©Copyright JASSS

JASSS

Flaminio Squazzoni and Claudio 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://jasss.soc.surrey.ac.uk/16/2/3.html>

Received: 02-Jul-2012 Accepted: 13-Oct-2012 Published: 31-Mar-2013

## Abstract

This paper investigates the impact of referee behaviour on the quality and efficiency of peer review. We focused on the importance of reciprocity motives in ensuring cooperation between all involved parties. We modelled peer review as a process based on knowledge asymmetries and subject to evaluation bias. We built various simulation scenarios in which we tested different interaction conditions and author and referee behaviour. We found that reciprocity cannot always have per se a positive effect on the quality of peer review, as it may tend to increase evaluation bias. It can have a positive effect only when reciprocity motives are inspired by disinterested standards of fairness.

Keywords:
Peer Review, Referees, Referee Behaviour, Reciprocity, Fairness

## Introduction

1.1 Peer review is a cornerstone of science. The process allows scientists to experimentally pursue new lines of research through a continuous, decentralised and socially shared process of trial and error and ensures the quality of knowledge produced. Whether directly or indirectly, peer review determines how the resources of the science system-including funding, positions, and reputation-are allocated. Despite its importance, peer review remains dramatically under-investigated (e.g., Campanario 1998a, 1998b; Godlee and Jefferson 2003; Kassirer and Campion 1994). Certain authors have argued, with little supporting evidence, that it is nothing but a "black-box" (Horrobin 2001) and that it has no "experimental base" (Smith 2006).

1.2 One of the main challenges is to understand referee behaviour and how to increase commitment and reliability for all parties (e.g., Squazzoni 2010; Squazzoni and Takács 2011). While journal editors and submitting authors can benefit from reputational rewards, understanding the incentives to and motivations of referees is more difficult. This is not a trivial matter, either, as it has been recently acknowledged that referees are dramatically overexploited, a fact which could undermine their commitment to the process (e.g., Neff and Olden 2006; Ware 2007). A recent survey estimated that more than 1 million journal articles per year are subjected to peer review, not to mention the innumerable conference proceedings, research proposals, fellowships and university-, department-, and institute-wide productivity evaluations (Björk, Roos and Lauri 2009). Serious doubt about the possibility of peer review continuing on in its present form has even appeared in the influential columns of *Science* (e.g., Alberts, Hanson and Kelner 2008).

1.3 Recent cases of misconduct and fraud have contributed to calls for a reconsideration of the rigour and reliability of the peer review process. In 1997, the editors of the *British Medical Journal* asked referees to spot eight errors intentionally inserted in a submission. Out of 221 referees, the median number spotted was two (Couzin 2006). Another example was the stem cell scandal, able to be traced back to a group of scientists from South Korea, who published an article in *Science* in 2005 which was based on falsified data. The myopic attitudes of certain editors, influenced by "aggressively seeking firsts", and by nine referees dazzled by the novelties of the paper, implied that review time had been dramatically shortened: the referees took just 58 days to recommend the publishing of the article, as compared against the average of 81 days typical for this influential journal (Couzin 2006). More recently, the Stapel scandal-in which data for numerous studies conducted over a period of 15-20 years and published in many top journals in the field of psychology were found to have been fabricated (Crocker and Crooper 2011)-gained public notoriety in newspapers and on social media. It is worth noting, first, that these cases have caused a misallocation of reputational credit in the science-publishing ecosystem, with negative consequences for competition and resource allocation. Second, such cases also carry serious consequences for the credibility of science in the perceptions of external stakeholders.

1.4 It is worth mentioning that these problems have recently been addressed in *Science*, where Alberts, Hanson and Kelner (2008) have suggested the need to subject the peer review process itself to a serious review in order to improve its efficiency and guarantee its sustainability. All current attempts at reform, however, which have insisted on the importance of referee reliability and the need for measures to improve said reliability in particular, have followed a trial and error approach which is unsupported by experimental investigation. Although some 'field experiments' concerning peer review have been performed by certain journals or funding agencies (e.g., Jayasinghe, Marsh and Bond 2006; Peters and Ceci 1982), it is widely acknowledged that sound experimental knowledge would be needed concerning essential peer review mechanisms; such knowledge would lend much needed support to any prescribed policy measures (e.g., Bornmann 2011).

1.5 A select few studies in behavioural sciences have attempted to understand the behaviour of the figures involved and the consequences of said behaviour for the quality and efficiency of the evaluation process. One of the few topics studied in depth has been the reviewing rate. Engers and Gans (1998) have suggested, for instance, a standard economic analytic model which examined the interaction of editors and referees. They aimed to understand why referees accepted the responsibility of ensuring sound quality in reviewing without receiving any material incentives and whether improving this latter point would act to increase the reviewing rate. They showed that payment could potentially motivate more referees to review author submissions, although raising the review rate could lead referees to underestimate the negative impact of their refusal, as they could come to believe that other referees had readily accepted a given work. This could in turn motivate journals to increase the payment given to reviewers in an effort to compensate for this effect while reducing the need for referees to incur private costs in order to enhance the quality of reviewed works. Finally, this could contribute to an escalation of compensation which would eventually prove unsustainable for journals.

1.6 Chang and Lai (2001) also studied reviewing rates and arrived at different conclusions. They suggested that, when reciprocity is present as a motive influencing the relationship between journal editors and referees, by providing room for reputation building for referees, the referee recruitment rate could significantly increase. They showed that this effect could significantly improve the review quality if accompanied by material incentives. This finding was confirmed by Azar (2008), who studied the response time of journals. He suggested that shorter response times of journals in specific communities were attributable to the strength of social norms-especially the mutual respect of good standards of evaluation, towards which referees were extremely sensitive (see also Ellison 2002).

1.7 The importance of social norms in peer review has been confirmed by recent experimental findings (Squazzoni, Bravo and Takács 2013). These results have shown that indirect reciprocity motives, as opposed to material incentives, can increase the commitment and level of reliability of referees. By manipulating incentives in a repeated

investment game which had been modified to mirror peer review mechanisms, the authors found that allocating material incentives to referees undermined pro-social motivations without generating higher evaluation standards. These results are in line with game theory-oriented experimental behavioural studies which have acknowledged the importance of reciprocity-both direct and indirect-in facilitating cooperation in situations of information asymmetries and potential cheating temptations, which could represent a typical situation of the peer review process (e.g., Bowles and Gintis 2011; Gintis 2009).

1.8 It is therefore possible to argue that referees would cooperate with journal editors in ensuring the quality of evaluation, as they are invested in protecting the prestige of the journal as a means of protecting their own impact-this is especially so in cases where the reviewer has previously published work in the target journal. On the other hand, reviewers could also be motivated to cooperate with authors-cooperation here meaning the providing of fair evaluation and constructive feedback-as they are interested in establishing good standards of reviewing as a potential benefit when they are themselves subject to the reviewing process as authors. In considering peer review as a cooperation problem, referees would pay a significant cost-i.e., the time and effort needed to conduct a review-to generate a considerable benefit to authors-i.e., publications, citations, and a higher academic reputation-in order to protect the quality of peer review as a public good, from which they expect to benefit themselves in the future.

1.9 Our paper is an attempt to contribute on this point by proposing a modelling approach (e.g., Martins 2010; Roebber and Schultz 2011; Thurner and Hanel 2011; Allesina 2012). Empirical research encounters serious problems when attempting to consider essential aspects of peer review and when investigating complex mechanisms of interaction. Following Squazzoni and Gandelli (2012), we have modelled a population of agents interacting as authors and referees in a competitive and selective science system. We extended the previous model to understand the impact of various agent strategies on the quality and efficiency of peer review and to test the influence of reciprocity between authors and referees.

1.10 The structure of this paper is as follows. In the second section, we introduce the model. In the third section, we present various simulation scenarios and the simulation parameters. In the fourth, we illustrate our simulation results, while in the concluding section we present a summary of results and highlight certain implications germane to the debate surrounding peer review.

## The Model[1]

2.1 We assumed a population of N scientists (N = 200) and randomly selected each to fill 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 $p$).[2]

2.2 We gave each agent a set of resources which were initially homogeneous ($R_\alpha$ (0)). 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.

2.3 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.

2.4 We assumed that the quality of submissions $\mu$ varied and was dependent on agent resources. Each agent had resources $R(\alpha) \in N$, from which we derived an expected submission quality as follows:

$$\mu = \frac{v * R_a}{v * R_a + 1} \quad (1)$$

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 $R_\alpha$ = 10 to reach a medium-sized quality submission (μ = 0.5).

2.5 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 $\sigma$ 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.

2.6 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.

2.7 If not published, following the "winner takes all" rule characterizing science, we assumed that authors lost all resources invested prior to submitting. This meant that, at the present stage, we did not consider the presence of a stratified market for publication, where rejected submissions could be submitted elsewhere, as happens in reality (e.g., Weller 2001).

2.8 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:

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

where $R_r$ was the referee's resources, $Q_\alpha$ was the real quality of the author's submission and $\mu_r$ 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.

2.9 We assumed that authors and referees were randomly matched 1 to 1 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 are average scientists. They will lose more resources than average scientists, however, because their time is more valuable than the latter.

2.10 We assumed two types of referee behaviour, namely *reliable* and *unreliable*. Reliability was here taken to connote the ability of referees to provide a consistent and unequivocal opinion which truly reflected the quality of the submission. In the case of reliability, referees did the best they could to provide an accurate evaluation and spent all needed resources for reviewing. In this case, we assumed a normal distribution of the referees' expected quality and a narrow standard deviation of their evaluation score from the real value of the submission ($\sigma=R(_{\alpha})/100$). This meant that the evaluation scores by reliable referees were likely to approximate the real value of author submissions.

2.11 We also assumed, however, that in the case of referee reliability there was a chance for some evaluation bias ($b = 0.1$), and that $b$ increased in proportion to the difference between referees' expected quality and author submission quality. This step was undertaken to represent the knowledge and information asymmetries between authors and referees which characterize peer review in science. To measure the quality of peer review, we measured 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 (see evaluation bias in Tables 2, 3 and 4).

