Chapter 11
How Many Times Should One Run a Computational Simulation?

Raffaello Seri and Davide Secchi

Abstract This chapter is an attempt to answer the question "how many runs of a computational simulation should one do," and it gives an answer by means of statistical analysis. After defining the nature of the problem and which types of simulation are mostly affected by it, the chapter introduces statistical power analysis as a way to determine the appropriate number of runs. Two examples are then produced using results from an agent-based model. The reader is then guided through the application of this statistical technique and exposed to its limits and potentials.

Why Read This Chapter?

To understand and reflect on the importance of determining an appropriate number of runs for a simulation of a complex social system, especially agent-based simulation models. Also the chapter guides readers through (a) the issues surrounding this determination, (b) the use of statistical power analysis to identify the number of runs, and (c) two examples to practice the computation.

11.1 Introduction

This chapter explores the issue of how many times a simulation should run. This is an often neglected issue (Ritter et al. 2011) that, sooner or later, all modelers dealing with simulations of complex systems encounter. The literature takes an agnostic stance on how many runs—per configuration of parameters or, as economists put it, *ceteris paribus*—a simulation is to be run. In fact, the focus has mostly been on

R. Seri
University of Insubria, Varese, Italy

D. Secchi (✉)
University of Southern Denmark, Slagelse, Denmark
e-mail: secchi@sdu.dk

B. Edmonds, R. Meyer (eds.), *Simulating Social Complexity*,
Understanding Complex Systems, https://doi.org/10.1007/978-3-319-66948-9_11

defining the "steps," the time, or the interactions within each run through sensitivity and convergence analysis, for example (Mungovan et al. 2011; Robinson 2014; Shimazoe and Burton 2013).

The central assumption of what is proposed in this chapter is that the number of runs in a simulation is often crucial for results to bear some meaning. Of course, this is not true for all simulations and it depends on scope, nature of the simulated phenomenon, purpose, and level of abstraction. We specify these aspects in the following section. For now, it suffices to write that for social simulations with a strong stochastic component, where emergence and complexity cause results to differ even within the same configuration of parameters, knowing how many runs are enough for differences to emerge (or not) becomes an extremely relevant information. This is where this chapter positions itself.

We first try to indicate—very broadly—what type of simulations this approach may apply to. Then, mediating from research on sample size determination for the behavioral sciences (Cohen 1988; Liu 2014), we introduce statistical power analysis and testing theory. The chapter also takes an agent-based model (ABM) with a strong stochastic component and provides two examples that show how crucial the issue is. At the same time, the chapter offers a practical guide on how to conduct the computation. Implications and concluding remarks follow.

11.2 Scope and Nature of Agent-Based Models

In this chapter we identify a particular sub-group of agent-based models that are fit for hypothesis testing. In order to frame the following discussion, we propose a classification of the aims of ABM, with the caveat that the following discussion may not be general or exhaustive. For a more general classification of the types of simulation, one may refer to Chap. 3 of this handbook (Davidsson and Verhagen 2017).

Some agent-based models have the purpose of studying the emergent properties of a system (Anderson 1972; Fioretti 2016). These properties arise when the system as a whole displays a behavior that is not explicit in its single components, in this case, the agents. When this is aimed at establishing whether an outcome is possible, hence the simulation has an exploratory purpose that reflects on theory, then the visual inspection of the trajectories of the simulated system or the computation of some descriptive statistics is sufficient to illustrate the existence of an emergent behavior. For example, Heckbert's (Heckbert 2013) model of the socio-economic system in which the ancient Maya civilization developed and disappeared can be thought of as a simulation of this kind. As a descriptive model, it establishes whether the conditions set in the model offer reasonable explanations of historical facts.

The study of emergent properties is also linked to another objective of some agent-based models, namely hypothesis generation (Bardone 2016; Secchi 2015). A researcher may run a model just to assess whether it is reasonable to suppose that some variables have an impact on a given outcome. The hypotheses obtained

in this way may be subject to empirical testing in a future laboratory experiment or through real data. An example of this type of ABM can be the simulation of a team of doctors and nurses working in the emergency room of a hospital, to isolate those socio-cognitive attitudes that may lead to increased performance (Thomsen 2016). Another example comes from political science (de Marchi and Page 2014) and it concerns the model of incumbent advantage in elections proposed by Kollman et al. (1992), that was first studied by simulation and then successfully tested empirically.

The techniques presented in this chapter are not necessarily pertinent to these first two model types described above, because they call for an exploratory approach in which the configurations of parameters and the number of runs are not rigidly chosen in advance, but they may be modified by trial and error while the researcher explores the potential outcomes of the model.[1]

A third objective of ABM is measurement, namely providing a numerical value for a quantity of interest. Since most agent-based models in the social sciences are too simplified a representation of reality to provide accurate estimates of real-world quantities, the models that pursue this objective are generally constrained to specific disciplines in which the rules of behavior of the agents are simple or particularly well known (for an example in biology see Sect. 1.1.1 in Railsback and Grimm (2011); for examples in transportation research, see Maggi and Vallino 2016). In this case, even if statistical tests can still be of interest, the researcher may direct his/her statistical analysis towards different tools. On the one hand, data from an ABM may be compared, through a distance (e.g., Lamperti 2015), with real time series to assess whether the two are similar enough. On the other hand, the researcher may settle on a sample size that guarantees a certain precision in the value computed for the quantity of interest rather than a certain level of power (see Sect. 11.5.1 below).

Finally, a fourth objective of ABMs is to test hypotheses in a controlled environment often emulating, with simplified rules, a real-world situation. The advantage of ABMs in this respect is that they allow the researcher to analyze a realistic situation by removing all the confounding factors arising in the observation of the real-world phenomenon. In this case, agent-based models can be considered the computational equivalent of laboratory experiments (Gilbert and Terna 2000). In the following pages, we explain how the parallel can be established. Note, however, that this is not the only possible setting. It is customary that an ABM has several parameters entering its formulation. The aim of a model can be that of exploring whether these parameters bear any impact on a quantity of interest, obtained as an outcome of the simulations of the model. The usual way is to identify some configurations of parameters that would correspond to different alternative treatments in an experiment, and to run several simulations of the model under each configuration. Each run of the model corresponds to an observation (e.g., a subject) in an experiment: the measured outcome can either be the terminal value of the series or a value computed on (a part of) the trajectory. The presumed independence

[1]Note, moreover, that the researcher should not test a hypothesis on the data that have been used to generate it.

of the simulation outcome on the configurations of parameters can then be tested in an ANOVA framework. An example of this type of ABM can be a model of intra-organizational bandwagons (Secchi and Gullekson 2016), in which authors run the simulation multiple times in order to test propositions as guide for future, probably empirical, research.

Another distinction that might be helpful when considering the number of runs can be drawn between models that strive at defining abstract and simple rules of behavior for their agents and those that are more concerned with describing a particular aspect of reality with fine degrees of details. This is the divide between the KISS ("Keep it Simple, Stupid!") and the KIDS ("Keep it Descriptive, Stupid!") principles (Edmonds and Moss 2005). While advocates of the first approach are in line with modeling efforts of the past (Troitzsch 2017; Coen 2009), those who indicate that ABM opens a new way stand with a more descriptive and complex approach to modeling (Edmonds and Moss 2005). If we take that the extreme for a KISS model is a system of deterministic equations and, on the other side, the bound for a KIDS model is the attempt to replicate reality, there is an entire spectrum of models (and ABMs) falling between these two extremes. However, considerations on the number of runs are more likely to become relevant as modelers tend toward increased complexity, without reaching the extreme of full description.

