Research Policy 42 (2013) 287–294

Contents lists available at SciVerse ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Does incentive provision increase the quality of peer review? An experimental study

Flaminio Squazzoni[a], Giangiacomo Bravo[b,c,*], Károly Takács[d]

[a] Department of Social Sciences, GECS Research Group, University of Brescia, Via San Faustino 74B, 25122 Brescia, Italy
[b] University of Torino, Via Sant' Ottavio 50, 10124 Torino, Italy
[c] Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy
[d] Institute of Sociology and Social Policy, Corvinus University of Budapest, Közraktár u. 4–6, 1093 Budapest, Hungary

ARTICLE INFO

Article history:
Received 28 March 2011
Received in revised form 26 April 2012
Accepted 26 April 2012
Available online 22 May 2012

Keywords:
Science policy
Peer review
Cooperation
Trust
Reputation

ABSTRACT

Although peer review is crucial for innovation and experimental discoveries in science, it is poorly understood in scientific terms. Discovering its true dynamics and exploring adjustments which improve the commitment of everyone involved could benefit scientific development for all disciplines and consequently increase innovation in the economy and the society. We have reported the results of an innovative experiment developed to model peer review. We demonstrate that offering material rewards to referees tends to decrease the quality and efficiency of the reviewing process. Our findings help to discuss the viability of different options of incentive provision, supporting the idea that journal editors and responsible of research funding agencies should be extremely careful in offering material incentives on reviewing, since these might undermine moral motives which guide referees' behavior.

1. Introduction

Although peer review is crucial for innovation and experimental discoveries in science, it is poorly understood in scientific terms. Peer review is not just important for scientists, but also for institutional agencies to allocate efficiently funds and research grants and for policy makers to guarantee that taxpayer money is well invested into a credible and well functioning system. The decisive role of peers opinion is what guarantees that scientific innovation can be experimentally pursued by scientists through a continuous, decentralized and distributed trial and error process and that science can endogenously self-regulate (although influenced by external constraints and policy guidelines) by determining scientists payoffs (Squazzoni and Takács, 2011).

With origins which dates back to 1752 when the Royal Society of London obtained responsibility for the "Philosophical Transactions", this mechanism is now under increasing strain, because of the growth of scientific publishing, the increasing complexity of research technologies and interdisciplinary collaboration in each work (Alberts et al., 2008; Grainger, 2007). Not only peer review is pivotal for scientific publications (e.g., journals and books), permitting an average of about 1,400,000 ISI journal articles published yearly (Björk et al., 2009). It is also used to allocate research funds and grants, decide about scientists recruitment and promotion and evaluate universities and research institutes productivity, when standard bibliometric criteria do not hold.

Recently, many journal editors and observers have come to the conclusion that some reform of peer review is needed and that the main problem is to increase the reliability and commitment of referees (Alberts et al., 2008; Hauser and Fehr, 2007). The problem is that, although numerous studies of sociology and economics of science have investigated certain principles and mechanisms of the reward structure in science, with important implications of peer review (e.g., Stephan, 1996), few studies have specifically investigated referee behavior and how to increase commitment. A notable exception was Engers and Gans (1998), which suggested a standard economic analytic model that looked at the interaction between editors and referees. Their aim was to understand why referees were willing to perform their task without payment and whether increasing payments to referees could improve journal quality. They showed that any improvements were so costly that they made such incentives unprofitable by generating an escalation of compensation. Indeed, although payment could potentially motivate more referees to agree to review a submission, raising the review rate meant that referees could expect to impose lower costs on the journal by refusing to review a submission. While payment raised the referees' benefit of reviewing, the effect on quality could lower the costs of declining. This implied that payment should increase

* Corresponding author at: University of Torino, Via Sant' Ottavio 50, 10124 Torino, Italy.
E-mail addresses: squazzon@eco.unibs.it (F. Squazzoni), giangiacomo.bravo@unito.it (G. Bravo), karoly.takacs@uni-corvinus.hu (K. Takács).

0048-7333/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.respol.2012.04.014

to compensate for this effect, but this reduced the need for referees to incur private costs in enhancing the quality.

On the other hand, Chang and Lai (2001) followed a similar approach to understand why certain economics journals decided to give referees some kind of rewards, such as a 1-year subscription or a discount submission fee. They concluded that, if reciprocity or reputation motives were present that influenced the relationship between journals and referees, a possible snowballing effect could emerge that increased the referee recruitment rates. If accompanied by payment, this effect could even increase the review quality.

To explore empirically this problem, we have developed an innovative experiment designed to reproduce peer review dynamics under different incentive conditions. Our findings suggest that journal editors and responsible of research funding agencies should be extremely careful in offering material incentives on reviewing, since these might undermine moral motives guiding referees' behavior. On the one hand, as there is no way for editors to dig into details about the referees' effort in due course, a problem of moral hazard by referees may arise even if material incentives are present. On the other hand, and more importantly in our view, following the motivation crowding theory, the presence of material incentives might undermine intrinsic pro-social motivations of individuals by transforming reviewing into a self-interest decision problem (e.g., Bowles, 2008; Frey and Jegen, 2001). This confirms certain arguments of the sociology and economics of science about the peculiarity of the reward structure of science and its normative foundations (e.g., Stephan, 1996) and is consistent with more recent studies on the importance of social norms for reviewing, which also emphasize the irreducible heterogeneity of norms in various scientific domains (e.g., Azar, 2008; Ellison, 2002).

The rest of the paper is organized as follows. Section 2 presents a literature review that revolves around the institutional foundations of peer review, as reflected in the sociology and economics of science literature. As we will see, although these studies are important to look at the public nature of scientific knowledge, a more focused outlook on cooperation problems at the micro level of peer review is needed to understand that reputational and material incentives might be different for the figures involved and to look at how scientists ensure the quality of the knowledge produced in this situation.