2.12 In the case of unreliability, 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). More specifically, unreliable referees spent fewer resources than did reliable referees ( $s$ = 0.5), and under- or over- estimated author submissions (see the parameters $u$ and $o$, respectively, in Table 1). To avoid the possibility that referees assigned the real value to submissions by chance we assumed that, when they underrated a submission, the evaluation score took a standard deviation of approximately -90% of the real quality of the submission ($u$ = 0.1). The opposite sign was assigned in the case of overrating (i.e., + 90%, or $o$ = 1.9).

2.13 It is worth noting that certain empirical studies have shown that these types of errors are more frequent than expected, especially in grant applications (e.g., Bornmann and Daniel 2007; van den Besselaar and Leydesdorff 2007). Bornmann, Mutz and Daniel (2008), for instance, examined EMBO selection decisions and found that 26-48 percent of grant decisions showed such errors, with underrating being more frequent (occurring in 2/3 of cases). A general estimate of the percentage of errors, for peer review in journals in particular, which could have been used to calibrate the model, was unfortunately not available.

2.14 Finally, all simulation parameters are shown in Table 1. At the beginning of the simulation agent resources were set to 0 for all ($R_{\alpha}$ (0)). At the first tick, 50% of agents were published randomly. Subsequently, everyone had a fixed amount of resources $F$ for each tick. When selected as authors, agents invested all available resources in conducting research and producing a good submission ($i$ = 1) (see the next section for some manipulation of this parameter). If accepted for publication, author agents had their resources multiplied by $m$ [1, 1.5], as explained in equation (3), and so their resources grew accordingly. This meant that the quality of their subsequent submission was presumably higher.

Table 1: Simulation parameters.

| Parameters | Description | Value |
|---|---|---|
| $N$ | Number of agents | 200 |
| $R_a(0)$ | Initial agent resources | 0 |
| $f$ | Fixed amount of resource | 1 |
| $p$ | Publication rate | [0.25, 0.50, 0.75] |
| $m$ | Publication multiplier | [1, 1.5] |
| $b$ | Evaluation bias by default | 0.1 |
| $i$ | Author investment | 1 |
| $s$ | Reviewing expenses for unreliable referees | 0.5 |
| $u$ | Underrating by unreliable referees | 0.1 |
| $o$ | Overrating by unreliable referees | 1.9 |
| $v$ | Velocity of submission quality increase | 0.1 |

## Simulation Scenarios

3.1 We 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). This was a restrictive definition of the various functions that peer review fulfils in the sciences. Here we considered only the screening function. Neither the role of peer review in helping authors add value to their submission via referee feedback (e.g., Laband 1990) nor its role in deciding the reputation of journals and their respective position in the market were considered here (e.g., Bornmann 2011). 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.

3.2 In the first scenario, called "*no reciprocity*", we assumed that agents had a random probability of behaving unreliably when selected as referees; this probability was constant over time and was not influenced by past experiences. When selected as authors, agents invested all available resources in publication ( $i$ = 1), irrespective of positive or negative past experiences with the submission and review process. In this case, there was no room for reciprocity strategies between authors and referees. In the second scenario, called "*indirect reciprocity*", we assumed that agents were influenced by their past experiences as authors when selected as referees. In cases in which their past submission has been previously accepted for publication, they reciprocated by providing reliable evaluations when selected as referees. Note that in this case, authors were self-interested and did not consider the pertinence of the referee evaluation, only their publication success or failure in their previous submission. This meant that they reciprocated negatively if they experienced rejection and positively when they had been successfully published even if they knew that their submission wasn't worthy of publication.

3.3 In the third scenario, called "*fairness*", author agents formulated a pertinent judgment of the referee evaluation of their submission. They measured the fairness of the referee's opinion by comparing the real quality of their submission and the evaluation rate received by the referees. If the referee evaluation approximated the real value of their submission (i.e., ≥ -10%), they concluded that the referee was reliable and had done a good job. In this case, when selected as referees, agents reciprocated positively irrespective of their past publication or rejection history. This meant that indirect reciprocity was now not based on the pure self-interest of agents but on normative standards of conduct.

3.4 The final two scenarios, "*self-interested authors*" and "*fair authors*", extended the previous two scenarios by examining author behaviour in conjunction with reviewer behaviour. In the "*self-interested authors*" scenario, we assumed that authors reacted positively and continued to invest all available resources into their next submission when published ($i$ = 1). In the case of rejection, they reacted negatively and invested fewer resources in subsequent attempts at publication ($i$ = 0.1). This reaction was independent from the pertinence of the referee evaluation. In the "*fair authors*" scenario, in cases in which the agent had received a pertinent referee evaluation when themselves an author, they reinforced their confidence in the quality of the evaluation process and continued to invest heavily in producing quality submissions irrespective of the fate of their submission. In the case of non-pertinent evaluation (see above), they invested less in the subsequent attempt at publication ($i$ = 0.1) and accumulated resources for the subsequent round irrespective of their previous publication. In this case, agents therefore inferred the overall situation of peer review

standards through their experiences as authors and consequently acted in a strategic manner.

## Results

4.1 Table 2 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 on a 200-simulation run in any parameter condition. First, results showed that the reciprocity motives of referees did not have *per se* a positive effect on the quality and efficiency of peer review when the publication rate was more competitive. Second, when the publication rate was higher the quality of the peer review process improved only minimally, but at the expense of referees' resources. Although increased competitiveness in general implied increasing evaluation bias, "*fairness*" implied lower bias and fewer resources lost by authors, although reviewing expenses were generally higher. Furthermore, it ensured greater robustness to the changes in competition pressures. On the other hand, indirect reciprocity without fairness by authors implied higher evaluation bias and higher resource loss when the publication rate diminished.

4.2 It is worth noting that, in order to calculate the resource loss, we calculated the amount of resources wasted by (unpublished) authors compared with the optimal solution —i.e., where only the best authors were published. To calculate the reviewing expenses we measured the resources spent by agents for reviewing compared with the resources invested by submitting authors.

4.3 Table 3 shows the impact of the reciprocal behaviour of authors as the publication rate varies. Results showed that the reciprocity of authors improved peer review only when associated with fair criteria applied to the judgment of their submission. When authors reacted to referee evaluation only in relation to their self-interest-i.e., eventually being published-the quality and efficiency of peer review drastically declined. Moreover, in case of authors' fairness, peer review dynamics improved even with increased competition.

Table 2: The impact of referee behaviour on the quality and efficiency of peer review in various selective environments (values expressed as percentage).

| Scenario | Evaluation bias | Resource loss | Reviewing expenses |
|---|---|---|---|
| *75% of published submissions* | | | |
| No reciprocity | 14.10 | 5.69 | 23.47 |
| Indirect reciprocity | 12.58 | 6.51 | 44.16 |
| Fairness | 13.14 | 7.48 | 40.61 |
| *50% of published submissions* | | | |
| No reciprocity | 26.32 | 15.65 | 30.32 |
| Indirect reciprocity | 25.32 | 12.64 | 39.88 |
| Fairness | 15.68 | 8.60 | 38.68 |
| *25% of published submissions* | | | |
| No reciprocity | 28.00 | 15.01 | 29.47 |
| Indirect reciprocity | 43.12 | 16.92 | 33.39 |
| Fairness | 19.52 | 8.32 | 38.29 |

Table 3: The impact of author reciprocal behaviour on the quality and efficiency of peer review in various selective environments (values expressed as percentage).

| | Evaluation bias | Resource loss | Reviewing expenses |
|---|---|---|---|
| *75% of published submissions* | | | |
| Self-interested authors | 15.30 | 12.63 | 46.07 |
| Fair authors | 14.85 | 5.10 | 29.55 |
| *50% of published submissions* | | | |
| Self-interested authors | 30.52 | 25.63 | 45.74 |
| Fair authors | 15.44 | 3.88 | 23.32 |
| *25% of published submissions* | | | |
| Self-interested authors | 45.04 | 38.31 | 47.13 |
| Fair authors | 14.24 | 4.00 | 15.96 |

4.4 We next calculated the resources of all agents at the end of the simulation run. Considering the value of "*no reciprocity*" as a benchmark, "*indirect reciprocity*" implied a loss of 20% of system resources and "*fairness*" a loss of 7%, while "*self-interested authors*" doubled the amount of resources and "*fair authors*" determined an exponential growth of resources. Figures 1 and 2 compare system resource accumulation in extreme conditions-i.e., when publication rates varied from 75% to 25%. Results showed that in "*fair authors*" scenarios, stronger competition determined an exponential growth of resources. This means that a higher quality peer review process determined a greater accumulation of resources at the system level. The explanation is that the best scientists were published more frequently, had more resources and were thus able to increase the quality of their subsequent submissions, thereby exploiting increasing returns earned for publication.

