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

Giangiacomo Bravo[1,2] & Flaminio Squazzoni[3] & Károly Takács[3]

[1] Department of Social Sciences, University of Torino
[2] Collegio Carlo Alberto
[3] Department of Social Sciences, University of Brescia
[4] Institute of Sociology and Social Policy, Corvinus University of Budapest

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 reviewers 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 reviewers' behavior.

Keywords: science policy; peer review; cooperation; trust; reputation.

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 influence by external constraints and policy guidelines) by determining scientists payoffs.

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 journal articles published yearly (Bjrk, Roos and Lauri 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 reviewers (Hauser and Fehr 2007). Mainstream economic theory predicts that scientists, like all, are rational agents who follow material incentives, so the quality and efficiency problem of peer review could be viewed as simply a problem of incentive provisions for reviewers. Given that reviewing is not compensated, nor it is at the top of the list for reputation building, we could argue that the commitment of reviewers could be improved by adding material incentives.

To test this hypothesis, 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 reviewers behavior. On the one hand, as the true quality of submissions cannot be properly judged by editors or responsible of funding agencies and there is no way for them to dig into details about the reviewers effort in due course, a problem of moral hazard by reviewers 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).

The remaining of the paper is organized as follows: Section 2 presents the research methods, Section 3 illustrates the results, while Section 4 discusses them.

2 Methods

The design of our experiment aims to model the interaction of editors authors and reviewers as a trust problem under uncertainty, where conflicting interests, cheating and moral hazards are possible. 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.

First, to observe the added value of peer review and treatment effects, we designed a *Baseline* treatment where the Investment Game took place without reviewers. 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 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. As in our game, editors face knowledge uncertainty about the quality of submissions. On the other hand, authors, like B 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, and there was no sanction for unfair behavior, assuming rational choice B players had no incentive to return anything. Therefore, 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 reviewer. When selected as reviewers, subjects were informed of the amount received and returned by the B players the last time they played in the same position. Then, reviewers 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 reviewers 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 reviewers 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 reviewers were introduced, we varied the incentive schemes offered to them. 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 nor by authors, since reviewers lack motivation for their job.

In the *Fixed incentive* treatment, reviewers 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 reviewers are supported by fixed stakes (e.g., access to scholarly archives) and this could motivate them to reciprocate by increasing their effort.

In the *A incentive* treatment, reviewer earnings were equal to the payoff of A players. This alignment of interests could resolve the principal-agent problem between editors and reviewers, by motivating the agents (reviewers) 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.

In the *B incentive* treatment, reviewer earnings were equal to the payoff of B players. As each published article includes also the contribution of reviewers in terms of feedbacks and suggestions, it is reasonable to think about measures to share payoffs between authors and reviewers—e.g., reviewers' names included in the published article—although currently not explored in scientific journals. The alignment of authors' and reviewers' 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 approximatively 75 minutes. 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. The English translation of the instructions and the questionnaire is included in the Appendix.

3 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; 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,B and Tab. 1). The amounts exchanged in the first three periods of the game, when reviewers 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.[5]

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

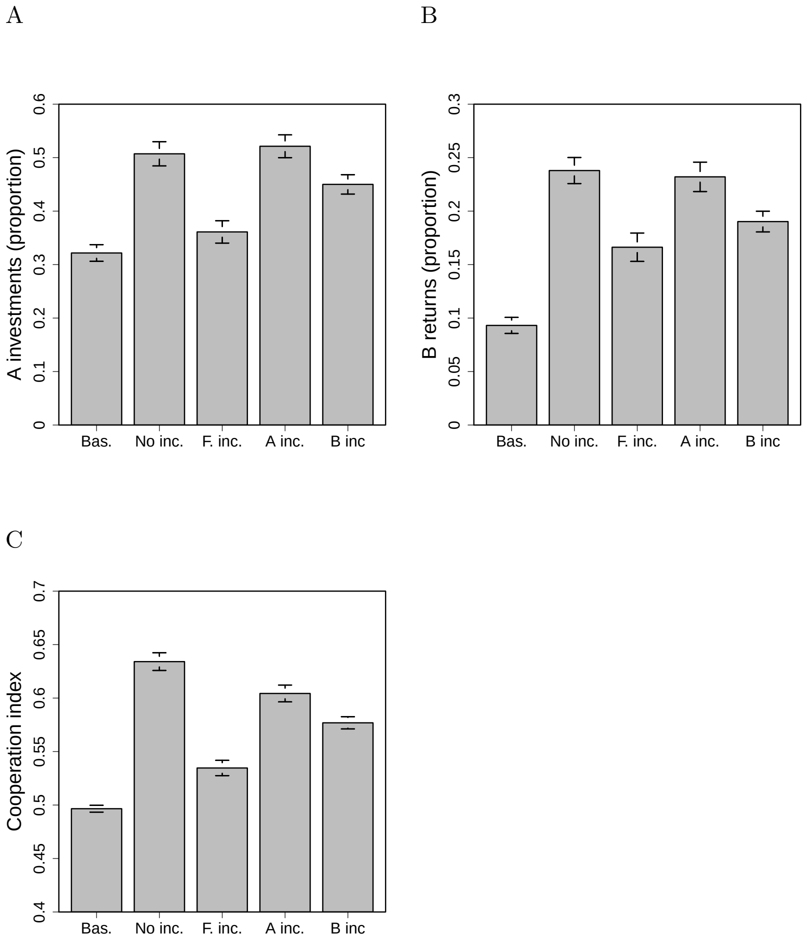

Fig. 1. Average investments (A), returns (B) and cooperation index (C) by treatment with standard error bars. Investments are represented in proportion of A endowment (10 MU). Returns are expressed as proportion of the overall B endowment (3× *amount received* +10 MU). The Cooperation Index varied from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

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 (Tab. 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*

	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 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.	
		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	No inc.	105.0	0.000						
(absolute)	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	No inc.	136.0	0.000						
(prop.)	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

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

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

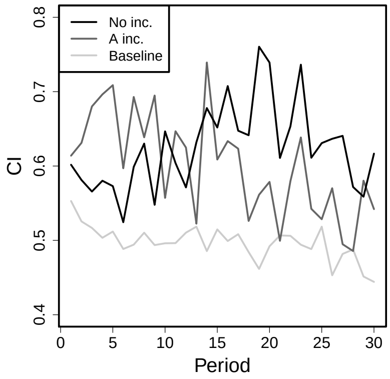

Fig. 2. Cooperation index dynamics in *No incentive* and *A incentive*. The *Baseline* curve at the bottom of the figure is inserted as reference.

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 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*, reviewers had no incentive to cooperate with A players. This indicated that material incentives, rather than guaranteeing higher reviewers' 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 treatment, A players largely used reviewers' 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

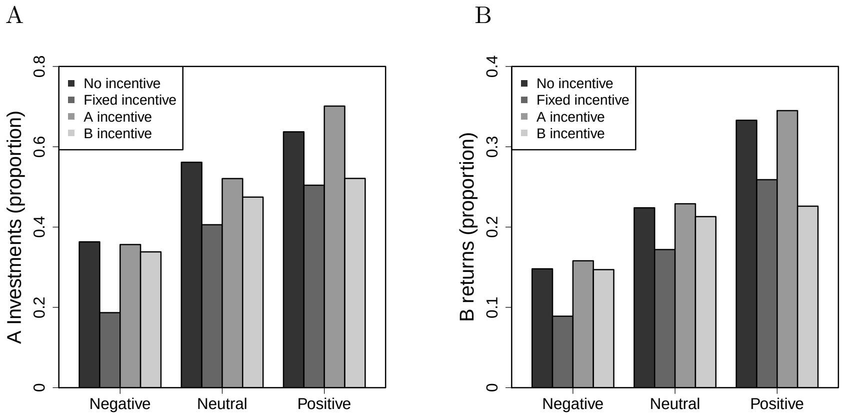

Fig. 3. Review effects. (A) Average investment proportion by review. (B) Average return proportions required by reviewers to award negative, neutral or positive evaluations.

that is crucial to understand cooperation differences among treatments. In *No incentive* and *A incentive* reviewers 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 (see Supporting Information). 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 reviewers 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* ($W = 376$, $p = 0.423$).

