6**(4).

Epstein, J. and R. Axtell (1996). Growing Artificial Societies: Social Science from the Bottom Up. Cambridge, MA, MIT Press.

Fogel, D. B., K. Chellapilla, et al. (1999). "Inductive Reasoning and Bounded Rationality Reconsidered." IEEE Transactions on Evolutionary Computation **3**(2): 142-146.

Galan, J. M. and L. R. Iquierdo (2005). "Appearances Can Be Deceiving: Lessons Learned Re-Implementing Axelrod's 'Evolutionary Approach to Norms'." Journal of Artificial Societies and Social Simulation **8**(3).

Hales, D., J. Rouchier, et al. (2003). "Model-to-Model Analysis." Journal of Artificial Societies and Social Simulation **6**(4).

Hammond, R. A. and R. Axelrod (2005). The evolution of ethnocentrism, University of Michigan.

Jones, C. (2000). Software assessments, benchmarks, and best practices. Boston, MA, Addison-Wesley Longman Publishing Co., Inc.

North, M. J. and C. M. Macal (2002). The Beer Dock: Three and a Half Implementations of the Beer Distribution Game. Swarmfest.

Parker, M. (2000). Ascape, The Brookings Institution.

Popper, K. R. (1959). The Logic of Scientific Discovery. New York, Harper & Row.

Rader, E. (2005). Ethnocentrism Model Validation, Personal Communication.

Rouchier, J. (2003). "Re-implementation of a multi-agent model aimed at sustaining experimental economic research: The case of simulations with emerging speculation." Journal of Artificial Societies and Social Simulation **6**(4).

Wilensky, U. (1999). NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.