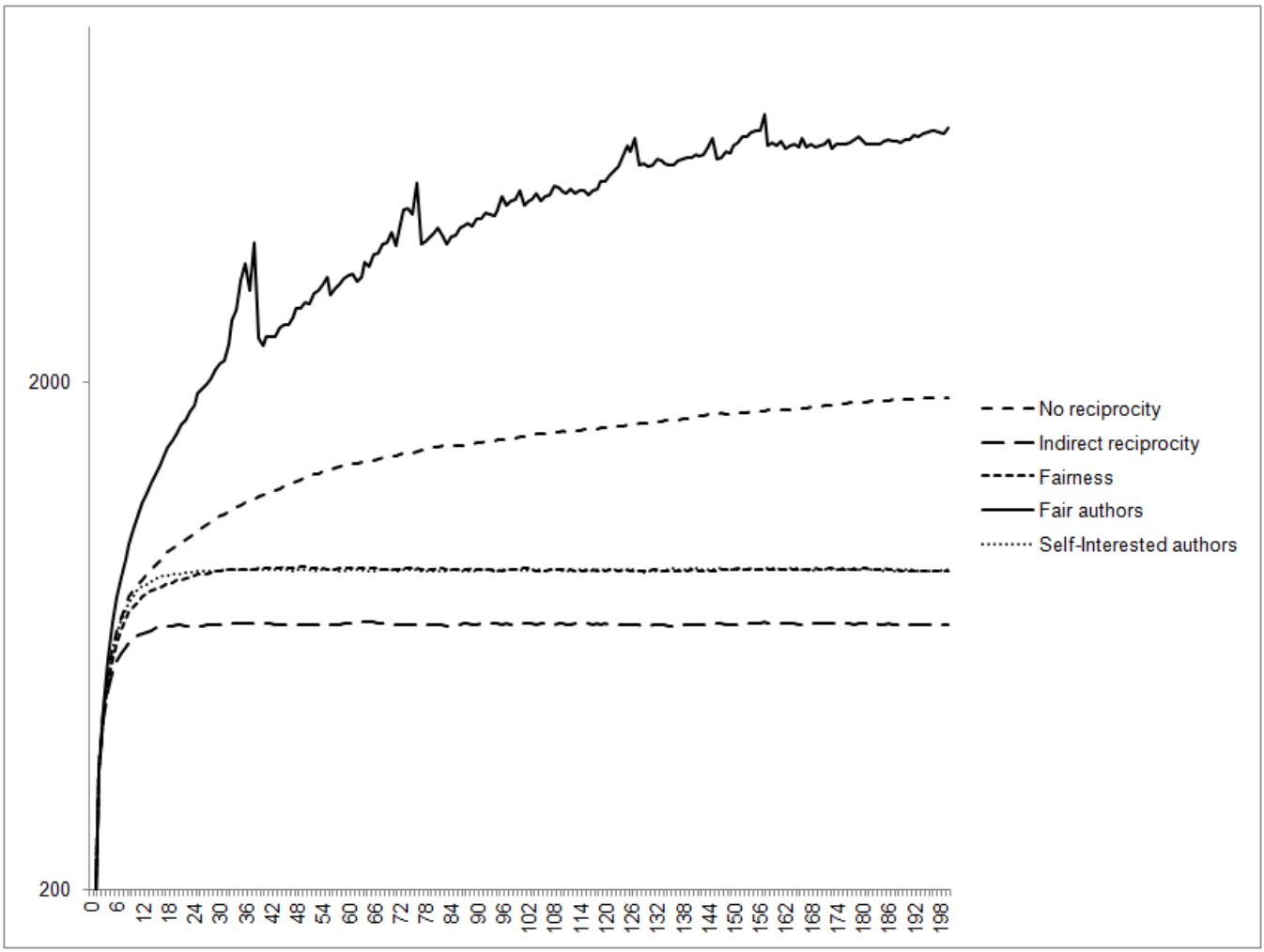


Figure 1. The impact of agent behaviour on system resource accumulation in weakly selective environments (75% of published submissions). The x-axis shows the number of the simulation run.

