Comput Math Organ Theory (2017) 23:94–121
DOI 10.1007/s10588-016-9218-0

MANUSCRIPT

Controlling for *false negatives* in agent-based models: a review of power analysis in organizational research

Davide Secchi[1] · **Raffaello Seri**[2,3]

Published online: 3 May 2016

Abstract This article is concerned with the study of statistical power in agent-based modeling (ABM). After an overview of classic statistics theory on how to interpret Type-II error (whose occurrence is also referred to as a *false negative*) and power, the manuscript presents a study on ABM simulation articles published in management journals and other outlets likely to publish management and organizational research. Findings show that most studies are underpowered, with some being overpowered. After discussing the risks of under- and overpower, we present two formulas to approximate the number of simulation runs to reach an appropriate level of power. The study concludes with the importance for organizational behavior scholars to perform their models in an attempt to reach a power of 0.95 or higher at the 0.01 significance level.

Keywords Statistical power · Agent-based modeling · Computational simulation · Effect size · Sample size · Organizational behavior research

1 Introduction

The last few years have seen a growing interest towards agent-based modeling (ABM) and its potentials to benefit management and organization studies (Fioretti 2013). As a technique to model complex adaptive social systems, it has been

✉ Davide Secchi
secchi@sdu.dk

Raffaello Seri
raffaello.seri@uninsubria.it

[1] COMAC Research Cluster, Centre for Human Interactivity (CHI), Department of Language and Communication, University of Southern Denmark, Sdr. Stationvej 28, 4200 Slagelse, Denmark

[2] Department of Economics, University of Insubria, Via Monte Generoso 71, 21100 Varese, Italy

[3] Center for Nonlinear and Complex Systems, Via Valleggio 11, 22100 Como, Italy

recently advocated that ABM is particularly well suited to team research (Secchi 2015) and, more broadly, to study organizational behavior (Secchi and Neumann 2016).

Research on organizational behavior and management is slowly taking ABM into consideration, with the help of a few specialized academic outlets such as *Computational and Mathematical Organization Theory* and the *Journal of Artificial Societies and Social Simulation*, and of increasing room at international management conferences such as EURAM and AOM. Given this expanding trend, we believe behavioral sciences and management studies have a lot to offer to the way computational simulations are performed. We refer to the typical toolkit of the management and organizational behavior researcher, the questions they ask, and the solutions they adopt when conducting a study. ABM simulations can be intended as experiments of a computational nature (Coen 2009a; Hoser 2013) consistent with experimental design and methods that flourish in our disciplines.

This article is concerned with an issue that is opposite to what the articles above point out (i.e. how ABM can be imported into management and organizational behavior studies). The present study aims at exporting one of the most relevant concerns of experimental methods to ABM. One of the issues that every experimental researcher deals with is statistical power and sample size determination (Cohen 1988, 1992). When performing any computer simulation, a researcher comes to the question of how many times the model should run. As known from the literature (e.g., Liu 2014), this is a problem of sample size determination that is usually addressed by power analysis.

By improving the way ABM research is conducted by the means of power analysis, we also improve the tools in the hands of those organizational behavior researchers that have embraced this new simulation technique. Moreover, by using statistical power analysis, the organizational researcher may feel more "at home," being able to apply tests that are more familiar to him/her.

It is clear that statistical power analysis is relevant only to ABM respecting some conditions. First of all, the models under scrutiny have to be stochastic (see North and Macal 2007, Sect. 2 for a comparison between deterministic and stochastic models). Second, the objective of the model should be testing of assumptions and not, say, observation of emergent behavior or detailed description of phenomena, as is often the case for exploratory ABM. Third, while our discussion of statistical power analysis can be applied to all kinds of tests, in the second part of the paper, we will focus on the following situation: one or more outcome measures are identified, and the hypothesis is to test that the expected outcomes under J configurations of parameters are the same; the expected outcome for each configuration is estimated through the mean over a certain number n of runs.

Although statistical power and Type-II and Type-I error are well known topics in statistics, some authors (e.g., Gigerenzer 2004; Friston 2012) highlight that many scholars get confused by the interpretation of some of the key elements and approaches involved. For this reason, we believe we should clarify what is the methodological backbone of the testing theory we are discussing. Hence, the following section features a description of the classic theory of power, setting the ground for the basic concepts used in this article. We then review ABM studies published between 2010 and 2013 and present calculations of power for all the 69 articles (whenever possible). A discussion of findings follows and we finally conclude the article with some recommendations for organizational and ABM researchers.

2 The traditional treatment of power

In the following, we consider the Neyman–Pearson framework, introduced by Jerzy Neyman and Egon Sharpe Pearson between the 20's and the 30's, where statistical tests start from the definition of a null hypothesis, H_0, and of an alternative hypothesis, H_1. This article is concerned with parametric models, i.e. statistical models defined through a set of parameters Θ, called *parameter space*. As an example, a normal distribution can be characterised by the mean μ, that can take any real value, and the standard deviation σ, that must be nonnegative. The parameter space is $\Theta = (-\infty, +\infty) \times [0, +\infty)$. The null hypothesis is generally characterized as a subset of the parameter space, say Θ_0. The most common hypotheses correspond to the nullity of a parameter (in the previous example, say, $\mu = 0$ and $\Theta_0 = \{0\} \times [0, +\infty)$) or a vector of parameters, or the equality of some parameters. The alternative hypothesis is given by the values of the parameters that are not in Θ_0; this set is generally called Θ_1 and is defined as the complement of Θ_0 in Θ (in the previous example, $\Theta_1 = (-\infty, 0) \times [0, +\infty) \cup (0, +\infty) \times [0, +\infty)$). It should therefore be clear, in what follows, that either the null or the alternative hypothesis is true.

The objective of a statistical test is to help the researcher make a decision as to which hypothesis between H_0 and H_1 is true. Usually this is done through a *test statistic* T and a subset $\mathcal{A}$ of the range of T, called *acceptance region*. In the sample the test statistic T takes the value t. It is customary to write that, if t does not fall inside the region $\mathcal{A}$, the test "rejects" the null hypothesis. On the other hand, if t falls inside the region $\mathcal{A}$, the test "fails to reject" or "does not reject" the null hypothesis. This apparently odd circumlocution is preferred instead of the more direct "accept." The widespread and uncontroversial use of this expression in the recent literature conceals a disagreement in earlier theoretical references, as witnessed by the contradiction between the use of "acceptance region" and of "fail to reject." Indeed, Neyman himself (Neyman 1950, p. 259) agreed with the use of the word "accept." The dichotomy between acceptance and rejection is coherent with the original purpose of tests in the Neyman–Pearson framework. Neyman, in particular, was extremely clear about the fact that tests should lead to decision based on acceptance or rejection and even to action (Neyman 1950, p. 259). Pearson, instead, was more reluctant (Pearson 1955, p. 206). It was Fisher (Fisher 1955, p. 73), in a different approach to testing, who strongly and consistently argued against the use of the word "accept" and this contributed to create the confusion on terminology. In fact, the prescription that "accept" should not be used inside the Neyman–Pearson approach is yet another part of what Gigerenzer (2004) calls the "null ritual." However, in this article, we follow Neyman and will use interchangeably "fail to reject" or "accept."

Two types of errors can be committed while testing the hypotheses. If we reject the null hypothesis when it is true, we commit a Type-I error (or *false positive*). This happens with a rate equal to the probability that T does not belong to $\mathcal{A}$ when H_0 is true. This probability is usually denoted as α. If we accept the null hypothesis when it is false, we commit a Type-II error (or *false negative*). This happens with rate β, the probability that T belongs to $\mathcal{A}$ when H_0 is false (and H_1 is true). The two probabilities, α and β, are linked by a trade-off: in order for α to decrease, one needs to increase $\mathcal{A}$ (because α is the probability that T does not belong to $\mathcal{A}$ under H_0), and therefore β increases too (because β is the probability that T belongs to $\mathcal{A}$ under H_1).[1] In the classic Neyman–Pearson approach, $\mathcal{A}$ is chosen in such a way that α is fixed and small (5 and 1 % are customary values). On the other hand, β is not directly (and rarely indirectly) controlled. However, when the sample size N increases, β generally tends to 0, so that, for N large enough, one can take α small and hope for β to be not too large.

2.1 An example on the relation between α and β

Consider the test for the nullity of the mean in a normally distributed population with mean μ and variance σ^2. The null hypothesis is $\mathsf{H}_0 : \mu = 0$, the alternative hypothesis is $\mathsf{H}_1 : \mu \neq 0$. Suppose to observe a sample $\{X_1, \ldots, X_N\}$ from the population. A test statistic for this hypothesis is

$$T = \frac{\sum_{i=1}^N X_i}{\sqrt{\sum_{i=1}^N X_i^2 - \frac{1}{N}\left(\sum_{i=1}^N X_i\right)^2}}.$$

If H_0 holds true, then its distribution is a Student's t with $N-1$ degrees of freedom, indicated as t_{N-1}. An acceptance region at level α will be given by the interval $\mathcal{A} = \left[t_{\frac{\alpha}{2},N-1}, t_{1-\frac{\alpha}{2},N-1}\right]$ (this is not the only possible one, but is the most common), where $t_{\gamma,n}$ is the $\gamma-$quantile of the distribution of a t_n random variable. This region will have a Type-I error rate equal to α by construction. Suppose now that H_0 does not hold true, i.e. $\mu \neq 0$ and H_1 holds. Then, T will be distributed as a noncentral Student's t with $N-1$ degrees of freedom and noncentrality parameter $\lambda_N = \sqrt{N}\mu/\sigma$, indicated as $t_{N-1}(\lambda_N)$. The Type-II error rate β is:

$$\begin{aligned}\beta &= \mathbb{P}\left\{t_{N-1}(\lambda_N) \in \left[t_{\frac{\alpha}{2},N-1}, t_{1-\frac{\alpha}{2},N-1}\right]\right\}\\ &= \mathbb{P}\left\{t_{\frac{\alpha}{2},N-1} \leq t_{N-1}(\lambda_N) \leq t_{1-\frac{\alpha}{2},N-1}\right\}.\end{aligned}$$

One can verify the properties we rapidly described above. When α decreases, $t_{\frac{\alpha}{2},N-1}$ and $t_{1-\frac{\alpha}{2},N-1}$ get far from 0, and β increases. When N increases, α is constant while β goes to 0, as λ_N gets further from 0.

[1] See van der Vaart (1998, p. 213) or Choirat and Seri (2012, Proposition 7, p. 285) for a quantitative version of this trade-off.

2.2 The power of a test

The power of a statistical test is the probability that it correctly rejects a false null hypothesis, namely one minus β. The previous reasoning lets one wonder which levels of power are supposed to be acceptable. The value that seems to be more commonly accepted is 80 % as it appears in several places in Cohen (1992) and also in Lehr (1992). Studies with lower values are often seen as underpowered. This value, corresponding to $\beta = 20\ \%$, does not seem very high, especially when $\alpha = 5\ \%$ or even 1 %. The profound asymmetry between the value required for α and the value deemed acceptable for β calls for a clarification.[2] We provide, in what follows, three different possible explanations, not necessarily alternative to each other.

First, it has been repeatedly indicated in the literature (Sedlmeier and Gigerenzer 1989; Hallahan and Rosenthal 1996; Cohen 1992) that many researchers do not perform formal power analysis but rely on sample size as an indicator of error. Indeed, when the sample size N increases, even if α is fixed, β tends to decrease to 0. Together with the increase in precision, this is probably the most important reason for which a large sample size is usually considered positively. Therefore, a researcher may feel dispensed with power considerations if he or she believes that the sample size is sufficiently large. However, we show below that this false sense of security often leads to underpowered studies. (We will see more on the role of sample size later.)

Second, suppose that we have two hypotheses, say H' and H'', such that either one or the other is true but both of them cannot be true at the same time. As in the Neyman–Pearson approach α is controllable while β isn't completely, it is often reasonable to choose as H_0 the hypothesis, say H', whose rejection is considered more serious. Indeed, in case H'' were chosen as the null, the rejection of H' when true (that would now correspond to a Type-II error rate β) would be out of control. This is related to the reason behind Cohen's choice (Cohen 1988, Sect. 2.4) of $\beta = 0.20$ when $\alpha = 0.05$ (see also Lakens 2013). Indeed, he explicitly states that the ratio β/α should be near to 4 when Type-I errors are about four times as serious as Type-II errors.

Third, the different emphasis on α and β is largely due to a misunderstanding between the Neyman–Pearson and Fisher approaches that is made particularly clear in Royall (1997, pp. 109–110). In the Fisher approach, the computation of the test statistic does not lead to any decision but to the determination of the p value: this is the probability, under the null hypothesis, of obtaining values that are as extreme as, or more extreme than the one that is observed, i.e. t. A small p value is considered evidence against the null hypothesis because, if the null is true, it is difficult to suppose that chance alone would lead to such a small probability of observing a sample as extreme as the one we have observed. Despite Fisher himself took

[2] As an interesting variation on the traditional choice of a fixed significance level, Arrow (1960) describes a procedure to compute α that starts from setting $\alpha = \beta$ for a value of the parameters under the alternative hypothesis.

position against this choice (Fisher 1956, p. 42)[3] the extraordinariness of the p value is often evaluated comparing it with α (if smaller, rejection ensues). Therefore α takes the further role of a gauge of extraordinariness and it is smaller than β because of the double meaning of which it is charged, i.e. error in a Neyman–Pearson framework and p value threshold in a Fisher framework.

This is complicated by the fact that the alternative hypothesis is generally composed of more than one possible value for the parameters. This means that for any possible value of the parameters respecting the alternative hypothesis H_1 it is possible to define a different value of β and power. However, it is generally the case that the parameters enter into the power function through a single number *ES*, called *effect size* and identified with μ/σ in the previous section. There are two kinds of power analysis that can be performed involving *ES*. The first one is called *a priori* power analysis and is generally used, before data is collected, to evaluate the sample size needed to obtain a certain value of β for fixed α under an hypothesized value for *ES*. This procedure is customarily performed imputing a value to *ES* on the basis of the evidence collected in previous similar studies (see Lenth 2001, Sect. 2; Lakens 2013). Cohen (1992) has compiled tables of *ES* values indicated as small, medium and large that can be used for this task. This kind of power analysis is universally considered as an important and statistically sound tool (despite the use of "canned" effect sizes has been subject to critique, see Lenth 2001, Sect. 6).

The second is called *post hoc* power analysis and is performed after estimation in order to obtain an a posteriori estimate of the power. This is achieved using the value of *ES* in which the parameters are replaced by their estimated values. There is some evidence that this technique has several drawbacks (for further details on the issues related to this analysis, see Korn 1990; Hoenig and Heisey 2001).

2.3 The importance of power for ABM

In this subsection we briefly discuss the role of Type-I and Type-II error rates for ABM as a theoretical preamble for the review performed in the section below. Indeed, in this case, the researcher makes decisions on the parameters of the simulation and on how many times the simulation should be performed. The interest of this class of models is often to show if and how a certain outcome measure varies with the parameters of the simulated model. Agent-based simulations are particularly useful in the social sciences for their ability to model complex adaptive systems (e.g., Miller and Page 2007). This makes "emergence" one of the main features of these models (e.g., Fioretti 2013; Secchi 2015), sometimes as a result of complexity. This points right at the core of the use of power for this class of models,

[3] The attitude of Fisher towards fixed thresholds was more ambivalent than this source suggests. As an example, Fisher (1926) advocated the comparison of the p value with a threshold chosen by the researcher according to his or her experience (2, 5 or even 10 %). It is therefore ironic that this paper is often considered as the origin of the fixed 5 % threshold because this is the number that Fisher used more frequently in it. A more nuanced use of p values is in Fisher (1925, p. 80 and elsewhere), where the 5 % threshold is used alongside other values, such as 1 %.

that is to make sure that the occurrence that results of simulations are mostly affected by random effects is avoided.[4]

ABM researchers may be running the simulation for different purposes. Some of them may be interested in the mean values of the outcome variable, some others may focus on extreme values, while other simulations may be descriptive or tied to a particular set of empirical data. In this article, we assume that the researcher considers the average value of the outcome variable to be informative, hence relevant to one's data analysis. This assumption may cut some of the ABM simulations off but we believe most approaches are covered. The modeler will generally identify, for example, a certain number J of parameter configurations to be experimentally tested in a computational simulation. The theoretical expectations of the outcome measure for each of these combinations are given by μ_j, $j = 1, \ldots, J$. Several null hypotheses can be tested, but most of them require that some of these means are equal and, in the extreme case, all of them are equal ($\mathsf{H}_0 : \mu_1 = \ldots = \mu_J$). We suppose that the simulation is balanced across configurations so that each mean is estimated through n runs, and the total number of simulations to be run is $N = n \cdot J$.

In ABM and more generally in simulations, most of the reasons for which, in real-world experimental studies, a large β can be tolerated suddenly cease to be acceptable. On the one hand, the statistical tests are quite standard and their power analysis is easily performed. On the other hand, both α and β could be reduced with respect to the values in use in most statistical practice. However, before we can elaborate further on the adequate level of power for ABM research, the next section shows not only that a formal power analysis—neither *a priori* nor *post hoc*—is by no means common in ABM but also that the values that can be reconstructed from the papers show that most studies are strongly underpowered.

3 ABM and power: a review

Once we have clarified what is the theoretical need for statistical power analysis, and before explaining the consequences of ignoring it, we may ask whether ABM research is actually exempt from these issues. One may claim that, for simulation studies, all it takes to avoid Type-II error is to increase the number of runs, or conduct convergence analysis, for example.

The simplest action to obtain high power, thus having a low probability that the null hypothesis is accepted when it is false, is that of increasing the number of runs. This would be equal to increasing the number of subjects in an experiment, bearing a positive and strong effect on power (Liu 2014). However, given the nature of agent-based simulations, even one more run can be sometimes particularly hard to perform. This is due to the fact that some of these models can be complex. Some advocate a KIDS ("Keep It Descriptive, Stupid") principle as opposed to the classic KISS ("Keep It Simple, Stupid") to signal that ABM can be very detailed

[4] This is made very clear by Morris (1987) whose example shows unequivocally that β is by far a more reasonable measure of reliability than the estimated *ES*.

representations of reality (Edmonds and Moss 2005). These are models that some consider “expensive” (Ritter et al. 2011), because each run may take a significant amount of time to complete. Of course, not all ABM are complex and expensive so, in principle, the strategy of increasing the number of runs may pay off although it can be difficult to achieve in practice.

The other claim—i.e. perform convergence analysis—tackles with a different issue that may affect the test power but it is not directly related to it. In fact, convergence or sensitivity analyses are usually performed to understand whether a given simulation reaches some sort of equilibrium around a given pattern of results (Robinson 2014). Clearly, this is a very important check to be run on a simulation but, as far as ABM is concerned, it deals with time rather than with runs. In other words, it provides information on when a given configuration of parameters provide meaningful results within a single run. It may help with Type-I error, because if data are of low quality (e.g., they have not reached the above-quoted equilibrium) the test is applied to a set of data intrinsically different from the one that the researcher would like to submit to test, and the size of the test may be incorrect.

Even whether power analysis is deemed unnecessary because Type-II error can be avoided easily, researchers should have a benchmark, a point of reference. For this reason, it is strongly advocated by the authors of this paper that power analysis should always be performed (Secchi 2014).

Given the importance of power analysis for the social and behavioral sciences (Cohen 1988; Liu 2014), it is not uncommon to find publications indicating the scarce use for empirical studies (e.g., Mone et al. 1996; Sedlmeier and Gigerenzer 1989). As far as our knowledge is concerned, a review of power has not been conducted for simulation studies. Given the prominence that ABM is gaining in the social and behavioral sciences, we have reviewed some of the publications featuring a model and calculated power where appropriate and possible. In the following, we describe the method of the review study and comment on its results.

3.1 Methods

The study was conducted on articles published over a limited period of time when ABM-related publications seemed to increase. We considered the four years from 2010 to 2013. Since our interest lies in the management and organizational behavior literature, we screened the two simulation journals that have the closest ties with our

Table 1 Summary of results for power calculations performed (number of models)

Journal	IP	OP	NC	Fine	Total	Articles
CMOT	17	7	5	–	29	24
JASSS	21	1	14	4	40	39
Other	4	2	–	–	6	6
TOTAL	42	10	19	4	75	69

IP insufficient power, *OP* overpower, *NC* not computable, *Fine* good power

Other: Organization Science, *MIS Quarterly*, *Journal of Management Studies*, *Strategic Management Journal*

Table 2 Power analysis estimations on ABM articles published between 2010 and 2013

Article	*ES* =	0.1	0.3	0.1	0.3	CoP	Runs	ER
	α =	0.01	0.01	0.05	0.05			
Computational and Mathematical Organization Theory								
Arroyo et al. (2010)		NC	NC	NC	NC	9, 6	NC	NC
Fioretti and Lomi (2010)		0.161	0.997	0.355	0.999	4	100	562
Fridman and Kaminka (2010)								
——exp 1		0.015	0.074	0.066	0.207	2	30	875
——exp 2		0.012	0.036	0.058	0.124	2	15	875
——exp 3		0.037	0.558	0.127	0.780	2	30[d]	875
——exp 4		0.014	0.060	0.063	0.180	4	25	68
Zhang and Gao (2010)								
——exp 1		0.999[a]	0.999[a]	0.999[a]	0.999[a]	24	1000	178
——exp 2		0.03	0.318	0.108	0.560	2	100	875
Zhang and Leezer (2010)		0.029	0.689	0.110	0.871	36[d]	10[d]	138
Zappala and Logan (2010)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	1024	5000	16
Ahrweiler et al. (2011)		0.017	0.133	0.073	0.324	5[d]	10	487
Grow and Flache (2011)		0.326	0.999[a]	0.572	0.999[a]	36	50	138
Hirshman et al. (2011)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	100	1000	72
Mungovan et al. (2011)		0.160	0.997	0.355	0.999	4	100	562
Sharpanskykh and Stroeve (2011)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	4446	8000	6
Zou and Yilmaz (2011)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	1000[d]	1000[d]	16
Castellani and Rajaram (2012)		NC	NC	NC	NC	NC	NC	NC
Cassell and Wellman (2012)		NC	NC	NC	NC	NC	NC	NC
Cioffi-Revilla et al. (2012)		0.048	0.826	0.155	0.940	8	30	360
Wang and Hu (2012)		NC	NC	NC	NC	NC	NC	NC
Abbas (2013)		0.012	0.031	0.057	0.120	4	4[d]	562
Bausch (2013)		NC	NC	NC	NC	NC	NC	NC
Dugundji and Gulyás (2013)		NC	NC	NC	NC	NC	NC	NC
Hoser (2013)		0.474	0.999[a]	0.708	0.999[a]	16	100	231
Fairchild et al. (2014)		NC	NC	NC	NC	NC	NC	NC
Shimazoe and Burton (2013)		0.266	0.999	0.497	0.999	8	100	360
Villarroel et al. (2013)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	9	1000	334
Udayaadithya and Gurtoo (2013)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	768	200	19
Yamanoi and Sayama (2013)		0.234	0.999	0.462	0.999	25	50	174
Journal of Artificial Societies and Social Simulation								
Altaweel et al. (2010)		0.366	0.999[a]	0.608	0.999[a]	2	1000	875
Boero et al. (2010)		0.397	0.999[a]	0.637	0.999[a]	7	1000[d]	393
Bosse and Gerritsen (2010)		NC	NC	NC	NC	NC	NC	NC
Cecconi et al. (2010)		0.033	0.692	0.119	0.871	16[d]	16	231
Dunn and Gallego (2010)		NC	NC	NC	NC	NC	NC	NC
Lee (2010)		NC	NC	NC	NC	NC	NC	NC
Miodownik et al. (2010)		0.016	0.115	0.072	0.294	4	10	562

Table 2 continued

Article	*ES* =	0.1	0.3	0.1	0.3	CoP	Runs	ER
	α =	0.01	0.01	0.05	0.05			
Quera et al. (2010)		NC	NC	NC	NC	NC	NC	NC
Radax and Rengs (2010)[b]		0.04	0.550	0.147	0.772	2	164	69[c]
Sobkowicz (2010)		0.999	0.999	0.999	0.999	42[d]	8000	125
Savarimuthu et al. (2010)		NC	NC	NC	NC	NC	NC	NC
Xianyu (2010)		0.037	0.558	0.127	0.781	4	30	562
Ballinas-Hernández et al. (2011)		NC	NC	NC	NC	NC	NC	NC
Nongaillard and Mathieu (2011)		0.565	0.999[a]	0.781	0.999[a]	20	100	201
Siebers and Aickelin (2011)		0.026	0.330	0.099	0.578	4	20	562
Wildman and Sosis (2011)		0.565	0.999[a]	0.781	0.999[a]	20	100	201
Fonoberova et al. (2012)		0.656	0.999[a]	0.846	0.999[a]	36	80	138
Grazzini (2012)		0.161	0.997	0.355	0.999	4	100	562
Letia and Slavescu (2012)		NC	NC	NC	NC	NC	NC	NC
Meadows and Cliff (2012)		NC	NC	NC	NC	NC	NC	NC
Montes (2012)		0.857	0.999[a]	0.953	0.999[a]	40	100	129
Patel et al. (2012)		0.543	0.999[a]	0.765	0.999[a]	19	100	207
Schindler (2012)		0.099	0.995	0.259	0.999	12	43	278
Sioson (2012)		0.014	0.058	0.063	0.174	2	24	875
Sutcliffe and Wang (2012)		0.424	0.999[a]	0.669	0.999[a]	48	50[d]	114
Xianyu (2012)		NC	NC	NC	NC	NC	NC	NC
Cockburn et al. (2013)		NC	NC	NC	NC	NC	NC	NC
Demarest et al. (2013)		0.137	0.987	0.318	0.997	3	100	675
Dubois et al. (2013)		0.999	0.999[a]	0.999	0.999[a]	4, 6	1000	562, 433
Gulden (2013)		NC	NC	NC	NC	NC	NC	NC
Heckbert (2013)		NC	NC	NC	NC	NC	NC	NC
Jansson (2013)		0.497	0.999[a]	0.728	0.999[a]	17	100	223
Kim et al. (2013)		0.548	0.999[a]	0.766	0.999[a]	6	200	433
Lee et al. (2013)		0.991	0.999[a]	0.998	0.999[a]	80	100	83
Nye (2013)		0.565	0.999[a]	0.781	0.999[a]	20	100	201
Schindler (2013)		0.024	0.372	0.095	0.622	8	15	361
Shiba (2013)		0.016	0.099	0.071	0.265	3[d]	10	675
Waldeck (2013)		0.046	0.757	0.148	0.905	6	32	433
Wijermans et al. (2013)		NC	NC	NC	NC	NC	NC	NC
Other journals								
Miller and Lin (2010)		0.225	0.999	0.451	0.999[a]	24	50	178
Aggarwal et al. (2011)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	16	10,000	231
Coen and Maritan (2011)		0.914	0.999[a]	0.976	0.999[a]	48[d]	100	114
Nan (2011)		0.179	0.999	0.392	0.999[a]	50	30	111
Levine and Prietula (2012)		0.048	0.867	0.154	0.958	12	25	278
Miller et al. (2012)		0.999[a]	0.999[a]	0.999[a]	0.999[a]	288	100	36

Table 2 continued

Article	*ES* = α =	0.1 0.01	0.3 0.01	0.1 0.05	0.3 0.05	CoP	Runs	ER
Mean		0.415	0.783	0.526	0.842			
SD		0.395	0.346	0.373	0.284			

[a] Overpower

[b] The article uses power tests to determine how many runs need to be performed and we have adapted *ES* to fit the calculation

[c] *ES* based on the information from their article

[d] Our estimate, not explicit in the paper

NC not computable, *CoP* how many configurations of parameters J in the study, *Runs* number of runs n per CoP performed in the study, *ER* estimated number of runs according to our formula (2) with $\alpha = 0.01$, $\beta = 0.05$, $ES = 0.1$

discipline (Meyer et al. 2009, 2011): *Computational and Mathematical Organization Theory* and the *Journal of Artificial Societies and Social Simulation*. Then, we also screened articles published in a list of top management journals (based on ISI Thompson's Impact Factor) and decided to include the four presenting ABM-related publications (Table 1): *Organization Science*, *Journal of Management Studies*, *Strategic Management Journal*, and *MIS Quarterly*. The total sample of articles selected for this study is 69, for a total of 75 experiments (some articles feature more than one computational experiment or model).

The criteria for the selection of articles to include in our study were very straightforward. We checked for publications built around an agent-based simulation or where the model was a significant part of the study. We did not screen for models that were more or less descriptive, nor we did check whether the article made enough information available for us to allow (or replicate) power calculations. The reasons for keeping all the ABM we could possibly find (in the time period considered) was that of being able to have a look at all model types. Some of the models reported as NC in Table 2 may be descriptive or of the kind mentioned earlier in this paragraph.

Once the articles were selected, data on power or Type-II error was extracted whenever possible. If no data or calculations were found in the article, we attempted to gather the information needed to compute statistical power. Given that the most difficult information to gather is the effect size *ES*, we hypothesized two worst case scenario, with a small- (0.1) and a medium-size (0.3) *ES à la* Cohen (1992). Since we treat ABM simulations as experiments, we calculated the statistical power of the test hypothesizing that an ANOVA was performed to test the differences provided by results from the different configurations of parameters. According to Cohen (1992) the medium *ES* for the case of ANOVA is 0.25. We decided to go a bit above that with 0.30 to reach a more significant impact on measurements. When *ES* is large, high power is reached with a limited number of runs and we deem that this is the case that may not present particular concerns. Another variable that requires careful consideration is the significance level α at which power calculations should be referred to. There is no consensus over its value and we have decided to take the standard $\alpha = 0.05$ and also a more stringent criterion of $\alpha = 0.01$ as a reference for our calculations.

3.2 Findings

Table 2 shows our calculations for 75 ABM studies found in the 69 articles selected for the present study in years 2010–2013 (see above for details). All calculations are performed using the ANOVA test with $\alpha = 0.01$, $\alpha = 0.05$, and for small (0.1) and medium (0.3) *ES*. Table 2 then presents how many configurations of parameters (CoP), J, the model in each paper uses, together with the number of runs n actually performed—as declared by the authors. The final column is the calculation of the recommended number of runs resulting from our formula (2) below when $1 - \beta = 0.95$, $\alpha = 0.01$, and $ES = 0.1$, although those simulations with $J = 2$ CoP only could have benefited from using a t formula, like Lehr's—we acknowledge this limitation and slight imprecision in the calculations. Power calculations appearing in columns 2 to 5 are obtained using Cohen's (1988) formulas as they appear in the `pwr` package for `R`—an open source software for statistical analysis. The information provided in Table 2 allows full replication of our study. A particularly sensible quantity is the number of configurations of parameters (CoP), i.e. J, that we computed, as far as possible, from the original articles, where the interested reader can check this information. Some of the studies employ a full factorial design so that every possible parameter assumes multiple values and the model is simulated a number n of times for every possible combination of parameter values *ceteris paribus*. When this is the case, it is relatively easy to calculate J by multiplying the various numbers as they appear in the text of the article.[5] Most studies do not perform any calculation to estimate the robustness of the simulation. In the selected period, only one article (Radax and Rengs 2010) presents statistical power analysis with the intention to determine the appropriate number of runs.

There are multiple strategies to determine either the number of runs or steps. Among the latter, some authors (e.g., Mungovan et al. 2011; Shimazoe and Burton 2013) report convergence analysis to estimate the steady state. Instead, among the former, Siebers and Aickelin (2011) refer to Robinson (2004) to justify the choice of 20 runs per each configuration of parameters. This is an approach that uses confidence intervals but it does not seem to specify to what these numbers are sufficient for. Another strategy for justifying the number of runs is that of Chebyshev's theorem (Shannon 1975), indicated in Lee et al. (2013). The logic seems to be similar to what found in Siebers and Aickelin (2011) in that it is based on 95 % confidence intervals for the performance measure (outcome).

[5] For example, in the article by Grow and Flache (2011), authors identify "36 experimental conditions" (p. 213), and this simplifies our job. Instead, in articles such as in Hoser (2013), the author indicates there are 3 parameters, each taking respectively 2, 4, and 2 values (p. 267). This gives $J = 2 \times 4 \times 2 = 16$. In some other articles such as Cioffi-Revilla et al. (2012), we had to estimate the number of parameters and their values because the authors were less explicit on the various configurations of the simulation or, at least, it was unclear to us.

In the panel of data, we also found 19 published experiments (26 %) where it was impossible (for us, at least) to understand how to calculate power. This signals that the information on methods was not easily accessible from just reading the paper. This may not be a significant problem, given that most ABM are made available in open-source platforms and this may eventually lead to access all information needed. However, this time we limited our analysis to what was available in the published article that is the piece of information with the largest diffusion among academics.

From Table 2, it is apparent that when *ES* is small (i.e. 0.1), mean power for the studies in our sample is 0.415 (SD = 0.395) at the more restrictive significance level of $\alpha = 0.01$, and it is 0.526 (SD = 0.373) when $\alpha = 0.05$. Both values are well below any known standard, indicating most studies are significantly underpowered. When *ES* is medium (i.e. 0.3), on average, the test is above the threshold recommended for power in empirical studies (i.e. 0.80; see Cohen 1988; Liu 2014) with $mean = 0.842$ (SD = 0.284) at the less stringent significance level of $\alpha = 0.05$. This power threshold is, on average, still not met when $\alpha = 0.01$, with $mean = 0.783$ (SD = 0.346).

Figures 1 and 2 are a graphical reorganization of the information in Table 2. In these two figures, only papers and experiments on which we performed the calculations were shown (55 observations). Also, experiments are sorted by publication outlet, using a different color and mark: red dot for JASSS, blue dot for CMOT, green other shapes for the other journals. The logic behind the two figures is to map what happens to power in the same study when significance level α is relaxed, respectively when the assumed *ES* is small (Fig. 1) and medium (Fig. 2). This exercise is interesting because it shows how the assumptions on stringency of a seemingly irrelevant element of power—i.e. the significance level α—affect power. Note that we transformed logarithmically the axes in both figures, to help make sense of the distribution of results.

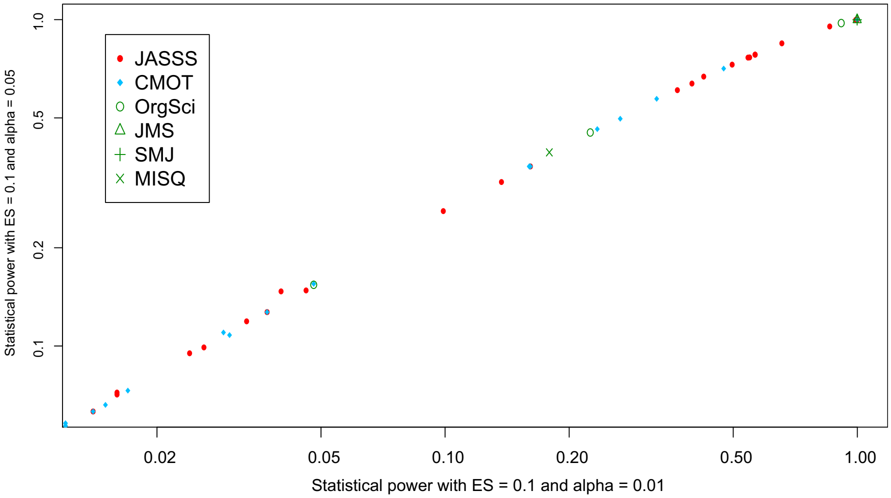

Fig. 1 Impact of the coefficient alpha and small effect size on power in selected studies (2010–2013; 55 observations; both axes are in logarithmic coordinates)

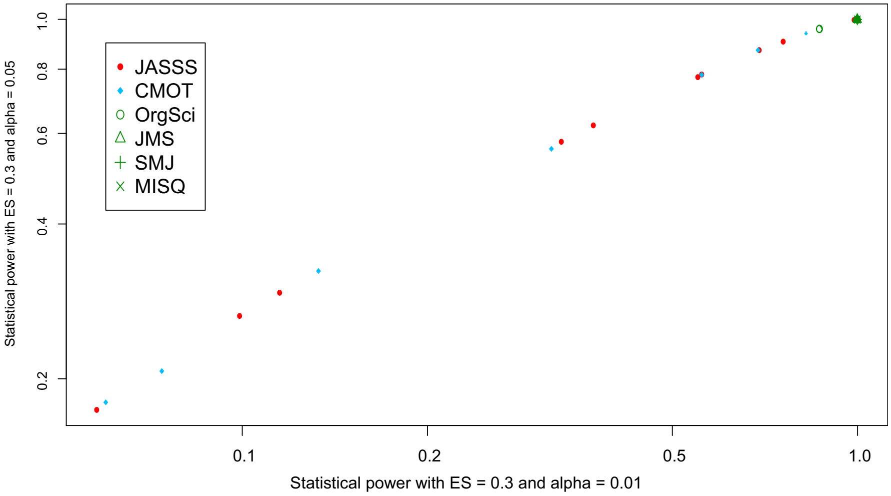

Fig. 2 Impact of the coefficient alpha and medium effect size on power in selected studies (2010–2013; 55 observations; both axes are in logarithmic coordinates)

Figure 1 intuitively shows that some studies have power below 0.50 and the majority seems to appear below 0.90—the actual numbers are 51 % below 0.50 and 73 % below 0.90. The change in the significance level does not seem to affect power in the ABM studies reviewed, when *ES* is small. For overpowered studies—i.e. those studies that have excessively high power (see below for an overview of the risks this entails)—a change in significance levels does not bear any effect at all. For other underpowered studies, there is some effect in that it seems power levels double when α is relaxed for $1 - \beta \leq 0.30$. Instead, for $1 - \beta > 0.30$, the impact of $\alpha = 0.05$ is never enough for the study to reach sufficient power. Hence, Fig. 1 makes it even more apparent that, when *ES* is small, a higher significance level α does not bear meaningful results. This implies that the most sensible strategy for researchers would be to increase the number of runs performed in the simulation. Of course, this requires power analysis to be taken into consideration.

Figure 2 shows the impact of significance levels when *ES* is medium (0.3). With larger *ES*, there are only 14 % of studies with power that is below 0.50 under both conditions. Instead, studies with power below 0.90 are 29 % of the total. The distribution is skewed towards higher levels of power, highlighting two interesting facts. On the one hand, when *ES* is relatively large, underpowered studies do not benefit significantly from a relaxation of α levels. On the other hand, there is a very limited number of borderline studies that would pass the threshold and reach $1 - \beta > 0.90$ (from just below 0.90) although none would reach 0.95 solely because of an α effect. This means, once again, that higher *ES* impacts power levels more effectively but the only viable way to tackle with low power is to increase the number of runs.

In short, both figures substantiate what is in Table 2 and highlight the importance of assuming a reasonable value of *ES* and running the simulation an appropriate number of times. The following section discusses these results further.

4 Discussion of results

The results of the review presented in the previous section points out a few issues with ABM research. Before discussing the results of the review and presenting two formulas for sample size determination, we need to specify what the threshold for power analysis should be in the case of computer simulation.

We have claimed above and elsewhere (Secchi 2014) that computer simulation studies cannot be subjected to the same standards to which empirical studies are. Not only computer simulation—especially ABM—is, obviously, different from empirical study but it is the nature of the difference that supports the need for different standards. The diversity of computational modeling from other scientific methods has been advocated by many scholars (e.g., Gilbert 2008; Coen 2009b; Miller and Page 2007), and we believe it is particularly relevant in the case of power analysis. ABM simulation studies are based on a simplification of reality where a given phenomenon is analyzed according to rules, environmental and agent characteristics. The control exercised on this artificial micro-world is much higher than that exercised, for example, in a lab experiment. For this reason, it is possible to structure the ABM in order to make sure errors are not plaguing or fogging results. More than a possibility, this should be the aim of every modeler. Any simplification of reality carries the risk of being too imprecise, lax, unfocused. Thus, given the assumptions, errors should be brought down to the bare minimum so that unclear findings may be directly identified as coming from the model's theoretical framework, not from its statistical shortcomings. We suggest the reference for every ABM should be to reach power of 0.95 and higher at a 0.01 significance level. More rigorous simulation studies have more potential to contribute to the advancement of our field because they would appear robust and more consistent within the range of assumptions.

It is fair to note that several authors (Johnson 2013; Colquhoun 2014) have recently advocated similarly stringent standards for Type-I error, pushing α to 0.001. However, these authors did not accompany this suggestion with an analogous one concerning a decrease in β. This creates a paradoxical situation, because none of the reasonings in Sect. 2.2 above is compatible with such a huge difference between α and β. This fact has been remarked by other authors (Fiedler et al. 2012; Lakens 2013; Lakens and Evers 2014) that have stressed the relevance of statistical power and Type-II errors for statistical inference as well as the need to balance the two errors. Our proposal of reducing both α and β in the stated proportions embraces the suggestion of decreasing the frequency of Type-I errors while making the ratio of the two probabilities quite near to the original value proposed by Cohen (1988, Sect. 2.4). While both these reductions can be difficult to accomplish in laboratory experiments, we think that most simulations are compatible with them.

4.1 The current norm: under-powered studies

Once we have clarified the threshold for power in ABM and simulation studies should be 0.95 or higher, we can interpret results more clearly. Most studies appear to be underpowered (medium *ES*) or strongly underpowered (small *ES*). This is consistent with the constituents of power, the *ES* being one of the elements affecting power the most. Even small increases in *ES* lead to higher power. Surprisingly enough, the increase we used in our analysis (+0.2) seems not to be enough in most cases. If we take this four-year sample to be representative of ABM research published in the social sciences, we obtain a very meagre picture. With small or even medium *ES*, studies published in most of these articles are not able to tell whether Type-II error is under control. Given that large *ES* depends on the characteristics of the phenomenon under analysis, it seems unlikely that all these studies can claim to have large *ES*. Hence, ABM research needs power calculations to make sure results are sound enough.

To further elaborate on the issues surrounding underpowered studies, we provide four arguments that lean on the variables entering the formula of statistical power $1-\beta$ of an ANOVA test, namely *ES*, *J*, *n* and α. We start with the problems associated with *ES*. The first implication is that low power may be symptom of faulty design, that becomes apparent by discarding effects that are, in fact, relevant to understand the dynamics of a simulation model. In addition to that, low power may depend on the fact that the researcher is testing configurations of parameters that are irrelevant (i.e. too close to each other). Low power may also derive from insufficient number of groups/runs *J* and *n*, so that results are more or less significant at random. Finally, one may have lax testing standards, on the belief that setting a more stringent α for simulation studies is not an issue and it does not affect power. Are faulty design, testing irrelevant differences, insufficient number of groups or runs, or lax standards a problem for ABM research? Indeed, we think they are. We will take on each one of these in the following.

First, faulty design may affect power in that the simulation model is not capable of discriminating significantly enough between parameter configurations (this is reflected in a low *ES*). This may depend from coding the impact of parameters on the outcome variable in a way that fails to make differences apparent and it may be related to coding or equation errors, parameter misspecification, etc. If the simulation is affected by these errors (that we call “faulty design”), with the given number of groups and runs, power remains low and some effects may remain hidden. Hence, although in this case power would not fix poor simulation design, checking for power would help the modeler control the model further. Of course, the *ES* may be low because that is the nature of the simulated relation among different configurations of parameters and, in that case, power analysis would only suggest the appropriate number of runs for that effect to become apparent.

Second, results may be relevant but *ES* is so small that it needs more runs to become apparent. Sometimes this may affect the interpretation of the model, hence making the contribution to academic discussion less relevant than it could have been. Take, for example, the simulation by Fioretti and Lomi (2010) that is based on the famous “garbage can” model (Cohen et al. 1972). In that very fine article, Fioretti and Lomi show that the agent-based version of the model confirms some of the results and discards some others. If we follow one of our hypotheses above, and consider that the *ES* for the parameter configurations is small, power is always insufficient (Table 2), independently of the significance level. This means that findings of such a fine piece of modeling are not accurate and, potentially, we cannot either confirm nor reject any of the features that are in the "garbage can" model. In particular, we cannot discard some of the effects that Fioretti and Lomi (2010) did not find. Yes, we may end up confirming some of the features of the model and rejecting some others although having more runs may surely cast clarity among results.

Third, insufficient number of runs n and/or groups (parameter configurations) J are the most common cause of low power as per our review. This is a very important issue because it undermines results. Not only low power makes researchers discard results corresponding to nonnull effect sizes—this is the very concept of *false negative*—but it raises questions on conditions that are accepted as significant too. As the number of runs in underpowered studies is low (less than needed to reach a certain value of power), it is likely that at least some large *ES* is just a random occurrence. Stated differently, we cannot confirm that the large *ES* will remain large when more runs are performed. One of the characteristics of good agent-based modeling is that simulations can be made to vary significantly, so that every run is different from another with the *same* configuration of parameters. Of course, these differences should be less relevant than those with runs from *different* configurations of parameters. However, in order for this to happen, the modeler needs to make sure that each configuration runs a number of times that is sufficient to exclude that the similarities (or the differences) are not product of random variation. This is why appropriately powered studies in agent-based simulation research are absolutely key. Discarding this issue on the claim that one is being conservative equals to stating that one does not know whether results are coming off a random effect or a stable, reliable, replicable effect. The least runs or groups one has in the simulation, the more one is exposed to the fact that results are inconsistent and/or anchored to a random occurrence. This is, we believe, the strongest argument for the need of power analysis in ABM.

Finally, agent-based modelers and simulation researchers in general are particularly keen on transposing the standards of empirical research to computational simulated environments. In this article, we advocate for more stringent standards for computational simulation models (see above). From that stance, what matters is that, for example, significance reference for modelers is still $\alpha = 0.05$, or $1 - \beta = 0.80$. These are very lax standards in simulated environments. However, as we show in the article, on average, ABM simulation studies do not even match them. Hence, there probably is an issue with lax standards in the community or, maybe, with lack of academic discussion on these issues. As we wrote above, these standards may still affect the interpretation of results although they are probably a secondary issue compared to the complete absence of power testing. We hope this article contributes to start a discussion on these important topics.

In short, all the four aspects above point at the fact that power analysis is a tool to make results more robust and reliable. There is no shortcut around these two aspects as we believe they are much needed in simulations (as well as in any other scientific analysis). Disregarding power may be considered as a conservative move when, in fact, it may just be that one is leaning on random effects reflected in the simulation results. Additional runs may end up changing the "face" of results, hence making them more robust and reliable. This, we believe, is a very important aspect that has the potential to make simulations more palatable to other organizational scholars.

4.2 The subtle risks of overpowered studies

Some of the studies in Table 2 appear to be overpowered, i.e. calculations show a number that is very close to 1.00 within the range of computational precision. This means that the number we show in the table (0.999) is practically undistinguishable from, although it can never be, 1.00. What happened in these cases is that researchers overran their project, performing an astonishingly high number of runs reaching an incredibly low probability for Type-II error to appear under the hypothesized effect size. For example, our estimation from information available in Sharpanskykh and Stroeve (2011) and Sobkowicz (2010) shows that they performed 8000 runs while Zappala and Logan (2010) did 5000 runs. The most over-performed model we found is Aggarwal et al. (2011) with 10,000 runs performed. These are researchers that showed some awareness of the issues related to low power and decided to produce a number of runs so high that the problem would not appear to be relevant any more. This can only happen when the simulation is not time consuming or, in the case it is, when researchers have supercomputers available. However, is this approach sound? What are the risks of overpower? Although a full article is needed to analytically show what are the actual risks of overpower, we can discuss a few points here as they seem particularly relevant to our results.

In other disciplines, such as medicine, overpowering studies bears high financial costs (Girard 2005). Luckily enough, the decrease in the cost of computing power over the last decades has been so steady that the cost of most ABM is nowadays negligible with respect to more traditional experiments. However, there are risks of overpower besides waste of time.

In particular, the risk is that overpowered studies may lead modelers to notice effects so small that are not worth considering. Mone et al. (1996, p. 115) clearly state that "Excessively large samples [...] raise a serious concern [...] of oversensitivity to trivial or irrelevant findings." This happens when secondary or marginal elements appear to be statistically significant, just because of very large samples. What we are trying to convey may appear clearer when one fixes α and β and looks at the relation between sample size n and effect size (*ES*). The larger the *ES* between two configurations of parameters the least runs are needed to reach the stated values of α and β; conversely, the smaller the *ES* the larger the number of runs. This implies that, when the number of runs increases for fixed α and β, hypothesis testing procedures associated with a very small *ES* will reach the stated values of α and β.

Consequently, researchers may end up not being able to distinguish between more or less important effects because both of them appear statistically significant.[6]

[6] The point we are going to make is similar to what Friston (2012) calls the "fallacy of classical inference," although we do not necessarily advocate his solution. We believe that a clear statement of the significance threshold and of the required power under an hypothesized effect size is always better than a ritual bound on the number of observations.

This consequence of excessive power is rarely stressed in statistical textbooks, but notable exceptions are DeGroot (1986, p. 497), Bickel and Doksum (2001, p. 231) and Larsen and Marx (2012, p. 383).[7] The topic is more often brought to the fore in applied statistics (Hochster 2008; McPhaul and Toto 2012, p. 61).

All in all, overpowered simulations end up being less reliable than appropriately powered simulations. This is not to state that results are to be discarded completely but they are less sound than better calibrated simulations. This property can be turned to good account in order to test how a model performs under extreme or boundary conditions. After obtaining the number of runs via power analysis and testing different parameter configurations, researchers have a first set of results. The following step would be to indiscriminately increment the number of runs to reach overpower with the purpose of testing when previously irrelevant (insignificant) results become statistically significant, if they do. This procedure would give modelers two pieces of information at least: (a) it is a "stress" test for the model and, as such, it may reveal modeling inaccuracies or faults (referred to as 'faulty design' for underpowered studies above), and (b) it allows researchers to have a better understanding of how/when a particular set of conditions is meaningful to the modeling effort. Of course, this is feasible only when the simulation is not time consuming.[8]

As our results seem to suggest, this risk of overpower is very significant for ABM research, where the ease of producing additional runs of the model may affect how "clean" and "relevant" results are (Chase and Tucker 1976; Lykken 1968). Appropriate sample size chosen in accordance with a prescribed level of power may be the answer to get clean data. Another implication of our results seems to suggest that there is no clear indication on how to implement statistical power analysis in ABM research. This may be at the basis of most studies not reporting power or misunderstanding the importance of number of runs determination. The following subsection is dedicated to this specific point and it shows two formulas we derived for sample size calculations for agent-based models.

5 Two new formulas for the determination of the number of runs

In ABM, the researcher has direct control over more factors than in most traditional data collection situations (e.g., real experiments, surveys, etc.), because parameter values have been chosen by the researcher and the incremental cost of adding further observations to the sample is generally low. Despite this, the previous sections delineated a situation in which most papers fail to achieve the most elementary power requirements. On the one hand, this is probably due to the fact that most researchers are unaware of the concept of power and of its importance in sample size determination. On the other hand, formulas helping researchers in the

[7] Pericchi and Pereira (2016, Sects. 1.3 and 1.4) go a bit further and present a (rather artificial) example in which the accumulation of information apparently in favor of an hypothesis leads to its rejection.

[8] We owe this very interesting consideration to one of the reviewers of this paper, whom we thank very much.

determination of the sample size as a function of power are not readily available in the literature.

Apart from the classic and computationally-demanding method of numerically inverting the formula for the power, popularized by Cohen and embodied in the package `pwr` in R, it is customary (Norman and Streiner 1998, pp. 214–215) to approach the multivariate case by reducing it to the univariate, covered in Lehr's formula (Lehr 1992).

In this section, we provide and discuss two formulas for sample size determination (runs, in the case of ABM) that explicitly take into account the multivariate nature of the comparisons.[9]

5.1 A general formula for *n*

Let α be the Type-I error rate, and β the Type-II error rate that one wants to achieve. We consider an ANOVA test of the null hypothesis $\mathsf{H}_0 : \mu_1 = \ldots = \mu_J$. Let ES be the effect size of the test (Cohen 1988, 1992). In this case the formula of ES is more complex than the one seen in Sect. 2.2 for the t test, but the general interpretation is similar, i.e. ES is a measure of the distance of the real values $\mu_1, \ldots, \mu_J$ with respect to the null hypothesis H_0. It turns out that in this case n asymptotically behaves as $n^\star$ where:

$$n^\star = \frac{1}{J \cdot ES} \left(\sqrt{\chi^2_{1-\alpha, J-1}} + \sqrt{\ln \frac{\left(\chi^2_{1-\alpha, J-1}\right)^{\frac{J-2}{2}}}{2\pi\beta^2}} \cdot \left(1 - \frac{J}{4} \cdot \frac{\ln \ln \frac{\left(\chi^2_{1-\alpha, J-1}\right)^{\frac{J-2}{2}}}{2\pi\beta^2}}{\ln \frac{\left(\chi^2_{1-\alpha, J-1}\right)^{\frac{J-2}{2}}}{2\pi\beta^2}} \right) \right)^2 . \qquad (1)$$

This formula shows several facts. First, when the effect size ES is small, a larger sample size is required. Second, when β decreases, n increases: in particular, when β is near to 0, n behaves like $\frac{2}{J \cdot ES} \cdot |\ln \beta|$. Third, when α decreases, n increases: in this case too, n increases like $\frac{2}{J \cdot ES} \cdot |\ln \alpha|$.

5.2 An empirical formula

The previous formula is valid for fixed α and ES, and is accurate for not too large J and small β. In this section, the task is to find an accurate formula for $n = n(J, ES)$, valid for a wider range of J and effect sizes ES, but restricted to the values $\alpha = 0.01$ and $\beta = 0.05$ (see above). We have resorted to a response surface analysis (Box and Wilson 1951). More details on the derivation are in the Appendix.

The proposed formula is:

$$n(J, ES) \simeq 14.091 \cdot J^{-0.640} \cdot ES^{-1.986}. \qquad (2)$$

A related formula for N can be obtained through the equality $N = n \cdot J$.

[9] Details of how formula (1) is derived are presented in Seri and Secchi (2014). How to apply this formula to agent-based models and simulation is shown in a previous study (Seri and Secchi 2014), where the R code to perform the test is also made available.

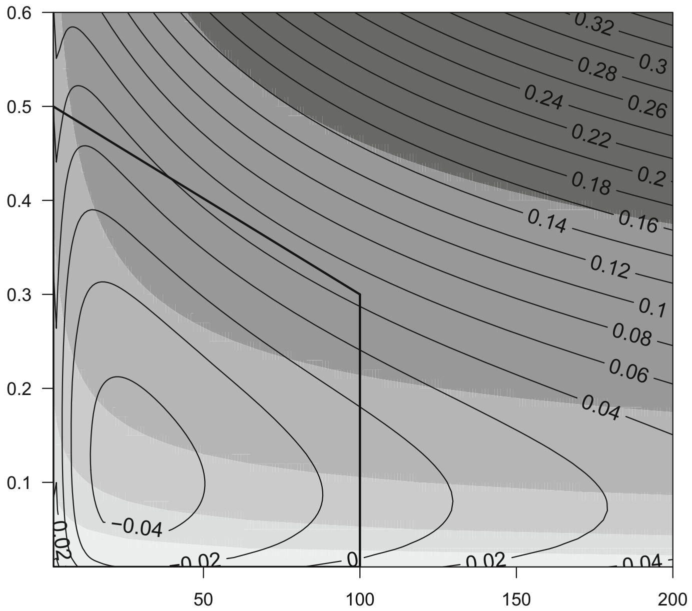

Fig. 3 Accuracy of empirical formula (2): the value of the function $\frac{n(J,ES)-14.091\cdot J^{-0.640}\cdot ES^{-1.986}}{n(J,ES)}$ is displayed through the level curves; the value of $n(J,ES)$ is displayed in shades of grey (each hue corresponds respectively, from darker to lighter, to $n < 4$, $4 < n < 16$, $16 < n < 64$, $64 < n < 256$, $256 < n < 1024$, $1024 < n$); the area on which the function has been calibrated is displayed as a trapezium

A graphical representation of the accuracy is obtained in Fig. 3 that displays the function $\frac{n(J,ES)-14.091\cdot J^{-0.640}\cdot ES^{-1.986}}{n(J,ES)}$ for J varying between 2 and 200 and ES between 0.01 and 0.6. The value of the function is displayed through the level curves. The value of $n(J,ES)$ is displayed in shades of grey (each hue corresponds respectively, from darker to lighter, to $n < 4$, $4 < n < 16$, $16 < n < 64$, $64 < n < 256$, $256 < n < 1024$, $1024 < n$). The area on which the function has been calibrated is displayed as a trapezium. It is clear from the figure that the accuracy deteriorates rapidly for large J and ES, but when ES is moderate and J is large, the formula is overall quite accurate. Moreover, note that the percentage error is higher where n is smaller, so that the error is comparatively less serious.

It is fair to note that, while formula (1) is a theoretical result, formula (2) is empirical. As such, it is not backed by a rigorous mathematical derivation but offers a guaranteed percentage error for certain values of the parameters.

6 Conclusions

In this article, we have described the importance of statistical power analysis for ABM research, especially when applied to the field of management and organizational behavior. We have then reviewed the literature on ABM from selected outlets in the social and behavioral sciences, years 2010–2013, and found that most studies are underpowered or do not provide any indication on *ES*, α levels, number of runs, or power. This is a very surprising and worrying result because it points at the reliability and significance of ABM research. Most importantly, it points at the need that every ABM researcher at least asks the question on how to avoid Type-II error, making results more robust and consistent. In the previous section, we have derived some implications and presented formulas for sample size (i.e. number of runs) determination of particular interest for agent-based and simulation research. Although the focus of this article is on ABM, the question "how many runs" a simulation should run is not strange to other techniques and we cannot exclude this approach can be successfully adopted in other areas of computational simulation. It is very likely that what we suggest may be useful to those running fitness landscape (or NK) models as well, although NK models may be considered "close relatives" to ABM. Given the scope of the current article, we leave further considerations on this possibility to future research.

ABM is a very promising technique, and it is spreading among the many disciplines of the social and behavioral sciences. Management and organizational behavior seem to lag behind this "new wave" of simulation research (Secchi 2015; Neumann and Secchi 2016; Fioretti 2013). However, this can be a strength more than a weakness. The first years of ABM research have been years of experimentations and challenge to find appropriate and sound methods. Although these are ongoing, our field can step in simulation research from a more solid ground, thanks to what has been done in the last twenty years. This may put management and organizational behavior on a more advanced ground, ready to develop the next generation of ABM simulation and research. Power analysis is part of this toolkit of the advanced simulation modeler.

Another aspect of the use of power is clearly related to the type of results that come out of ABM models. If results of any given simulation are not solid enough, there is the risk that scholars may go back to old prejudices on simulation studies. In the recent past, computer simulation suffered from the excessive simplification of assumptions, abstraction (i.e. distance from reality), and complicated design. Results were often deemed very difficult to grasp and practical implications were lacking. Inappropriate power may face the risk of seeing these prejudices come back and undermine what is the most promising advancement in computer simulation we have seen in decades. This is particularly important in management and organization studies because ABM use has just started.

As we argue in the article, not only we need to encourage researchers to be more precise in the determination of the number of runs for their simulations, but we also need to establish thresholds that are meaningful for ABM research. Our proposal is that of defining a power of 0.95 at a 0.01 significance level.

There are a few limitations of this article and we mention a couple. First of all, we do not know what the *ES* of the selected studies actually is, and our review may be based on misjudgement, if one was to show that the *ES* of those articles is higher than hypothesized. However, we cannot do science hoping that data and results are sound enough. On the contrary, we should develop scenarios that allow us to make informed decisions on possibly unfavorable as well as more favorable occurrences.

Another limitation is that our proposed thresholds—i.e. power of 0.95 at a 0.01 significance level—may reveal to be inadequate or too restrictive. More research is needed to assure modelers that these are reasonable levels for producing sound and clean results.

Despite these limitations, the article indicates that there are some important reasons why statistical power analysis is particularly important for ABM research per se and for the diffusion of this technique in management and organization studies.

Appendix

We have considered all integer values of J ranging from 2 to 100, for a total of 99 values, and all effect sizes between 0.01 and 0.5 with a mesh size of 0.01, for a total of 50 values. This constitutes a rectangle in the plane (J, ES).

Then we have removed the combinations (J, ES) such that $J + 490 \cdot ES \geq 247$: this is equivalent to remove the upper right corner of the rectangle. We came up with this solution because removing it does not alter the behavior of the function for large J and ES, but provides a much better approximation for small values of J and ES. This choice is reasonable because it is sensible to suppose that, when J is increased, this happens adding new combinations of parameters near to the previous ones, so that ES witnesses an overall decrease.

For each J and ES, for a total of 3909 values, we have computed the exact value of $n = n(J, ES)$ with $\alpha = 0.01$ and $\beta = 0.05$. We have then approximated n through the function $\gamma_0 \cdot J^{\gamma_1} \cdot ES^{\gamma_2}$ choosing $(\gamma_0, \gamma_1, \gamma_2)$ to minimize the function:

$$\max_{(J,ES)} \left| \frac{n(J, ES) - \gamma_0 \cdot J^{\gamma_1} \cdot ES^{\gamma_2}}{n(J, ES)} \right|$$

on the grid of J and ES values (the effect of the granularity of the grid on the solution has been investigated in Still 2001; Seri and Choirat 2013). This means that we are minimizing the maximum absolute percentage error in the approximation of $n(J, ES)$ over the grid. The present formula gives a maximum absolute percentage error of 4.83 % on the grid. The formula for $N = n \cdot J$ obtained from the previous one has the same maximum absolute percentage error on the same grid.

References

Abbas SMA (2013) An agent-based model of the development of friendship links within Facebook. Comput Math Organ Theo 19(2):232–252

Aggarwal VA, Siggelkow N, Singh H (2011) Governing collaborative activity: interdependence and the impact of coordination and exploration. Strat Manag J 32(7):705–730

Ahrweiler P, Gilbert N, Pyka A (2011) Agency and structure: a social simulation of knowledge-intensive industries. Comput Math Organ Theo 17(1):59–76

Altaweel M, Alessa LN, Kliskey AD (2010) Social influence and decision-making: evaluating agent networks in village responses to change in freshwater. J Artif Soc Soc Simul 13(1):15

Arrow KJ (1960) Decision theory and the choice of a level of significance for the *t*-test. In: Contributions to probability and statistics. Stanford University Press, Stanford, pp 70–78

Arroyo J, Hassan S, Gutiérrez C, Pavón J (2010) Re-thinking simulation: a methodological approach for the application of data mining in agent-based modelling. Comput Math Organ Theo 16(4):416–435

Ballinas-Hernández AL, Munoz-Meléndez A, Rangel-Huerta A (2011) Multiagent system applied to the modeling and simulation of pedestrian traffic in counterflow. J Artif Soc Soc Simul 14(3):2

Bausch AW (2013) Evolving intergroup cooperation. Comput Math Organ Theo 20(4):1–25

Bickel PJ, Doksum KA (2001) Mathematical statistics. Basic ideas and selected topics, vol 1. Prentice Hall, Upper Saddle River

Boero R, Bravo G, Castellani M, Squazzoni F (2010) Why bother with what others tell you? An experimental data-driven agent-based model. J Artif Soc Soc Simul 13(3):6

Bosse T, Gerritsen C (2010) Social simulation and analysis of the dynamics of criminal hot spots. J Artif Soc Soc Simul 13(2):5

Box GEP, Wilson KB (1951) On the experimental attainment of optimum conditions (with discussion). J R Stat Soc Ser B 13:1–45

Cassell BA, Wellman MP (2012) Asset pricing under ambiguous information: an empirical game-theoretic analysis. Comput Math Organ Theo 18(4):445–462

Castellani B, Rajaram R (2012) Case-based modeling and the SACS toolkit: a mathematical outline. Comput Math Organ Theo 18(2):153–174

Cecconi F, Campenni M, Andrighetto G, Conte R (2010) What do agent-based and equation-based modelling tell us about social conventions: the clash between ABM and EBM in a congestion game framework. J Artif Soc Soc Simul 13(1):6

Chase LJ, Tucker RK (1976) Statistical power: derivation, development, and data-analytic implications. Psychol Record 26:473–486

Choirat C, Seri R (2012) Estimation in discrete parameter models. Stat Sci 27(2):278–293

Cioffi-Revilla C, De Jong K, Bassett J (2012) Evolutionary computation and agent-based modeling: biologically-inspired approaches for understanding complex social systems. Comput Math Organ Theo 18(3):356–373

Cockburn D, Crabtree SA, Kobti Z, Kohler TA, Bocinsky RK (2013) Simulating social and economic specialization in small-scale agricultural societies. J Artif Soc Soc Simul 16(4):4

Coen C (2009a) Contrast or assimilation: choosing camps in simple or realistic modeling. Comput Math Organ Theo 15(1):19–25

Coen C (2009b) Simple but not simpler. Introduction CMOT Special Issue-simple or realistic. Comput Math Organ Theo 15(1):1–4

Coen CA, Maritan CA (2011) Investing in capabilities: the dynamics of resource allocation. Organ Sci 22(1):99–117

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

Cohen J (1992) A power primer. Psychol Bull 112(1):155–159

Cohen MD, March JG, Olsen HP (1972) A garbage can model of organizational choice. Adm Sci Q 17(1):1–25

Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 1(3):140216

DeGroot MH (1986) Probability and statistics, 2nd edn. Addison-Wesley Publishing Co, Reading

Demarest J, Pagsuyoin S, Learmonth G, Mellor J, Dillingham R (2013) Development of a spatial and temporal agent-based model for studying water and health relationships: The case study of two villages in Limpopo, South Africa. J Artif Soc Soc Simul 16(4):3

Dubois E, Barreteau O, Souchére V (2013) An agent-based model to explore game setting effects on attitude change during a role playing game session. J Artif Soc Soc Simul 16(1):2

Dugundji ER, Gulyás L (2013) Structure and emergence in a nested logit model with social and spatial interactions. Comput Math Organ Theo 19(2):151–203

Dunn AG, Gallego B (2010) Diffusion of competing innovations: the effects of network structure on the provision of healthcare. J Artif Soc Soc Simul 13(4):8

Edmonds B, Moss S (2005) From KISS to KIDS: an ‘anti-simplistic’ modelling approach. In: Davidson P (ed) Multi agent based simulation, vol 3415. Lecture notes in artificial intelligence. Springer, New York, pp 130–144

Fairchild G, Hickmann KS, Mniszewski SM, Del Valle SY, Hyman JM (2014) Optimizing human activity patterns using global sensitivity analysis. Comput Math Organ Theo 20(4):394–416

Fiedler K, Kutzner F, Krueger JI (2012) The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. Perspect Psychol Sci 7(6):661–669

Fioretti G (2013) Agent-based simulation models in organization science. Organ Res Methods 16(2):227–242

Fioretti G, Lomi A (2010) Passing the buck in the garbage can model of organizational choice. Comput Math Organ Theo 16(2):113–143

Fisher RA (1955) Statistical methods and scientific induction. J R Stat Soc Ser B 17(1):69–78

Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh

Fisher RA (1926) The arrangement of field experiments. J Minist Agric GB 33:83–94

Fisher RA (1956) Statistical methods and scientific inference. Oliver and Boyd, Edinburgh

Fonoberova M, Fonoberov VA, Mezic I, Mezic J, Brantingham PJ (2012) Nonlinear dynamics of crime and violence in urban settings. J Artif Soc Soc Simul 15(1):2

Fridman N, Kaminka G (2010) Modeling pedestrian crowd behavior based on a cognitive model of social comparison theory. Comput Math Organ Theo 16(4):348–372

Friston K (2012) Ten ironic rules for non-statistical reviewers. NeuroImage 61(4):1300–1310

Gigerenzer G (2004) Mindless statistics. J Socio-Economics 33(5):587–606

Gilbert N (2008) Agent-based models, Quantitative applications in the social sciences, vol 153. Sage, Thousand Oaks

Girard P (2005) Clinical trial simulation: a tool for understanding study failures and preventing them. Basic Clin Pharmacol Toxicol 96(3):228–234

Grazzini J (2012) Analysis of the emergent properties: stationarity and ergodicity. J Artif Soc Soc Simul 15(2):7

Grow A, Flache A (2011) How attitude certainty tempers the effects of faultlines in demographically diverse teams. Comput Math Organ Theo 17(2):196–224

Gulden TR (2013) Agent-based modeling as a tool for trade and development theory. J Artif Soc Soc Simul 16(2):1

Hallahan M, Rosenthal R (1996) Statistical power: concepts, procedures, and applications. Behav Res Theo 34(5–6):489–499

Heckbert S (2013) MayaSim: an agent-based model of the ancient Maya social-ecological system. J Artif Soc Soc Simul 16(4):11

Hirshman B, St Charles J, Carley K (2011) Leaving us in tiers: can homophily be used to generate tiering effects? Comput Math Organ Theo 17(4):318–343

Hochster HS (2008) The power of “P”: On overpowered clinical trials and “positive” results. Gastrointest Cancer Res 2(2):108–109

Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55(1):19–24

Hoser N (2013) Public funding in the academic field of nanotechnology: a multi-agent based model. Comput Math Organ Theo 19(2):253–281

Jansson F (2013) Pitfalls in spatial modelling of ethnocentrism: a simulation analysis of the model of Hammond and Axelrod. J Artif Soc Soc Simul 16(3):2

Johnson VE (2013) Revised standards for statistical evidence. Proc Natl Acad Sci 110(48):19313–19317

Kim Y, Zhong W, Chun Y (2013) Modeling sanction choices on fraudulent benefit exchanges in public service delivery. J Artif Soc Soc Simul 16(2):8

Korn EL (1990) Projecting power from a previous study: maximum likelihood estimation. Am Stat 44(4):290–292

Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 4:863

Lakens D, Evers ERK (2014) Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. Perspect Psychol Sci 9(3):278–292

Larsen R, Marx M (2012) An introduction to mathematical statistics and its applications, 5th edn. Prentice Hall, Boston

Lee K, Kim S, Kim CO, Park T (2013) An agent-based competitive product diffusion model for the estimation and sensitivity analysis of social network structure and purchase time distribution. J Artif Soc Soc Simul 16(1):3

Lee S (2010) Simulation of the long-term effects of decentralized and adaptive investments in cross-agency interoperable and standard IT systems. J Artif Soc Soc Simul 13(2):3

Lehr R (1992) Sixteen S-squared over D-squared: a relation for crude sample size estimates. Stat Med 11(8):1099–1102

Lenth RV (2001) Some practical guidelines for effective sample size determination. Am Stat 55(3):187–193

Letia IA, Slavescu RR (2012) Logic-based reputation model in e-commerce simulation. J Artif Soc Soc Simul 15(3):7

Levine SS, Prietula MJ (2012) How knowledge transfer impacts performance: a multilevel model of benefits and liabilities. Organ Sci 23(6):1748–1766

Liu XS (2014) Statistical power analysis for the social and behavioral sciences. Routledge, New York

Lykken DT (1968) Statistical significance in psychological research. Psychol Bull 70(3):151–159

McPhaul MJ, Toto RD (2012) Clinical research: from proposal to implementation. Lippincott Williams & Wilkins, Philadelphia

Meadows M, Cliff D (2012) Reexamining the relative agreement model of opinion dynamics. J Artif Soc Soc Simul 15(4):4

Meyer M, Lorscheid I, Troitzsch KG (2009) The development of social simulation as reflected in the first ten years of JASSS: a citation and co-citation analysis. J Artif Soc Soc Simul 12(4):12

Meyer M, Zaggl MA, Carley KM (2011) Measuring CMOT's intellectual structure and its development. Comput Math Organ Theo 17(1):1–34

Miller JH, Page SE (2007) Complex adaptive systems. An introduction to computational models of social life. Princeton University Press, Princeton

Miller KD, Lin SJ (2010) Different truths in different worlds. Organ Sci 21(1):97–114

Miller KD, Pentland BT, Choi S (2012) Dynamics of performing and remembering organizational routines. J Manag Stud 49(8):1536–1558

Miodownik D, Cartrite B, Bhavnani R (2010) Between replication and docking: "adaptive agents, political institutions, and civic traditions" revisited. J Artif Soc Soc Simul 13(3):1

Mone MA, Mueller GC, Mauland W (1996) The perceptions and usage of statistical power in applied psychology and management research. Pers Psychol 49(1):103–120

Montes G (2012) Using artificial societies to understand the impact of teacher student match on academic performance: the case of same race effects. J Artif Soc Soc Simul 15(4):8

Morris CN (1987) Testing a point null hypothesis: the irreconcilability of p values and evidence—comment. J Am Stat Assoc 82(397):131–133

Mungovan D, Howley E, Duggan J (2011) The influence of random interactions and decision heuristics on norm evolution in social networks. Comput Math Organ Theo 17(2):152–178

Nan N (2011) Capturing bottom-up information technology use processes: a complex adaptive systems model. MIS Q 35(2):505–532

Neumann M, Secchi D (2016) Exploring the new frontier: computational studies of organizational behavior. In: Secchi D, Neumann M (eds) Agent-based simulation of organizational behavior. New frontiers of social science research. Springer, New York, pp 1–16

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

Nongaillard A, Mathieu P (2011) Reallocation problems in agent societies: a local mechanism to maximize social welfare. J Artif Soc Soc Simul 14(3):5

Norman GR, Streiner DL (1998) Biostatistics: the bare essentials. B. C. Decker, Hamilton

North MJ, Macal CM (2007) Managing business complexity: discovering strategic solutions with agent-based modeling and simulation. Oxford University Press Inc, New York

Nye BD (2013) The evolution of multiple resistant strains: an abstract model of systemic treatment and accumulated resistance. J Artif Soc Soc Simul 16(4):2

Patel A, Crooks A, Koizumi N (2012) Slumulation: an agent-based modeling approach to slum formations. J Artif Soc Soc Simul 15(4):2

Pearson ES (1955) Statistical concepts in their relation to reality. J R Stat Soc. Ser B 17(2):204–207

Pericchi L, Pereira C (2016) Adaptative significance levels using optimal decision rules: balancing by weighting the error probabilities. Braz J Probab Stat 30(1):70–90

Quera V, Beltran FS, Dolado R (2010) Flocking behaviour: agent-based simulation and hierarchical leadership. J Artif Soc Soc Simul 13(2):8

Radax W, Rengs B (2010) Prospects and pitfalls of statistical testing: insights from replicating the demographic prisoner's dilemma. J Artif Soc Soc Simul 13(4):1

Ritter FE, Schoelles MJ, Quigley KS, Klein LC (2011) Determining the numbers of simulation runs: treating simulations as theories by not sampling their behavior. In: Rothrock L, Narayanan S (eds) Human-in-the-loop simulations: methods and practice. Springer, London, pp 97–116

Robinson S (2004) Simulation: the practice of model development and use. Wiley, Chicester

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

Royall RM (1997) Statistical evidence: a likelihood paradigm. Monographs on statistics and applied probability, vol 71. Chapman & Hall, London

Savarimuthu BTR, Cranefield S, Purvis MA, Purvis MK (2010) Obligation norm identification in agent societies. J Artif Soc Soc Simul 13(4):3

Schindler J (2012) Rethinking the tragedy of the commons: the integration of socio-psychological dispositions. J Artif Soc Soc Simul 15(1):4

Schindler J (2013) About the uncertainties in model design and their effects: an illustration with a land-use model. J Artif Soc Soc Simul 16(4):6

Secchi D (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 (2015) A case for agent-based model in organizational behavior and team research. Team Perform Manag 21(1/2):37–50

Secchi D, Neumann M (eds) (2016) Agent-based simulation of organizational behavior. New frontiers of social science research. Springer, New York

Sedlmeier P, Gigerenzer G (1989) Do studies of statistical power have an effect on the power of studies? Psychol Bull 105(2):309–316

Seri R, Choirat C (2013) Scenario approximation of robust and chance-constrained programs. J Optim Theo Appl 158(2):590–614

Seri R, Secchi D (2014) Sample size determination in multivariate problems. Working paper, unpublished

Shannon RE (1975) Systems simulation: the art and science. Prentice-Hall, Englewood Cliffs

Sharpanskykh A, Stroeve SH (2011) An agent-based approach for structured modeling, analysis and improvement of safety culture. Comput Math Organ Theo 17(1):77–117

Shiba N (2013) Analysis of asymmetric two-sided matching: agent-based simulation with theorem-proof approach. J Artif Soc Soc Simul 16(3):11

Shimazoe J, Burton RM (2013) Justification shift and uncertainty: why are low-probability near misses underrated against organizational routines? Comput Math Organ Theo 19(1):78–100

Siebers PO, Aickelin U (2011) A first approach on modelling staff proactiveness in retail simulation models. J Artif Soc Soc Simul 14(2):2

Sioson E (2012) Flora: a testbed for evaluating the potential impact of proposed systems on population wellbeing. J Artif Soc Soc Simul 15(3):6

Sobkowicz P (2010) Dilbert-Peter model of organization effectiveness: computer simulations. J Artif Soc Soc Simul 13(4):4

Still G (2001) Discretization in semi-infinite programming: the rate of convergence. Math Prog 91(1, Ser. A):53–69

Sutcliffe A, Wang D (2012) Investigating the relative influence of genes and memes in healthcare. J Artif Soc Soc Simul 15(2):1

Udayaadithya A, Gurtoo A (2013) Governing the local networks in Indian agrarian societies: a MAS perspective. Comput Math Organ Theo 19(2):204–231

van der Vaart A (1998) Asymptotic statistics, vol 3., Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge

Villarroel JA, Taylor JE, Tucci CL (2013) Innovation and learning performance implications of free revealing and knowledge brokering in competing communities: Insights from the Netflix prize challenge. Comput Math Organ Theo 19(1):42–77

Waldeck R (2013) Segregated cooperation. J Artif Soc Soc Simul 16(4):14

Wang M, Hu X (2012) Agent-based modeling and simulation of community collective efficacy. Comput Math Organ Theo 18(4):463–487

Wijermans N, Jorna R, Jager W, van Vliet T, Adang O (2013) Cross: modelling crowd behaviour with social-cognitive agents. J Artif Soc Soc Simul 16(4):1

Wildman W, Sosis R (2011) Stability of groups with costly beliefs and practices. J Artif Soc Soc Simul 14(3):6

Xianyu B (2010) Social preference, incomplete information, and the evolution of ultimatum game in the small world networks: An agent-based approach. J Artif Soc and Soc Simul 13(2):7

Xianyu B (2012) Prisoner's dilemma game on complex networks with agents' adaptive expectations. J Artif Soc Soc Simul 15(3):3

Yamanoi J, Sayama H (2013) Post-merger cultural integration from a social network perspective: a computational modeling approach. Comput Math Organ Theo 19(4):516–537

Zappala J, Logan B (2010) Effects of resource availability on consensus decision making in primates. Comput Math Organ Theo 16(4):400–415

Zhang G, Gao R (2010) Modularity and incremental innovation: the roles of design rules and organizational communication. Comput Math Organ Theo 16(2):171–200

Zhang Y, Leezer J (2010) Simulating human-like decisions in a memory-based agent model. Comput Math Organ Theo 16(4):373–399

Zou G, Yilmaz L (2011) Dynamics of knowledge creation in global participatory science communities: open innovation communities from a network perspective. Comput Math Organ Theo 17(1):35–58

Davide Secchi (PhD, University of Pavia) is an Associate Professor of organizational cognition at the COMAC Research Cluster, and the Centre for Human Interactivity (CHI), Department of Language and Communication, University of Southern Denmark in Slagelse. His current research efforts are on socially based decision making, agent-based modeling, rational processes in organizations and individual social responsibility. He is author of more than 50 articles and book chapters, the monograph *Extendable Rationality* (2011), and recently edited *Agent-Based Simulation of Organizational Behavior* with M. Neumann (2016).

Raffaello Seri (PhD, Politecnico di Milano and Université Paris Dauphine) is an Associate Professor in Econometrics at the Department of Economics, Università degli Studi dell'Insubria (Italy). His publications range from econometrics (Journal of Applied Econometrics, Journal of Econometrics), statistics (Statistical Science), and probability (The Annals of Probability) to decision making (Management Science), mathematical psychology (Journal of Mathematical Psychology), and numerical analysis (Mathematics of Computation), as well as other social and behavioral science disciplines.