Section 3 introduces our idea that the quality of peer review depends on a cooperation problem between editors, authors and referees where conflicting interests, cheating and moral hazard are all possible. Following game-theory literature on cooperation in experimental behavioral sciences (e.g., Gintis et al., 2005; Gintis, 2009), we have focused on trust, incentives and social norms. We have proposed a modified version of the investment game—i.e., a standard experimental framework (see Berg et al., 1995)—which looks at the triadic interaction between editors, authors and referees and allows us to test various incentive schemes. More specifically, our aim is to test whether material incentive provision can increase cooperation between everyone involved in peer review. While existing literature on peer review mostly takes an empirical, case-based approach (e.g., Bornmann, 2011), our idea is to look at the essential mechanisms of peer review through an abstract model that can be tested in the laboratory. This also makes a difference with the few existing economic studies on referee behavior mentioned above, which did not consider realistic and testable behavioral foundations (Chang and Lai, 2001; Engers and Gans, 1998). Moreover, this approach allows us to disentangle peer review mechanisms and to verify the impact of various interaction conditions. This also allows us to evaluate certain measures frequently recurring in the debate on peer review reform among editors of top journals (e.g., Alberts et al., 2008).

Finally, Section 4 illustrates the results of our "peer-review game" while Section 5 discusses them.

2. Science institutions and peer review

The idea that scientific knowledge is a public good and that scientists developed a normative system particularly suitable for its production, which is different from typical market and technology incentives, was suggested by Merton (1942, 1957), Nelson (1959) and Arrow (1962). In general, these classical studies argued that competitive markets provide poor incentives for scientific knowledge production as providers cannot appropriate the benefits derived from use. Moreover, being puzzle solving and discovery so intrinsically rewarding for scientists, the behavior of scientists cannot be understood as a typical maximization problem as the price of the good "knowledge" strongly depends on the preferences of the producer (Pollak and Watcher, 1975).

Dasgupta and David (1994) David (2004) followed this starting point and suggested an institutional perspective by arguing that science and technology should be seen as alternative knowledge production systems based on distinctive social institutions, i.e., distinct values, social norms and rewards. In their view, the "Realm of technology" was inhabited by secrecy, privatization and protection of knowledge, which ensured that knowledge could intercept market rewards. On the other hand, the values of openness, communitarism, disinterestedness and universalism were functional to the development of the so-called "Republic of science", where competition was based on priority and rewards followed reputational credit accumulation in the public sphere.

In an influential review of the economics literature on science, Stephan (1996) argued that these institutional features explain why a reward structure based on "non-market-based incentives" evolved in science that encouraged the production of the public good "knowledge". Her argument was that as scientists compete for priority in a context of possible mutual discoveries, they are pushed to share knowledge in a timely fashion. This generates positive externalities, such as the appropriability of knowledge by others and its growing value through multiple uses, which give rise to reputational credits for knowledge providers, such as publications and citations, which in turn fuel new knowledge production, e.g., access to new research funds for highly reputed scientists. Therefore, in this view, the fact that scientists' careers depend on reputational credits, which are built upon publications and citations, explains the public nature of knowledge and ensures a solution of the appropriability dilemma inherent in the creation of any public good.

Furthermore, these studies suggested that science can avoid the classical tragedy of the commons thanks to the strength of the intrinsic motivation of scientists (e.g., Dasgupta and David, 1994; Stern, 2004). This seems to be even corroborated by recent empirical findings. Using data survey from over 400 science and engineering PhD students in North Carolina, a recent study emphasized that students who opted for an academic career differed from those who followed a career in the private sector. Indeed, the former showed a stronger "taste for science" and a weaker concern for salary and access to resources than the latter (Roach and Sauermann, 2010; see also Lacetera, 2009). This confirmed certain results of an influential empirical survey on multiple job offers to post-doctoral biologists in the US, where it was found that wages and science were negatively correlated, so that scientists seemed to "pay" to become scientists (Stern, 2004). Similar differences of motivations and attitudes between academic scientists and private researchers were recently found also by Häussler (2011), who built a dataset of 1353 academic and 341 industry-based bio-scientists. Her results showed that academic scientists conformed to the norm of open science and shared their knowledge even when sharing

could decrease the economic value of knowledge and so their material payoffs. This was not true for private researchers.

The problem is that although these studies have provided insights to understand crucial problems related to the public nature of scientific knowledge, they did not sufficiently consider the importance of understanding what is the mechanism that ensure the quality of the knowledge produced by scientists. The fact that the quality of scientific knowledge is inherently determined through peer review and its value is co-produced even through anonymous peer collaboration implies that micro conditions and mechanisms that preside over the peer production of the public good "knowledge" should be seriously considered.

This is for two reasons. First, peer review plays a twofold role in science. It is a screening mechanism to avoid that low quality research is published ensuring that the allocation of reputational credits and rewards can self-regulate appropriately. On the other hand, it also has a knowledge generation function as anonymous peers contribute to increase the knowledge value of any author submission at their own expenses (e.g., Bornmann, 2011; Hamermesh, 1994; Laband, 1990). Secondly, even if the reward structure postulated by the economics of science studies can account for the knowledge production and accumulation at the macro level, this does not reflect peer review at the micro level, where the reward structure is more ambiguous, especially in case of referees.

Some recent cases of misconduct and fraud in science, such as the stem cell in Science 2005 or the more recent Stapel scandal, casted doubt upon the normative peculiarity of science and the efficiency of the reward structure that presides over peer review, especially in times of increasing competitiveness among scientists (e.g., Couzin, 2006; Crocker and Cooper, 2011; Squazzoni, 2010). In these cases, unfair behavior by self-interested authors combined with unreliable behavior of referees.

In this respect, the few economic studies on peer review have shown that one of the main challenges of peer review is to understand referee behavior, which is far from being fully understandable in rational strategic terms. Some authors argued that referees might be motivated to take their job seriously by direct or indirect reciprocity (e.g., Azar, 2008; Chang and Lai, 2001; Engers and Gans, 1998). In this view, referees would cooperate with journal editors in ensuring the quality of evaluation as they are concerned about protecting the prestige of the journal as a means to protect their own impact in case of previous publication. However, this does not explain why they would cooperate with authors, by providing feedback that helps to improve the submission quality. Another explanation is that referees would cooperate with authors for indirect reciprocity, that is, by establishing good standards of review in prospect of benefiting from other referees when they will be authors.

If this is true, we also must consider that experimental game theory and behavioral literature showed that any reciprocity strategy is extremely sensitive to interaction situations as it strongly refers to behavior of others and its contribution to optimal cooperation outcomes depends on the co-presence and mutual balance of other strategies, such as free-riding and altruism (e.g., Gintis, 2009; Gintis et al., 2005). In the case of peer review, this means that the challenge is to explain why and how scientists escape from free-riding temptations and anonymously collaborate via peer review to protect the quality of journals and increment the knowledge value of submissions at the benefit of authors, even if reputational and material rewards are ambiguous and difficult to predict for one of the most important figures involved, that is, the referee.

To fill this gap, we believe that sociology and economics of science should be integrated with experimental research capable of looking at peer review in detail. Indeed, peer review is a typical cooperation problem between editors, authors and referees. In our view, looking at this interaction is essential to understand how science works as the concatenation of the strategies of everyone involved determines the quality of peer review, the reliability and value of the scientific knowledge and consequently the efficient self-regulation of reputational credits and rewards in science. Following this approach, we suggest to view peer review as an investment game played by editors, authors and referees in condition of trust and information asymmetries problems. In our view, editors are aimed to maximize their investment (e.g., time dedicated to referee search and selection and review management, initiatives to maintain the prestige of their journal) to protect their journal reputation and attract good author submissions, by sharing the burden of this with referees. Referees are needed for competent evaluations that help to reduce the knowledge and information asymmetries between the other figures involved and build mutual trust. Authors aim to maximize their chances of being published by deciding a level of research effort commensurate with the prestige of the journal, while they may be tempted to cheat by submitting research of lower quality than actually claimed. Referees are called to express reliable and fair evaluations commensurate both with the prestige of the journal and the quality of the author submission, but operate under ambiguous incentives.

Our assumption is that the quality of peer review is the outcome of this complex triadic interaction, with multiple interests potentially misaligned which should be carefully examined. The goal of the experimental research presented in this paper is to illuminate this interaction and to test various incentive schemes, so as to look at implications of referee behavior for the quality of the peer review process.

3. Methods

We started from a standard experimental framework, known as the "Investment Game" (Berg et al., 1995), which we modified to look at the most important peer review mechanisms so as to test the efficiency of different incentive schemes.[1] First, to observe the added value of peer review and treatment effects, we designed a *Baseline* treatment where the investment game took place without referees. Subjects were randomly paired to play in A and B positions. In each pair, both subjects received an initial endowment (d) of 10 monetary units (MU). First, A players decided how much of their endowment to "invest" (i) with B players. The amount not invested remained as part of A earnings. Investments were then tripled[2] and sent, in addition to the endowment, to B players, who chose an amount to return (r) to A. The amount returned was summed with A earnings, while the part kept by B players represented their payoff.

The investments of A players are analogous to the time and effort invested by editors to attract articles that increase or at least maintain the reputation of their journals. This follows the findings of previous studies which have emphasized the role of editors in ensuring the quality of peer review (e.g., Neff and Olden, 2006). Not only should editors invest time and money to manage the whole evaluation process and improve the quality and accountability of their journal evaluation policy, they should also be competent in many fields to ensure an appropriate referee selection.

Moreover, as in our game, editors face knowledge uncertainty about the quality of submissions. On the other hand, authors, like B

[1] Full instructions of the experiment and the ex-post questionnaire to be filled by participants are available upon requests to the corresponding author.

[2] While tripling the amount invested was not necessary to look at the trust problem of peer review (any multiplier greater than one was sufficient), we used this coefficient, which is standard in investment game literature (e.g., Berg et al., 1995; Keser, 2003), to compare our results with previous experiments.

players in the experiment, could honor the editors' investment by providing work with true and original scientific quality. Pressurized by the publish or perish rule, authors may be tempted to cheat, e.g., by submitting research findings of lower quality than actually claimed.

Considering that interactions were one-shot, couples were randomly assigned each round, there was no sanction for unfair behavior and assuming rational choice B players had no incentive to return anything, the only rational strategy for A players was to keep their whole endowment. This led to the only subgame perfect equilibrium of the game, where both investments and returns are zero and all players earn 10 MU. This outcome was sub-optimal since any sum invested by A was tripled by the experimenter, therefore increasing the total amount to share. Pareto optimality was given by A players investing their whole endowment, while an outcome both optimal and fair was possible for $i = 10$ and $r = 20$, with all players earning 20 MU.

Then, we introduced a third player into the game (player C) in the role of the referee. When selected as referees, subjects were informed of the amount received and returned by the B players the last time they played in the same position. Then, referees were asked to rate B players' behavior as "negative", "neutral" or "positive". Reviews were displayed to A players before the subsequent investment decisions. As C players, the referees should guarantee the editors' investment by writing reliable evaluations of authors' submissions. The fact that C players knew both A investments and B returns mirrors the typical situation of referees who should express an evaluation matching both the journal's quality (i.e., the amount of the A investment) and the quality of the contribution (i.e., the amount of B returns).

Once referees were introduced, we varied the incentive schemes offered to them. This was crucial as the payoffs of referees were the real challenge to investigate. In the *No incentive* treatment, subjects did not receive any reward for reviewing. This treatment mimics peer review as it is now. When applied to this interaction scheme, the incentive-based rational choice perspective predicts that reviews should not be seriously taken into account either by editors or by authors, since referees lack motivation for their job.

In the *Fixed incentive* treatment, referees received a fixed payoff of 10 MU, equal to A and B endowments. Fixed incentives mirror the present situation at certain journals (e.g., the British Medical Journal), where referees are supported by fixed stakes (e.g., money or access to scholarly archives) and this could motivate them to reciprocate by increasing their effort (e.g., Chang and Lai, 2001).

In the *A incentive* treatment, referees' earnings were equal to the payoff of A players. This alignment of interests could resolve the principal-agent problem between editors and referees, by motivating the agents (referees) to act on behalf of the principals (editors) guaranteeing that the self-interest of the latter coincides with the objectives of the former. This treatment is therefore expected to lead to more reliable reviews and higher efficiency. Although this does not mimic a real situation, this treatment is important as most editors would like to increase referee commitment.

In the *B incentive* treatment, referee earnings were equal to the payoff of B players. This follows Laband (1990) argument that, as each published article includes also the contribution of referees in terms of feedbacks and suggestions, it is reasonable to think about measures to share payoffs between authors and referees—e.g., referees' names included in the published article—although currently not explored in scientific journals. The alignment of authors' and referees' interests was expected to determine an exploitation of the goodwill of editors and therefore to produce less reliable reviews and lower editors' investment.

Subjects ($N = 136$) participated in the experiment held at the University of Brescia at the end of November 2010. Participants were students recruited across the different university faculties using the online system ORSEE (Greiner, 2004). They played in groups of 27 subjects (28 in the *Baseline*) in one of the above treatments for 30 periods. Couples in the *Baseline* and triplets in peer review treatments were randomly rematched after each period to avoid the use of reciprocal strategies. Subjects interacted anonymously through a computer network using the experimental software z-Tree (Fischbacher, 2007). Each session, including reading of instructions, playing the game for 30 periods and filling in an ex-post questionnaire, took approximately 75 min. In all treatments, we used virtual monetary units with an exchange rate of 1 MU = 2.5 Euro Cents. Participants were paid immediately after the experiment in cash and earned an average of 14.90 Euros.

4. Results

Previous experiments using the investment game showed that A players invested on average between one third to half of their endowments. Returns were slightly lower than investments, making trustful behavior not particularly profitable on average (Berg et al., 1995; Johnson and Mislin, 2011; Ortmann et al., 2000). Our study replicated these results and, consistently with previous studies which introduced reputational motives in the investment game (Boero et al., 2009; Keser, 2003), showed that peer review improved both efficiency and cooperation dramatically. Both investments and returns were higher in peer review treatments, with investments increasing from an average of 3.22 MU in the *Baseline* up to 5.21 MU in *A incentive* and returns rising from 2.00 in the *Baseline* to 6.87 in *No incentive* (Fig. 1A and B and Table 1). The amounts exchanged in the first three periods of the game, when referees had no previous information to evaluate, and in the last three periods, when B players knew that no further review would take place, were not included in the analysis.[3]

Differences with the *Baseline* for both investments and returns were significant for all treatments except *Fixed incentive*, where the difference was significant only for returns (Table 2). However, significant differences also existed between peer review treatments, especially for B returns. Both *No incentive* and *A incentive* led to higher returns than *Fixed incentive* (Wilcoxon rank sum tests on individual averages, $W = 531.0$, $p = 0.002$ and $W = 199.0$, $p = 0.002$ respectively). There were no significant differences between *No incentive* and *A incentive* ($W = 385.0$, $p = 0.365$). Differences in investments were smaller, but still remained statistically significant at 5% between *No incentive* and *Fixed incentive* ($W = 508.0$, $p = 0.006$) and between *A incentive* and *Fixed incentive* ($W = 176.5$, $p = 0.001$).

To better describe the dynamics of cooperation in the peer review game, we built a concise indicator that summarized the results of the game in a single measure. The fundamental reason in doing this arose from the fact that, in the investment game, Pareto optimality depends only on A investments, but we should also take B's behavior into account as a critical element that determines scientific quality. Nevertheless, Pareto optimality remains an important indicator of the overall system efficiency in the different treatments. This is indicated by $E = i/d$ where i represents A's investment and d is the endowment. This indicator is clearly zero when A invests zero and one when A invests the whole endowment. Following previous research (Almås et al., 2010; Fehr and Schmidt, 1999; Nowak et al., 2000; Rabin, 1993), we took B returns into account by adopting a fairness criterion favoring outcomes where both players obtained equal payoffs $F = 1 - (|P_A - P_B|)/(P_A + P_B)$ where P_A and P_B are the payoffs earned by A and B players respectively. This is

[3] Our dataset may be accessed upon request to the corresponding author.

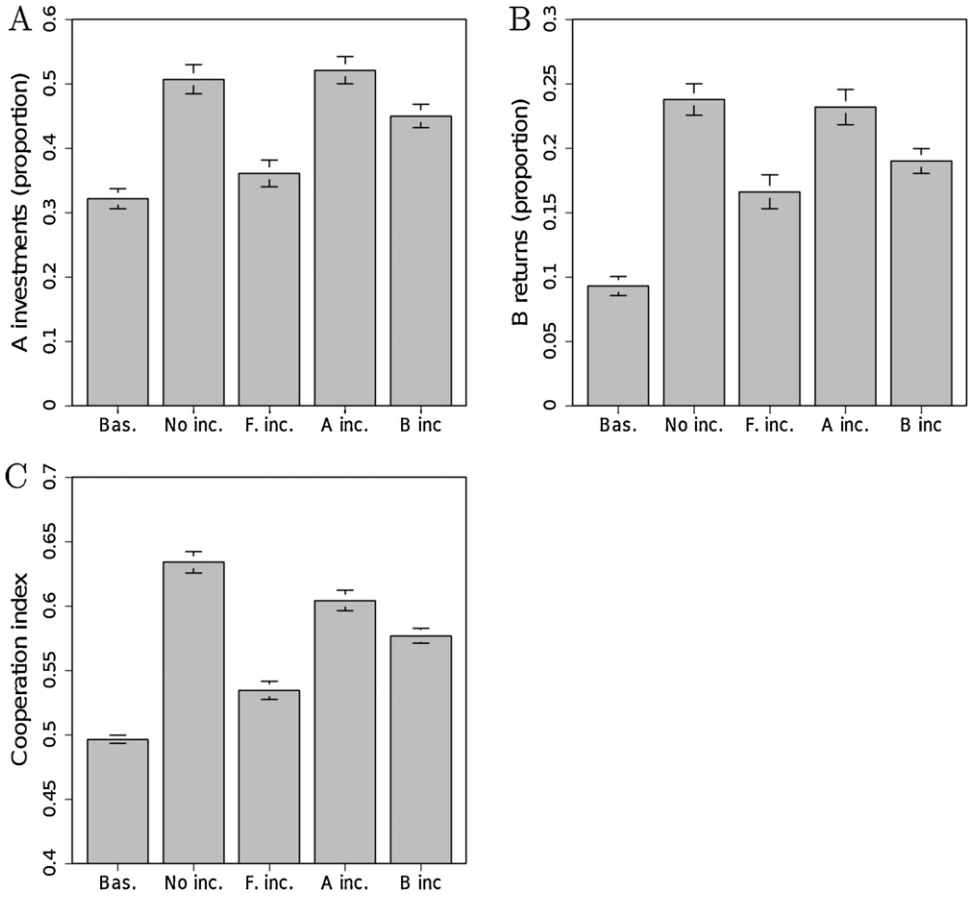

Fig. 1. Average investments (A), returns (B) and cooperation index (C) by treatment with standard error bars. Investments are represented in proportion of A endowment (10 MU). Returns are expressed as proportion of the overall B endowment (3× *amount received* +10 MU). The cooperation index varied from zero for highly inefficient and inequitable outcomes to one for efficient and equitable outcomes.

Table 1
Average investments, returns and cooperation index by treatment. Returns are showed both as absolute figures and as proportion of B endowment.

	Baseline		No inc.		Fixed inc.		A inc.		B inc.	
	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM
A inv. (MU)	3.22	0.16	5.07	0.23	3.61	0.21	5.21	0.21	4.50	0.18
B ret. (MU)	2.00	0.16	6.87	0.42	3.75	0.30	6.42	0.45	4.75	0.28
B ret. (pr.)	0.09	0.01	0.24	0.01	0.17	0.01	0.23	0.01	0.19	0.01
CI	0.50	0.00	0.63	0.01	0.54	0.01	0.60	0.01	0.58	0.01

Table 2
Wilcoxon rank sum tests on differences between treatments with one tailed *p* values.

	Baseline		No inc.		Fixed inc.		A inc.	
	W	*p*	*W*	*p*	*W*	*p*	*W*	*p*
A invest.								
No inc.	212.5	0.003						
Fixed inc.	341.0	0.269	508.0	0.006				
A inc.	189.0	0.001	359.5	0.469	176.5	0.001		
B inc.	245.5	0.013	418.0	0.180	248.0	0.022	439.5	0.099
B returns (absolute)								
No inc.	105.0	0.000						
Fixed inc.	238.0	0.009	531.0	0.002				
A inc.	90.0	0.000	385.0	0.365	199.0	0.002		
B inc.	150.5	0.000	480.0	0.023	287.0	0.091	467.5	0.038
B returns (prop.)								
No inc.	136.0	0.000						
Fixed inc.	258.0	0.022	487.0	0.017				
A inc.	120.0	0.000	380.0	0.398	241.0	0.017		
B inc.	173.0	0.000	450.5	0.070	301.0	0.138	438.0	0.104
CI								
No inc.	70.0	0.000						
Fixed inc.	228.0	0.006	583.0	0.000				
A inc.	20.0	0.000	452.0	0.067	133.0	0.000		
B inc.	62.0	0.000	509.0	0.006	206.0	0.003	457.0	0.056

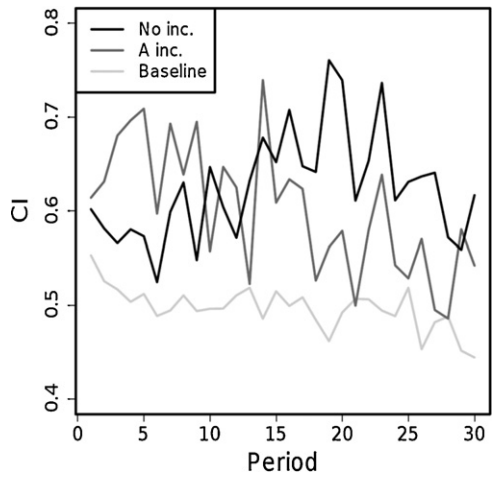

Fig. 2. Cooperation index dynamics in *No incentive* and *A incentive*. The *Baseline* curve at the bottom of the figure is inserted as reference.

zero when one of the players obtains the whole amount at stake and the other receives zero, while it becomes one when both players obtain the same payoff. Averaging the two criteria, we defined the cooperation index as $CI = (E + F)/2$. This is zero when A players invest zero and B players return all their endowments, grows with both A investments and a fairer distribution of final payoffs, and becomes one when As invest d and Bs return half of their overall endowment.

The treatment with the highest *CI* was *No incentive*, which led to more cooperative outcomes than any other treatment (Fig. 1C). Differences were statistically significant at 10% with *A incentive* and at 5% with the other treatments. The high *CI* value in *No incentive* was especially important since, unlike *A incentive*, referees had no incentive to cooperate with A players. This indicated that material incentives, rather than guaranteeing higher referees' commitment, were superfluous and might even backfire by eroding the reliability of the entire review process.

It is worth noting that the most cooperative treatments in our experiment performed differently in the first and in the final part of the game (see Fig. 2). In periods 4–15, the *CI* was 0.60 ± 0.01 in *No incentive* and 0.64 ± 0.01 in *A incentive*, while these figures were 0.67 ± 0.01 and 0.56 ± 0.01 respectively in periods 16–27. The differences were significant ($W = 252$, $p = 0.026$ for periods 4–15 and $W = 558$, $p = 0.000$ for periods 16–27), suggesting that material disinterest guaranteed more robust cooperation in the long run.

In all peer review treatments, A players largely used referees' ratings for their investment decisions and systematically invested more when they received positive reviews (Fig. 3A). There were also differences in the average return proportion that induced negative, neutral, or positive reviews (Fig. 3B), a fact that is crucial to understand cooperation differences among treatments. In *No incentive* and *A incentive* referees were more selective, requiring an average return proportion of about one third of B overall endowment to award positive reviews. In *Fixed incentive* and *B incentive* this proportion declined instead to one quarter or less, leaving more room for the authors' opportunistic behavior.

A questionnaire at the end of the experiment focused on the participants' perception of other subjects' behavior. Participants rated B players as more trustworthy in all peer review treatments than in the *Baseline* ($W = 1790.5$, $p = 0.061$) and in *No incentive* than in *A incentive*, although these differences were significant at only 10% ($W = 447.5$, $p = 0.068$). Also referees were rated as most reliable in *No incentive*. Differences were significant between *No incentive* and both *Fixed* and *B incentive* ($W = 457$, $p = 0.052$ and $W = 457$, $p = 0.050$, respectively), whereas they were not significant between *No incentive* and *A incentive* ($W = 376$, $p = 0.423$).

5. Discussion

Our findings show that the most effective peer review scheme is the one currently in use where referees are not supported by material incentives. Its maintenance avoids that peer review undergoes a frame effect motivating also well disposed referees to behave selfishly in turn. Questionnaire answers further confirmed that higher trust and cooperation were guaranteed by the reviewing scheme set up in *No incentive*. This is consistent with previous studies showing that people were less committed when material incentives were added to social interactions that were usually driven by intrinsic, materially disinterested motivations (Bowles, 2008; Heyman and Ariely, 2004; Vohs et al., 2006). A recent theory called "motivation crowding theory" has been elaborated that accounts for a broad range of phenomena where incentives undermine intrinsic prosocial motivations of individuals so as to dominate the traditional relative price effect (Frey and Jegen, 2001). As material interests and moral motives cannot be separated, incentives could transform interactions into a self-interest decision problem. This would make self-interest the appropriate behavior (Bowles, 2008) and peer review would not be an exception.

The *A incentive* scheme, where referees had incentives aligned with editors, was similarly productive, but less robust than the former. It also was less equitable in cooperation outcomes. This is important as only a positive equilibrium between editor investments and author returns can ensure effective quality coordination in science. Indeed, as suggested by Merton (1968) and more recently by van Dalen and Henkens (2005), unequal allocation of attention is functional to the good working of science, especially in times of increasingly complex publishing market, where figures have become impressive. This means that intelligible quality signals are important to drive scientists' attention toward a given publication, such as the prestige of a journal. As such, these signals help to determine a rational allocation of reputational credit for authors and journals. In this sense, a more equitable outcome of our game reflects similar coordination issues in the science market. Indeed, when our cooperation index increased, it meant that investment and returns were positively concatenated so that the higher prestige of a journal, which was dependent on higher editor investments, corresponded to an author contribution of higher level (e.g., higher author returns) and vice-versa. It is worth noting that one of the most crucial coordination problems between producers and users of scientific knowledge is to avoid that while good quality journals publish low quality contributions, good quality contributions are hidden in low quality journals, thereby increasing the cost of users to find them and the probability of suboptimal credit allocation.

Moreover, aligned incentive provision is extremely difficult to implement in journals, as it requires incentives which are sensitive to interaction outcomes. This means that the scientific value of a published article should be completely assessable within peer review interaction, as well as the effort needed for reviewing it. Unfortunately, we know that the former can be evaluated only ex-post and in the long run while the latter differs from subject to subject and is practically impossible to measure. The only feasible way to add material incentives to peer review is introducing fixed rewards, but our experiment showed that this scheme was the worst in promoting cooperation.

Our experiments explain why the current practice of peer review based on voluntary contributions is so pervasive and efficient. It is likely that this is so because the current practice fully exploits the reciprocity motives that typically drive human

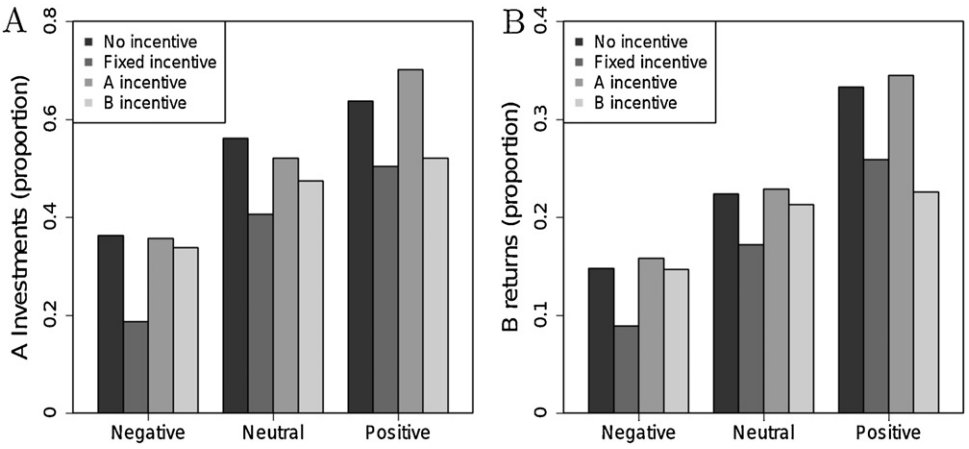

Fig. 3. Review effects. (A) Average investment proportion by review. (B) Average return proportions required by referees to award negative, neutral or positive evaluations.

behavior in many social interactions (Gintis, 2000; Ostrom and Walker, 2003; Sigmund, 2010). Most of us take seriously reviewing and do their best to return useful and detailed reports to authors, as we know that our peers will do the same in turn to our benefit. On the other hand, the credibility of referees is essential to motivate editor investments and reduce author free riding temptations and this explains why the quality of peer review is higher when referees are disinterested. This would confirm some recent studies on the importance of social norms and learning for reviewing, where referee behavior seems to depend less from incentives and more from social imitation and norm compliance (e.g., Azar, 2008). In this respect, also the role of editors should be investigated in more detail.

This point requires us to touch upon certain crucial questions, which also involves considering certain limitations of our experimental research. Being general and abstract, our experiment did not account for the presence of various incentive and reward structures, nor it seriously reflected possible disciplinary heterogeneity and difference in social norms in science (e.g., Lamont, 2009; Whitley, 1984). On the one hand, it is widely acknowledged that scientific institutions have recently become "stretched" institutions that encounter competitive and even contradictory pressures, such as also responding to industry needs and favoring technology transfer (Bonaccorsi and Daraio, 2010; Nowotny et al., 2005). If this is true, it is reasonable to expect that various incentive and reward structures might co-exist in science that make generalization difficult. For instance, in fields where research is closely connected with industry and market applications, payoffs of scientists could be less related to academic reputation and more to entrepreneurial achievements. This means that, in such fields, material incentives could have a more positive impact on cooperation in peer review. Therefore, in future developments of our research more attention should be paid to disciplinary differences, such as those between traditional academic communities and communities more oriented toward applied research and more sensitive toward marketization. Field experimental studies capable of looking at real behavior of scientists would help to understand this.

On the other hand, academic communities could have developed specific social norms that make cooperation and trust strongly embodied in codes of conduct of individuals, so that cooperative outcomes would less depend from agent rational strategies and more from field specific institutions (Lamont, 2009). For instance, in case of refereeing time, certain features of normative heterogeneity have been empirically found even between brother disciplines such as economics and finance (e.g., Azar, 2006, 2008). This would make us suppose that in reality scientists are even less responsive to incentives than in our experiment. On the one hand, this calls for considering institutional specificities, which only field experiments could help to illuminate (Helbing and Balietti, 2011), providing also an important external validity test for our lab results. On the other hand, the fact that social norms had a strong influence on individual behavior in a "cold", abstractly modeled peer review lab interaction, with students who were not-socialized to a scientific ethos, can even strengthen the generalizability of our results toward real science system.

Finally, suggesting to avoid material incentives does not mean that journals, academic associations and research agencies could revel in doing nothing. In our view, there are two possible lines for improving the present situation. The first one has to do with the attempt of valuing more the reviewing activity of scientists for their professional recognition and reputation. The second one has more to do with improving the normative foundations of science.

As regards to the first, the value and the payoff of each publication embody comments, ideas and efforts by referees but are capitalized just by authors, as the former do not have any concrete reputational benefit from authors' publications. Certain studies have even shown that peer review is probably more useful for improving author contributions than for sorting "the wheat from the chaff" (e.g., Pierie and Walvoort, 1996; Smith, 1999). Our suggestion here is that journals could improve the way referees' contributions are presently acknowledged by establishing symbolic awards for referees, including referees' names in each published articles and, more importantly, defining clear rules that link the admission and turnover of peers into their editorial boards also to excellence in reviewing. Research agencies could similarly find ways to value the reviewing experience of applicants when evaluating applications. These types of initiatives would exploit reputational motives rather than material self-interest, and consequently would improve cooperation without deteriorating the moral dimension behind peer review.

As regards to the second point, initiatives by scientific associations and research agencies which could promote intrinsic motivations and the moral dimension of science, by emphasizing the relevance of reviewing should be taken. An example could be teaching reviewing and its moral importance in science in PhD courses. Obviously, given that our findings help to establish what should not be done, further research is needed that examines which initiatives need to be taken to improve peer review.

Acknowledgments

We gratefully acknowledge help provided in the design and realization of the experiment by Riccardo Boero, Vincent Buskens, Niccolò Casnici and Marco Castellani, along with the

Intra-European Fellowship Program of the European Union, GA-2009-236953. A preliminary version of this paper was presented at the conferences on "Game Theory and Society", held at ETH Zurich, on 27–30 July 2011 and at the Third ICORE Conference on Reputation, held at Agropolis International, Montpellier on 19 September 2011. We would like to thank conference referees and participants for useful remarks. We would also like to thank two anonymous journal referees for important suggestions and for having directly confirmed our findings on the importance of referees. Funding was provided by the Collegio Carlo Alberto and the University of Brescia. A special thanks to Robert Coates, Centro Linguistico Bocconi University Milan, for his linguistic revision. Usual disclaimers apply.

References

Alberts, B., Hanson, B., Kelner, K.L., 2008. Reviewing peer review. Science 321, 15.
Almås, I., Cappelen, A.W., Sorensen, E.O., Tungodden, B., 2010. Fairness and the development of inequality acceptance. Science 328, 1176–1178.
Arrow, K.J., 1962. Economic welfare and the allocation of resources for invention. In: National Bureau of Economic Research (Ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton University Press, Princeton, NJ, pp. 609–625.
Azar, O.H., 2006. The academic review process: how can we make it more efficient? American Economist 50, 37–50.
Azar, O.H., 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, 420–435.
Berg, J., Dickhaut, J., McCabe, K.A., 1995. Trust, reciprocity and social history. Games and Economic Behavior 10, 122–142.
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.
Boero, R., Bravo, G., Castellani, M., Laganà, F., Squazzoni, F., 2009. Pillars of trust: an experimental study on reputation and its effects. Sociological Research Online 14 (5), 5, http://www.socresonline.org.uk/14/5/5.html.
Bonaccorsi, A., Daraio, C., Geuna, A., 2010. Universities in the new knowledge landscape: tensions, challenges, change: an introduction. Minerva 48, 1–4.
Bornmann, L., 2011. Scientific peer review. Annual Review of Information Science and Technology 45, 199–245.
Bowles, S., 2008. Policies designed for self-interested citizens may undermine "the moral sentiments": evidence from economic experiments. Science 320, 1605–1609.
Chang, J., Lai, C., 2001. Is it worthwhile to pay referees? Southern Economic Journal 68, 457–463.
Couzin, J., 2006. . . . and how the problems eluded peer reviewers and editors. Science 311, 614–615.
Crocker, J., Cooper, M.L., 2011. Addressing scientific fraud. Science 334 (6060), 1182.
Dasgupta, P., David, P., 1994. Towards a new economics of science. Research Policy 23, 487–521.
David, P.A., 2004. Can 'open science' be protected from the evolving scheme of IPR protections? Journal of Institutional and Theoretical Economics 160 (1), 9–34.
Ellison, G., 2002. The slowdown of the economics publishing process. Journal of Political Economy 110, 947–993.
Engers, M., Gans, J., 1998. Why referees are not paid (enough). American Economic Review 88, 1341–1349.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114, 817–868.
Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.
Frey, B., Jegen, R., 2001. Motivation crowding theory. Journal of Economic Surveys 15, 589–611.
Gintis, H., Bowles, S., Boyd, R., Fehr, E. (Eds.), 2005. Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life. The MIT Press, Cambridge, MA.
Gintis, H., 2000. Strong reciprocity and human sociality. Journal of Theoretical Biology 206, 169–179.
Gintis, H., 2009. The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences. Princeton University Press, Princeton.
Grainger, D.W., 2007. Peer review as professional responsibility. A quality control system only as good as the participants. Biomaterials 28, 5199–5203.
Greiner, B., 2004. An online recruitment system for economic experiments. In: Kremer, K., Macho, V. (Eds.), Forschung und Wissenschaftliches Rechnen 2003. Ges. für Wiss, Datenverarbeitung, Göttingen, pp. 79–93.
Häussler, C., 2011. Information-sharing in academia and the industry: a comparative study. Research Policy 40, 105–122.
Hamermesh, D., 1994. Facts and myths about refereeing. The Journal of Economic Perspectives 8, 153–163.
Hauser, M., Fehr, E., 2007. An incentive solution to the peer review problem. PLoS Biology 5, e107.
Helbing, D., Balietti, S., 2011. How to create an innovation accelerator. The European Physical Journal 195, 101–136.
Heyman, J., Ariely, D., 2004. Effort for payment: a tale of two markets. Psychological Science 15 (11), 787–793.
Johnson, N.D., Mislin, A.A., 2011. Trust games: a meta-analysis. Journal of Economic Psychology 32, 865–889.
Keser, C., 2003. Experimental games for the design of reputation management systems. IBM System Journal 42, 498–506.
Laband, D.N., 1990. Is there value-added from the review process in economics? Preliminary evidence from authors. The Quarterly Journal of Economics 105 (2), 341–252.
Lacetera, N., 2009. Different missions and commitment power in R&D organizations: theory and evidence on industry–university alliances. Organization Science 20 (3), 565–582.
Lamont, M., 2009. How Professors Think: Inside the Curious World of Academic Judgment. Harvard University Press, Cambridge, MA.
Merton, R.K., 1942. The normative structure of science. Journal of Legal and Political Sociology 1, 115–126 (Reprinted in The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press, 1973, 267–278).
Merton, R.K., 1957. Priorities in scientific discovery: a chapter in the sociology of science. American Sociological Review 22 (6), 635–659.
Merton, R.K., 1968. The Matthew effect in science. Science 159 (3810), 56–63.
Neff, B.D., Olden, J.O., 2006. Is peer review a game of chance? BioScience 56 (4), 333–340.
Nelson, R.R., 1959. The simple economics of basic scientific research. Journal of Political Economy 67 (3), 549–583.
Nowak, M.A., Page, K.M., Sigmund, K., 2000. Fairness versus reason in the ultimatum game. Science 289, 1773–1775.
Nowotny, H., Trute, H.H., Schmidt, A. (Eds.), 2005. The Public Nature of Science Under Assault: Politics, Markets, Science and the Law. Springer Verlag, Berlin.
Ortmann, A., Fitzgerald, J., Boeing, C., 2000. Trust, reciprocity, and social history: a re-examination. Experimental Economics 3, 81–100.
Ostrom, E., Walker, J. (Eds.), 2003. Trust & Reciprocity: Interdisciplinary Lessons from Experimental Research. Russel Sage Foundation, New York.
Pierie, J., Walvoort, H., Overbeke, A.J., 1996. Readers' evaluation of effect of peer review and editing on quality of articles in the Nederland's tijdschrift voor geneeskunde. The Lancet 348, 1480–1483.
Pollak, R.A., Watcher, M.L., 1975. The relevance of the household production function and its implications for the allocation of time. Journal of Political Economy 83 (1), 255–277.
Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.
Roach, M., Sauermann, H., 2010. A taste for science? PhD scientists' academic orientation and self-selection into research careers in industry. Research Policy 39, 422–434.
Sigmund, K., 2010. The Calculus of Selfishness. Princeton University Press, Princeton.
Smith, R., 1999. Opening up BMJ peer review. British Medical Journal 318 (4), 4–5.
Squazzoni, F., 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.
Squazzoni, F., 2010. Peering into peer review. Sociologica 3. http://www.sociologica.mulino.it/doi/10.2383/33640.
Stephan, P.E., 1996. The economics of sciences. Journal of Economic Literature 34, 1199–1235.
Stern, S., 2004. Do scientists pay to be scientists? Management Science 60 (6), 835–853.
van Dalen, H., Henkens, K., 2005. Signals in science: on the importance of signaling in gaining attention in science. Scientometrics 64 (2), 209–233.
Vohs, K.D., Mead, N.L., Goode, M.R., 2006. The psychological consequences of money. Science 314, 1154–1156.
Whitley, R., 1984. The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford.