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

Secchi, Davide; Seri, Raffaello

Published in:
Computational & Mathematical Organization Theory

DOI:
10.1007/s10588-016-9218-0

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Document license
CC BY-SA

Citation for published version (APA):
Controlling for false negatives in agent-based models: a review of power analysis in organizational research

Davide Secchi¹ · Raffaello Seri²,³

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
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 $\Theta$, called parameter space. As an example, a normal distribution can be characterised by the mean $\mu$, that can take any real value, and the standard deviation $\sigma$, 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 $\Theta_0$. 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 $\Theta_0$; this set is generally called $\Theta_1$ and is defined as the complement of $\Theta_0$ in $\Theta$ (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 $A$ 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 $A$, the test “rejects” the null hypothesis. On the other hand, if $t$ falls inside the region $A$, 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 $\alpha$. If we accept the null hypothesis when it is false, we commit a Type-II error (or false negative). This happens with rate $\beta$, the probability that $T$ belongs to $\mathcal{A}$ when $H_0$ is false (and $H_1$ is true). The two probabilities, $\alpha$ and $\beta$, are linked by a trade-off: in order for $\alpha$ to decrease, one needs to increase $\mathcal{A}$ (because $\alpha$ is the probability that $T$ does not belong to $\mathcal{A}$ under $H_0$), and therefore $\beta$ increases too (because $\beta$ 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 $\alpha$ is fixed and small (5 and 1% are customary values). On the other hand, $\beta$ is not directly (and rarely indirectly) controlled. However, when the sample size $N$ increases, $\beta$ generally tends to 0, so that, for $N$ large enough, one can take $\alpha$ small and hope for $\beta$ to be not too large.

### 2.1 An example on the relation between $\alpha$ and $\beta$

Consider the test for the nullity of the mean in a normally distributed population with mean $\mu$ and variance $\sigma^2$. The null hypothesis is $H_0 : \mu = 0$, the alternative hypothesis is $H_1 : \mu \neq 0$. Suppose to observe a sample $\{X_1, \ldots, X_N\}$ from the population. A test statistic for this hypothesis is:

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

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 $\alpha$ will be given by the interval $\mathcal{A} = \left[t_{\alpha,N-1}, t_{1-\alpha,N-1}\right]$ (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 $\alpha$ 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 $\beta$ is:

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

One can verify the properties we rapidly described above. When $\alpha$ decreases, $t_{\alpha,N-1}$ and $t_{1-\alpha,N-1}$ get far from 0, and $\beta$ increases. When $N$ increases, $\alpha$ is constant while $\beta$ goes to 0, as $\lambda_N$ 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.
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 $\beta$. 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 $\alpha$ and the value deemed acceptable for $\beta$ calls for a clarification.\textsuperscript{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 $\alpha$ is fixed, $\beta$ 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 $H'_0$ and $H''_0$, 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 $\alpha$ is controllable while $\beta$ isn’t completely, it is often reasonable to choose as $H_0$ the hypothesis, say $H'_0$, whose rejection is considered more serious. Indeed, in case $H''_0$ were chosen as the null, the rejection of $H'_0$ when true (that would now correspond to a Type-II error rate $\beta$) 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 $\beta/\alpha$ should be near to 4 when Type-I errors are about four times as serious as Type-II errors.

Third, the different emphasis on $\alpha$ and $\beta$ 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

\textsuperscript{2} As an interesting variation on the traditional choice of a fixed significance level, Arrow (1960) describes a procedure to compute $\alpha$ 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 \(z\) (if smaller, rejection ensues). Therefore \(z\) takes the further role of a gauge of extraordinariness and it is smaller than \(\beta\) 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 \(\beta\) 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 \(\mu/\sigma\) in the previous section. There are two kinds of power analysis that can be performed involving \(ES\). The first one is called \textit{a priori} power analysis and is generally used, before data is collected, to evaluate the sample size needed to obtain a certain value of \(\beta\) for fixed \(z\) 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 \textit{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.\footnote{This is made very clear by Morris (1987) whose example shows unequivocally that $\beta$ is by far a more reasonable measure of reliability than the estimated $ES$.}

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 $\mu_j$, $j = 1, \ldots, J$.

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 ($H_0 : \mu_1 = \ldots = \mu_J$). 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 $\beta$ 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 $\alpha$ and $\beta$ 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 \textit{a priori} nor \textit{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.

\section{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
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

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

<table>
<thead>
<tr>
<th>Article</th>
<th>ES</th>
<th>0.1</th>
<th>0.3</th>
<th>0.1</th>
<th>0.3</th>
<th>CoP</th>
<th>Runs</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control for false negatives…</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Computational and Mathematical Organization Theory</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arroyo et al. (2010)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>9, 6</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Fioretti and Lomi (2010)</td>
<td>0.161</td>
<td>0.997</td>
<td>0.355</td>
<td>0.999</td>
<td>4</td>
<td>100</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>Fridman and Kaminka (2010)</td>
<td>0.151</td>
<td>0.074</td>
<td>0.066</td>
<td>0.207</td>
<td>2</td>
<td>30</td>
<td>875</td>
<td></td>
</tr>
<tr>
<td>——-exp 1</td>
<td>0.012</td>
<td>0.036</td>
<td>0.058</td>
<td>0.124</td>
<td>2</td>
<td>15</td>
<td>875</td>
<td></td>
</tr>
<tr>
<td>——-exp 2</td>
<td>0.037</td>
<td>0.558</td>
<td>0.127</td>
<td>0.780</td>
<td>2</td>
<td>30d</td>
<td>875</td>
<td></td>
</tr>
<tr>
<td>——-exp 3</td>
<td>0.015</td>
<td>0.060</td>
<td>0.063</td>
<td>0.180</td>
<td>4</td>
<td>25</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>Zhang and Gao (2010)</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>24</td>
<td>1000</td>
<td>178</td>
<td></td>
</tr>
<tr>
<td>Zhang and Leezer (2010)</td>
<td>0.029</td>
<td>0.689</td>
<td>0.110</td>
<td>0.871</td>
<td>36d</td>
<td>10d</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>Zappala and Logan (2010)</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>1024</td>
<td>5000</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Ahrweiler et al. (2011)</td>
<td>0.017</td>
<td>0.133</td>
<td>0.073</td>
<td>0.324</td>
<td>5d</td>
<td>10</td>
<td>487</td>
<td></td>
</tr>
<tr>
<td>Grow and Flache (2011)</td>
<td>0.326</td>
<td>0.999a</td>
<td>0.572</td>
<td>0.999a</td>
<td>36</td>
<td>50</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>Hirshman et al. (2011)</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>100</td>
<td>1000</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Mungovan et al. (2011)</td>
<td>0.160</td>
<td>0.997</td>
<td>0.355</td>
<td>0.999</td>
<td>4</td>
<td>100</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>Sharpanskykh and Stroeve (2011)</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>4446</td>
<td>8000</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Zou and Yilmaz (2011)</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>1000d</td>
<td>1000d</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Castellani and Rajaram (2012)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Cassell and Wellman (2012)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Cioffi-Revilla et al. (2012)</td>
<td>0.048</td>
<td>0.826</td>
<td>0.155</td>
<td>0.940</td>
<td>8</td>
<td>30</td>
<td>360</td>
<td></td>
</tr>
<tr>
<td>Wang and Hu (2012)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Abbas (2013)</td>
<td>0.012</td>
<td>0.031</td>
<td>0.057</td>
<td>0.120</td>
<td>4</td>
<td>4d</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>Bausch (2013)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Dugundji and Gulyás (2013)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Hoser (2013)</td>
<td>0.474</td>
<td>0.999a</td>
<td>0.708</td>
<td>0.999a</td>
<td>16</td>
<td>100</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>Fairchild et al. (2014)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Shimazoe and Burton (2013)</td>
<td>0.266</td>
<td>0.999</td>
<td>0.497</td>
<td>0.999</td>
<td>8</td>
<td>100</td>
<td>360</td>
<td></td>
</tr>
<tr>
<td>Villarroel et al. (2013)</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>9</td>
<td>1000</td>
<td>334</td>
<td></td>
</tr>
<tr>
<td>Udayaditya and Gurtoo (2013)</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>0.999a</td>
<td>768</td>
<td>200</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Yamanoi and Sayama (2013)</td>
<td>0.234</td>
<td>0.999</td>
<td>0.462</td>
<td>0.999</td>
<td>25</td>
<td>50</td>
<td>174</td>
<td></td>
</tr>
<tr>
<td><strong>Journal of Artificial Societies and Social Simulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Altaweel et al. (2010)</td>
<td>0.366</td>
<td>0.999a</td>
<td>0.608</td>
<td>0.999a</td>
<td>2</td>
<td>1000</td>
<td>875</td>
<td></td>
</tr>
<tr>
<td>Boero et al. (2010)</td>
<td>0.397</td>
<td>0.999a</td>
<td>0.637</td>
<td>0.999a</td>
<td>7</td>
<td>1000d</td>
<td>393</td>
<td></td>
</tr>
<tr>
<td>Bosse and Gerritsen (2010)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Cecconi et al. (2010)</td>
<td>0.033</td>
<td>0.692</td>
<td>0.119</td>
<td>0.871</td>
<td>16d</td>
<td>16</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>Dunn and Gallego (2010)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Lee (2010)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Miodownik et al. (2010)</td>
<td>0.016</td>
<td>0.115</td>
<td>0.072</td>
<td>0.294</td>
<td>4</td>
<td>10</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>Article</td>
<td>ES 0.1</td>
<td>ES 0.3</td>
<td>ES 0.1</td>
<td>ES 0.3</td>
<td>CoP</td>
<td>Runs</td>
<td>ER</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>-----</td>
<td>------</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>Quera et al. (2010)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Radax and Rengs (2010)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.04</td>
<td>0.550</td>
<td>0.147</td>
<td>0.772</td>
<td>2</td>
<td>164</td>
<td>69c</td>
<td></td>
</tr>
<tr>
<td>Sobkowicz (2010)</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>42d</td>
<td>8000</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>Savarimuthu et al. (2010)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Xianyu (2010)</td>
<td>0.037</td>
<td>0.558</td>
<td>0.127</td>
<td>0.781</td>
<td>4</td>
<td>30</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>Ballinas-Hernández et al. (2011)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Nongaillard and Mathieu (2011)</td>
<td>0.565</td>
<td>0.999a</td>
<td>0.781</td>
<td>0.999a</td>
<td>20</td>
<td>100</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>Siebers and Ackelin (2011)</td>
<td>0.026</td>
<td>0.330</td>
<td>0.099</td>
<td>0.578</td>
<td>4</td>
<td>20</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>Wildman and Sosis (2011)</td>
<td>0.565</td>
<td>0.999a</td>
<td>0.781</td>
<td>0.999a</td>
<td>20</td>
<td>100</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>Fonoberova et al. (2012)</td>
<td>0.656</td>
<td>0.999a</td>
<td>0.846</td>
<td>0.999a</td>
<td>36</td>
<td>80</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>Grazzini (2012)</td>
<td>0.161</td>
<td>0.997</td>
<td>0.355</td>
<td>0.999</td>
<td>4</td>
<td>100</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>Letia and Slavescu (2012)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Meadows and Cliff (2012)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Montes (2012)</td>
<td>0.857</td>
<td>0.999a</td>
<td>0.953</td>
<td>0.999a</td>
<td>40</td>
<td>100</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>Patel et al. (2012)</td>
<td>0.543</td>
<td>0.999a</td>
<td>0.765</td>
<td>0.999a</td>
<td>19</td>
<td>100</td>
<td>207</td>
<td></td>
</tr>
<tr>
<td>Schindler (2012)</td>
<td>0.099</td>
<td>0.995</td>
<td>0.259</td>
<td>0.999</td>
<td>12</td>
<td>43</td>
<td>278</td>
<td></td>
</tr>
<tr>
<td>Sioson (2012)</td>
<td>0.014</td>
<td>0.058</td>
<td>0.063</td>
<td>0.174</td>
<td>2</td>
<td>24</td>
<td>875</td>
<td></td>
</tr>
<tr>
<td>Sutcliffe and Wang (2012)</td>
<td>0.424</td>
<td>0.999a</td>
<td>0.669</td>
<td>0.999a</td>
<td>48</td>
<td>50d</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>Xianyu (2012)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Cockburn et al. (2013)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Demarest et al. (2013)</td>
<td>0.137</td>
<td>0.987</td>
<td>0.318</td>
<td>0.997</td>
<td>3</td>
<td>100</td>
<td>675</td>
<td></td>
</tr>
<tr>
<td>Dubois et al. (2013)</td>
<td>0.999</td>
<td>0.999a</td>
<td>0.999</td>
<td>0.999a</td>
<td>4</td>
<td>6</td>
<td>562, 433</td>
<td></td>
</tr>
<tr>
<td>Gulden (2013)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Heckbert (2013)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
<tr>
<td>Jansson (2013)</td>
<td>0.497</td>
<td>0.999a</td>
<td>0.728</td>
<td>0.999a</td>
<td>17</td>
<td>100</td>
<td>223</td>
<td></td>
</tr>
<tr>
<td>Kim et al. (2013)</td>
<td>0.548</td>
<td>0.999a</td>
<td>0.766</td>
<td>0.999a</td>
<td>6</td>
<td>200</td>
<td>433</td>
<td></td>
</tr>
<tr>
<td>Lee et al. (2013)</td>
<td>0.991</td>
<td>0.999a</td>
<td>0.998</td>
<td>0.999a</td>
<td>80</td>
<td>100</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Nye (2013)</td>
<td>0.565</td>
<td>0.999a</td>
<td>0.781</td>
<td>0.999a</td>
<td>20</td>
<td>100</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>Schindler (2013)</td>
<td>0.024</td>
<td>0.372</td>
<td>0.095</td>
<td>0.622</td>
<td>8</td>
<td>15</td>
<td>361</td>
<td></td>
</tr>
<tr>
<td>Shiba (2013)</td>
<td>0.016</td>
<td>0.099</td>
<td>0.071</td>
<td>0.265</td>
<td>3d</td>
<td>10</td>
<td>675</td>
<td></td>
</tr>
<tr>
<td>Waldeck (2013)</td>
<td>0.046</td>
<td>0.757</td>
<td>0.148</td>
<td>0.905</td>
<td>6</td>
<td>32</td>
<td>433</td>
<td></td>
</tr>
<tr>
<td>Wijermans et al. (2013)</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td></td>
</tr>
</tbody>
</table>

*Other journals*

| Miller and Lin (2010)                | 0.225  | 0.999  | 0.451  | 0.999a | 24  | 50   | 178|
| Aggarwal et al. (2011)               | 0.999a | 0.999a | 0.999a | 0.999a | 16  | 10,000 | 231|
| Coen and Maritan (2011)              | 0.914  | 0.999a | 0.976  | 0.999a | 48d | 100  | 114|
| Nan (2011)                           | 0.179  | 0.999  | 0.392  | 0.999a | 50  | 30   | 111|
| Levine and Prietula (2012)           | 0.048  | 0.867  | 0.154  | 0.958  | 12  | 25   | 278|
| Miller et al. (2012)                 | 0.999a | 0.999a | 0.999a | 0.999a | 288 | 100  | 36|
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 did we 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 \( \alpha \) at which power calculations should

<table>
<thead>
<tr>
<th>Article</th>
<th>( ES = 0.1 )</th>
<th>( 0.3 )</th>
<th>( 0.1 )</th>
<th>( 0.3 )</th>
<th>CoP</th>
<th>Runs</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.415</td>
<td>0.783</td>
<td>0.526</td>
<td>0.842</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.395</td>
<td>0.346</td>
<td>0.373</td>
<td>0.284</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 \( \alpha = 0.01, \beta = 0.05, ES = 0.1 \)
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. 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).

---

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 $\alpha$ 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 $\alpha$—affect power. Note that we transformed logarithmically the axes in both figures, to help make sense of the distribution of results.
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 \( \alpha \) 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 \( \alpha \) 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 \( \alpha \) 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 \( \alpha \) 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 \( \alpha \) to 0.001. However, these authors did not accompany this suggestion with an analogous one concerning a decrease in \( \beta \). This creates a paradoxical situation, because none of the reasonings in Sect. 2.2 above is compatible with such a huge difference between \( \alpha \) and \( \beta \). 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 \( \alpha \) and \( \beta \) 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 $\alpha$. 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 $\alpha$ 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 indistinguishable 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 $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$; conversely, the smaller the $ES$ the larger the number of runs. This implies that, when the number of runs increases for fixed $\alpha$ and $\beta$, hypothesis testing procedures associated with a very small $ES$ will reach the stated values of $\alpha$ and $\beta$.

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 over power 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 \texttt{pwr} in \texttt{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 \(\alpha\) be the Type-I error rate, and \(\beta\) the Type-II error rate that one wants to achieve.
We consider an ANOVA test of the null hypothesis \(H_0 : \mu_1 = \ldots = \mu_J\). 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
\(n^*\) where:

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

This formula shows several facts. First, when the effect size \(ES\) is small, a larger
sample size is required. Second, when \(\beta\) decreases, \(n\) increases: in particular, when \(\beta\)
is near to 0, \(n\) behaves like \(\frac{2}{7ES} \cdot |\ln \beta|\). Third, when \(\alpha\) decreases, \(n\) increases: in this
case too, \(n\) increases like \(\frac{2}{7ES} \cdot |\ln \alpha|\).

5.2 An empirical formula

The previous formula is valid for fixed \(\alpha\) and \(ES\), and is accurate for not too large
\(J\) and small \(\beta\). 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:

\[
n(J, ES) \approx 14.091 \cdot J^{-0.640} \cdot ES^{-1.986}.
\]

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.
A graphical representation of the accuracy is obtained in Fig. 3 that displays the function \( \frac{n(J, ES) - 14.091 - 0.640 \cdot ES^{-1.986}}{n(J, ES)} \) 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.

A graphical representation of the accuracy is obtained in Fig. 3 that displays the function \( \frac{n(J, ES) - 14.091 - 0.640 \cdot ES^{-1.986}}{n(J, ES)} \) 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 experiments 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.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

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 $n \equiv n(J, ES)$ with $\alpha = 0.01$ and $\beta = 0.05$. We have then approximated $n$ through the function $n(J, ES) = n_0 \cdot J^{c_1} \cdot ES^{c_2}$ choosing $(c_0, c_1, c_2)$ to minimize the function:

$$\max_{(J, ES)} \left| \frac{n(J, ES) - n_0 \cdot J^{c_1} \cdot ES^{c_2}}{n(J, ES)} \right|$$

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


Controlling for false negatives...


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


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


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


© Springer

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.