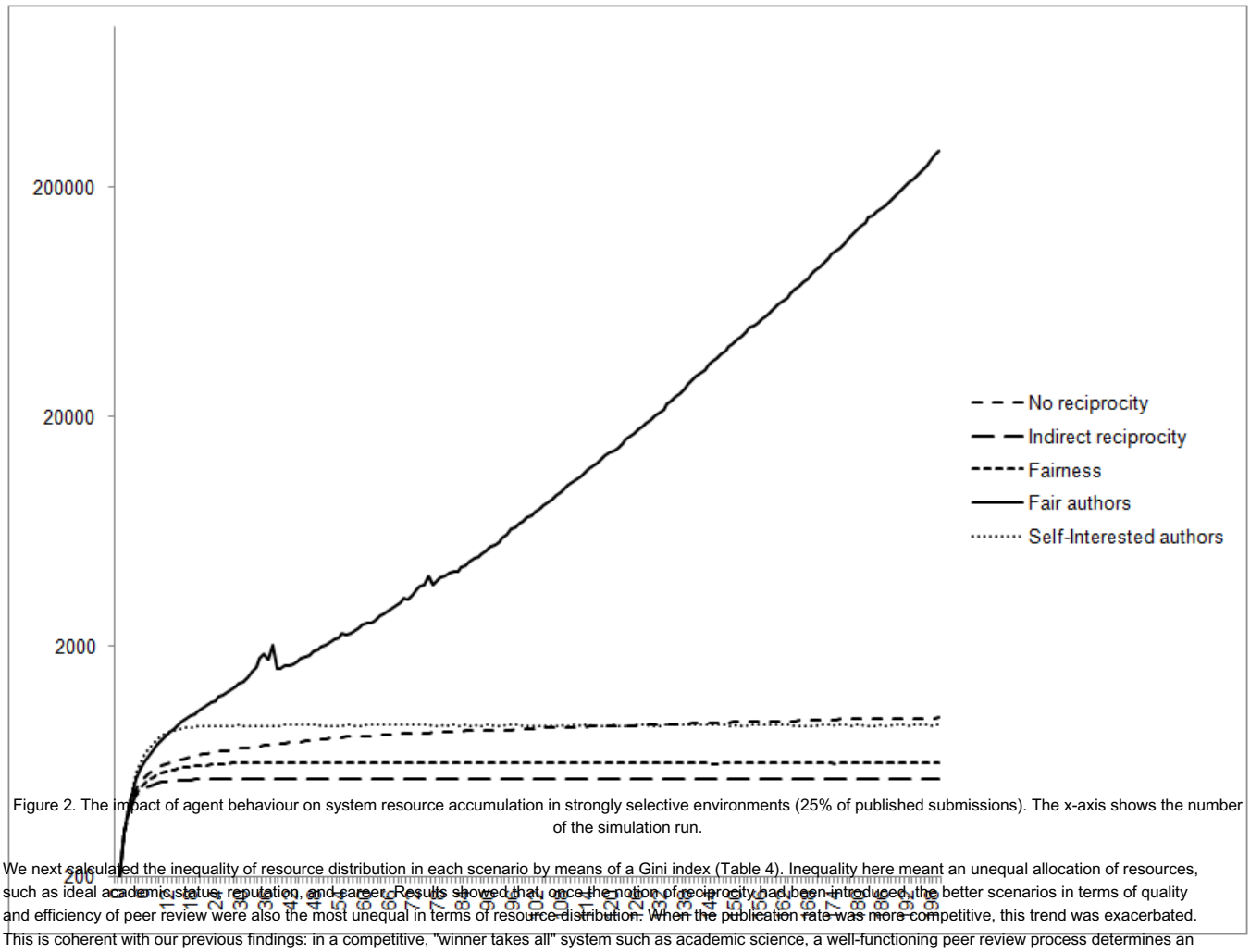


Figure 2. The impact of agent behaviour on system resource accumulation in strongly selective environments (25% of published submissions). The x-axis shows the number of the simulation run.

4.5 We next calculated the inequality of resource distribution in each scenario by means of a Gini index (Table 4). Inequality here meant an unequal allocation of resources, such as ideal academic status, reputation, and career. Results showed that, once the notion of reciprocity had been introduced, the better scenarios in terms of quality and efficiency of peer review were also the most unequal in terms of resource distribution. When the publication rate was more competitive, this trend was exacerbated. This is coherent with our previous findings: in a competitive, "winner takes all" system such as academic science, a well-functioning peer review process determines an unequal resource distribution as advantages accrue to the best scientists (Squazzoni and Gandelli 2012). 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 and reliability of referees. This would replicate a well-known stylized fact of the prevailing science-publishing system: science outputs, resources, reputation, and talents are concentrated around a few star scientists and institutions (Merton 1973).

Table 4: The Gini index for all the scenarios in weakly and strongly selective environments. The index was 0 when there was complete equality in resource distribution among agents and 1 when a single agent controlled all resources.

| Scenario | Gini index |
|---|---|
| *75% of published submissions* | |
| No reciprocity | 0.55 |
| Indirect reciprocity | 0.34 |
| Fairness | 0.36 |
| Self-interested authors | 0.34 |
| Fair authors | 0.74 |
| *25% of published submissions* | |
| No reciprocity | 0.47 |
| Indirect reciprocity | 0.34 |
| Fairness | 0.45 |
| Self-interested authors | 0.35 |
| Fair authors | 0.88 |

## Conclusions

5.1 One of the most convincing explanations as to why referees tend to cooperate with editors and authors in ensuring good quality and efficiency of peer review has been reciprocity (e.g., Chang and Lai 2001; Squazzoni, Bravo and Takács 2013). By conceptualising peer review as a cooperation game, we can argue that referees could rationally bear the cost of reviewing so as to establish good standards of reviewing with the prospect of benefiting from cooperation by other referees when they are subsequently cast as submitting authors. This is a typical mechanism through which reciprocity can sustain cooperation in repeated interactions, as it can transform the cost for referees into an investment in a potential future benefit.

5.2 Although highly abstracted and greatly simplified compared to in-depth empirical or experimental research, our results have shown that reciprocal behaviour of scientists engaged in peer review cannot always have *per se* a positive effect on the quality of peer review, as it may tend to increase evaluation bias. The positive effect is ensured only when the reciprocity motives of scientists are inspired by disinterested standards of fairness-i.e., when authors pertinently judge the work of referees independent of their own self-interest. A possible implication is that the scientific community should strengthen social norms of disinterestedness which reflect ethical, cultural and competence-based standards of conduct for scientists (e.g., Lamont 2009).

5.3 Neither theoretical generalisation nor serious policy implications can be drawn from our simulation study. We did not pretend to convey realism here. Our results were

based on a highly abstracted model and thus every conclusion should be considered with caution. For instance, peer review in reality is not equally distributed among the population, and editors are of course also important in providing room for reputation building and reciprocity motives in referees. These are certainly points which can inform future research efforts.

5.4 A crucial challenge facing any future work, however, will be an effort to address the gap between theory and empirical observation (e.g., Watts and Gilbert 2011). Although it is difficult to obtain empirical data which point to agent behaviour affecting peer review-especially at the scale needed to examine general aspects of the process-one possible means of development could be to empirically test referee behaviour in highly representative journals.

5.5 Certain empirical measures which have already been developed could be applied to test our findings. Laband (1990), for instance, examined referee reliability by measuring the lines of the report text sent to submitting authors, assuming that the longer the text, the higher the quality of the referee comments and the more reliable the final score assigned to the submissions. Although it is difficult to measure peer review effectiveness (e.g., Jefferson, Wager and Davidoff 2002), this is a brilliant idea to build an *ex-ante* measure which could complete the most common ex-post measures of peer review validity, such as citation indices or the fate of rejected submissions (e.g., Weller 2001). An alternative would be to exploit, where available, the ratings of referees assigned by journal editors as a proxy of the quality of reviews or to indirectly derive such measures by considering the number of reports in which a given referee was involved, assuming that the more often a referee was involved in the peer review process by a journal editor the higher the quality of his/her reviewing work.

5.6 Let us suppose that we can select a set of representative journals, possibly comprised of different scientific communities and audiences, and that we have access to both the list of referees and authors and to the referee reports. Let us suppose we apply one of the measures mentioned above to assess the *ex-ante* validity of peer review and measure the *ex-post* validity-e.g., by collecting data on subsequent citations of published articles or by analysing the fate of rejected submissions. This would allow us to build a statistical measure of the reliability of referee evaluation. By measuring the link between referees and authors in these journals and looking at the fate of past referee submissions to the journal we could thereby test whether space was given for reciprocity and fairness which might have influenced the quality of the evaluation.

## Acknowledgements

We gratefully acknowledge the financial support provided by the Dipartimento di Studi Sociali, University of Brescia. We thank three anonymous JASSS referees for their useful comments. A preliminary version of this paper was presented at the ECMS 2012 conference in Koblenz, May 29th - June 1st, 2012. We thank the conference referees and audience for their interesting remarks. The usual disclaimers apply.

## Notes

[1] The model was built in NetLogo, version 4.1 (Wilensky 1999). Codes and instructions to rerun the simulation can be accessed at: http://www.openabm.org/model/3145/version/2/view.

[2] For the sake of simplicity, as the focus of the paper was on referee behaviour, we excluded the role of editors and assumed that authors and referees were matched randomly.

## References

ALBERTS, B, Hanson, B, Kelner, KL (2008) Reviewing Peer Review. *Science*, 321, p. 15 [doi:10.1126/science.1162115]

ALLESINA, S (2012) Modeling Peer Review: An Agent-Based Approach. *Ideas in Ecology and Evolution*, 5(2), pp. 27-35. [doi:10.4033/iee.2012.5b.8.f]

AZAR, OH (2008) Evolution of Social Norms with Heterogeneous Preferences: A General Model and an Application to the Academic Review Process. *Journal of Economic Behavior and Organization*, 65, pp. 420-435 [doi:10.1016/j.jebo.2006.03.006]

BJÖRK, B-C, Roos, A, Lauri, M (2009) Scientific Journal Publishing - Yearly Volume and Open Access Availability. *Information Research*, 14(1): http://informationr.net/ir/14-1/paper391.html

BORNMANN, L. (2011) Scientific Peer Review. *Annual Review of Information Science and Technology*, 45, pp. 199-245 [doi:10.1002/aris.2011.1440450112]

BORNMANN, L and Daniel HD (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), pp. 204-213 [doi:10.1016/j.joi.2007.01.002]

BORNMANN, L, Mut, R and Daniel, H-D (2008) How to Detect Indications of Potential Sources of Bias in Peer Review: A Generalized Latent Variable Modeling Approach Exemplified by a Gender Study. *Journal of Informetrics*, 2(4), pp. 280-287 [doi:10.1016/j.joi.2008.09.003]

BOWLES, S and Gintis, H (2011) *A Cooperative Species. Human Reciprocity and Its Evolution* Princeton, NJ: Princeton University Press

CAMPANARIO, JM (1998a) Peer Review for Journals as It Stands Today - Part I. *Science Communication*, 19(3), pp. 181-211 [doi:10.1177/1075547098019003002]

CAMPANARIO, JM (1998b) Peer Review for Journals as It Stands Today - Part II. *Science Communication*, 19(4), pp. 277-306 [doi:10.1177/1075547098019004002]

CASATI, F, Marchese, M, Ragone, A, and Turrini, M (2009) Is Peer Review Any Good? A Quantitative Analysis of Peer Review. DISI University of Trento, Technical Report # DISI-09-045, : http://eprints.biblio.unitn.it/archive/00001654/

CHANG, J, Lai, C (2001) Is it worthwhile to pay referees? *Southern Economic Journal,* 68, pp. 457-463 [doi:10.2307/1061605]

COUZIN, J (2006) … and how the Problems Eluded Peer Reviewers and Editors. *Science*, 311, pp. 614-615. [doi:10.1126/science.1124948]

CROCKER, J, Cooper, ML (2011) Addressing Scientific Fraud. *Science*, 334 (6060), p. 1182 [doi:10.1126/science.1216775]

ELLISON, G (2002) The Slowdown of the Economics Publishing Process. *Journal of Political Economy*, 110, pp. 947-993 [doi:10.1086/341868]

ENGERS, M and Gans, J (1998) Why Referees Are Not Paid (Enough). *American Economic Review*, 88, pp. 1341-1349

GINTIS, H (2009) *The Bounds of Reason. Game Theory and the Unification of the Behavioral Sciences* Princeton: Princeton University Press

GODLEE, F and Jefferson, T (Eds.) (2003) *Peer Review in Health Sciences*. Second Edition. London: BJM Publishing Group

HORROBIN, DF (2001) Something Rotten at the Core of Science? *Trends in Pharmacological Sciences,* 22(2), pp. 51-52 [doi:10.1016/S0165-6147(00)01618-7]

JAYASINGHE, UW, Marsh, HW and Bond, N (2006) A New Reader Trial Approach to Peer Review in Funding Research Grants: An Australian Experience. *Scientometrics*, 69(3), pp. 591-606 [doi:10.1007/s11192-006-0171-4]

JEFFERSON, T, Wager, E, Davidoff, F (2002) Measuring the Quality of Editorial Peer Review. *JAMA*, 307(6), pp. 539-628 [doi:10.1001/jama.287.21.2786]

KASSIRER, JP, Campion, EW (1994) Peer Review: Crude and Understudied. *Journal of American Medical Association,* 272, pp. 96-97 [doi:10.1001/jama.1994.03520020022005]

LABAND, DN (1990) Is There Value-Added From the Review Process in Economics? Preliminary Evidence from Authors. *The Quarterly Journal of Economics*, 105(2), pp. 341-252 [doi:10.2307/2937790]

LABAND, DN and Piette, JM (1994) Favoritism Versus Search of Good Papers. Empirical Evidence Regarding the Behavior of Journal Editors. *Journal of Political Economy*, 102, pp. 194-203 [doi:10.1086/261927]

LAMONT, M (2009) *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press [doi:10.4159/9780674054158]

MARTINS, A (2010) Modeling Scientific Agents for a Better Science. *Advances in Complex Systems*, 13(4), pp. 519-533 [doi:10.1142/S0219525910002694]

MERTON, RK (1973) *The Sociology of Science. Theoretical and Empirical Investigations*. Chicago: University of Chicago Press

NEFF, BD and Olden JO (2006) Is Peer Review a Game of Chance? *BioScience*, 56(4), pp. 333-340 [doi:10.1641/0006-3568(2006)56[333:IPRAGO]2.0.CO;2]

PETERS, DP and Ceci, SJ (1982) Peer-Review Practices of Psychological Journals - The Fate of Accepted, Published Articles, Submitted Again. *Behavioral and Brain Sciences*, 5(2), pp. 187-195 [doi:10.1017/S0140525X00011183]

ROEBBER, PJ, Schultz, DM (2011) Peer Review, Program Officers and Science Funding. *Plos One*, 6(4), e18680: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0018680 [doi:10.1371/journal.pone.0018680]

SMITH, R (2006) Peer Review. A Flawed Process at the Heart of Science and Journals. *Journal of the Royal Society of Medicine*, 99, pp. 759-760 [doi:10.1258/jrsm.99.4.178]

SQUAZZONI, F (2010) Peering into Peer Review. *Sociologica*, 3: http://www.sociologica.mulino.it/doi/10.2383/33640

SQUAZZONI, F, Bravo, G, and Takács, K (2013) Does Incentive Provision Increase the Quality of Peer Review? An Experimental Study. *Research Policy*, 42(1), pp. 287-294 [doi:10.1016/j.respol.2012.04.014]

SQUAZZONI, F and Gandelli, C (2012) Saint Matthews Strikes Again. An Agent-Based Model of Peer Review and the Scientific Community Structures. *Journal of Informetrics*, 6, pp. 265-275. [doi:10.1016/j.joi.2011.12.005]

SQUAZZONI, F and Takács, K (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 Hanel, R (2011) Peer Review in a World with Rational Scientists: Toward Selection of the Average. *The European Physical Journal B*, 84, pp. 707-711 [doi:10.1140/epjb/e2011-20545-7]

VAN den BESSELAAR, P and Leydesdorff, L (2007) Past Performance as Predictor of Successful Grant Applications. A Case Study. Den Haag, The Netherlands: Rathenau Institute. Available: http://www.leydesdorff.net/magw/magw.pdf.

WELLER, AC (2001) *Editorial Peer Review: Its Strengths and Weaknesses*. Medford, NJ: Information Today, Inc

WARE, M (2007) Peer Review in Scholarly Journals: Perspective of the Scholarly Community_ An International Study. Bristol: Mark Ware Consulting. Available: http://www.publishingresearch.net/documents/PeerReviewFullPRCReport-final.pdf

WATTS, C. and Gilbert, N (2011) Does Cumulative Advantage Affect Collective Learning in Science? An Agent-Based Simulation. *Scientometrics*, 89, pp. 437-463 [doi:10.1007/s11192-011-0432-8]

WILENSKY, U (1999) NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL