WHEN COMPETITION IS PUSHED TOO HARD. AN AGENT-BASED MODEL OF STRATEGIC BEHAVIOUR OF REFEREES IN PEER REVIEW

Juan Bautista Cabotà Departament d'Informàtica Universitat de València Avinguda de la Universitat, s/n 46100 Burjassot-València Email: juan.cabota@uv.es Francisco Grimaldo Departament d'Informàtica Universitat de València Avinguda de la Universitat, s/n 46100 Burjassot-València Email: francisco.grimaldo@uv.es Flaminio Squazzoni Department of Economics and Management University of Brescia, Via San Faustino 74/B 25122 Brescia E-mail: squazzon@eco.unibs.it

KEYWORDS

Peer review; referees; competition; rational cheating; fairness; agent-based model.

ABSTRACT

This paper examines the impact of strategic behaviour of referees on the quality and efficiency of peer review. We modelled peer review as a process based on knowledge asymmetry and subject to evaluation bias. We built two simulation scenarios to investigate largescale implications of referee behaviour and judgment bias. The first one was inspired by "the luck of the reviewer draw" idea. In this case, we assumed that referees randomly fell into Type I and Type II errors, i.e., recommending submissions of low quality to be published or recommending against the publishing of submissions which should have been published. In the second scenario, we assumed that certain referees tried intentionally to outperform potential competitors by underrating the value of their submissions. We found that when publication selection increased, the presence of a minority of cheaters may dramatically undermine the quality and efficiency of peer review even compared with a scenario purely dominated by "the luck of the reviewer draw". We also found that peer review outcomes are significantly influenced by differences in the way scientists identify potential competitors in the system.

INTRODUCTION

Pressures towards competition have recently increased in science, with scientists harshly competing for funds and reputation at a national and international level (e.g., Fanelli 2010). The digitalization of scientific publications and the development of scientometrics now permit to measure the scientist's performance through a variety of indicators, such as the impact factor, the "*h* index" and so on. By objectively ranking everyone, these measures allow us to compare our respective achievements and better identify potential competitors. Although specificities of standards exist even between sister disciplines, the strength of quantitative rankings is especially strong in the realm of the so-called "hard sciences", whereas it is certainly less true for the humanities, where objective measures are hardly systematically applicable.

The increasing importance of performance indicators requires to understand whether competitive spirits of scientists could influence the peer review process. For instance, referees could be tempted to exploit strategically their important gatekeepers' position to outperform potential competitors (Thurner and Hanel 2011; Grimaldo and Paolucci 2013; Callahan 2004). Indeed, while the debate on misbehaviour followed by submission authors has gained momentum, even in the media, less is known about the possible effect of referee behaviour (e.g., Bornmann 2011). On the one hand, as shown by the recent Stapel scandal, unfair scientists might profit from information asymmetry of referees and editors by manipulating experimental data and overselling their results. In doing so, they gain competitive advantages against fair colleagues (Crocker and Crooper 2011). On the other hand, referees could be induced to deliberately underrate potential competitors by providing unfair judgment, especially when possible publication of work by these colleagues does not have positive effects on their own reputation (e.g., in terms of citations), includes results that challenge referees' work or increases reputation of competing research groups for funding.

Our paper aims to examine these problems by proposing a modeling approach that looks at scientist behaviour in the peer review process to understand complex macro implications (e.g., Edmonds *et al.* 2011; Roebber and Schultz 2011; Thurner and Hanel 2011; Allesina 2012; Payette 2012; Squazzoni and Gandelli 2012). We modelled a population of agents interacting as authors and referees in a selective science system. Following Thurner and Hanel (2011) and Grimaldo and Paolucci (2013), we tested macro implications of unfair referee behaviour for the quality of peer review. Unlike these studies, we also considered efficiency problems and focused on aggregate consequences in terms of system's resource allocation and growth inequality. The rest of the paper is as follows. In the second section, we introduce the model and the simulation parameters, while in the third we present our simulation scenarios. In the fourth one, results are illustrated, while, in the concluding section, we present a summary of results and draw some implications for the current debate on peer review.

THE MODEL

Following Squazzoni and Gandelli (2012; 2013), we assumed a population of N scientists (N = 200) randomly selected each to playfill one of two roles: author or referee. The task of an author was to submit an article with the goal of having it accepted to be published. The task of a referee was to evaluate the quality of author submissions. As informed by the referees' opinion, only the best submissions were published (i.e., those exceeding the publication rate).

We gave each agent a set of resources which were initially homogeneous (R_a =1). Resources were a proxy of academic status, position, experience, and scientific achievement. The guiding principle was that the more scientists published, the more resources they had access to, and thus the higher their academic status and position.

We assumed that resources were needed both to submit and review an article. With each simulation step, agents were endowed with a fixed amount of resources f, equal for all (e.g., common access to research infrastructure and internal funds, availability of PhD. students, etc.). They then accumulated resources according to their publication score.

We assumed that the quality of submissions varied and was dependent on agent resources. Each agent had resources $R_a \in \mathbb{R}$, from which we derived an expected submission quality μ as follows:

$$\mu = \frac{\nu * R_a}{\nu * R_a + 1} \tag{1}$$

where *v* indicated the velocity at which the quality of the submission increased with the increase of author resources. For instance, this means that for v = 0.1 each agent needed $R_a = 10$ to reach a medium-sized expected quality submission ($\mu = 0.5$).

We assumed that authors varied in terms of the quality of their output depending on their resources. More specifically, the quality of submissions by authors followed a standard deviation σ which proportionally varied according to agent resources and followed a normal distribution $N(\mu, \sigma)$. This means that, with some probability, top scientists could write average or low quality submissions, and average scientists had some chance to write good submissions.

We assumed that successful publication multiplied author resources by a value m, which varied between 1.5 for less productive published authors and 1 for more productive published authors. We assigned a heterogeneous value of m after various explorations of the parameter space. This was seen as mimicking reality, where publication is crucial in explaining differences in scientists' performance, but is more important for scientists at the initial stages of their academic careers and cannot infinitely increase for top scientists. Thus, the resources of published authors grew accordingly, leading to subsequent submissions of presumably higher quality. If not published, following the "winner takes all" rule characterizing science, we assumed that authors lost all resources invested prior to submitting.

The chance of being published was determined by evaluation scores assigned by referees. The value of author submissions was therefore not objectively determined (i.e., it did not perfectly mirror the real quality of submissions), but was instead dependent on the referees' opinion. We assumed that reviewing was a resource-intensive activity and that agent resources determined both the agent's reviewing quality and the cost to the reviewer (i.e., time lost for publishing their own work). The total expense S for any referee was calculated as follows:

$$S = \frac{1}{2}R_r [1 + (Q_a - \mu_r)]$$
(2)

where R_r was the referee's resources, Q_a was the real quality of the author's submission and μ_r was the referee's expected quality. This last was calculated as in equation (1). It is worth noting that, when selected as referees, agents not only needed to allocate resources toward reviewing but also potentially lost additional resources as a result of not being able to publish their own work in the meantime.

We assumed that authors and referees were randomly matched one to one so that multiple submissions and reviews were not possible and the reviewing effort was equally distributed among the population. We assumed that reviewing expenses grew linearly with the quality of authors' submissions. We assumed that, if referees were matched with a submission of a quality close to a potential submission of their own, they allocated 50% of their available resources toward reviewing. They spent fewer resources when matched with lower quality submissions, more when matched with higher quality submissions. Reviewing expenses, however, were proportionally dependent on agent resources, meaning that top scientists would be expected to spend less time reviewing in general, as they have more experience and are better able to evaluate sound science than average scientists are. They will lose more resources than average scientists, however, because their time is more valuable than the latter.

SIMULATION SCENARIOS

We first built various simulation scenarios to test the impact of referee behaviour on the quality and efficiency of the peer review process. By quality, we meant the ability of peer review to ensure that only the best submissions were eventually published (e.g., Casati *et al.* 2009). By efficiency, we meant the ability of peer review to achieve quality by minimizing the resources lost by authors and the expenses incurred by referees.

In the first scenario, called "random behaviour", we assumed that referees had a constant probability of being biased in their judgment. When fair, referees had the ability to provide a consistent and unequivocal opinion which truly reflected the quality of the submission. In this case, they did their best to provide an accurate evaluation and spent all needed resources for reviewing. We assumed that referees estimated the authors' resources following a normal distribution of the actual authors' resources and a narrow standard deviation ($\sigma = R_a/10$). Then, they estimated the author submissions' quality as in (1). This meant that the evaluation scores by fair referees were likely to approximate the real value of author submissions, but we assumed that there was a chance for some bias in order to mimic typical knowledge and information asymmetries between authors and referees which characterize peer review.

In the case of unfairness, referees fell into type I and type II errors: recommending submissions of low quality to be published or recommending against the publishing of submissions which should have been published (e.g., Laband and Piette 1994; Bornmann and Daniel 2007). More specifically, when unfair, referees spent half of the resources spent by fair referees, and under- or over- estimated author submissions. To avoid the possibility that referees assigned the real value to submissions by chance we assumed that, when they underrated a submission, they estimated the authors' resources as described for fair referees, but applying an underrating factor to the actual authors' resources (u =0.1). An opposite factor was applied in the case of overrating (i.e. o = 1.9).

In the second scenario, called "*cheating*", we assumed that referees intentionally outperformed potential competitors by systematically underrating their submission, even at their own expenses (e.g., resources spent for reviewing). More specifically, we assumed that referees were capable to estimate submission authors' resources (R_a) and identify each author with an expected R_a similar or higher than his/her R_r as a competitor.

To measure the quality of peer review, we considered the percentage of errors made by referees by calculating the optimal situation, in which submissions were published according to their real value, and by measuring the discrepancy with the actual situation in each simulation step (e.g., see evaluation bias in Tab. 1).

RESULTS

Tab. 1 shows the impact of referee behaviour on the quality and efficiency of peer review under various conditions of the publication rate (25%, 50%, and 75% of published submissions). Data were averaged across 10 simulation runs on 200 simulation steps. Results indicated that cheating increased evaluation bias

particularly when the selection rate of publication was stronger. This also caused higher productivity loss, in terms of resources wasted by unpublished authors who deserved publication and more considerable reviewing expenses. We measured *productivity loss* as the percentage of resources wasted by unpublished authors who deserved to be published, while *reviewing expenses* were the percentage of resources spent by referees compared with the resources invested by submitting authors. Obviously, less competitive publication rates decreased the negative impact of cheating.

Tab. 2 shows the number of cheaters in the population in situations of more or less selective publication rate. Results indicated that, in case of stronger selection for publication, a percentage of 27% of cheaters was sufficient to generate 70.86% of low quality published authors who should not deserve publication (see also Tab. 1). In case of less selective publication rates, even higher probability of cheating caused less biased allocation of publications (e.g, 35% of cheaters produce 20.07% of evaluation bias under the weak selection environment).

Table 1: The impact of referee behaviour on the quality
and efficiency of peer review in various selective
environments (values in percentage).

Scenario	Evaluation	Productivity	Reviewing
	bias	loss	expenses
Weak selection (75% published submissions)			
Random	16.51	7.68	25.98
behaviour			
Cheating	20.07	4,91	21.34
Medium-level selection (50% published submissions)			
Random	25.27	14.98	30.77
behaviour			
Cheating	56.63	28.02	32.21
Strong selection (25% published submissions)			
Random	29.42	15.00	29.42
behaviour			
Cheating	70.86	34.72	35.24

Table 2: Percentage of cheaters among the referees in the "cheating scenario" in various selection rate environments (values of cheaters in percentage on the total number of referees).

Selection rage	Cheaters
Strong selection	0.27
Medium selection	0.28
Weak selection	0.35

We next calculated the resources of all agents at the end of the simulation run in each scenario. Figures 1 and 2 compare system productivity accumulation in weakly and strongly publication selection rate. Results showed that, in case of stronger competition for publication, cheating implied less concentration of resources.

We also considered inequality of resource distribution by calculating a Gini index, which typically

measures the inequality among values of a frequency distribution. In our case, inequality meant an unequal allocation of resources, such as ideal academic status, reputation, and career (see Tab. 3). Results showed that, in case of stronger competition for publication, cheating caused less inequality of resource allocation in the system. The situation was different in case of less competition for publication.



Figure 1: The impact of agent behaviour on system resource accumulation when competition for publication

was weak. The "random behaviour" scenario is in dotted grey, the "cheating" scenario is in solid black. In the x-axis, the resources. In the y-axis, the percentage of agents.



Figure 2: The impact of agent behaviour on system resource accumulation when competition for publication was strong. The "random behaviour" scenario is in dotted grey, the "cheating" scenario is in solid black. In the x-axis, the resources. In the y-axis, the percentage of agents.

Table 3: The Gini index in weak and strong competition for publication (values calculated at the end of the simulation). The index takes 0 when there was complete equality in resource distribution among agents and 1 when a single agent had everything.

Weak competition for publication	Gini index
Random behaviour	0.54
Cheating	0.57
Strong competition for publication	Gini index
Random behaviour	0.47
Cheating	0.28

This is coherent with previous findings (e.g., Squazzoni and Gandelli 2012): in a competitive, "winner takes all" system such as academic science, a better functioning peer review process determines an unequal resource distribution as advantages accrue to the best scientists. This can be attributed to the fact that the best published authors gain access to more resources and more chances to be re-published by taking advantage of the fairness of certain referees. It is worth noting that this kind of "Saint Matthew effect" (i.e., "the rich get richer, the poor get poorer") has been widely acknolwedged in science, from the classical contribution by Merton (1973) to recent findings (e.g., Barabási, Song and Wang 2012; Tol 2013).

The next step was to consider that competitive behaviour of scientists could be influenced by different possible ways of identifying potential competitors. For instance, in certain scientific communities, especially among the so-called "hard sciences", widely shared objective measures exist that help everyone to precisely commensurate his/her respective performance with that of others. This means that competitors might be precisely identified across the whole population. This does not hold in other scientific communities, especially those revolving around the humanities, where these standards do not exist and are even widely contested. In these cases, the definition of potential competitors depends on the stratification of scientists in local groups, with the prevalence of disciplinary or group specificities of standards (e.g., Laudel and Gläeser 2006; Lamont 2009).

We created two supplementary scenarios where we modified the way in which cheaters identified their competitors. Unlike the previous "*cheating*" scenario, which followed a threshold function to detect possible competitors, we tested a "*local competition*" scenario, where this function followed a Gaussian shape and a "*glass ceiling*" scenario, where competitors' detection function followed a logistic shape.

In the first case ("local competition"), we assumed that scientists detected possible competitors only in their own performance neighbourhood. This was to mimic certain fragmented scientific communities where scientists tend to compete locally. More specifically, we assumed that competitor's detection followed a normal distribution $N(R_r, \sigma_2)$ where R_r was the referee's resources and σ_2 was the standard deviation calculated as a proportion of R_r .

In the second case ("*glass ceiling*"), we assumed that scientists tried to similarly outperform the less and the more productive colleagues. This was to protect against upstart and outperform superior scientists. The shape of the logistic function is shown in Fig. 3.



Figure 3: The logistic function for competitor detection in the *glass ceiling* scenario.

More specifically, the probability C_r of a referee to cheat a submission author followed equation (3), where d indicated the respective distance of author's and referee's resources (i.e., $d = R_a - R_r$).

$$C_r = \frac{1}{(e^{-(b_1 * d + b_0)} + 1)}$$
(3)

The constants b_0 and b_1 determined the shape of the curve shown in Fig. 3 and were calculated as in the equations (4) and (5). These equations were related to the following three parameters (also shown in Fig. 3): k_1 indicated the probability of cheating (C_r) when both referees and authors had the same amount of resources (i.e. d=0), and k_2 indicated the probability of cheating when the distance between the author's and the referee's resources was equal to k_3 .

$$b_{0} = -\ln\left(\frac{1-k_{1}}{k_{1}}\right)$$
(4)
$$b_{1} = \frac{\ln\left(\frac{1-k_{2}}{k_{2}}\right) + b_{0}}{k_{1}}$$
(5)

 k_3

Compared with the baseline scenario (i.e., threshold function), results showed that "*local competition*" scenario ensured less bias independently of the publication selection rates. In case of higher selection rate for publication, this scenario also guaranteed higher efficiency in terms of minimization of wasted resources for authors and reviewing expenses.

On the other hand, the "glass ceiling" scenario determined levels of bias and unefficiency similar to the baseline scenario. This means that the logistic function for competitor's detection performed similarly to the threshold one.

Table 4: The impact of different competition detection scenarios on the quality and efficiency of peer review in various selective environments (values in percentage).

Scenario	Evaluation	Productivity	Reviewing
	bias	loss	expenses
Weak selection (75% published submissions)			
Cheating	20.07	4.91	21.34
Local	14.65	4.21	22.56
competition			
Glass ceiling	16.67	3.68	21.66
Medium-level selection (50% published submissions)			
Cheating	56.63	28.02	32.21
Local	33.86	18.39	30.97
competition			
Glass ceiling	43.06	15.30	23.16
Strong selection (25% published submissions)			
Cheating	70.86	34.72	35.24
Local	31.04	15.63	30.13
competition			
Glass ceiling	70.35	34.70	34.56

Tab. 5 shows the number of cheaters in the population. For shortage of space, we reported only the case of strong competition for publication. Results showed, first, that the "glass ceiling" scenario determined higher number of cheaters. Secondly, they showed that although "local competition" scenario included a considerable percentage of cheaters (i.e., 20%), this did not have a negative effect on the quality of peer review. In case of strong competition for publication, 20% of cheaters caused only 31% of biased judgment.

Table 5: Percentage of cheaters among the referees in various competitor detecting scenarios with strong competition for publication (values of cheaters in percentage on the total number of referees).

Scenario	Cheaters
Cheating	0.27
Local competition	0.20
Glass ceiling	0.34

Figure 4 and Tab. 6 show the impact of various cheating mechanisms on the growth and distribution of resources in the case of strong competition for publication. Results showed that "cheating" and "glass ceiling" scenarios generated similar outcomes, whereas "local competition" implied higher inequality of resource distribution.



Figure 4: The impact of cheating on system resource accumulation in strongly selective environments. The "cheating behaviour" scenario is in dotted grey, the "local competition" scenario is in solid black and the "glass ceiling" scenario is in solid gray. In the x-axis, the resources. In the y-axis, the percentage of agents.

Table 6: The Gini index in the "local competition" and the "glass ceiling" scenarios in strongly selective environments (values calculated at the end of the simulation). The index takes 0 when there was complete equality in resource distribution among agents and 1 when a single agent had everything.

Scenario with strong selection	Gini index
Cheating	0.28
Local competition	0.46
Glass ceiling	0.29

CONCLUSIONS

Previous computational studies have indicated that even a small proportion of referee cheating may dramatically distort the publication quality (Thurner and Hanel 2011). Our results confirmed these findings but also permitted us to consider implications of cheating for the system's resource allocation in terms of growth dynamics and distribution inequality. Furthermore, we found that peer review outcomes are sensitive to differences in the way scientists identify their competitors. Certain mechanisms, such as the stratification of scientists in local competing groups and the presence of niches of competition, might reduce the negative effect of cheating on the quality of the peer review process as scientists can develop more competent judgment.

It is important to outline that making any strong generalisation on peer review from these simulation studies is inappropriate. Developments that integrate theoretical models and empirical data are fundamental both to test aggregate findings and realistically calibrate important model parameters. However, it must be also said that computational models of peer review allow us to pinpoint the importance of considering microbehavioural and interaction aspects, which are unfortunately difficult to look at empirically (Squazzoni and Takács 2011).

ACKNOWLEDGEMENTS

This work has been jointly supported by the Spanish MICINN and the European Commission FEDER funds, under grant TIN2009-14475-C04. Usual disclaimers apply.

REFERENCES

- Allesina, S. 2012 "Modeling Peer Review: An Agent-Based Approach". *Ideas in Ecology and Evolution*, 5, 2, 27-35.
- Barabási, A.-L., Song, C. and Wang, D. 2012. "Handful of Papers Dominates Citation". *Nature* 491, 40.

- Bornmann, L. 2011 "Scientific Peer Review". Annual Review of Information Science and Technology, 45, 199-245.
- Bornmann, L., Daniel, H.-D. 2007. Convergent Validation of Peer Review Decisions Using the *H* Index: Extent of and Reasons for Type I and Type II errors. *Journal of Informetrics*, 1(3), 204-213.
- Callahan, D. 2004. "Rational cheating: Everyone's doing it". Journal of Forensic Accounting, 5, 577-580.
- Casati, F., M. Marchese, A. Ragone and M. Turrini. 2009. "Is Peer Review Any Good? A Quantitative Analysis of Peer Review". DISI University of Trento, Technical Report # DISI-09-045, accessible at: <http://eprints.biblio.unitn.it/archive/00001654/>.
- Crocker, J., M. L. Cooper. 2011. "Addressing Scientific Fraud". *Science*, 334 (6060), 1182.
- Edmonds, B., Gilbert, N., Ahrweiler, P., Scharnhorst, A. 2011. Simulating the Social Processes of Science. Journal of Artificial Societies and Social Simulation 14(4) 14: http://jasss.soc.surrey.ac.uk/14/4/14.html.
- Fanelli, D. 2010. "Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data". *PLoS ONE* 5(4): e10271.
- Grimaldo, F. and Paolucci, M. 2013. "A Simulation of Disagreement for Control of Rational Cheating in Peer Review. *Advances in Complex Systems*, forthcoming.
- Laband, D. N. and J. M. Piette. 1994. "Favoritism Versus Search of Good Papers. Empirical Evidence Regarding the Behavior of Journal Editors". *Journal of Political Economy*, 102, 194-203.
- Lamont, M. 2009. How Professors Think. Inside the Curious Worlds of Academic Judgment. Cambridge, MA: Harvard University Press.
- Laudel, G. and J. Gläeser. 2006. Tensions between Evaluations and Communication Practices. *Journal of Higher Education Policy and Management*, 28: 289-295.
- Merton, R. K. 1973. The Sociology of Science. Theoretical and Empirical Investigations. Chicago: University of Chicago Press.
- Payette, N. 2012. "Agent-Based Models of Science". In A. Scharnhorst, K. Börner & P. van den Besselaar (Eds.), Models of Science Dynamics, Complexity Series. Springer.
- Roebber, P. J. and D. M. Schultz, 2011. "Peer Review, Program Officers and Science Funding". *Plos One*, 6, 4, e18680, accessible at: http://www.plosone.org/article/info:doi%2F10.1371%2Fjo urnal.pone.0018680.
- Squazzoni, F. and C. Gandelli. 2012. "Saint Matthews Strikes Again. An Agent-Based Model of Peer Review and the Scientific Community Structures". *Journal of Informetrics*, 6: 265-275.
- Squazzoni, F. and C. Gandelli. 2013. "Opening the Black Box of Peer Review. An Agent-Based Model of Scientist Behaviour. Journal of Artificial Societies and Social Simulation, 16(2) 3: http://iacsg.com.gurgu.com/16/2/2.html

http://jasss.soc.surrey.ac.uk/16/2/3.html.

- Squazzoni, F. and K. Takács. 2011. "Social Simulation that 'Peers into Peer Review". *Journal of Artificial Societies and Social Simulation*, 14(4) 3: http://jasss.soc.surrey.ac.uk/14/4/3.html.
- Thurner, S. and R. Hanel. 2011. "Peer Review in a World with Rational Scientists: Toward Selection of the Average". *The European Physical Journal B*, 84, 707-711.
- Tol, R. S. J. 2013. The Matthew Effect for Cohorts of Economists. *Journal of Informetrics*, 7(2), 522-527.

AUTHOR BIOGRAPHIES

JUAN BAUTISTA CABOTÀ is PhD. Student at the Universitat de València in the Networks and Virtual Environments Group (GREV, http://grev.uv.es). His research interests include agent-based modelling and simulation and intelligent decision-making support systems. His email address is: juan.cabota@uv.es.

FRANCISCO GRIMALDO is lecturer at the Universitat de València and member of the Networks and Virtual Environments Group (GREV). His research is focused on applied artificial intelligence, agent-based

modelling/simulation and computational social choice. He is member of the HiPEAC network of excellence and the IEEE Systems, Man & Cybernetics Society. His email is francisco.grimaldo@uv.es and his webpage can be found at: http://www.uv.es/grimo/.

FLAMINIO SQUAZZONI is assistant professor of economic sociology at the University of Brescia, where he leads the GECS-Research Group on Experimental and Computational Sociology. He is President of *ESSA* The European Social Simulation Association. His email is: squazzon@eco.unibs.it and his web-page can be found at: http:www.eco.unibs.it/gecs/Squazzoni.html