4 Discussion

Our findings show that the most effective peer review scheme is the one currently in use where reviewers are not supported by material incentives. Its maintenance avoids that peer review undergoes a frame effect motivating also well disposed reviewers 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 pro-social 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 reviewers had incentives aligned with editors, was similarly productive, but less robust than the former. 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 behavior in many social interactions (Gintis 2000; Ostrom and James 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. Moreover, we showed that adding material incentives is difficult and in most cases deteriorates the present situation.

This 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, everyone knows that the value and the payoff of each publication embody comments, ideas and efforts by reviewers but are capitalized just by authors, as the former do not have any concrete reputational benefit from authors publications. Our suggestion here is that journals could improve the way reviewers' contributions are presently acknowledged by establishing symbolic awards for reviewers, including reviewers name 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 Authors gratefully acknowledge help provided in the design and realization of the experiment by Riccardo Boero, Vincent Buskens, Niccol Casnici and Marco Castellani, along with the Intra-European Fellowship Program of the European Union, GA-2009-236953. 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.

Bibliography

Alberts, B., Hanson, B. and Kelner, K. L.: 2008, Reviewing peer review, *Science* **321**, 15.

Almås, I., Cappelen, A. W., Sorensen, E. O. and Tungodden, B.: 2010, Fairness and the development of inequality acceptance, *Science* **328**, 1176–1178.

Berg, J., Dickhaut, J. and McCabe, K. A.: 1995, Trust, reciprocity and social history, *Games and Economic Behavior* **10**, 122–142.

Boero, R., Bravo, G., Castellani, M., Lagan, F. and Squazzoni, F.: 2009, Pillars of trust: An experimental study on reputation and its effects, *Sociological Research Online* **14**(5), 5.
URL: *¡http://www.socresonline.org.uk/14/5/5.html¿*

Bowles, S.: 2008, Policies designed for self-interested citizens may undermine the moral sentiments: Evidence from economic experiments, *Science* **320**, 1605–1609.

Fehr, E. and 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 Economi* **10**, 171–178.

Frey, B. and Jegen, R.: 2001, Motivation crowding theory, *Journal of Economic Surveys* **15**, 589–611.

Gintis, H.: 2000, Strong reciprocity and human sociality, *Journal of Theoretical Biology* **206**, 169–179.

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* K. Kremer and V. Macho (eds), *Forschung und Wissenschaftliches Rechnen 2003*, Ges. für Wiss. Datenverarbeitung, Göttingen, pp. 79–93.

Hauser, M. and Fehr, E.: 2007, An incentive solution to the peer review problem, *PLoS Biology* **5**, e107.

Heyman, J. and Ariely, D.: 2004, Effort for payment: A tale of two markets, *Psychological Science* **15**(11), 787–793.

Keser, C.: 2003, Experimental games for the design of reputation management systems, *IBM System Journal* **42**, 498–506.

Nowak, M. A., Page, K. M. and Sigmund, K.: 2000, Fairness versus reason in the ultimatum game, *Science* **289**, 1773–1775.

Ortmann, A., Fitzgerald, J. and Boeing, C.: 2000, Trust, reciprocity, and social history: A re-examination, *Experimental Economics* **3**, 81–100.

Ostrom, E. and James (eds): 2003, *Trust & Reciprocity: Interdisciplinary Lessons from Experimental Research*, Russel Sage Foundation, New York.

Rabin, M.: 1993, Incorporating fairness into game theory and economics, *American Economic Review* **83**, 1281–1302.

Sigmund, K.: 2010, *The Calculus of Selfishness*, Princeton University Press, Princeton.

Vohs, K. D., Mead, N. L. and Goode, M. R.: 2006, The psychological consequences of money, *Science* **314**, 1154–1156.