In summary, the determination of the number of runs in a simulation is warranted every time the researcher is seeking to measure—with some degree of confidence—whether different configurations of parameters are more or less likely to affect the outcome.

11.3 Testing Theory: Controlling for Alpha and Beta

In this section, we provide an introduction to testing theory that can be read independently from the rest of the paper. As an example, in this section, the term *parameter* denotes an unknown characteristic of a population, and not a quantity whose value is fixed before data are collected, as customary in ABMs. Therefore, we suppose that the researcher has identified some parameters describing the behavior of the population from which the data have been sampled (as a trivial example, the mean and the variance of the population). We also assume that he/she has formulated a null hypothesis H_0, i.e. an assertion about the value of the parameters.

The original approach to testing, pioneered by K. Pearson and theorized by R.A. Fisher, looks for a statistic T with the following property: when the hypothesis H_0 is verified, the value t that the statistic T assumes in the sample is near to a fixed value, generally identified with 0. Therefore, small values of t appear to bring support to the null hypothesis H_0, while extreme values of t are seen as witnessing a possible violation of H_0. This explains why the most sensible summary of the test, in Fisher's approach, is the p-value, i.e. the probability of observing, under H_0, values of T that

Table 11.1 Table of possible outcomes for a Neyman–Pearson test

	H_0 true	H_1 true
H_0 chosen	True negative	Type-II error or false negative
H_1 chosen	Type-I error or false positive	True positive

are as extreme or more extreme than the one computed on the sample. The *p*-value is sometimes (erroneously) perceived as a measure of strength of support in the null hypothesis. It is however clear that this method does not have the possibility to offer anything more than mild support to H_0, especially because of the absence of an hypothesis that holds true when H_0 does not.

This approach to testing was amended by J. Neyman and E.S. Pearson, who modified it to allow for the possibility of decision and action. The new theory starts with the introduction of the null hypothesis, H_0, and the alternative hypothesis, i.e. H_1, that is supposed to be true when H_0 is not. The hypothesis H_0 is generally, but not always, associated with the absence of an effect (of one variable on another, for example), while H_1 is generally associated with its presence. The researcher is uncertain as to whether H_0 or H_1 holds true. The decision between these two hypotheses is performed, as in a trial, on the basis of the available data (we will see later how).[2] This leads to a table of possible outcomes, see Table 11.1. The use of *positive* and *negative* to denote respectively the choice of H_1 and H_0 comes from the medical use of the same terms, where they indicate the positive or negative result of a medical test. A negative, i.e. a result in which the disease is not detected, can be either true or false, when the unobserved true hypothesis coincides or not with the choice of the procedure; the same holds true for a positive. A *false positive* is also called, with a more statistical term, a Type-I error, while a *false negative* is also called a Type-II error. These two "sources of error" (Neyman and Pearson 1928, p. 177) exist whichever method is used to choose between H_0 and H_1.

The standard procedure to decide between H_0 and H_1 is to consider a statistic T whose distribution is known under H_0 (let us denote the probability as $\mathbb{P}_{\mathsf{H}_0}$). The researcher builds an *acceptance region* $\mathcal{A}$ such that, when t belongs to $\mathcal{A}$, then H_0 is chosen as the true hypothesis. The possible values of t that are not contained in $\mathcal{A}$ form a *rejection region* $\mathcal{R}$. Therefore $\mathcal{A}$ and $\mathcal{R}$ make up the entire space in which T varies and are generally chosen in such a way that[3]:

$$\mathbb{P}_{\mathsf{H}_0}\{T \in \mathcal{A}\} = 1 - \alpha$$
$$\mathbb{P}_{\mathsf{H}_0}\{T \in \mathcal{R}\} = \alpha$$

[2] The metaphor of the trial has been introduced in Neyman and Pearson (1933, p. 296) but has been criticized as misleading in Liu and Stone (2007).

[3] Here $\in$ means "belongs to," so that $T \in \mathcal{A}$ means "T belongs to $\mathcal{A}$."

where $\alpha \in [0, 1]$ is called *Type-I error rate*, i.e. the probability of rejecting the null hypothesis when it is true, or *significance level*.[4]

Suppose now that the alternative hypothesis is verified and let $\mathbb{P}_{\mathsf{H}_1}$ be the probability distribution under the alternative hypothesis. If the null hypothesis is associated with the absence of an effect (of a variable on the outcome), the alternative hypothesis implies generally that there is an effect. It is generally the case that this effect can be measured through a quantity d called *effect size*,[5] that enters the formulation of $\mathbb{P}_{\mathsf{H}_1}$. If we suppose that the alternative hypothesis is true, the probability that T belongs to $\mathcal{A}$ or to $\mathcal{R}$ is:

$$\mathbb{P}_{\mathsf{H}_1}\{T \in \mathcal{A}\} = \beta(d)$$
$$\mathbb{P}_{\mathsf{H}_1}\{T \in \mathcal{R}\} = 1 - \beta(d)$$

where β, belonging to $[0, 1]$, is the *Type-II error rate*, i.e. the probability of accepting the null hypothesis when it is false. Note that β depends on the effect size d. The quantity $1 - \beta$, especially when seen as a function of the effect size d, is called *(statistical) power* of the test and measures the probability that the test correctly identifies the presence of an effect when there is one.

In a hypothetical simulation, for example, one may want to study how decision making makes employee motivation more effective under conditions of more or less organized corporate structures (Herath et al. 2017). The null hypothesis may be that the average motivation does not vary under alternative corporate structures, and this can be tested using, say, ANOVA. The Type-I error rate or significance level α, that is usually set at 5%, can be used to obtain an acceptance region $\mathcal{A}$. If the null hypothesis is false, the effect size d—i.e. the "strength" of the effect—measures the impact that the different conditions exercise, on average, on the outcome. The probability that the statistic takes a value inside $\mathcal{A}$ under the alternative hypothesis is β and depends on d. If the effect as measured by d is small, the alternative hypothesis is near to the null hypothesis, and the probability β that the statistic T falls inside $\mathcal{A}$ under H_1 is near to the probability $1 - \alpha$ under H_0. If d increases, β decreases.

It is clear that there are several degrees of freedom in the choice of the probability α, of the statistic T for testing H_0, and of the acceptance region $\mathcal{A}$. Now, while the test statistic is often suggested by the problem under scrutiny, the probability α is chosen routinely from a set of possibilities that have been determined by tradition more than by reflection. As to the choice of $\mathcal{A}$ and $\mathcal{R}$, it is often the case that $\mathcal{R}$

[4]We note that Neyman (1950, p. 259) used the term "accept" where most modern treatments propose to use "fail to reject" or "do not reject." The original choice of the author is in line with his idea of testing as leading to decision, while the modern use appears to be incorrectly borrowed from Fisher's approach (Fisher 1955, p. 73). However Pearson was more cautious (Pearson 1955, p. 206) and this even suggested to some authors the idea that he had rejected the approach pioneered with Neyman (Mayo 1992).

[5]More generally, the effect size d measures the distance of the true distribution from the distribution under the null hypothesis, and is generally a function of the parameters.

contains the most extreme values of T, i.e. the tails of its distribution. Most users of statistics stop here, and perform a test verifying whether t belongs to $\mathcal{R}$ or to $\mathcal{A}$ and, as a consequence, respectively reject H_0 or fail to reject it, as part of a ritual (Gigerenzer 2004). In this situation, an alternative way of reaching the same result is to compare the p-value, whenever defined, to a fixed threshold α: if the p-value is smaller than α, we reject H_0, otherwise we fail to reject it.

It is interesting to review the relations among the quantities seen until now. We saw before that the effect size d has an impact on β. Since d is a measure of how easy it is to discriminate between H_0 and H_1, it is generally the case that power, $1-\beta$, increases with d when α is fixed.[6] Another factor affecting α and β is the sample size N. In this case too, $1-\beta$ generally increases with N, when α is fixed. At last, the formulas $\mathbb{P}_{\mathsf{H}_0}\{T \in \mathcal{A}\} = 1-\alpha$ and $\mathbb{P}_{\mathsf{H}_1}\{T \in \mathcal{A}\} = \beta$ show that there is a trade-off between α and β. Indeed, when $\mathcal{A}$ gets larger, α decreases while β increases, and vice versa. This explains why, when N and d are fixed, it is not possible to reduce α without consequences on the Type-II error rate β.[7]

This is the reason why one cannot make α as small as possible, that is because this inflates β. This fact suggests that good results could be achieved by balancing the two error rates. This was indeed proposed by Neyman and Pearson in 1933,[8] and has been revived several times since then. A more recent attempt in this direction is the *compromise power analysis* of Erdfelder (1984). However, the most common approach is to consider the two sources of error differently.

A first approach completely disregards β: a value for α is rigorously fixed (often as $\alpha = 0.05$), and the test checks whether t belongs to $\mathcal{A}$ or not using a sample whose size N has been selected without reference to β. This approach is the one that most closely resembles the original Fisher paradigm, as the alternative hypothesis has practically no role in it. It is based on the fact that, as N increases, β goes to 0, so that a large sample size guarantees that β will be small enough. A second approach supplements this part of the analysis with the computation of power using a value of d estimated on the basis of the data, a procedure called *post hoc power analysis*. Because of the large variability of the estimated effect size, this approach is generally regarded with suspicion by statisticians (Korn 1990; Hoenig and Heisey 2001). In the third approach, the researcher fixes α and β, hypothesizes a value of d, and chooses $\mathcal{A}$ and N so that both $\mathbb{P}_{\mathsf{H}_0}\{T \in \mathcal{A}\} = 1-\alpha$ and $\mathbb{P}_{\mathsf{H}_1}\{T \in \mathcal{A}\} = \beta(d)$ hold true. This procedure, called *a priori power analysis*, guarantees that, if d is correctly guessed, the desired values of α and β will be achieved.

[6] This also explains why in some cases it is possible to increase the power of a test by designing an experiment in which it is expected that the effect size d, if not null, is large. As an example, in ABM this could be done by setting some of the quantities entering the model to their extreme values.

[7] See also van der Vaart (2000, p. 213) or Choirat and Seri (2012, Proposition 7, p. 285).

[8] The authors say: "The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator" (Neyman and Pearson 1933, p. 296).

It is important to recall that the Neyman–Pearson theory of testing is essentially designed to provide the researcher with a decision rule guaranteeing, in the long run, a specified error probability under the null hypothesis. The decision rule equating the rejection of H_0 with the occurrence of a value of t inside $\mathcal{A}$ makes sure that when a large number of tests are performed, the null hypothesis is incorrectly rejected 100α percent of the times, but does not guarantee a good performance in the case of the single test. Otherwise stated, in the Neyman–Pearson approach a controlled long-run performance is obtained if the researcher chooses α and $\mathcal{A}$ and decides on the basis of the fact that t belongs to $\mathcal{A}$ or not (or, equivalently, on the basis of the fact that the p-value is larger than α or not). In general, it is also expected that the researcher sets a value of β and chooses, on the basis of experience or pilot runs, a value of d, and computes N on the basis of these values.

However, this is not the way in which tests are generally performed in practice. Indeed, it is customary that the researcher computes the test statistic t and the p-value and uses the latter as a measure of the support in the null hypothesis. For example, it is quite common that a p-value just under 5% is treated differently than a p-value under 1%, the latter providing a stronger evidential value against the null hypothesis. This is so widespread that some researchers do not report the p-value but only $p < 1\%$ or $p < 5\%$. From the point of view of the Neyman–Pearson theory of testing this is nonsensical. However this has entered common practice and has evolved into an approach of its own, different from the Fisher and the Neyman–Pearson approaches, yet gathering aspects of both, and called *Null-Hypothesis Significance Testing* (NHST). This approach takes from the Fisher approach the emphasis on the p-value and its disregard for power; from the Neyman–Pearson theory, the approach emphasizes the threshold values of α.

In this chapter, we follow more closely the original Neyman–Pearson theory than the NHST. The elements of this approach are the two probabilities of error α and β, a measure of the effect under scrutiny or of the distance between the alternative and the null hypothesis d, and the sample size N. These quantities are linked by some equations. We will see below that determining a value N amounts at choosing some values for the quantities α, β and d, whose interpretation is generally simpler than the one of N.

11.4 The Use of Power in Practice: Two Examples

In order to show how power analysis can help to determine the number of runs in a simulation, we decided to select a model and to proceed with some calculations. The simulation we selected for this computational exercise is an agent-based model that was developed by Fioretti and Lomi (2008, 2010) on the basis of the famous "garbage can" model (Cohen et al. 1972), hereby GCM.

There are several reasons that led to the selection of this ABM. One of the obvious reasons is that it describes a very well-known model that informed the

decision making literature and had an extremely significant impact.[9] As a result of this, the basic assumptions of the model should be easy to understand for most scholars. Moreover, the agent-based implementation by Fioretti and Lomi attracted some attention because it does not support all the conclusions of the original model. Another reason—and this is not a secondary reason—is that authors made the code available so that anyone interested could download and run the simulation in `NetLogo`, an ABM software (Wilensky 1999). Finally, the work of Cohen, March, and Olsen is very much in line with the legacy of Simon (1976, 1978, 1997), thus consistent with the introduction to this handbook (Edmonds and Meyer 2017).

The two examples that follow are both hands-on cases that should inform readers on how to determine the number of runs in an agent-based simulation.[10] In Example 1, the model runs a limited number of times so that insufficient power leads to the risk of not rejecting hypotheses that should be rejected. In Example 2, the model is run a very high number of times to produce over-powered results, reducing to a minimum the likelihood not to make any effect statistically significant.

11.4.1 Short Description of the Model

The "garbage can" is a model of decision making in organizations (Cohen et al. 1972). There are four types of agents: (a) problems, (b) opportunities, (c) solutions, and (d) participants. The overall goal of the model is to determine whether a formal (hierarchic) organizational structure provides the institutional backbone for problem solving that is better than an informal (anarchic) organizational structure or not. In the first case, the four types of agents interact following a specified sequence while in the other they interact at random.

The aim of the model is to match the four elements mentioned above to study the most effective way for an organization to make decisions. Originally, the model was designed to understand whether opportunities become more available to decision makers when organizations relax hierarchical and structural ties. This is what the ABM simulation attempts to study as well. Figure 11.1 shows a screenshot of the model interface; each agent has a different shape and they move on the black environment.

There are two ways in which participants make decisions in the organization. One type of decision is called *by resolution* and it happens when problems are solved once participants match opportunities to the right solutions (Cohen et al. 1972). This happens graphically when the right combination of the four agents are on the same position at the same time (i.e., they overlap, see Fig. 11.1). Another type is

[9]The number of citations of the original paper (Cohen et al. 1972) in Google Scholar amounts at 9196 and those from Thomson's Web of Science are 1864.

[10]Even though we use this method for ABM, it may reveal to be useful for any simulation with emergent properties derived from a relevant stochastic component.

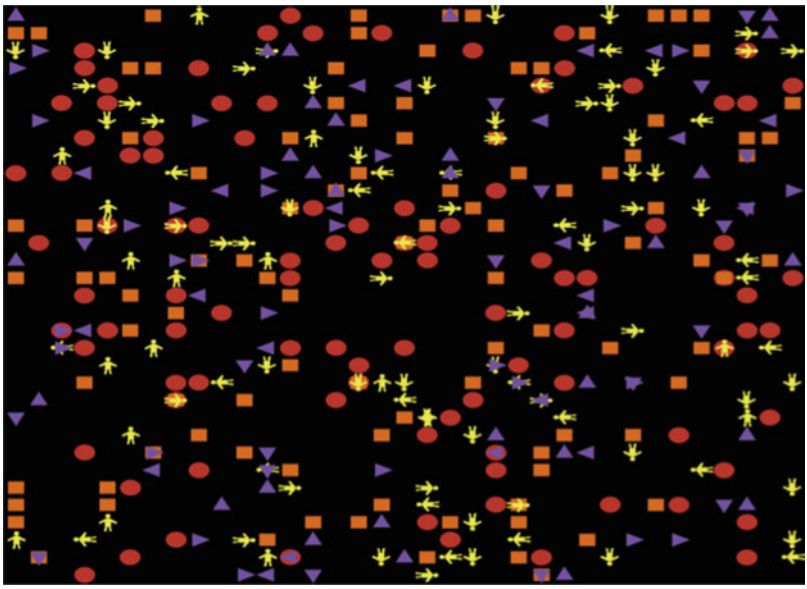

Fig. 11.1 `NetLogo` screenshot for the "Garbage Can" model by Fioretti and Lomi (2010)

when decisions are made *by oversight* and it is when solutions and opportunities are available to participants but no problems are actually solved (Cohen et al. 1972). Not all problems are solved automatically, just by having opportunities, participants, and solutions available. In fact, all problems have difficulty levels, participants have abilities, and solutions have a certain degree of efficiency. The problem is solved if the match of the participant with an opportunity and a solution is greater than the difficulty of the problem (Fioretti and Lomi 2010).

In the agent-based version of the model, there are three types of structure:

- **Anarchy**. There is no hierarchy so that abilities, efficiencies, and difficulties are randomly distributed among agents.
- **Hierarchy-competence**. The hierarchical structure is such that abilities, efficiencies, and difficulties increase as one moves up the hierarchical ladder.
- **Hierarchy-incompetence**. The hierarchical structure is such that abilities, efficiencies, and difficulties decrease as one moves up the hierarchical ladder.

Finally, the model implements two modes of (not) dealing with problems. One is called *buck passing*, and it happens when one participant has the alternative of passing the decision on a problem to another participant. The other mode is *postpone*, and it refers to problems that are kept on hold by participants and eventually solved at an unspecified future time.

For the purpose of this chapter, we calculate the ratio of decisions made by resolution on those made by oversight in the three cases of anarchy, hierarchy with competence, and hierarchy with incompetence.

11.4.2 Example 1

We performed a simulation for the ABM version of the GCM (Fioretti and Lomi 2010), using the second version of the two models uploaded on the `NetLogo` community platform. The model has three overall conditions—anarchy, hierarchy-competence, and hierarchy-incompetence—and each of these has four parameter configurations, with *buck passing* [*true*, *false*], and *postpone* [*true*, *false*]. We decided to test a simple case, setting both parameters to *false*. This gives a design of 3 configurations of parameters (CoP). Each run had 5000 steps as per the original simulation (Fioretti and Lomi 2010).

Power analysis should be performed before obtaining data from the model to choose how many times a simulation should be run. To do that, there are a few elements to determine. First of all, the researcher should choose a certain number G of configurations of parameters (also called groups). Then, considering the nature of the model or previous simulations one should guess a value of the effect size d that, in the case of ANOVA, is identified by the letter f (Cohen 1988; Liu 2014). At last, one should choose a level for α and a corresponding goal for the level of power—i.e. $1 - \beta$—to be achieved. Although the power threshold of $1 - \beta$ for empirical research is set at 0.80, some (Secchi and Seri 2014, 2017) argue that it can be set at 0.95 for simulations, because the control exerted on variables and parameters is much higher than that usually in place in empirical research. Consistently with this, also the threshold for α can be set at the more stringent level of 0.01 (Secchi and Seri 2017).[11]

As explained above, the dependent variable is the ratio r_{ro} of decisions by resolution in relation to those made by oversight. The differences in its average value across the three CoP can be easily explored by performing a one-way ANOVA with the null hypothesis that the expected value is the same across conditions. We set some notation. If G denotes the number of groups/CoP and n the number of observations per CoP, the sample size N turns out to be $N = n \cdot G$.

[11] In an interesting exchange with Bruce Edmonds, we came to realize that this approach might raise some important issues. One of the concerns is that thresholds do not usually adjust because the experiment is so well planned that results come out to be extremely clear; that is to say that good experimental work still accepts or rejects hypotheses at the level $\alpha < 0.05$ with $1-\beta \approx 0.80$. This implies that adjustments of these levels for simulation work appears to be arbitrary. Our position on this critique is that thresholds actually change as it happens in some medical studies, where $1 - \beta$ raises to 0.90 (Lakatos 2005), or when we listen to the calls *not* to interpret the traditional choices of α levels as absolute from either social scientists (Gigerenzer 2004) or statisticians (Wasserstein and Lazar 2016). While a complete review of the reasons leading to the traditional choices of α and β is in Secchi and Seri (2017), the introduction to testing theory above should have made clear that the fathers of this theory thought of α and β as quantities to be chosen according to the problem at hand. This justifies our proposals as long as we cannot compare artificial computational experiment to real-life experiments because of different variability of observations, observer's control and role, and the usual difficulty of increasing sample size for empirical experiments.

