# *Building Better Crime Simulations: Systematic replication and the introduction of incremental complexity.*

Paper to be submitted to: Journal of Experimental Criminology "Simulated Experiments in Criminology and Criminal Justice"

Michael Townsley[1] and Daniel J Birks[2]
UCL Jill Dando Institute of Crime Science

1.
School of Criminology and Criminal Justice
Mt Gravatt Campus
Griffith University
Qld 4111
Australia
E: m.townsley@griffith.edu.au
T: +61 7 3735 1025
F: +61 7 3735 5608

2.
UCL Jill Dando Institute of Crime Science
University College London
Second Floor Brook House 2 - 16 Torrington Place
London WC1E 7HN
Phone +44 (0)20 7679 0818
Fax +44 (0)20 7679 0828
Email d.birks@ucl.ac.uk

## Abstract

Computer simulation models have changed the ways in which researchers are able to observe and study social phenomena such as crime. The ability of researchers to replicate the work of others is fundamental to a cumulative science, yet this rarely occurs in computer simulations. In this paper, we argue that for computer simulations to be seen as a legitimate methodology in social science and for new knowledge to be generated, serious consideration needs to be given to how simulations could or should be replicated. We develop the concept of *systematic replication*, a method for developing simulation experiments which move towards generalisable inference that is *directed*, *explicit* and incorporates complexity *incrementally*. Finally, we outline how the discrete parts of this process might be carried out in practice using a simple simulation model.

## Key words

Simulation modelling, crime, validation, replication, systematic

# Introduction

John Eck (2007) likened the numerous yet somewhat limited interjections of other disciplines into criminology to the attacks on Europe by Barbarians during the Middle Ages. The Barbarians would land, pillage a few towns and villages, burn down a few buildings and, whilst in-situ, strike fear into the hearts of the indigenous population. Ultimately though, after plundering the easier targets, they would tire, retreat and leave an area to recover, ultimately leaving it unchanged by the attack.

Over the past few decades, advances in a number of disciplines have lead several such invasion forces to the shores of criminology. However, only a handful actually settled and provided more than a fleeting diversion to the general populace. The recent emergence of computer simulation models[1], whose origins lie in the disciplines of computer science and artificial intelligence, could hail the arrival of another invading force, which may empower those interested in crime with an additional tool aimed at increasing understanding. On the other hand, it may prove to be a temporary sortie that criminologists simply need to put up with until interest wanes.

Recent research (see Liu and Eck (2008) for the most complete and up to date treatment of the potential for simulation models in criminology and crime reduction) has conceived, discussed and presented several criminological applications of simulation and described how the methodology potentially offers researchers the ability to observe and study the crime phenomenon in ways that were not previously possible for logistic, moral or economic reasons. Through the implementation of *in silico* experiments, which, unlike traditional analyses, allow for perfect observation and measurement, it aims to produce models of behaviour or systems providing a framework around which existing criminological theories can be examined, tested and, where applicable, refined.

Social scientists should be excited by the prospect of simulation models as a methodology for a number of reasons:

- Simulation offers social scientists an analogue to controlled experiments for examining social phenomena. In a simulation researchers can alter factors normally beyond their control, implement interventions perfectly and explore dose-response relationships beyond logistic and financial constraints. By extension, criminological theories can be made testable.
- Simulation allows for perfect observation and measurement. Researchers have the ability to record *all* interactions between components of the model, as well as knowing the outcome of these with absolute precision. Obviously, this is limited to the validity of the rules programmed, but this is a separate matter. In the real world, even with a perfect measurement instrument (which we do not have), there are many factors which strive to compromise the precision of outcome measures (e.g. sampling errors).

- It is possible, but not necessary, to model processes according to a 'bottom-up' approach by creating and following individuals overtime with a view to examine how specific turning points of experiential trajectories can be examined. DeAngelis and Gross, (1992)□ provide an authoritative source on the advantages and disadvantages of both individual and population level models. While individual level models are currently in vogue, under certain circumstances (large populations of relatively homogeneous units, availability of appropriate data for corroboration; see DeAngelis and Rose (1992)□ for more detail) population level models possess profoundly attractive properties.
- Simulation models can be performed en masse relatively easily and quickly. Once the model is built, minor adjustments are simple to perform.
- The above factors circumvent what Holland (1986) calls the 'fundamental problem of causal inference', the counterfactual problem. Our inability to observe the effect of two rival treatments on the same experimental unit has dogged researchers since the days of R. A. Fisher. Two main solutions present themselves (assuming homogeneity among experimental units and aggregation over treatment groups), but these vary in applicability according to the research question. Simulation models allow repeated experiments under identical conditions save for differences selected by the researcher.
- As implied from the previous point, conducting simulations that are equivalent to randomised controlled trial is far simpler and cheaper than in the real world.
- Additionally, aside from the potential application of simulation models themselves, the actual process of decomposing theories into simulation formalisms, such as the rules which govern agent behaviour, is useful in that it provides researchers greater insight, and demands that they specify theories and concepts in explicit terms. This can highlight potential inconsistencies or shortcomings and, by doing so, contributes to the subsequent strengthening of theory.

We will not devote any more space to explaining the promise of simulation models for criminology, as the other articles in this special edition are testament to this. Instead, the focus of this paper is on an aspect of the methodology which is given little attention. With the great potential of simulation methods comes an even greater obligation; one which, if overlooked, may result in simulation becoming yet another methodological fad that promised much and delivered little. For simulation models to live up to their potential, methods for directing their development and validating their findings need to be established. If this does not occur, we predict the methodology of simulation in criminology is at risk of declining into relative obscurity.

## Research Problem

The aim of this article is to outline the conditions under which the results of *in silico* experimentation can be thought of as valid. The principle method by which empirical findings are validated with established experimental designs is through replication. So, it follows that if simulations are a form of experiment, replicating *in silico* simulations will reveal their validity.

Before proceeding, the term 'replication' needs to be explicitly defined. We can specify two types of replication; *pure replication* occurs when the same methods are used under identical conditions (same laboratory, sample, research staff, etc.) and *practical replication* when an experiment or study is designed to imitate an initial experiment or study as closely as possible but contains some differences either by accident or design. Almost all replications in social science falls into this second type of replication. Pure replication occurs very rarely, but is possible to achieve in computer simulation models as long as random seeds were recorded for initial runs to be incorporated in successive replications. Indeed, the computer science task of verification is similar in nature to this. The closest related concept in social science is internal validity (Shadish, et al. 2002).

In terms of practical replication, these replications are typically performed so they closely resemble the original study, but with slight differences. These minor differences (different lab, sample, etc.) allow researchers to probe whether such disparities influence the size and magnitude of the experimental effect. The precise experimental effect may not be observed, but hopefully the differences will be small. Subsequently, researchers may purposely alter an independent variable (for example, in drug trials) to encapsulate representations more akin to the designated target construct. The collective goal of such replications is to produce a picture of the extent to which a generalised causal inference will hold over a number of differing configurations. Clearly, this type of replication has links to external validity (Shadish, et al., 2002). Other benefits of replication exist, as are other types of validity required to demonstrate causal relationships, but these will not be outlined here. Townsley and Johnson (2008) describe these in detail with special attention paid to simulation models.

Both types of replication are required to demonstrate valid experimental result. Yet, replicating *in silico* experiments is admittedly a different task to replicating real world experiments. Especially in the social sciences, researchers have an onerous time controlling the myriad factors which might influence the causal relationship. Even with the most carefully designed study and substantial agency and practitioner commitment, it seems unlikely that real world replications will approach anything similar to pure replication. The advantage of simulation models is that they empower the researcher with absolute control, which helps considerably in understanding and subsequently validating causal inference.

To date there have been remarkably few attempts to replicate simulations. The most insightful recent example is probably Hales et al.'s (2003) model-to-model analysis. They comment that the noticeably increased interest in applications of simulation techniques to a wide variety of social and physical

problems has been accompanied by a paucity of attempts at replication. They lament the fact that “[simulation] researchers tend to work in isolation, designing all their models from scratch and reporting their results without anyone else reproducing what they found” (1.2). Takadama et al. (2003) set out a method of ‘cross-element validation’, whereby comparisons are made between two simulations identical save for one element. By altering a small component of the system, the impact of system settings can be scrutinised. On the other hand, Klüver and Stoica (2003) tested different algorithms for the same research problem and found quite high levels of correspondence. Edmonds and Hales (2003) developed and tested two independent replications of a published simulation and found consistency (of results) between the two replications, but not with the original simulation. They comment “that, almost certainly, the vast majority of published social simulations do not completely comply with their authors' intentions” (1.5).

Axtell et al. (1996) developed an analogous concept to replication which they call model alignment. They argue that two models can be considered aligned, or ‘docked’, if they produce equivalent results under equivalent conditions. This allows objective and transparent comparisons between distinctly different models to be made. In their case study, Axtell et al. used two models of cultural transmission (the well known Sugarscape simulation by Epstein and Axtell (1995) and Axelrod’s Culture Model (ACM) described in Axelrod (1995)). The study is interesting for three reasons. Firstly, the authors admit that the docking exercise was underpinned by the experimental method in that their procedures were “roughly analogous to those used when a second investigator in a laboratory science is attempting to reproduce results obtained in a first investigator's laboratory” (p127). Secondly, the authors describe how they wrestled with inferring equivalence in experimental effects and found that simple summary statistics of experimental effects were not sufficient. They conclude that “distributional information about reported measurements is necessary if statistical methods to test equivalence are to be employed by a later investigator” (p135). Thirdly, the virtue of modular programming approaches was demonstrated in a compelling fashion. The authors report that the entire docking exercise required an estimated 60 researcher hours (excluding report writing).

Investing programming effort in the development of the simulations prior to publication meant that many of the modifications to the *in silico* experimental settings were analogous to ‘throwing switches’ (p 133).To explain briefly, using modular programming approaches such as Object-Orientated Design (OOD), researchers can minimise the work involved in maintaining, modifying and analysing their existing models. OOD allows the logical segmentation of program components into objects, which hold state and behaviour information. This offers considerable flexibility to those modifying existing simulation components, as internal functions, algorithms and data manipulations can be altered, refined and/or replaced without the need for the model to be redesigned. This inherent modularity also provides OOD with a distinct advantage over other more traditional approaches for researchers wishing to analyse complex program interactions and states, the likes of which are often seen in simulation models.

Ultimately, Axtell et al. (1996) conclude that efforts to demonstrate model alignment in the future is only likely when “a precise, detailed statement of how the model works” is available (p135). Unfortunately, it seems that simulation replications are rare and opportunistic, or at least not systematic. Researchers rarely commence development of a simulation with consideration of how it might be replicated by others. For this, we blame the disconnect between the two varieties of researchers involved in simulation modelling: computer scientists and domain experts. Each group predominantly focuses on different aspects of modelling. Domain experts are typically interested in *using* the model, whereas computer scientists focus on *building* the model. There is, of course, inevitable overlap in their respective agendas, but despite the penetration into the social sciences, much of the heavy cognitive lifting in simulation development is still carried out by computer scientists[2].

The field of model replication seems currently underdeveloped, especially by those with potentially most to gain, i.e. those who might use simulations to direct action or policy rather than computer scientists. The complex non-linear nature of crime interactions dictates that the processes of replication and evaluation that simulations provide can substantially aid in increasing our understanding of causal inference, thus maximising the utility of any model tempered by a number of simulation findings. Simulation replications, much like their traditional counterparts, can be used to increase construct robustness by gradually introducing greater levels of intricacy aimed at moving from *in vitro* inference to *in situ* inference; investigating the degree to which causal relationships hold over a variety of different settings[3]. This process of directed incremental complexity allows us to examine our model through numerous cross sections of the system simulated as we move towards sufficiency (Eck and Liu, this volume). However, if we are to present a cogent argument for the validity and subsequent utility of our simulation models, then the implementation of replication and its use in the introduction of incremental complexity will obviously require an investigator to utilise a systematic and logical approach.

Therefore, we now describe a methodology of systematic replication which aims to move towards generalisable inference, suggesting that, whilst originally devised for traditional experiments, its key components might just as well relate to *in silico* experiments. Before doing so, a few qualifiers are necessary. First, experimentation is not the only method of conducting science. Other hard science fields (e.g. astronomy) advance without exclusive adherence to experiments. Experiments are merely the *simplest* method from which to draw inferences. Second, simulation models only get us so far on the spectrum of understanding. Experiments can reveal causal description, but can only offer candidate causal explanation by demonstrating generative sufficiency (Epstein 2006). Thus, we may demonstrate that a phenomena is reproducible given a theorised mechanism, but we can never be sure that such a mechanism is necessary in the real world. Third, strictly speaking, a simulation model can never be mis-specified. The system operates exactly as the components and their interaction rules were designed. Its output is valid.

However, in the real world, if models are mis-specified, there is a natural feedback mechanism to indicate this mis-specification. Excluding an important factor results in poor fit or predictions by default, but *in silico* we will never know if an omitted variable is important. Worse, even if we do include it, but get the interaction rules wrong, we are unlikely to know. This problem can only be solved by developing simulation models in a transparent and explicit manner, to be held up for peer scrutiny and replication.

The remainder of this article is organised in the following way: having described the research focus of the paper, we present a simple simulation model for the express purpose of illustrating how a series of simulation replications can be developed to increase the robustness and subsequent utility of a simulation model. Next, systematic replication is defined, first in conceptual terms, then by an explicit methodology to methodically approach causal inference. In this section the case study and a number of real world crime studies are used to illustrate various components of systematic replication. The final section of the article deals with the means to document systematic replications.

## A Case Study: *Cops & Robbers*

In this section we briefly describe a simple simulation model as a means to illustrate the nuances of replicating *in silico* experiments in a systematic fashion. Cops & Robbers, initially described by Birks et al. (2007), was developed using NetLogo[4], a "cross-platform multi-agent programmable modelling environment"[5]. Briefly, the model's aim is to provide insight into the victim-offender-location interaction described in routine activity theory (Cohen and Felson, 1979). In order to do so, it examines several simplistic theories of both victimization and prevention, allowing exploration of the effects of variations in several micro-level properties upon macro-level outputs, such as simulated crime and prevention rates. In the most abstract version, Cops & Robbers imagines a simulation world that is inhabited by three distinct groups of individuals: potential victims, offenders and law enforcement. This abstract model of offending initially theorizes:

> *that all individuals move around the environment in a random fashion; that a crime occurs when an offender comes into the same location at the same time as a target; and that a detection/prevention occurs when an offender, target and guardian all come together at the same point in space and time.*

The central premise of Cohen and Felson's theory is that the incidence of crime is not determined by the number of offenders or victims per se, but the *rate* at which they intersect without a guardian. Examining the guardianship element of our theory, a series of experiments can be designed to investigate this mechanism, implementing various configurations of guardianship and exploring the direction and magnitude of any observed causal relationships.

These hypothetical experiments will form the basis of our example simulation used to illustrate the key concepts to be discussed.

In building the simulation, we initially formalise two key mechanisms that allow the experiment to take place: a guardianship behaviour, the algorithmic equivalent of the guardianship element of our initial theoretical statement, which we incorporate into the decision calculus of our law enforcement agents; and a similar offending construct, which is bestowed upon our offender agents. Our outcome measures are the crime and prevention rates (the latter is the number of crimes denied by the presence of a guardian – an interim metric impossible to measure outside *in silico* experimentation). These measures allow us to describe the experimental effect by observing the relative difference in magnitude and direction of both crime and prevention occurrence across simulations with varying properties and representations of units, treatments and settings.

Our aim is to investigate whether we can produce a generative model which can demonstrate candidate causal explanation for the spatial and temporal distribution of crime. To do so, we begin with an extremely simplistic model of routine activity, then, through a series of systematically directed experiments/trials, introduce model complexity by incorporating moderator variables additional moderator constructs relating to different aspects of the model, whilst maintaining sufficient outcome measures to capture the experimental effect studied. To be clear, the overarching aim of this series of experiments will be to explore the guardianship construct and probe the strength and extent of its relationship with crime levels. For each simulation replication model, a distinct hypothesis is generated. These hypotheses relate to the incremental introduction of moderator variables. Throughout this process we advocate the application of an explicit and transparent method of describing simulation experiments in order to aid in the incremental replication of a model by other researchers to further validate its findings.

As we will later discuss in more detail, in order to promote the process of systematic replication, initial simulation experiments should be both explicit and transparent in their documentation. In practice, this dictates that investigators should denote both the experimental aim, in our case, the study of the guardianship mechanism hypothesised by routine activity theory, and *utos* configuration of each *in silico* experiment.

## Systematic Replication

Systematic replication is proposed as a method for developing simulation experiments which move towards generalisable inference that is *directed*, *explicit* and incorporates complexity *incrementally*. It consists of two distinct components:

i) Cronbach's (1982) formulation of generalisation; and
ii) what Shadish et al. (2002) call the 'phased model of increasingly generalisable studies'.

Systematic replication is underpinned by the generalisation formulation conceived by Cronbach (1982). He stated that any study can be considered a unique configuration of units (subjects), treatments (manipulable causes), outcomes/observations (measurement of effect) and settings (social and physical environment of study). Thus, a single study is denoted as *utos*. Replications of this *utos* configuration are pure replications as all conditions are the same (as long as identical random seeds are utilised). If the observed experimental effect is reproducible through a series of pure replications, then that unique configuration of *utos* can be said to produce valid outcomes. We explain later in detail a process for conducting this type of validation task.

Using Forrester et al.'s (1990) description of the well-known Kirkholt burglary prevention project as an example, the *utos* is comprised of burgled households (units); a suite of tactics, including careful problem diagnosis, coin meter removal, promoting interagency cooperation, cocoon neighbourhood watch, prevention of repeat victimisation (treatment); recorded repeat victimisation, recorded burglary, victim and neighbour interviews (outcome); and a high crime, clearly defined housing estate in the North of England (setting). The observed impact of the Kirkholt *utos* was an approximate 75 percent reduction in the annual burglary count three years after implementation (with roughly 42%, 57%, 25% and 21% reductions year on year).

Obviously, policy makers and researchers are interested in the degree to which the observed experimental impact of burglary reductions under the Kirkholt *utos* might hold over other *utos* configurations. Tilley (1993) described three replications of the Kirkholt study. Decomposing the replication that closest resembled the Kirkholt *utos* (Tilley's so called '?R1'), the Rep1 *utos* could be described as a configuration of burgled households (units); a suite of tactics not limited to problem diagnosis, target hardening for social housing tenants, limited interagency cooperation, traditional neighbourhood watch with modest coverage, prevention of repeat victimisation (treatment); police burglary records and victim interviews (outcome); and neither a high crime area, nor a clear demarcation between the action area and the surrounding environment (setting). The observed impact of the Rep1 *utos* was an increase in burglary of about 22 percent in the first year of implementation, followed by a further 43 percent increase the next.

The differences in *utos* configurations and outcomes are stark and decisions about generalisability are difficult to make with confidence (was it the different setting, the different tactics, or both that led to the different observed impact?). Cronbach (1982) argued that generalised causal inference was only possible if one was able to demonstrate that the experimental effect held over many *utos* configurations. He defined *UTOS* as the population of units, treatments, outcomes and settings over which a causal relationship holds. Lower case letters, therefore indicate sampled values of *u*, *t*, *o* and *s*. Capital letters indicate the population for that component. Thus, the generic task of generalisation is to go from *utos* (one particular realisation) to *UTOS* (a causal relationship that holds over many types of units, treatments, outcomes and settings).[6]

To illustrate, Table 1 describes the simple Cops & Robbers model in *utos* terminology alongside other putative configurations. These additional configurations make sense in terms of assessing the generalisability of experiments/models. As well will explain later, in our view investigators should also provide a list of other potentially viable *utos* configurations with a view to being included in subsequent replications. Thus, Table 1 outlines the aim of our example experiment, the relevant *utos* configurations of our initially described model and several other candidate *utos* configurations[7].

[insert Table 1 about here]

Before discussing the applicability of the 'phased model of increasingly generalisable studies' in allowing us to move towards *UTOS*, two points concerning the applicability of *utos* for demonstrating *in silico* experimentation generalisation need to be made:

1. Cronbach was primarily interested in intervention evaluation, reflected by the inclusion of the treatment component in *utos*. Our purpose is broader than this, encompassing the falsification of theories, as well as theory development. Real world interventions can be considered as an implementation of some policy which, in turn, is simply a theory of how particular constructs are related. Therefore, we take the position that interventions are just a special case of theory testing and by allowing the *in silico* treatment to describe the theory, our treatment *utos* is directly applicable for our purpose.

2. For *in silico* experimentation, generalising over the *o* dimension of *utos* does not completely transfer from classical experimental and quasi-experimental studies. This is because in the real world, measurement issues force researchers to compromise and capture abstract constructs using proxy measures. If our hypothesis relates increased guardianship with lower levels of crime, there are many ways to measure offending (the outcome). For instance, levels of crime can be captured using recorded crime, court records or self-report surveys, each having relative strengths and weaknesses. Cronbach argued that results derived from single measures could be compromised by particulars of that measurement type. That is, there are some elements of a given measure which inadvertently bias the results. If experimental effects are, therefore, consistent across different measurement types, we can be much more confident that those results are valid. Simulation models, however, possess perfect observation and measurement. Therefore, all we need do is ensure that sufficient outputs are observed to capture the experimental effect under study. Although, if simulation results are to be compared to those from empirical research methods, analogous outputs which capture the frailty of traditional recording methods may need to be engineered (this will be discussed briefly later in this paper). Randomised controlled trails are good for revealing whether an outcome measure has changed, but are not sufficient to say why. Fortuitously, *in silico* experiments allow for the measurement

> of numerous interim variables and system states between cause and effect which might allow some insight into causal explanation or mechanisms. Thankfully, unlike real world experiments, the implications of realising one has not recorded all the data required to describe an experimental effect *in silico* are relatively modest.

In discussing the utility of the *utos* methodology for developing systematic replication of simulation models, the next substantive issue which arises is concerned with how to determine the variety of units, treatments, outcomes and settings that might allow us to demonstrate a valid causal relationship, and how we might go about demonstrating this. With respect to the former, in the real world, when scientists manage a directed programme of studies, they do not randomly sample from the plethora of units, treatments, etc. available. Instead, they *purposively* sample, drawing on theory, experience and logic. By employing a systematic approach, as will be elaborated later, we envisage that there is tremendous power in researchers explicitly documenting the factors that have been tested and those yet to be included in a simulation model.

As for the way in which a move from *utos* to *UTOS* might be executed, consider what Shadish et al. (2002) call the 'phased model of increasingly generalisable studies'. This is an approach where several studies are conducted, some of which are nested within discrete study phases. These phases are distinct because the research objective changes from phase to phase. The rationale of this incremental approach allows research activity to develop only at the pace of the evidence. Further, the experimental conditions begin in 'perfect' controlled environments and are made realistic in a piecemeal fashion so that a cumulative causal picture emerges. We suggest that such an approach would well suit the incremental development of simulation models of crime activity. Table 2 shows this phased approach for generic science and *in silico* experimentation.

[insert Table 2 about here]

The first phase is about developing a hypothesis from first principles or highly abstract scenarios (theory). The second phase involves developing reliable and accurate procedures for conducting the study. This might mean using a piece of laboratory equipment or identifying a robust statistical method, as well as some form of data generating procedure (police recorded crime, interviews, etc.). The third phase involves testing the hypothesis under perfect, controlled conditions. In drug testing and development, say, this usually means using healthy males, presumably because they are a relatively homogeneous group with respect to the particular causal relationship being tested. In the social sciences, finding homogeneous units is very difficult and conducting highly controlled trials even more so. Typically, some compromise is made, either in the experimental design or treatment implementation for pragmatic reasons. These initial three phases are stylised versions of how individual studies are conducted.

The next two phases describe efforts to replicate studies whose hypotheses survive the first three phases. The fourth phase tests the limits of the causal relationship by introducing heterogeneity into the study. The principle here is that if a causal relationship can be observed within a heterogeneous sample, we can be more confident of its validity than if it has only been observed from a homogeneous group of experimental units. Of course, heterogeneity is introduced for a purpose; it represents the presence of a supposed moderator variable. Imagine a substance abuse education campaign is initially tested on a sample of children from a single school. Prior to widespread adoption, researchers may speculate that elements of the campaign may not enjoy the same level of effectiveness for children outside the social class of the pilot school (as schools usually limit their enrolment to children residing within a particular catchment area). Thus, the socio-economic status is identified as a potential moderator variable. Further testing in schools across the social class spectrum would determine whether this factor is indeed a moderator variable and, if so, its impact on the generality of the campaign's casual validity.

Generalising to *UTOS in silico* means to move from a single (sampled) observation to a population in the following way: a naive model ($utos_0$) involving homogeneous units and a homogeneous setting is defined. $utos_0$ is made incrementally more complex by introducing heterogeneity with respect to units, treatments and settings (*in silico utos* is already generalised for outcomes). Practically, moderator variables are incorporated by creating heterogeneity. For example, suppose a simulation is developed where agents move about and make decisions according to some predefined formalism. Whatever the outcome, the observed experimental effect is valid for these homogeneous units in this monoculture setting. This naive model ($utos_0$) could be made more generalisable and less abstract by incrementally introducing complexity. Suppose theory suggests that gender of units has a significant impact on target selection. A slightly less naive model would have a system populated with male and female agents, with a corresponding altered decision calculus.

An initial simulation (and almost any published experiment) can be epistemologically located at phase 3 and provides some indication of internal validity depending on the experimental design. The main thrust of this article is articulating a systematic method of carrying out phases 4 and 5. Even so, just as with a real world experiment, researchers attempting to replicate prior studies need to scrutinise phases 1 through 3 of the original study in order to ensure the parameters, constructs and formalisms of the *in silico* experiment are understood. This is the process of pure replication. Figure 1 shows an idealised version of how this cycle of incremental development might operate in practice.

[insert Figure 1 about here]

Space restrictions do not permit a thorough examination of how phases 1 to 3 can be conducted or scrutinised systematically. Instead, we point out that phases 1 to 3 require fundamental scientific research tasks, which a competent researcher ought to be able to perform. While the *in silico* version

is somewhat different, the nature and types of decisions made have much in common with mainstream research methodologies. For instance, regression modelling requires, among other things, the constructs of interest to be operationalised in an appropriate manner, careful consideration in model development, attention to data quality and meaning (does a drug arrest hot spot indicate drug dealing or police activity?), avoiding over-fitting the data and a considered interpretation of results. In short, despite our not focussing on phases 1 to 3, they are of critical importance. Additionally, replication and scrutiny, both internally and by other researchers at phases 4 and 5, will uncover threats to internal validity related to phases 1 to 3. Eck and Liu (this volume) comment on these issues and set out what this entails.

Recently, systematic reviews have gained popularity in the social policy research arena[8]. They differ from meta analyses and narrative reviews in that the criteria for conducting the review is transparent and explicit at all stages (proposal, inclusion of studies, analysis, reporting). In this aspect, the approach proposed in this article is a systematic, cumulative approach to simulation modelling. It requires that changes at each iteration be made explicit in order that impact on experimental effects can be observed and that other researches can scrutinise the model. The main difference between systematic reviews and systematic replications is that one is inherently retrospective and the other prospective. The disadvantage of the retrospective approach is that is the reviewer can only use those studies that satisfy the criteria for inclusion, which can result in small sample sizes. Clearly, this is not a problem for systematic replication, as simulation models can be modified in light of errors and flaws which become revealed during their study. Further, new research questions may be added opportunistically over time.

## Systematic Replication Criteria

### *Criterion 1: Testing for presupposed emergence as a rival hypothesis*

This first criterion deals with determining the internal validity of $utos_0$. Simply put, prior to any efforts to replicate a simulation, researchers should implement the model described and replicate the decisions made in phases 1 to 3. Of principle importance is whether any observed experimental outcomes are the result of programming decisions rather than relationships between constructs. We do not explore this criterion any further as it is outside the scope of this article. Eck and Liu discuss some of these issues in this volume.

### *Criterion 2: Generalising about experimental units (developing utos → Utos)*

The second criterion is concerned with moving from inferences on homogeneous experimental units (usually people) to inferences on the population of units (that are necessarily heterogeneous). In practice, this is about introducing and operationalising moderator variables, which effectively partition the $utos_0$ homogeneous units into strata of experimental units.

The Minneapolis Domestic Violence (DV) experiment described in Sherman and Berk (1984) provides a clear example of the importance of delineating between differing units when establishing causal relationships. The original study, which aimed to reduce domestic violence, examined the effect of three police responses (treatments) aimed at domestic violence offenders: mandatory arrest; advice (including mediation); and 'sending', where the offender was sent away from the home for a 'cooling-off' period. In the initial experimental setting, the observed experimental effect was substantially reduced recidivism for units receiving the 'mandatory arrest' treatment. However, when replicated and implemented in other areas, the relationship between mandatory arrest and DV recidivism was, in some cases, reversed. It has been hypothesised that a key factor influencing this effect was the employment status of offenders: in the initial study area, DV offenders predominantly held jobs and were shamed by arrest; whilst in certain replication areas, offenders were unemployed and not necessarily shamed by arrest. Therefore, the effect of offender employment was not adequately captured in the initial study, leading to incomplete causal inference. Formally, the experimental effect observed in the Minneapolis *utos* was not consistent for *utos* configurations where the unit moderator variable 'employment' took on different values to the initial Minneapolis *utos*.

*In silico* we suggest the unit construct denotes, much as it does in the real world, the characteristics and behaviours of our experimental subjects. Using the Cops & Robbers guardianship experiment as an example, we might wish to examine the effect that different offending rates have on the experimental effect. This is the first candidate moderator variable. Thus, we grant the offender agents a new characteristic – lambda. High values of this characteristic indicate offenders who are driven to offend more frequently than those with low values. By implementing this new parameter, we wish to observe whether the experimental effect is consistent, even with different types of offender agents.

So, phase 4 of the phased approach would be to determine how this moderator variable will be operationalised and what elements of the model this will impact upon. In this case, an agent characteristic lambda is created and the offender agent decision calculus modified so that when offenders encounter a potential victim, the likelihood of them choosing to offend is dictated by the value of lambda. If lambda has three potential values, 'high', 'medium' and 'low', we create three simulations, each with a homogeneous population of offending agents that relate to the different levels of lambda. Subsequently, the emergent properties of each simulation are examined through a series of pure replications and the direction and magnitude of the experimental effect observed over all three. If this process produces a plausible experimental effect in line with our existing knowledge base, we proceed to the next phase.

Phase 5 involves a new simulation model populated with offending agents who are heterogeneous with respect to the lambda characteristic. Therefore, each level of the 'offending rate' construct is present in this model. Again, we

run our simulation and examine the magnitude and direction of the experimental effect, this time over a heterogeneous group. In this scenario the proportion of agents represented by each level of the moderator variable may be determined by empirical evidence if so desired.

There are several advantages to completing the tasks in phase 4 prior to implementing a heterogeneous model described in phase 5. First, aggregation effects like Simpson's Paradox are avoided. Experimental effect in sub-populations may vary dramatically, only to be diluted in the aggregate. Second, the incremental nature of operationalisating a moderator variable forces the researcher to be explicit and aids in uncovering errors at as early a stage as possible.

Further replications may explore the impact upon the experimental effect of both the incorporation of other potentially viable units or the alteration of existing ones, initially in homogeneous groups (phase 4) and subsequently across heterogeneous ones (phase 5). For example, thus far, lambda values for each agent within the population have remained static over the entire simulation (i.e., the agent life-course). However, existing theory, experience and logic suggest this representation is not reflective of real offender characteristics. Therefore, further simulation iterations might introduce dynamic representations of lambda that change over the course of a simulation. An initial dynamic model may, following a simplistic interpretation of the age-crime curve, dictate that all offenders begin with lambda at low; then, after a given amount of simulation time, lambda becomes high, later declining to medium and then returning to low. Initial phase 4 replications would bestow all offender agents with this identical construct, dictating that the changes to lambda over time occur for the group as a whole. After examining the experimental effect of this configuration, subsequent phase 5 replications will allow for simulation-time changes in lambda to be localised to individual offenders creating a heterogonous population of offenders, all of which possess their own dynamic representation of lambda. Eventually, replications may develop this construct further by allowing lambda to represent a typology of life-course trajectories (Nagin et al., 1995).

### Criterion 3: Generalising about experimental treatments (developing utos → uTos

The third criterion is concerned with moving from inferences on homogeneous experimental treatments, such as specific crime prevention interventions, to inferences on a suite of treatments aimed at producing the desired experimental effect. As previously discussed, the authors consider such interventions to be special cases of theory testing, in that they are informed by policy (which is in effect a theory about some aspect of society) and attempt to manipulate some cause (albeit imperfectly compared to controlled experiments) to generate a desired effect.

Therefore, in generative simulation experimentation we suggest the treatment should delineate the mechanisms around which the current experimental aim

is aligned. In our Cops & Robbers example, our experimental aim is to study routine activity theory, and in particular, the guardianship mechanism it presents. Therefore, the treatment within our current experiment denotes the configuration of the guardianship mechanism within the simulation, its associated agent parameters and the rules which dictate how and which agents employ it. In addition, it should also be noted that any outcome measures should be based around the requirement for adequate measurement and understanding of the treatment mechanism being studied. This requirement is discussed further in criterion 4.

The initial model grants all law enforcement agents the absolute capability of guardianship. Drawing further from routine activity theory, our initial replication model may introduce a moderator variable which represents the capability of such guardianship. Again, we might initially delineate our new variable capability into three categories: high, medium and low; altering the respective agent calculus accordingly so that when encountered, the likelihood of an individual providing adequate guardianship to prevent a crime is dictated by their capability. Phase 4 will run simulation models, each with homogeneous populations at each level of capability, again observing the experimental effect. Phase 5 will with run simulation models with heterogeneous populations of guardians.

Moving further towards a more plausible representation of guardianship as defined by routine activity theory, further model iterations might introduce and examine the effect of other non-enforcement agents acting as guardians. Additionally, we may establish a sphere of influence around each guardian agent within which guardianship is provided, allowing guardians to prevent offences not only at their exact location, but also within their close vicinity. Initial (phase 4) experiments will allow for the scrutiny of such mechanisms in simplistic monocultures, where the experimental effect is more transparent and easier to decompose. Subsequent (phase 5) replications could incorporate and examine the effect heterogeneity of both capability and guardianship effect size has within our population of potential guardians, and, more generally, on the observed experimental effect[9].

This criterion strives to probe whether the initial treatment *t* is a valid representative of the construct treatment *T.* It seeks to determine whether the observed experimental result was generated by the treatment construct or other facets of *t* which are not a direct consequence of the treatment construct. For example Braga et al (1999) describe a randomised controlled trial where 12 pairs of violent crime hot spots are used to test the efficacy of problem-oriented policing. The question here is whether the observed experimental effect (reduced crime in the treatment areas with no displacement) was the result of the stated treatment construct 'problem-oriented policing' or was it, say, a Hawthorne effect induced by elevated agency attention and investment? Replications that implement different versions of the treatment construct help answer this question.

***Criterion 4: Generalising about experimental outcomes (developing utos → utOs)***

When performing traditional experiments, the investigator needs to identify the most appropriate outcome measure that will capture the experimental effect. This is made all the more difficult as measurement imprecision can cast doubts on the validity of experimental results. The common remedy is to identify multiple outcome measures in order to avoid relying too much on a single measure. For example, offending can be measured through self-report surveys, police arrests, court conviction or probation data. As we have previously discussed, the selection of varying outcome measures can lead to varying inferences about experimental effect. Faggiano et al. (2005) conducted a systematic review of education-based programmes to prevent drug consumption. They described three basic intervention types: skill-based (refusal skills, safety skills); knowledge-based (information about the consequences of drug abuse); and affective-oriented (self-esteem, motivation, personal development). When evaluated on the primary outcome measure of increased skills, all programmes were shown to be effective at increasing participant skills. However, when the outcome measure was changed to drug use, only skills training programmes showed an impact.

Greater confidence can be afforded inferences drawn in circumstances when multiple outcome measures produce similar findings, as the particular bias evident in each measure does not seem to be large enough to change the results in the aggregate. If all measures align in a similar fashion, we can be more confident about the generality of any causal inference made. However, in simulation experimentation, the investigator is provided a great advantage over the traditional experimenter, as he is given the ability to engineer and specify any number of outcome measures, which provide absolute measurement and can be positioned at a wide variety of locations and levels of granularity within a simulation. For example, we might record every offender agent's decision-making process individually, down to currently perceived risks and rewards, whilst also recording the overall number of crimes committed within an area by both an individual and the entire offending population.

A further advantage provided by ubiquitous measurement that should be exploited is its inherent temporality. Traditional experiments can often, for understandable reasons, only implement a limited number of measurement points from which to draw inferences throughout the implementation of some mechanism. Simulation, on the other hand, allows for *continual* monitoring of *all* selected outcome measures. Further, if lack of sufficient measurement emerges as an issue, an *in silico* experiment can easily be remedied, as additional outcome measures can be incorporated and experiments repeated where necessary.

Beyond the required adequate description of the experimental effect, simulation also provides a window, as it were, into our theory, allowing us to examine the micro, meso and macro ramifications of our theoretical assumptions. In doing so, simulation provides a distinct contrast to statistically based explanations, which provide limited causal description (i.e., how dependent variables relate to independent ones, but nothing in-between). The

generative nature of simulation allows for the production of *candidate* causal explanations of crime phenomena. The gathering of candidate causal explanations is aided significantly through the examination of the simulations' interim states, which are often invisible in traditional experiments. These measures allow us to more closely scrutinise the mechanism under study and, where possible, establish the significance of numerous potentially contributory elements within the observed experimental effect. For example, in a traditional experiment, a reduction in the number of burglaries may be observed after the implementation of high-visibility policing strategies. In this scenario, the experimenter cannot infallibly infer that the increased number of police officers prevented more crimes from occurring, as opposed to other potentially rival explanations, since there is no reliable metric for the number of crimes prevented. Again, the fundamental problem of causal inference (Holland, 1986), our inability to observe the counterfactual, bounds the confidence with which inferences are drawn. No such limitation exists for simulation models. Yet, this wealth of choice afforded to the investigator dictates that if assessment of simulations is to be facilitated, he or she should proceed in a systematic fashion.

In summary, the role of the *in silico* investigator, with respect to outcome measures, is to record sufficient data to capture the experimental effect and, where possible, introduce interim variables which capture the experimental effect at the micro, meso and macro level, providing greater insight into how candidate causal explanations may operate. Beyond this, as simulation models become more and more realistic, at some point, simulation outcomes may be contrasted with empirical findings. Empirical crime data are inherently flawed and often fraught with a myriad of potential dangers: the dark figure, under and over-reporting, deception, recording errors and political bias, to name but a few. Therefore, if one is to compare simulation and empirical findings, it would seem sensible to create simulation outputs which attempt to best capture existing recording practices and their inherent flaws, to allow for direct comparison. Issues such as the modelling of this noise in crime data are of great importance for models aimed at end users like policy makers. Although such techniques are currently underdeveloped, they are, unfortunately, beyond the scope of this article. However, the issue is discussed in more detail in this volume (see Eck & Liu).

### *Criterion 5: Generalising about experimental settings (developing utos → utoS)*

The *s* of *utos* denotes the setting; in traditional experimentation this is the cultural and physical environment in which an experiment takes place. Generalising about experimental settings involves the extrapolation from a single setting to many settings, therefore increasing the universal applicability of treatments implemented. This extrapolation is all but impossible in practice, as factors that operate at the neighbourhood or community level will take on only one value per study. Within a single study there are likely to be a host of potential moderator variables at play (e.g., deprivation level, social housing, inner city/suburban, extent of public transport, etc).

Tilley (1993) described three replications of the famous Kirkholt burglary prevention study. Even though there were a number of differences between the Kirkholt *utos* and the Rep1 *utos* (described earlier), assume that the only one of importance is the setting. Kirkholt is a relatively self-contained housing estate, with clearly demarcated borders – in short, '[a]nyone entering or leaving the estate would know that they were doing so' (p3). The action area in the Rep1 *utos* was different however. The scheme covered 3.5 times the number of properties as Kirkholt, and there were no clear boundaries to the geographic extent of the project. The different experimental effects observed in the two studies could be caused by some difference in the physical and social environments of the two areas. Perhaps Kirkholt worked because all local offenders knew the opportunity structure of the entire estate had changed; something that may not have been as obvious in the replication project[10].

For Cops & Robbers, our aim is to progress through a series of incrementally complex replications, from our initial simplistic setting to a more robust and plausible representation of 'setting' in which our agents are situated and our treatment acts. The critical aspects of *in silico* settings can be captured by two categories: environmental configuration and population configuration. The environmental configuration describes the simulation environment's morphology, such as world size, the presence, or lack thereof, of transport networks, zoning, physical barriers etc. The population configuration describes those conditions which occur as a result of interactions between agent (unit) populations and the environmental configuration; for instance, offender, guardian and potential victim densities or spatial distributions.

In the first iteration of Cops & Robbers ($utos_0$), the environmental configuration for our initial setting is a uniform size toroidal homogeneous world that allows for unconstrained movement in all directions by all agents within it. The population configuration describes a world where all agents begin the simulation in random locations, and where offender, guardian and potential victim population sizes remain static throughout a simulation, thus dictating that the aggregate density and distribution of agents remains fixed.

In advancing our simulation we may choose to examine the effect that changes to the environmental configuration, which introduce more purposeful and realistic agent movement mechanics, might have. Initial phase 4 experiments following routine activity theory may introduce simplistic zoning so that the simulation environment is divided into 3 distinct zones; residential, employment and entertainment. Using these three zones, our agents' decision calculus could be modified to allow agents to follow simplistic routines, which centre on a number of plausible trips between zone types. Again, in line with routine activity theory, phase 5 experiments might introduce individual routine activity nodes for each agent.

When manipulating population configurations, initial phase 4 replications may investigate the effect of varying the number of, and therefore (where the world size remains constant) densities of, offender, guardian and potential victim

populations. Again, for simplicity's sake, we may initially run three simulations, the first with low densities of all agents, the second with medium and the third high. Phase 5 configurations will allow for differing densities between agent groups for each of the three agent sets. This would allow examination of the interaction of differing population scales and the experimental effect.

## Documenting Systematic Replications

If systematic replications are to gain acceptance, investigators should consider the potential subsequent replications of their simulation when designing and describing the initial model. By being explicit and transparent in documenting all model configurations, parameters and constructs, investigators can maximise both their own and others' understanding of the model. In doing so, they increase the ability to perform valid and pertinent replications which move toward a cumulative goal. The point is that all accepted methodologies in science have criteria for reporting results, from laboratory research (the American Psychological Association sets out how an experiments should be reported) to systematic reviews.

Whilst we have provided a phased model of incremental development and examples for each systematic replication criterion, the tasks outlined in phases 4 and 5 in Table 2 are described in quite abstract terms. Therefore, we now outline a framework for investigators whose aim is to use simulation models, and the replication thereof, to move towards models of causal inference.

1) The first task of the replicator is to replicate steps 1-3 of the phased model (see Table 2) and generate repeatable findings.
2) The original experimenter should specify the *utos* configuration included in the original simulation. Table 1 provides an example.
3) In addition, the original researcher should include two other lists of variables (which, although not strictly necessary, would assist future replications). First, a list of other *utos* variants not included/tested which the researcher believes (through experience, theory or empirical evidence) may have some bearing on the causal inference. The second list is the set of *utos* variants presumed to be irrelevant to the causal inference.
4) Future replications should draw on constructs not yet incorporated into the simulation, as well as others that might be created and/or informed by new research findings.
5) The researcher should move toward causal inference by either determining irrelevancies (those constructs that do not vary the experimental effect) or making discriminations (those constructs that do vary or limit the experimental effect).

One potential criticism of the approach proposed is that it is unrealistic to impose these standards when the main channels of dissemination remain remarkably inflexible on matters of word length and style. Our view is that the virtues of *reproducible research* allow effective dissemination of research

findings without usurping conventional publication mechanisms and aid in technology transfer among colleagues. Reproducible research allows consumers of research to reproduce the analysis of studies precisely and quickly. In the simplest case, imagine a word processor file that is bundled with all the supporting data, algorithmic code and the means to re-run the analysis. The description of the research is identical to a standard journal article and the supporting files (data and commands) can be accessed through macros embedded throughout. Suppose there is a sentence like, 'No statistical differences were observed between the two groups ($p > 0.05$)'. A curious reader could access the commands and data utilised to compute the quoted statistic, or any of the figures, or any part of the analysis. Similarly, if someone had data in a similar format, they could generate equivalent findings to the study by using *the same commands used by the original researcher*. One example of documenting reproducible research is described by Schwab et al. (2000). There is no feasible reason why simulation modelling could not harness a similar process of dissemination with a view to promoting validation through replication and peer scrutiny. A side benefit of that if one knows that all the data and code that support the study are available to peers, a great deal more care is taken to report and interpret the results objectively.

It is in the spirit of reproducible research that the Unified Modelling Language (UML) was originally conceived and developed. Defined by a consortium of major IT companies known as the Object Management Group (OMG) including IBM, Sun Microsystems and Hewlett Packard the aim of UML is to provide software engineers with a general-purpose notation which allows for the abstract description of software solutions irrespective of their development platform. Focusing around object-orientated software development UML, and its most recent incarnation UML 2.0, enables developers to visually depict a program's respective objects, their attributes and the relationships between them. Thus providing software developers and analysts alike with a birds-eye view of a system's structure and function, and in turn increasing both understanding and the ease with which software can be reproduced and/or modified.

However, like most other developing standards, UML is not without its critics some of whom claim it has become increasingly cumbersome and unwieldy in its application due to its growing complexity as more and more elements are added. Furthermore, due to its attempts to be all things to all men UML has difficultly in providing sufficient functionality in a number of more specialized domains. To this end, researchers have made efforts to extend UML in order to more adequately capture the requirements of emerging technologies. With respect to subject of this paper there have been several efforts to make UML more suited to the depiction of simulations and in particular agent-based models (Odell et al. 2000; Cranefield et al. 2001; Huget and Odell, 2005)

Recalling the phased approach to increasing complexity set out in Table 2, and in particular the second phase which involves the development of methods and instruments to ensure accurate and valid means to test hypotheses, we suggest that the use of UML in describing simulations, or a

suitable extension thereof, may be of considerable importance, allowing for faster and more robust transfer of model specifications across research teams, thus increasing research productivity.

## Conclusion

In this paper we have discussed the potential of simulation methods for the social scientist and, more specifically, those interested in synthesising the interactions between explanatory variables. We have noted the relative scarceness of simulation replication (the methodology through which traditional experimental results are often validated), especially within the social sciences, a field that has potentially much to gain from the validation of *in silico* experimentation.

Implementing any existing methodology in a novel way within an established discipline is likely to split the research population into three groups: those who are unaware of or indifferent to the use of the methodology; those who are aware, but sceptical about its usefulness; and its proponents, who might be somewhat enamoured with the method. The existence of each group is necessary, especially the latter, without whom the approach may not have diffused across disciplines in the first place. Yet, whatever one's perspective, it is to the mutual benefit of all that new methods are systematically and objectively employed with a view to assessing and determining their validity. Therefore, to minimise the potential of methodological stagnation, we suggest that simulations need to be considered scientifically and their results validated with respect to purpose. To this end, we have presented an iterative development framework, inspired by traditional experimentation and analogous to clinical drug trials, which we hope will aid in the development and subsequent validation and verification of models of crime occurrence.

The role of *utos* can help us to logically and systematically introduce moderator variables in order to probe the extent of causal inference. The use of a phased model of increasing generalisability allows us to manage the incremental introduction of complexity. It serves to do what is impossible in the real world, i.e. permitting a single modification of the *utos* configuration each iteration, rather than being subject to the real world frailties of implementation failure and measurement error. Further, we promote the usage of systematic, explicit and transparent documentation, applying principles of reproducible research, in a hope to aid in the scrutiny and reproduction of simulations throughout the research community. Ultimately, we aim to produce models capable of providing greater insight into causal description and, where possible, candidate causal explanations.

It is also important to note that, although we advocate the use of simulation models, this is not to the detriment of any existing methodologies. Rather, we hope simulation models will provide an additional tool to the researcher interested in examining crime. Moreover, the *in silico* experiment should complement the empirical experiment and vice versa. Additionally, the roles and configurations of systematic replication techniques presented within this paper should only be considered our initial model and are likely to be

improved in subsequent treatments of simulation model validation. What is of fundamental importance is that, however generative models of crime interactions are built, the approach is a systematic and scientific one. The authors hope that the methods and metrics presented will evolve to facilitate their practical usage amongst those interested in examining the formation of crime patterns. While critics may argue with our precise formulation, it at least introduces many of the attributes (i.e., transparency, explicitness) of good science. The task may seem Herculean judging by the plethora of conditions listed that require satisfying, but this merely highlights the scepticism with which claims from single simulation studies need to be treated.

# Acknowledgements

A version of this paper was presented at a meeting of the Crime Simulation Network and feedback of numerous delegates has improved the paper. The authors would also like to thank the three anonymous reviewers for their constructive comments that have improved the scope of the arguments set out in the paper.

# References

Axtell, R., Axelrod, R., Epstein, J.M. &Cohen, M.D. (1996). Aligning simulation models: A case study and results. *Computational and Mathematical Organization Theory,* 1(2), 123-141

Axelrod, R. (1995). The Convergence and Stability of Cultures: Local Convergence and Global Polarization. Santa Fe Institute working paper 95-03-028.

Birks, D.J., Donkin, S., Wellsmith, M.J. (2008) “Synthesis over Analysis: Towards an Ontology for Volume Crime Simulation”. In L. Liu & J. Eck (Eds.), *Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems*. Hershey, PA: Idea Group Publishing.

Braga, A.A., Weisburd, D.L., Waring, E.J., Mazerolle, L.G., Spelman, W., & Gajewski, F. (1999). “Problem-Oriented Policing in violent crime places: A randomised controlled experiment”. *Criminology*, 37(3), 541-580.

Braga, A.A. and Kennedy, D.M. and Waring, E.J. and Piehl, A.M. (2001), 'Problem-Oriented Policing, Deterrence, and Youth Violence: An Evaluation of Boston's Operation Ceasefire', Journal of Research in Crime and Delinquency, 38(3), 195-225

Cohen, L.E. & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, 44(4), 588-608

Cranefield, S., Haustein, S., Purvis, M. (2001) UML-based ontology modelling for software agents. In: Proceedings of the Autonomous Agents 2001 Workshop on Ontologies in Agent Systems.

Cronbach, L.J. (1982) *Designing Evaluations of Educational and Social Programs*, San Francisco, CA: Jossey-Bass Inc Pub

DeAngelis, D. L., & Gross, L. J. (Eds.). (1992). *Individual-based models and approaches in ecology: populations, communities, and ecosystems* . New York: Chapman and Hall.

DeAngelis, D. L., & Rose, K. A. (1992). Which individual-based approach is most appropriate for a given problem? In D. L. DeAngelis & L. J. Gross (Eds.), *Individual-Based Models and Approaches in Ecology: Populations, Communities, and Ecosystems* (pp. 67-87). New York: Chapman and Hall.

Eck, J. (2007). Simulating Police Outcomes: A Framework for Understanding Policing Strategies. (Paper presented at Crime Hot Spots: Behavioral, Computational and Mathematical Models conference, Institute of Pure and Applied Mathematics, UCLA, Los Angeles, Feb 2007)

Edmonds, B. & Hales, D, (2003).Replication, Replication and Replication: Some Hard Lessons from Model Alignment. *Journal of Artificial Societies and Social Simulation*, 6(4). Retrieved 13 September, 2006, from<http://jasss.soc.surrey.ac.uk/6/4/11.html>

Epstein, J. M. & Axtell, R. (1995). *Growing Artificial Societies: Social Science From the Bottom Up*. The Brookings Institution: Washington, D.C.

Epstein, J. M. (2006). Generative Social Science: Studies in Agent-based Computational Modeling. Princeton, NJ: Princeton University Press.

Faggiano, F., Vigna-Taglianti, F.D., Versino, E., Zambon, A., Borraccino, A. & Lemma, P. (2005). School-based prevention for illicit drugs' use. *Cochrane Database of Systematic Reviews*, Issue 2. Art. No.: CD003020. DOI: 10.1002/14651858.CD003020.pub2.

Forrester, D., Frenz, S., O'Connell, M. & Pease, K. (1990). *The Kirkholt Burglary Prevention Project: Phase II,* Crime Prevention Unit Paper 23, London: Home Office.

Hales, D., Rouchier, J. & Edmonds, B. (2003). Model-to-Model Analysis. *Journal of Artificial Societies and Social Simulation*, 6(4). Retrieved 13 September, 2006, from<http://jasss.soc.surrey.ac.uk/6/4/5.html>

Huget, M. P., & Odell, J. (2005). Representing agent interaction protocols with agent uml, *Agent-Oriented Software Engineering V: 5th International Workshop, AOSE 2004, New York, NY, USA, July 19, 2004: Revised Selected Papers*.

Holland, P. (1986).Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945-960

Klüver, J. & Stoica, C. (2003). Simulations of Group Dynamics with Different Models. *Journal of Artificial Societies and Social Simulation*, 6(4). Retrieved 16 September 2006, from <http://jasss.soc.surrey.ac.uk/6/4/8.html>

Liu, L. & Eck, J. (Eds.) (2008). *Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems*. Hershey, PA: Idea Group Publishing.

Nagin, D.S, Farrington, D.P. and Moffitt, T.E. (1995) 'Life-Course Trajectories of Different Types of Offenders', *Criminology*, 33(1): 111-139

Odell, J., Parunak, H. and Bernhard, B. (2000) "Representing Agent Interaction Protocols in UML", Proceedings of the First International Workshop on Agent-Oriented Software Engineering (AOSE-2000), Limerick, Ireland.

Odell, J., Parunak, H. and Bernhard, B. (2000) Extending UML for agents. Proceedings Agent Oriented Information System Workshop at the 17th National Conference on Artificial Intelligence

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston, MA: Houghton Mifflin Co.

Schwab, M., Karrenbach, N., and Claerbout, J. (2000) "Making scientific computations reproducible",.*Computing in Science & Engineering*, 2(6):61-67

Sherman, L.W. and Berk, R.A. (1984) *The Minneapolis Domestic Violence Experiment*, Washington, DC: Police Foundation

Takadama, K., Suematsu, Y.L., Sugimoto, N., Nawa N.E. & Shimohara, K. (2003). Cross-Element Validation in Multiagent-based Simulation: Switching Learning Mechanisms in Agents. *Journal of Artificial Societies and Social Simulation*, 6(4). Retrieved 16 May 2007, from <http://jasss.soc.surrey.ac.uk/6/4/6.html>

Tilley, N. (1993). *After Kirkholt: Theory, Method and Results of Replication Evaluation*. Crime Prevention Unit Paper 47, London: Home Office

Townsley, M., & Johnson, S. D. (2008). The need for systematic replication and tests of validity in simulation In L. Liu & J. Eck (Eds.), *Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems* (pp. 1-18). Hershey, PA: Idea Group Publishing.

Wilson, D. B., MacKenzie, D. L, & Mitchell, F. N. (2005). *Effects of Correctional Boot Camps on Offending.* A Campbell Collaboration systematic review. Retrieved 20 May,2007, from http://www.aic.gov.au/campbellcj/reviews/titles.html

Table 1: Summary of initial Cops & Robbers in silico experiment to investigate Guardianship Mechanism of Routine Activity Theory.

| ***utos* Components** | **Initial *utos* Configuration** | **Other Potential *utos* Configurations** (These represent a suite of potential model iterations; in practice, these elements would be implemented in an incremental fashion.) |
|---|---|---|
| **Units** | ***Agent Parameters**:* None<br><br>***Agent Calculus**:*<br>Movement Mechanism (all agents) – random<br>Offending Mechanism (offenders) – fixed, homogeneous agent populations | ***Agent Parameters:*** 'Lambda' (offending rate - high, medium, low)<br><br>***Agent Calculus:***<br>Movement Mechanism (all agents) – deliberative (routine activities, zone/node-based)<br>Offending Mechanism (offenders) – lambda-based, heterogeneous offender populations |
| **Treatment** | ***Guardianship Mechanism:***<br>– Limited to law enforcement agents only<br>– Provided at guardian's exact location only. | ***Guardianship Mechanism:***<br>– Provided by all agents<br>– Provided in proximity to agent locations<br>– Varying degrees of guardianship capability |
| **Outcome** | ***For each potential victim**:* Number of victimisations and time occurred for each.<br><br>***For each offender:*** Number of crimes committed and time for each.<br><br>***For each law enforcement agent:*** Number of prevented crimes and time for each.<br><br>Aggregate measures for each population of units could be computed. Thus, a crime incidence rate, offending rate and a prevention rate could be produced and examined. These could be aggregated to whatever level of granularity is required. | ***For each potential victim**:* Number of victimisations and time occurred for each.<br><br>***For each offender:*** Number of crimes committed and time for each.<br><br>***For each law enforcement agent:*** Number of prevented crimes and time for each.<br><br>Aggregate measures for each population of units could be computed. Thus, a crime incidence rate, offending rate and a prevention rate could be produced and examined. These could be aggregated to whatever level of granularity is required. |
| **Setting** | ***Environmental Configuration:***<br>Homogeneous, static-size, torroidal, no backcloth<br><br>***Population Configuration:***<br>Agents begin at random locations.<br>Static agent population sizes. | ***Environmental Configuration:***<br>Heterogenous, Zoning, Physical Barriers, Transport Network<br><br>***Population Configuration:***<br>Non-uniform agent distributions & densities. |

Table 2: Phased studies of increasing complexity for real world and in silico experimentation

| Phase | Real World Experimentation | *In silico* Experimentation | Validation Stage |
|---|---|---|---|
| 1 | Review existing theoretical and applied research to identify a testable hypothesis. | Review existing theoretical and applied research to identify a testable hypothesis. | Initial simulation or single study |
| 2 | Develop methods and instruments to ensure accurate and valid procedures to test hypothesis; e.g. surveys, statistical models and data or laboratory procedure. | Develop attributes of the model to ensure accurate and valid means to test hypothesis; e.g., the formalism that operationalises the constructs. From a computer science perspective, this involves simulation component verification (are we building the model correctly?) and validation (are we building the correct model?) Use UML to document this phase. | |
| 3 | Conduct controlled trial under ideal/abstract conditions. This might entail sampling healthy young males for drug trials (homogeneous experimental units) or first year psychology students (abstract but controlled application). | Conduct simulation under ideal conditions. That is an initial simulation (a naive model) with homogeneous units of analysis within a monoculture environment with a low level of complexity. | Pure replications |
| 4 | Determine if hypothesis operates in particular sub-populations of interest that are similar to real world application. By purposively sampling different experimental units, some picture of the consistency of the causal relationship is obtained. | Determine if hypothesis operates in more complex scenarios. Introduce complexity incrementally. Moderator variables are operationalised by allowing variation of experimental unit attributes (or settings or treatments) and modify decision calculus accordingly. Each level of moderator variable generates a distinct simulation populated with fixed values for the moderator variable in question. | Practical replications or iterations |
| 5 | Large-scale study aimed at entire communities (maximum generalisability) in order to identify aggregate impact. | Implement a combined simulation where all values of moderator variable are present. Observe the sign and direction of the causal relationship in aggregate. | |

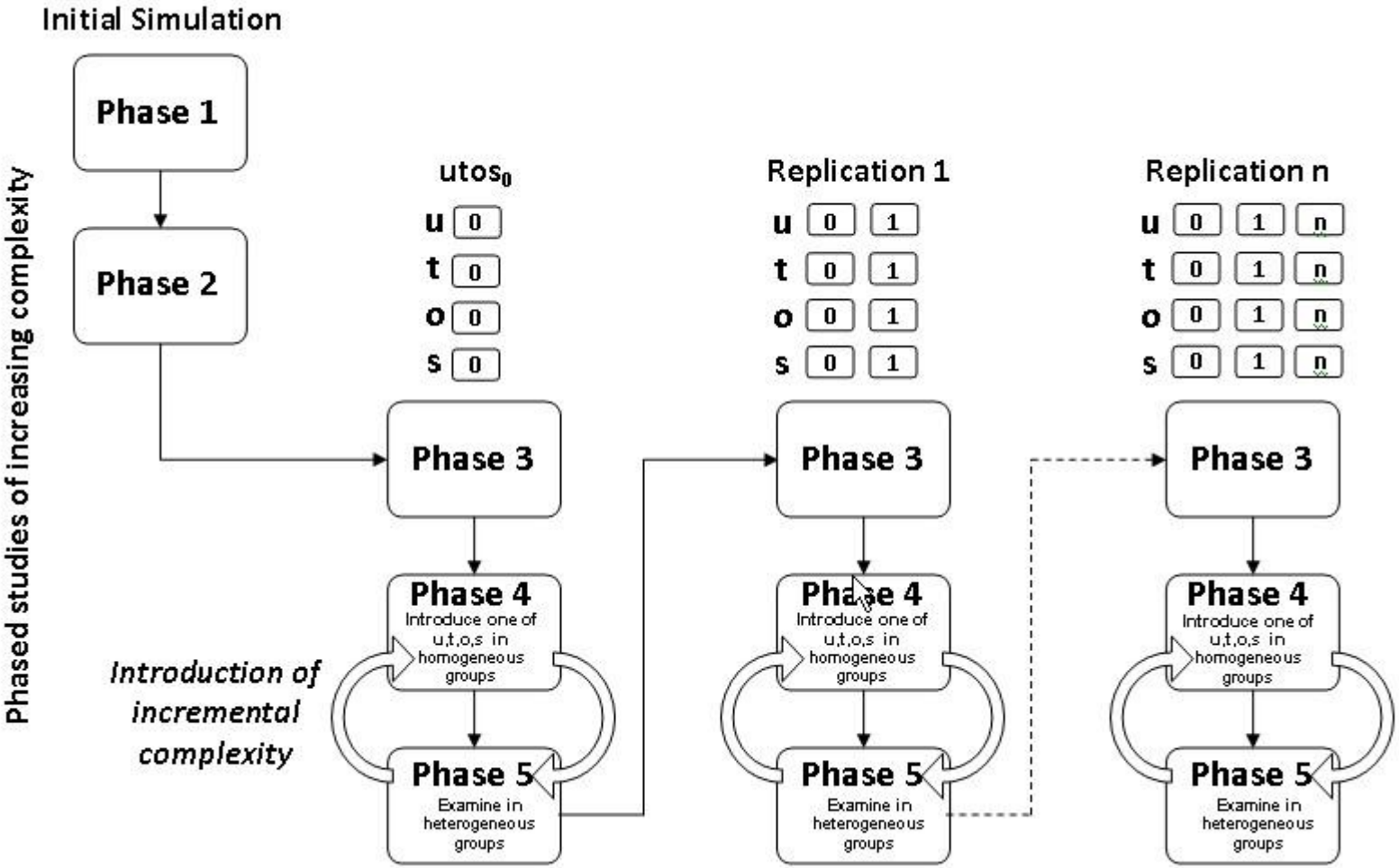


Figure 1: Representation of how the phased model is used in conjunction with *utos* notation.

[1] For the purposes of this article, we define a simulation model as a system defined by rules, some of which relate to interactions between components of the system. Allowing the system to 'run' the application of these rules generates output. All aspects of the system are defined and directly programmed by the researcher.

[2] These comments apply equally to any discipline which acts as a 'service' discipline to others. Obvious examples include applied statistics and mathematics. We do not believe all or even most computer scientists (or statisticians and mathematicians) are rapacious mercenaries obsessed with methodological gymnastics for the sake of it.

[3] Computer scientists use convergence and sensitivity test at this stage as well. Our focus in this article is primarily on the social scientists who we feel have the most to gain from testing the validity of models that focus on social phenomena.

[4] Freely available from http://ccl.northwestern.edu/netlogo/

[5] Ibid

[6] For similar reasons to Shadish et al. (2002) we represent Cronbach's notation in a modified form, which we believe makes it more interpretable.

[7] This list is by no means exhaustive and serves only to aid in the presentation of the case studies.

[8] See www.campbellcollaboration.org.

[9] Later simulations may also recreate treatments such as crime prevention interventions in a more analogous way to traditional uses of Treatment within utos. Importantly though, such intervention simulation relies heavily upon a sufficiently validated and verified base model of all the interactions associated with crime and its key elements. Therefore, the initial models we describe only strive towards theoretically, rather than operationally, relevant applications.

[10] Of course, in reality there were other differences between the two studies so it is unlikely that purely setting differences were responsible for the difference.