11.4.2.1 Identifying the Appropriate Effect Size

We are left with the problem of guessing a value for the effect size. In the following we propose some reasonings concerning the choice. For this purpose, we need some basic notation. For ANOVA, the effect size f is to be calculated as $f = \frac{\sigma_m}{\sigma}$ (Cohen 1992), where σ_m is the standard deviation of the group means and σ is the within-population (or pooled) standard deviation. The same quantity can be expressed using sums of squares. Let the subscript j, ranging from 1 to G, indicate the j-th group/CoP. Let n be the number of observations in each group/CoP and let i, ranging from 1 to n, denote the i-th observation. Therefore, x_{ij} is the i-th observation in the j-th group. Moreover, we consider the sample means $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $\bar{x} = \frac{1}{nG}\sum_{i=1}^{n}\sum_{j=1}^{G} x_{ij}$, where the latter is the grand mean of all the observations. Now, the variation of the observations between configurations (or groups) is simply the square root of the Sum of Squares Between (SSB), or $n\sum_{j=1}^{G}\left(\bar{x}_j - \bar{x}\right)^2$, divided by the Sum of Squares Within (SSW) observations, or $\sum_{j=1}^{G}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$.

A first possibility is to use "canned" effect sizes (Cohen 1988). In this case, verbal descriptions of the strength of the effect (e.g., small, medium and large; see below) are mapped onto numerical values, usually based on effect sizes retrieved from a review of the literature. As per the literature on the GCM, we can expect that many decisions are made by oversight and that the number of difficult problems solved increases under anarchy (Fioretti and Lomi 2008, 2010; Herath et al. 2015). Hence, the "distance" between conditions could be classified as *medium* or *large*, and we can set to an effect size of 0.25 or 0.4 (consistently with Cohen 1992). The computation of the sample size providing the desired level of power can use the formulas in Cohen (1988) as implemented in the `R` package `pwr` on power analysis (see Champely et al. 2016). For $f = 0.25$, this yields the result $n = 112$ (exactly 111.68). A simpler approximation, proposed in Secchi and Seri (2017), yields $n = 109$ (exactly 109.47).[12] The latter approach makes the relation between components of power more explicit. For $f = 0.40$, `R` package `pwr` yields $n = 45$ (more precisely 44.58) and our formula yields $n = 43$ (more precisely 43.04).

A second possibility is to guess a value for the effect size by having the simulation run for a pilot study and by calculating the estimated effect size from the results. We will deal below with some problems involved in this approach. For expository purposes, we have decided to run the model for $n = 10$ runs per condition. Taking the definition of f above, the numerator SSB is 0.000813 and the denominator is 0.004341 so that the estimated effect size for these three conditions is 0.43. If we calculate the number of runs reaching a power of 0.95 with such a large effect size as $f = 0.43$, we obtain $n = 38$ (more precisely 38.31) using Cohen's formulas and $n = 37$ (more precisely 36.81) using our approximation.

A third possibility is to use the results of former studies on the same topic. Here, we can use the results of Fioretti and Lomi (2010) in the case without buck

[12] This formula can be used in `R` with an ad hoc function taken from one of our previous publications (Secchi and Seri 2017). See the Appendix for the code for both formulas.

passing and postponement. In that source, the authors state that, based on 100 runs, the average number of decisions by resolution (resp., by oversight) is 43.90 (resp., 779.57) under anarchy (group 1), 24.82 (resp., 461.94) under competent hierarchy (group 2) and 7.71 (resp., 192.77) under incompetent hierarchy (group 3). We can approximate the average value of the ratio r_{ro} through the ratio of the averages, i.e. $\bar{x}_1 \simeq 0.0563$, $\bar{x}_2 \simeq 0.0537$ and $\bar{x}_3 \simeq 0.0400$. Therefore, we expect the difference between the average value of r_{ro} in competent hierarchy with respect to anarchy to be around $\bar{x}_2 - \bar{x}_1 \simeq -0.0026$, and in incompetent hierarchy with respect to anarchy to be around $\bar{x}_3 - \bar{x}_1 \simeq -0.016$. These coefficients are remarkably near to the ones obtained in the tables below. From the Appendix, we can see that:

$$f = \sqrt{\frac{\sum_{j=1}^{G} \left(\bar{x}_j - \bar{x}\right)^2}{\frac{1}{n} \sum_{j=1}^{G} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2}}.$$

The numbers above allow us to estimate the quantity $\sum_{j=1}^{G} \left(\bar{x}_j - \bar{x}\right)^2$ as 0.000153. Instead, $\frac{1}{n} \sum_{j=1}^{G} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$, i.e. SSW divided by n, cannot be estimated from Fioretti and Lomi (2010), but we can use the value from our pilot runs with $n = 10$, i.e. $0.004341/10 = 0.000434$. The final result is $f = 0.594$ that would lead to $n = 21$ (more precisely $n = 21.07$; $n = 19.60$ with our formula). While one should not give too much credit to these numbers, they suggest that the effect size f may be larger than expected.

Another consideration may provide some hints about how to interpret the values provided by the previous three techniques. The standard error associated with estimated effect sizes is generally quite large. Nothing guarantees that the estimated f is indeed equal or even near to the true value. A good idea is therefore to investigate what happens choosing a value f in a neighborhood around the estimate. As an example, if we suppose that f is 0.35 or 0.5, our formula yields respectively n equal to 56 or 28. We will see below that it is generally better to overshoot the correct sample size than to undershoot it. From this point of view, a possibility is to use the estimated effect size to choose a *smallest effect size of interest* (SESOI, see Lakens (2014) for its definition in a different context), i.e. a value of the effect size that is the smallest one for which we want to achieve the desired level of power.[13] This means that for f larger than the SESOI we will experience overpower while for f smaller we will be in underpower. This asymmetry is justified by the fact that values of the effect size under the SESOI are deemed to be improbable or uninteresting. The SESOI is then used in the computation of the sample size. Whether the researcher chooses to use the SESOI or not, the importance of these sensitivity analyses can hardly be exaggerated, as they shed light on the factors that impact the choice of the sample size.

[13] A possibility is to choose, as SESOI, the lower bound of a confidence interval on the effect size with a specified confidence probability, e.g., 0.95 or 0.90.

Table 11.2 OLS Regression Results (DV: decisions by resolution/decisions by oversight)

	Model 5	Model 40
(Intercept)	0.052***	0.056***
St. err.	(0.006)	(0.002)
t value	8.721	29.804
Type: HC/AR	−0.005	−0.007**
St. err.	(0.009)	(0.003)
t value	−0.542	−2.692
Type: HI/AR	−0.013	−0.012***
St. err.	(0.009)	(0.003)
t value	−1.481	−4.637
R-squared	0.157	0.156
F-statistic	1.123	10.842
Degrees of freedom	2, 12	2, 117
p-value	0.357	0.000
N	15	120

Note. *HC* hierarchy-competence, *HI* hierarchy-incompetence, *AR* anarchy
Signif. codes: 0 "***" 0.001 "**" 0.01 " " 1

On the basis of the previous reasonings, taking into account the expository nature of this example, we decided to take $n = 40$, consistently with a value of f around 0.4. In Table 11.2 we reproduce the estimation results for a model with 5 runs (i.e. Model 5), that is clearly under-powered, and for a model with 40 runs (i.e. Model 40), that is correctly powered under an effect size f equal to 0.40. We expect therefore the second model to provide a test of the effect of parameters on the number of decisions by resolution in comparison to those made by oversight, with the desired levels of α and β.

11.4.2.2 The Impact of Under-Power on Outcomes

The previous discussion shows that 5 runs should still be insufficient to provide reliable results. Let us see how. As stated above, we are interested in understanding whether the number of decisions by resolution on those by oversight change (decrease) as we move from anarchy to hierarchy. Hence, we can perform an OLS regression[14] and produce a table with results calculated on 5 and 40 runs, to compare findings from an under-powered to those from an appropriately-powered study. Table 11.2 shows these comparisons and refers to them as Model 5 for the under-powered and Model 40 for the balanced simulation.

[14]See the Appendix for details on how the effect size of the ANOVA and OLS regressions map onto each other.

From results in Table 11.2 it is immediately apparent that there are differences between the two models. The under-powered Model 5 is not able to detect some of the effects that are instead captured by the more balanced Model 40. In fact, Model 5 fails to identify the relation between hierarchy with competence (HC) and anarchy (AR) as statistically significant as well as the relation between hierarchy with incompetence (HI) and anarchy (AR). In other words, the null hypothesis was accepted when (probably) false, hence falling into Type-II error. And we know that this is the case because a very similar regression coefficient ($\beta_{\text{HC/AR}} = -0.007$, St. err. $= 0.003$) leads instead to the rejection of the null hypothesis—that the corresponding parameter is zero—in Model 40, where it is more reasonable to suppose that power requirements are met. The second coefficient—hierarchy with incompetence on anarchy—is also statistically significant in Model 40 ($\beta_{\text{HI/AR}} = -0.012$, St. err. $= 0.003$) as opposed to Model 5 ($\beta_{\text{HI/AR}} = -0.012$, St. err. $= 0.008$).

At last, note that in Model 5 the F-statistic for the joint nullity of both effects does not lead to the rejection of the null hypothesis, thus suggesting that there is no effect overall of the structure on problem solving. The conclusion is at odds with the one from Model 40, that leads to the strong rejection of the same hypothesis.

In short, the impact of some of the conditions fails to be acknowledged in the under-powered study with only 5 runs, leaving important and interesting implications out of the study.

11.4.3 Example 2

We also conduct a second example to illustrate the risks and problems of over-powering the simulation. In this example, we over-power the simulation and calculate results on 500 runs, with the same parameter specifications used in the example above.

Results of the two simulations are explored in Table 11.3, where we show the estimation outputs of two OLS regression models. In the table, Model 40 shows results for the correctly-powered simulation while Model 500 refers to the over-powered simulation. The beta coefficients are very close to each other, with a variation that is mostly reflected in the standard errors, that decrease in the case of the over-powered simulation. This leads to a different t value so that the respective probability (the p-value) becomes closer to zero for Model 500 than for Model 40.

From the perspective of accepting or rejecting results in the regression, there is little or no difference. In fact, most values are well below the threshold for statistically significant results. This points at the fact that, if one is interested in accepting or rejecting hypotheses, there is no particular difference between the two.

However, in another article (Secchi and Seri 2017), we warn modelers of the risks of over-power. There we write that over-power hides some dangers because it might be unnecessarily costly (time consuming, for example), it makes small effects as significant as larger ones, and destroys the balance between the two probabilities

Table 11.3 OLS Regression Results (DV: decisions by resolution/decisions by oversight)

	Model 40	Model 500
(Intercept)	0.056***	0.055***
St. err.	(0.002)	(0.001)
t value	29.804	100.12
Type: HC/AR	−0.007**	−0.005***
St. err.	(0.003)	(0.001)
t value	−2.692	−5.99
Type: HI/AR	−0.012***	−0.015***
St. err.	(0.003)	(0.001)
t value	−4.637	−19.18
R-squared	0.156	0.205
F-statistic	10.842	192.497
Degrees of freedom	2, 117	2, 1497
p-value	0.000	0.000
N	120	1500

Note. *HC* hierarchy-competence, *HI* hierarchy-incompetence, *AR* anarchy
Signif. codes: 0 "***" 0.001 "**" 0.01 " " 1

of error α and β,[15] thus decreasing the overall reliability of the model. However, all things considered, Example 2 shows that, in the case of large effect sizes such as this one, overpower does not bear particularly relevant problems besides accuracy. In fact, the two models present results that are close to each other and only differ in the granularity and reliability of details.

One last remark concerns the value of f as estimated from Model 500. In that case, we get $f = 0.51$. This confirms that our initial guess (f between 0.25 and 0.4) was probably an underestimation, and validates with hindsight our choice of focusing on the upper bound of the interval $[0.25, 0.40]$.

11.5 Implications and Conclusions

A few implications can be drawn from the two examples above. The first is that power analysis can guide researchers on establishing the number of times a simulation should run. The most immediate advice to modelers is that using power to compute the number of runs should help avoid under-powered studies. In that

[15] Over-power reduces β well below the chosen value of α. This is a problem because Type-I errors are generally perceived as more serious than Type-II errors, and when $\beta \ll \alpha$ we expect exactly a higher incidence of serious errors and a lower incidence of less serious ones. That is the reason why, at least in the intentions of Neyman and Pearson, α and β should have been chosen in a balanced way.

case, Example 1 shows that results are unreliable and one might discard effects that are, in fact, relevant to the study. At the same time, Example 2 shows that—for studies with large effect sizes—overpower does not pose too relevant threats to the overall reliability of a study.

In any case, knowing what makes the ABM more likely to produce reliable results is a relevant information for modelers. It seems more so when modelers perform their simulation a limited number of times per configuration of parameters. But also when too many runs are performed, the absence of power calculations may mislead one's judgement on the effects and actual meaning of the simulation. However, the asymmetry of the effects between under- and over-power suggests that power analysis can be used to provide, if not a guess, at least a lower guess on the number of runs (see the concept of SESOI introduced above). The value that is calculated with the aid of statistical power analysis is a number that—if not taken at face value—should inform the choice on the number of runs, and could at least work as a benchmark.

In a review of models published mostly in *Computational and Mathematical Organization Theory* (CMOT) and in the *Journal of Artificial Societies and Social Simulation* (JASSS) between 2010 and 2013 (Secchi and Seri 2017) it was found that most models are under-powered. If a small effect size $d = 0.1$ is hypothesized, then the average power is $1 - \beta \approx 0.41$, while if a medium effect size $d = 0.3$ is taken, then power becomes $1 - \beta \approx 0.84$ (with $\alpha = 0.01$). In both cases, the review shows that models are under-powered even by the milder standards of $1 - \beta = 0.90$ suggested in Ritter et al. (2011).

11.5.1 Comparing Statistical Power to Other Approaches

Using power is not the only way in which one can determine the number of runs in an experimental study and, in particular, in an ABM.

As an example, another approach sometimes called *accuracy in parameter estimation* (AIPE) (Maxwell et al. 2008) has been proposed. In this approach, first the researcher identifies a quantity of interest (a coefficient in a regression, a correlation, etc.) and chooses the desired width of a confidence interval around this value. Then, the researcher selects the sample size that allows one to reach this objective. The technique is already established, under different names, in medicine (Bland 2009), engineering (Hahn and Meeker 2011, Sect. 8.3), and psychology (Maxwell et al. 2008). A similar approach, putting together AIPE and power analysis, has also been proposed in the context of simulation models in Ritter et al. (2011).

However, we think that, in order to become a feasible option for ABM, this method should overcome some difficulties. First, AIPE may be surely of interest whenever the objective of the analysis is to obtain a precise enough measure of the effect of a treatment (see above for references). However, most ABM studies are not framed in this way (see the distinction between KISS and KIDS above).

The reason is that ABM studies are often simplified representations of reality. Therefore, the effect of a treatment is rarely their desired outcome, as it is clear that the value obtained from an ABM will generally not be the same value observed in reality. Second, even when the outcome of an ABM study is of interest in itself, it is rarely the case that one has a precise idea of what the width of a confidence interval should be. This may be different whenever the outcome variable is measured on a well-known scale, as it is often the case in the disciplines in which AIPE is an established alternative to power analysis. The paper (Schönbrodt and Perugini 2013) (see also Lakens and Evers 2014) provides an interesting example, based on Cohen (1988), of how to determine the width of an interval, but this seems difficult to generalize to other situations.

11.5.2 Concluding Remarks

The message of this article is that statistical power analysis can help modelers to refine their ideas on how many times their ABM simulation should be performed. In this chapter, we first wrote a few notes on the importance of determining the number of runs, and then turned our attention to the type of models that would benefit the most from this approach. The focus is then moved to testing theory so that we could provide an appropriate statistical background for this approach. Finally, some practical examples show the risks and perils of under- or over-estimating the number of runs in a simulation. The implications are then further discussed at the beginning of this section.

As a way to provide a summary of this chapter and, at the same time, help modelers clarify what under- and over-power imply, Table 11.4 shows calculations of power for $\alpha = 0.01$ and $1 - \beta = 0.95$, using the formula that we developed and also appearing in the Appendix.

The left column in Table 11.4 shows the hypothetical number of parameter configurations (or groups G) that a potential ABM could have. Knowing how to determine the appropriate number of configurations is a complex issue that falls beyond the scope of this chapter. However, sensitivity and steady state analyses can provide sound support (Thiele et al. 2015). The table calculates the number of runs that are necessary to reach $1 - \beta = 0.95$ at $\alpha = 0.01$ for five different effect sizes, respectively *ultra-micro* $= 0.01$, *micro* (0.05), *small* (0.1), *medium* (0.2), *large* (0.4), and *huge* (0.8). Results from these calculations confirm with more granularity of details that small simulations, with few configurations of parameters (up to 10) need to be performed many times unless the effect size is large or very large. As the number of configurations grows, the number of runs to perform clearly decreases significantly to the point where one run per configuration is enough when variability is spread to its limits (from 1000 and up) in the presence of large and very large effect sizes.

Table 11.4 A map of statistical power: Number of runs for $\alpha = 0.01$ and $1 - \beta = 0.95$

	Effect sizes f					
CoP (G)	*ultra-micro*	*micro*	*small*	*medium*	*large*	*huge*
2	84,777.89	3,468.39	875.55	221.02	55.79	14.08
3	65,400.97	2,675.65	675.44	170.51	43.04	10.87
4	54,403.07	2,225.71	561.85	141.83	35.80	9.04
5	47,162.95	1,929.51	487.08	122.96	31.04	7.84
10	30,265.08	1,238.19	312.57	78.90	19.92	5.03
20	19,421.49	794.56	200.58	50.63	12.78	3.23
50	10,804.41	442.02	111.58	28.17	7.11	1.79
100	6,933.33	283.65	71.60	18.08	4.56	1.15
200	4,449.21	182.02	45.95	11.60	2.93	0.74
500	2,475.15	101.26	25.56	6.45	1.63	0.41
1000	1,588.33	64.98	16.40	4.14	1.05	0.26
3000	786.29	32.17	8.12	2.05	0.52	0.13
5000	567.02	23.20	5.86	1.48	0.37	0.09
10,000	363.86	14.89	3.76	0.95	0.24	0.06

Note. Effect sizes: *ultra-micro* $= 0.01$, *micro* $= 0.05$, *small* $= 0.1$, *medium* $= 0.2$, *large* $= 0.4$, *huge* $= 0.8$. CoP (G): configuration of parameters (groups)

Clearly, Table 11.4 needs to be taken as an exemplification of how likely it is that a given number of configurations may lead to an under- or over-powered simulation, hence determining the likelihood to make Type-II error or to over-emphasize results. The table can be used as a first indication of how this approach to ABM runs can be applied. More fine grained results may vary depending on the circumstances of each simulation, including the levels of α, β, and the purpose of the model.

Further Reading

Details on several power measures can be found in Cohen (1988) and Liu (2014). Specific information on ABM and power are in Secchi and Seri (2017).

Appendix

Number of Runs Calculations

The following is the `R` code for a function that calculates the number of runs for the configuration of parameters (G, here `G`) and effect size (f, here `ES`), given $1 - \beta = 0.95$, $\alpha = 0.01$:

```
n.runs <- function(G, ES) {
  return(14.091 * G^(-0.640) * ES^(-1.986))
}
```

In the case discussed in Exercise 1 above, the numbers are:

```
n.runs(3, 0.25)
[1]  109.465
```

The same analysis using the exact function of the package `pwr` on power analysis (see Champely et al. 2016) is:

```
pwr.anova.test(f=0.25, k=3, power=0.95,
                    sig.level=0.01)
```

and yields $n = 111.677$.

Effect Size for ANOVA vs OLS Regression

In the text we have used a one-way ANOVA test to estimate the number of runs, taking $1 - \beta = 0.95$, $\alpha = 0.01$ and a given effect size f. However, we then used regression analysis to study the differences between under-, correctly-, and over-powered models.

Since there is transformation between the parameters of ANOVA and OLS regression, it is possible to connect the way effect size is calculated in the first to the second.

As mentioned in the text of the chapter, the effect size for ANOVA is:

$$f = \sqrt{\frac{n \sum_{j=1}^{G} \left(\bar{x}_j - \bar{x}\right)^2}{\sum_{j=1}^{G} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2}}$$

The quantity under the square root is the SSB divided by the Sum of Squares Within (SSW) or, in Cohen's terms, $f = \frac{\sigma_m}{\sigma}$ (Cohen 1992). The effect size for regression is, according to Cohen (1992), $f^2 = \frac{R^2}{1-R^2}$. It is easy to demonstrate that:

$$f^2 = \frac{R^2}{1 - R^2} = \frac{\text{SSB}}{\text{SSR}}$$

where the SSW in a one-way ANOVA is comparable to the Sum of Squares of Residuals (SSR) in an OLS regression with exactly the same dependent and independent variables.

References

Anderson, P. (1972). More is different. *Science, 177*(4047), 393–396.

Bardone, E. (2016). Intervening via chance-seeking. In D. Secchi & M. Neumann (Eds.), *Agent-based simulation of organizational behavior. New frontiers of social science research* (pp. 203–220). New York: Springer.

Bland, J. M. (2009). The tyranny of power: Is there a better way to calculate sample size? *BMJ, 339*, b3985.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., & Rosario, H. D. (2016). Pwr: Basic functions for power analysis.

Choirat, C., & Seri, R. (2012). Estimation in discrete parameter models. *Statistical Science, 27*(2), 278–293.

Coen, C. (2009). Simple but not simpler. Introduction CMOT special issue–simple or realistic. *Computational and Mathematical Organization Theory, 15*, 1–4.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: LEA.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cohen, M. D., March, J. G., & Olsen, H. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly, 17*(1), 1–25.

Davidsson, P., & Verhagen, H. (2017). Types of simulation. doi: https://doi.org/10.1007/978-3-319-66948-9_3.

de Marchi, S., & Page, S. E. (2014). Agent-based models. *Annual Review of Political Science, 17*(1), 1–20.

Edmonds, B., & Meyer, R. (2017). Introduction to the handbook. doi: https://doi.org/10.1007/978-3-319-66948-9_1.

Edmonds, B., & Moss, S. (2005). From KISS to KIDS — an 'anti-simplistic' modelling approach. In P. Davidson (Ed.), *Multi agent based simulation*. Lecture Notes in Artificial Intelligence (Vol. 3415, pp. 130–144). New York: Springer.

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des β-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle [The significance and control of the β-error during the inference-statistical examination of the log-linear models]. *Zeitschrift für Sozialpsychologie, 15*(1), 18–32.

Fioretti, G. (2016). Emergent organizations. In D. Secchi & M. Neumann (Eds.), *Agent-based simulation of organizational behavior. New frontiers of social science research* (pp. 19–41). New York: Springer.

Fioretti, G., & Lomi, A. (2008). An agent-based representation of the garbage can model of organizational choice. *Journal of Artificial Societies and Social Simulation, 11*(1), 1.

Fioretti, G., & Lomi, A. (2010). Passing the buck in the garbage can model of organizational choice. *Computational and Mathematical Organization Theory, 16*(2), 113–143

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological), 17*(1), 69–78

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*, 587–606.

Gilbert, N., & Terna, P. (2000). How to build and use agent-based models in social science. *Mind and Society, 1*, 57–72.

Hahn, G. J., & Meeker, W. Q. (2011). *Statistical intervals: A guide for practitioners*. Hoboken: Wiley.

Heckbert, S. (2013). MayaSim: An agent-based model of the ancient Maya social-ecological system. *Journal of Artificial Societies and Social Simulation, 16*(4), 11.

Herath, D., Secchi, D., & Homberg, F. (2015). Simulating the effects of disorganisation on employee goal setting and task performance. In D. Secchi & M. Neumann (Eds.), *Agent-based simulation of organizational behavior. New frontiers of social science research* (pp. 63–84). New York: Springer.

Herath, D., Costello, J., & Homberg, F. (2017). Team problem solving and motivation under disorganization – an agent-based modeling approach. *Team Performance Management, 23*(1/2), 46–65.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power. *The American Statistician, 55*(1), 19–24.

Kollman, K., Miller, J. H., & Page, S. E. (1992). Adaptive parties in spatial elections. *The American Political Science Review, 86*(4), 929–937.

Korn, E. L. (1990). Projecting power from a previous study: Maximum likelihood estimation. *The American Statistician, 44*(4), 290–292.

Lakatos, E. (2005). Sample size determination for clinical trials. In *Encyclopedia of biostatistics*. Hoboken: Wiley.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*(7), 701–710.

Lakens, D. & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*(3), 278–292.

Lamperti, F. (2015). *An Information Theoretic Criterion for Empirical Validation of Time Series Models*. LEM Papers Series 2015/02, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy.

Liu, X. S. (2014). *Statistical power analysis for the social and behavioral sciences*. New York: Routledge.

Liu, T., & Stone, C. C. (2007). *Law and statistical disorder: Statistical hypothesis test procedures and the criminal trial analogy*. SSRN Scholarly Paper ID 887964, Social Science Research Network, Rochester, NY.

Maggi, E., & Vallino, E. (2016). Understanding urban mobility and the impact of public policies: The role of the agent-based models. *Research in Transportation Economics, 55*, 50–59.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*(1), 537–563.

Mayo, D. G. (1992). Did pearson reject the neyman-pearson philosophy of statistics? *Synthese, 90*(2), 233–262.

Mungovan, D., Howley, E., & Duggan, J. (2011). The influence of random interactions and decision heuristics on norm evolution in social networks. *Computational and Mathematical Organization Theory, 17*(2), 152–178.

Neyman, J. (1950). *First course in probability and statistics*. New York: Henry Holt and Company.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*(1/2), 175–240.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289–337.

Pearson, E. S. (1955). Statistical concepts in the relation to reality. *Journal of the Royal Statistical Society. Series B (Methodological), 17*(2), 204–207.

Railsback, S. F., & Grimm, V. (2011). *Agent-based and individual-based modeling: A practical introduction* (59468th ed.). Princeton: Princeton University Press.

Ritter, F. E., Schoelles, M. J., Quigley, K. S., & Cousino-Klein, L. (2011). Determining the numbers of simulation runs: Treating simulations as theories by not sampling their behavior. In L. Rothrock & S. Narayanan (Eds.), *Human-in-the-loop simulations: Methods and practice* (pp. 97–116). London: Springer.

Robinson, S. (2014). *Simulation. The practice of model development and use* (2nd ed.). New York: Palgrave.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612.

Secchi, D. (2015). A case for agent-based model in organizational behavior and team research. *Team Performance Management, 21*(1/2), 37–50.

Secchi, D., & Gullekson, N. (2016). Individual and organizational conditions for the emergence and evolution of bandwagons. *Computational and Mathematical Organization Theory, 22*(1), 88–133.

Secchi, D., & Seri, R. (2014). 'How many times should my simulation run?' Power analysis for agent-based modeling. In *European Academy of Management Annual Conference, Valencia, Spain*.

Secchi, D., & Seri, R. (2017). Controlling for 'false negatives' in agent-based models: A review of power analysis in organizational research. *Computational and Mathematical Organization Theory, 23*(1), 94–121.

Shimazoe, J., & Burton, R. M. (2013). Justification shift and uncertainty: Why are low-probability near misses underrated against organizational routines? *Computational and Mathematical Organization Theory, 19*(1), 78–100.

Simon, H. A. (1976). How complex are complex systems. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 2, pp. 507–522). Baltimore: Philosophy of Science Association.

Simon, H. A. (1978). Rationality as process and a product of thought. *American Economic Review, 68*, 1–14.

Simon, H. A. (1997). *Administrative behavior* (4th ed.). New York: The Free Press.

Thiele, J., Kurth, W., & Grimm, V. (2015). Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. *Journal of Artificial Societies and Social Simulation, 17*(3), 11.

Thomsen, S. E. (2016). How docility impacts team efficiency. An agent-based modeling approach. In D. Secchi & M. Neumann (Eds.), *Agent-based simulation of organizational behavior. New frontiers of social science research* (pp. 159–173). New York: Springer.

Troitzsch, K. G. (2017). Historical introduction. doi:https://doi.org/10.1007/978-3-319-66948-9_2.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *American Statistician, 70*(2), 129–133.

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