Industriens Utredningsinstitut
THE INDUSTRIAL INSTITUTE FOR ECONOMIC AND SOCIAL RESEARCH

A list of Working Papers on the last pages

No. 309, 1991

The Perils of Peer Review in Economics and Other Sciences

by

Stefan Fölster

November 1991

Postadress
Box 5501
114 85 Stockholm

Gatuadress
Industrihuset
Storgatan 19

Telefon
08-783 80 00
Telefax
08-661 79 69

Bankgiro
446-9995

Postgiro
19 15 92-5

11.89 10.000

THE PERILS OF PEER REVIEW

IN ECONOMICS AND OTHER SCIENCES

by

Stefan Fölster

The Industrial Institute for Economic

and Social Research

Storgatan 19

114 85 Stockholm

Abstract

The quality of researchers' work in economics and other sciences is generally evaluated through a system of peer review. In an experimental test it is shown here that the peer review system can be very inefficient by creating a bias towards incremental development of existing methods and against exploration of new methods. This can occur when researchers have strategic uncertainty about the extent to which referees reject methods that they do not use themselves. If such a risk is perceived then researchers have an incentive to respond in the same manner, also rejecting methods they do not use themselves. As a consequence few researchers may dare to try new methods. The experiment also shows that the bias generated by peer review can be alleviated by shifting some quality evaluation to non-researchers, even if these are poor at discerning quality.

All of these problems could cause referees to make biased decisions in favor of the method they use themselves or other established methods. In the following this is called "inherent method bias". Inherent method bias does not depend on what other researchers do.

In contrast to inherent method bias there may be another problem which can be called "induced method bias". This is caused by the fact that referee A who suspects the existence of inherent method bias has an incentive to accept papers that use his own method and reject those that use a different method. This affects the composition of referees in the future such that more referees use the same method that referee A is using. As a consequence referee A faces a reduced risk of having a paper rejected by a referee with inherent method bias.

Even in the absence of inherent method bias strategic uncertainty can conceivably spark off induced method bias. The game bears some resemblance to the one presented by Van Huyck et al. (1990) in which it is shown experimentally that strategic uncertainty can lead human subjects to settle on Nash equilibria that are far from the pareto optimal equilibrium. In that game, as in the peer review game analyzed here, standard equilibrium refinements do not reduce the set of equilibria. Therefore the experimental method provides a useful way of studying equilibrium selection.

In the experiments reported here each subject represents a researcher and submits one "paper" to the referees in each time period. A paper consists of a choice of method. If the referees decide to publish a researcher's paper he becomes a referee in the next period. The number of publications, and thereby referees, is held constant by a rationing mechanism. Subjects are rewarded for each publication. The size of the reward increases with the social value of all papers published in the current time period. In some experimental treatments researchers have varying degrees of "competence" which determines the social value of the papers they submit.

From the viewpoint of current referees the pareto-optimal Nash equilibrium is that referees publish each other in a tacit coalition. If this coalition becomes unstable due for example to strategic uncertainty, an alternative could be to try to establish a stable coalition using the choice of method as a focal signal.

The experiment abstracts in some respects from real peer review systems. One abstraction is that there is no editor in the experiments. An editor could conceivably mitigate any method bias or coalition formation by overriding the referees' judgement. If the editor is a researcher himself however he would be subject to the same incentives as other researchers and thus make method biased judgements when other researchers do. In fact one can even think of one of the referees in the experiment as the editor, thus assuming that the editor is a researcher himself and personally evaluates merely with a smaller fraction of the submitted papers.

A related point is that real editors often select referees that use the same method as the submitted paper. In the experiment however papers are distributed randomly to referees. This makes it more difficult in the experiment for a referee to implement method bias since there is a smaller chance that a paper submitted to the referee uses the same method.

A further assumption in the experiment is that there are diminishing marginal social and private returns to each method. If one assumed that all methods have increasing marginal returns, then method bias would be likely to lead to the same outcome that unbiased referees would choose. It would seem implausible however that all methods have increasing returns forever. Many methods presumably reach a stage of decreasing returns sooner or later. The experiment can therefore be seen as a test of whether method bias can persist even in situations when both individual and social returns are negatively affected.

The experiments show that peer review does succeed in selecting researchers with higher than average levels of competence. Nevertheless induced method bias can lead to a drastic deterioration in the effectiveness of the peer review system even when no inherent method bias actually exists. Effectiveness can sometimes be improved by letting some papers be refereed by referees that are not researchers themselves and that choose papers among non-established methods, even when these external referees are poor at selecting quality.

Even though this paper concentrates on peer review in sciences, the results could be generalized to careers in firms and public organisations. In most organisations career advancement depends crucially on how one is judged by ones' peers. A more or less effective "business culture" in a firm may for example be the outcome of inherent or induced method bias. Interestingly, firm managers often express an awareness of the advantages of filling some high level positions with outsiders.

Section 2 describes the model. Section 3 presents the design of the experiment, and section 4 the results.

2 The model

To focus the analysis the following game can be defined. Assume that there are n competing researchers. Each researcher submits one paper in each of a sequence of discrete time periods. Researchers are rewarded for having papers accepted for publication by referees. The reward is 0 if the paper is not accepted for publication.

In case of publication the reward, termed P, contains two components. One is that a published paper increases the chances

of getting a better paid position. For simplicity this is modelled as a fixed reward A per publication.

In addition researchers are also affected by the extent to which they generate insights that society finds useful. This determines the amount of resources society is willing to invest in the science. Greater investment by society raises demand for scientists and thereby wages even for those already in secure positions. Further scientists can often work as consultants and sell not only their own research results but the professions' collective results. Again scientists earnings increase in proportion to results that society finds useful.

Suppose that the benefits to society are q (quality) for any particular paper and Q is the total social benefit of insights generated in the current period t. Then the rewards to publishing are

(1) $P = A + c\,Q$

Researchers choose among a finite set of methods. Here the term method is defined broadly as any category of research activity that is commonly distinguishable and the fact that it is distinguishable is common knowledge.

It is assumed that researchers can freely and costlessly switch method. In this regard the experiments abstract from real life where some choices of method involve long-term investments in specific human capital.

If one assumed that all methods exhibit increasing marginal returns then method bias, if it occurs, may imply the same outcome that rational referees would have chosen anyway. It seems unlikely, however, that all methods always have increasing marginal returns. A more reasonable assumption would be that some methods exhibit increasing marginal returns some of the time. In the experiment, however, the aim is to test whether

method bias can persist even when it involves an unambiguous social and private loss. Therefore it is assumed that there are diminishing marginal returns to research for all methods. Let $R_{t,m}$ be the sum of all q using a certain method from the beginning of the game to the current time period. Then $q = f(R_{t,m})$ is decreasing in $R_{t,m}$. Papers that are not published are assumed to generate no q since the results are not spread.

The players have complete information about the payoff function and strategy space and know that the payoff function and strategy space are common knowledge. They also know the identity of those that publish and the q achieved. The referees know the identity of those they referee but researchers do not know by who they are being refereed.

In most sciences publication of papers is rationed. In the short term there exist a certain number of journals with fixed budgets and space. Presumably journal space is roughly related to the number of scientists with permanent positions in such a way that publication can be used as a criteria for eligibility for permanent positions.

In many sciences there rankings of journals allowing papers that are not published in a high-ranking journal to be published in a low-ranking one. Even in these sciences, however, the number of journals in which publication by common agreement serves as criteria for achievement is limited. Therefore it is assumed here that the journal space is fixed at p published papers in each time period with $p < n$.

A central element of the peer review process is that those who publish are also chosen as referees. This is modelled here in a somewhat stylized form by assuming that those who publish in period t are referees in period t+1. Thus there are p referees.

An editor could conceivably have a large impact on the selection of papers. In practice, however, editors tend to get involved

only with evaluation of a small fraction of the submitted papers. Also, editors are usually researchers themselves and are selected from the group of current referees. Within the framework of the model one can therefore think of the editor as being any one of the referees. The editor's administrative functions are modelled as random selection rules. All submitted papers are split randomly among the referees. The referees can then recommend publication or rejection of each paper they receive. Papers are randomly chosen for publication among those that referees recommend. If there are fewer than p recommendations papers are also chosen randomly from non-recommended submissions.

If this game lasted only one period the payoff for a player would not vary with the choice made as a referee. The choice of method could have an effect but players only have their prior beliefs to determine which method could give a greater chance of acceptance. Thus in the period game any outcome could be a Nash equilibrium.

In the repeated game however referees have incentives to form a coalition. If referees could explicitly coordinate their actions, the decision problem would be trivial. The referees would agree to publish each other. Unlike games with incentive problems the first best outcome for the referees is then a self-enforcing Nash equilibrium point in the repeated game.

In the peer review system however explicit agreements are both illegal and face practical difficulties. Since the referees are anonymous the agreement cannot be enforced. Also, since papers are randomly assigned to referees the agreement would have to encompass a large number of potential referees.

Nevertheless the gains to forming coalitions are considerable so that incentives are strong to attempt to form a tacit coalition. In this tacit coordination problem the central questions is which

signals serve as focal points for the coalition.[4]

The most obvious information that can be used to solve the tacit coordination problem is who is currently referee. Given that the tacit coalition among current referees holds, referees have incentives to submit papers using methods that maximize Q. This is the payoff dominant equilibrium for the referees in the repeated game.

When players have strategic uncertainty, however, it is not certain that the payoff dominant equilibrium will result. Van Huyck, Battalio and Beil (1990) give an example of game with a large number of self-enforcing Pareto ranked Nash equilibria. Due to strategic uncertainty however players usually end up with the least pareto efficient equilibrium solution after a number of repetitions of the game.

In the peer review game strategic uncertainty is an uncertainty about whether other players will break the coalition. If referee A suspects that a referee in the current coalition is less reliable than a researcher outside of the coalition then referee A has a motive for rejecting current referees and accepting outsiders.

This bears some resemblance to the gunman game where two gunmen face each other and each has incentives to shoot first if he believes the other will shoot. Whether the other shoots depends, however, on what he believes about the first.

When the current coalition crumbles or is feared to crumble then players may rationally attempt to establish other coalitions with the help of available focal signals. The choice of method could serve as such an alternative focal signal. A referee may therefore perceive a researcher using the same method as a

---

[4] The role of learned focal points as solutions to coordination games is analysed for example in Shelling (1960) and Crawford and Haller (1990).

"safer" coalition memeber than one of the existing referees.

As soon as one referee begins to oust current referees and replaces them with researchers using the same method then other referees may feel compelled to either switch to that method or try to establish a coalition around their own method. Thus induced method bias can conceivably arise even if no referee has inherent method bias. If some referees have inherent method bias one would expect other referees to respond with induced method bias.

Varying levels of capacity

The main purpose of the referee system is presumably to ensure that the most competent researchers are published and become referees. It would therefore be unjust to test peer review without considering its ability to detect competence. In an extension of the simple model above competence is modelled by randomly varying the level of competence C prior to the start of the game. It is assumed that the q a researcher achieves is a function of R and the C assigned to him.

(2) $q = f(R_{t,m}, C)$

The introduction of C in the model gives players additional information that can be used as a focal signal. A stable coalition with high competence as the focal signal is clearly desirable from a social point of view. If the distribution of competence was known to all players then such a coalition could probably maintain stability.

Assuming a known distribution of competence is not a particularly plausible assumption however. In the real world the distribution of competence is constantly changing as new generations of researchers enter. Also, since competence is not well-defined in the real world, researchers presumably have some degree of uncertainty about their own as well as others' competence. In

the model it is therefore assumed that players have no forehand knowledge about the distribution of competence, but they observe competence of those that publish. Under that assumption a tacit coalition of high-competence researchers may not be stable. Those referees in the coalition that at any time have the lowest C face a high risk of being ousted and therefore have incentives to build coalitions using other focal signals.

Avoiding method bias

The principal agent dilemma that permits method bias to occur is that researchers are not directly rewarded for the q they produce. A natural solution would therefore be to reward high q rather than publication. This solution is in fact attempted in some instances where financing and even employment decisions are not taken by peers but by interdisciplinary committees or even public authorities.

Rewarding high q will only work unambiguously, however, when those non-researchers deciding on who to reward have the competence to judge quality. An interesting question is therefore whether method bias can be avoided even when non-researcher referees cannot distinguish between high and low quality. This is modelled by assuming that one of the referees is replaced by a mechanism that randomly accepts a paper among those that do not use an established method, viz. a method that has been used by a published paper in the current period.

Non-researcher referees make it harder to establish a tacit coalition based on the choice of method. Thus method bias may be reduced. At the same time the efficiency of the system is also hampered by the fact that the external referees do not distinguish among papers whith high and low q or researchers with high and low C. The interesting experimental question is whether method bias can be significantly reduced by introducing such a small fraction of non-researcher referees that selection of competence and quality is not seriously affected.

3 The experiment

Table 1 outlines the design of the experiments reported in this paper. Between 14 and 17 subjects participated in each experimental treatment, playing the peer review game for 25 periods. In each period subjects chose a method, submitted a "paper" to the referees, and had it accepted or rejected. A paper consisted of a choice of method and the subject's identity number.

In total there were 14 experimental treatments. The first six, labelled 1.1 to 1.6, represent the basic model with different numbers of referees. Treatments 2.1 to 2.4 allow for capacity variation. Finally, treatments 3.1 to 3.4 test the effects of introducing non-researcher referees.

In each treatment instructions were both handed out and read aloud to ensure that the description of the game was common information. After reading the instructions, but before the experiment began, the students filled out a questionnaire to determine that they understood the payoff formula. In two treatments a subject did not respond correctly and the instructions were reread. Subjects were not allowed to communicate directly either before or during the game.

After each period the following information was displayed on a projected computer display: The identity number of each researcher that published, the method used, the q achieved, the capacity C implied by the q for every published paper and the sum of papers published for each method. All the displayed information was recorded. In addition method choices made by subjects who did not publish were recorded.

Subjects increase their payoff by P after every period of the

game in which they publish. P was paid in Swedish kronor (where one US dollar roughly equals 6 kronor). P is calculated as follows.

(3) $P = 2 + Q/p$

The sum of q in the current period, Q, depends on the number of published papers per period, p. To prevent P from varying with the number of published papers permitted, Q is divided by p in calculating P. The quality q of the paper is calculated as follows.

(4) $q = 2 - 0.1$ (number of previously published papers using same method)

In each treatment publications in the first period were randomly assigned to p subjects. The assigned publications also used assigned methods. As shown in table 1 the number of assigned initial methods was given one smaller and one larger value for each type of treatment. For example, in treatment 1.1 the five subjects who published initially were assigned one of two different methods, while in treatment 1.2 those five subjects were assigned one of four different methods.

The submitted papers were divided as evenly as possible among referees. Each referee could then accept or reject any of the papers submitted to him. If the total number of accepted papers exceeded the number of referees some randomly selected papers were rejected. If the number of referees exceeded the number of accepted papers then additional papers were randomly accepted. All researchers who published papers in one period became referees in the next period.

In experiments 1.1 to 1.6 three levels of p were tested, namely 3, 5, and 8 published papers. In experiments 1.1, 1.2, 1.3 and 1.4 method bias was artificially induced by replacing one randomly chosen referee's judgement with acceptance of one paper

submitted to that referee using the same method as the referee used. All other papers submitted to that referee were rejected.

In experiment 2.1 to 3.4, where competence is allowed to vary, q is calculated as

(5) $q = 2 + 2\,C - 0.1$ (number of previously published papers using same method)

C was distributed from 1 to n such that one subject had a C of 1, another subject a C of 2 and so forth. Assignment of C to subjects was random. The distribution of C was not revealed to subjects.

In experiments 3.1 to 3.4 one randomly chosen referee in every time period was not permitted to judge papers submitted to him. Instead one of the papers submitted to that referee using a method not published in the previous period was randomly chosen and accepted. The other papers submitted to that referee were rejected.

The subjects were undergraduate students attending Stockholm university. They were recruited from sophomore economics courses. A total of 208 students participated in the experiments.

4 Experimental results

Table 2 reports the experimental results for treatments 1.1 to 1.6. For each time period the table shows the number of methods used in published papers and the number of referees that are replaced by new referees.

When the number of replaced referees is small the coalition is

stable. This means that referees are accepting papers by other referees and rejecting papers submitted by non--referees. In treatments 1.1 and 1.2 there are fairly long periods of stability. In treatments 1.3 and 1.4 where 8 referees were used there is considerably less stability. In contrast treatments 1.5 and 1.6 with only three referees are nearly always stable. Thus stability clearly decreases as the number of referees increases.

The striking results in table 2 are that when a coalition is stable referees choose a wider spectrum of methods. When a coalition breaks down however the choice of method appears to become the alternative organising principle. Immediately after break-downs the number of methods used decreases rapidly, often to the extent that all use the same method. At that point the coalition often becomes stable again.

In four of the treatments method bias was artificially induced in period 20 by faking one referee's decision so that he replaced a current referee with a researcher using the same method. This simulates the case where a referee has inherent method bias. In the treatments with 5 referees this clearly sparks off induced method bias while in the treatment with 3 referees the effects are minimal.

Table 3 shows the pattern for treatment 2.1 to 2.4 in which competence varies. Compared to treatment 1.1 to 1.4 there are fewer periods with stable coalitions. The referees with relatively low C often reject referees and replace them with referees using the same method. As a result method bias seems more pronounced than in treatments without capacity distribution.

Table 4 shows the pattern for treatments 3.1 to 3.4 where one referee has been replaced by a non-researcher referee. Here coalitions rarely succeed in maintaining stability and the number of methods used is generally greater than in treatments 2.1 to 2.4. Thus it appears that the use of an external referee has inhibited method bias and that this effect outweighs the poorer

selection of competence.

Table 5 summarizes the results by showing the "scientific progress" in terms of the average Q generated per time period relative to the maximum that could have been achieved. Treatments 1.1 to 1.6 show that induced method bias leads to a greater deterioration of efficiency as the number of referees increases. Treatments 2.1 to 3.4 show that scientific progress was significantly higher in the treatment with non-researcher referees.

Table 5 also shows to what extent the peer review system succeeds in selecting those with high competence. In treatments 2.1 to 2.4 the average competence of referees is clearly higher than for non-referees. Thus peer review does succeed in selecting talent but achieves a low degree of effectiveness nevertheless.

In treatments 3.1 to 3.4 where a non-researcher referee is introduced competence selection is somewhat poorer. Yet efficiency is higher due to reduced method bias.

5 Conclusion

The experimental results indicate that method bias in the peer review system can seriously undermine the effectiveness of a science. This problem may be alleviated by inviting som non-researcher referees to to participate in evaluating submissions to scientific journals.

In some instances this problem has been recognized and has led to a shift of management responsibility in scientific development. In Britian for example dissatisfaction with the relevance of research in natural sciences has led to the establishment of the Advisory Council on Science and Technology (ACOST) which under guidance of managers from high-tech

industries rather than scientists exerts an influence on the distribution of funds for science.

Some journals insist that referees do not learn the identity of those submitting papers. The results presented here tend to suggest that this would indeed inhibit coalitions of current referees. The cost of this practice could be, however, to encourage the choice of method as a focal signal, leading to more pronounced method bias.

Even though this paper has concentrated on peer review in sciences, similar processes presumably steer careers in firms and public organisations. A more or less effective "business culture" in a firm may for example be the outcome of inherent or induced method bias. Interestingly firm managers often express an awareness of the advantages of filling some high level positions with outsiders.

References

Blaug, Mark, The methodology of economics or how economists explain. Cambridge: Cambridge University Press, 1980.

Colander, David and A.W. Coats, The spread of economic ideas. Cambridge: Cambridge University Press, 1989.

Crawford, V. P. and Haller, H., Learning how to cooperate: optimal play in repeated coordination games, Econometrica, 58, 1990, 571-595.

Kuhn, T.S. The structure of scientific revolutions. Chicago: University of Chicago Press, 1962.

Lakatos, I., The methodology of scientific research programmes. Philosophical papers. J Worrall and G. Currie (eds.). Cambridge: Cambridge University Press, vols.1,2.

OECD, Evaluation of research: A selection of current practices. Paris, 1987.

Popper, K. The logic of scientific discovery. New York: Harper Torchbooks, 1959.

Schelling, T. The strategy of conflict. Cambridge, MA: Harvard University Press, 1990.

Sherif, M. and Sherif, C. Social psychology, New York: Harper & Row, 1969, 208-209.

Tsukahara, S. and Yamada, K. A note on the time lag between the life cycle of a discipline and resource allocation in Japan. Research Policy, 1982, 133-140.

Tversky and Kahneman, Cognitive psychology, 5, 1973, 207-232.

Van Huyck, J., Battalio, R. and Beil, R. Tacit coordination games, strategic uncertainty, and coordination failure, American Economic Review, 1990, 80, 235-248.

Table 1. Experimental design

Treatment	Size	Number of referees	Number of initial methods
1. Basic experiment			
1.1	14	5	2
1.2	15	5	4
1.3	14	8	2
1.4	16	8	5
1.5	15	3	1
1.6	14	3	3
2. Varying competence			
2.1	16	5	2
2.2	14	5	4
2.3	15	8	2
2.4	15	8	5
3. Varying competence and non-researcher referees			
3.1	15	5	2
3.2	14	5	4
3.3	17	5	2
3.4	14	5	4

Table 2 Experimental results terms of number of methods published and new referees in each time period.

	Treatments					
Period	1.1	1.2	1.3	1.4	1.5	1.6
1	2/5	4/5	2/8	5/8	1/3	3/3
2	3/2	3/3	5/4	7/3	1/0	3/1
3	2/2	4/0	6/5	6/5	2/0	2/1
4	2/0	5/2	5/5	2/2	3/0	2/0
5	2/0	3/2	4/5	1/0	3/0	3/0
6	3/0	3/1	3/4	3/0	3/0	3/0
7	4/0	2/0	3/5	4/1	3/0	1/0
8	4/1	2/0	2/3	4/1	2/0	3/0
9	2/3	2/0	2/3	5/1	3/0	3/0
10	2/0	3/0	1/2	5/2	* 3/1	* 2/1
11	2/0	3/0	1/0	5/3	2/1	2/0
12	3/0	3/0	2/1	5/5	2/0	3/0
13	3/0	4/0	1/4	2/5	3/0	3/0
14	3/0	3/0	1/1	2/6	3/0	3/0
15	4/0	3/0	1/0	2/4	3/0	2/0
16	4/0	4/0	2/0	1/3		
17	3/0	4/0	3/1	1/0		
18	3/0	3/0	3/1	1/0		
19	4/0	3/0	3/2	1/0		
20	* 4/1	* 3/1	3/3	2/0		
21	3/3	3/2	3/5	2/1		
22	3/2	2/2	3/6	2/0		
23	2/2	2/3	3/3	2/0		
24	1/2	2/0	2/3	2/1		
25	1/0	2/0	2/2	2/0		

* Artificially induced method bias

Table 3 Experimental results in terms of number of methods published and new referees.

	Treatments			
Period	2.1	2.2	2.3	2.4
1	2/5	4/5	2/8	5/8
2	4/3	5/2	6/3	7/3
3	4/2	5/3	6/4	4/3
4	2/2	5/2	7/5	2/4
5	2/2	3/2	5/5	1/5
6	1/1	2/0	3/3	1/0
7	1/0	3/0	2/3	2/0
8	3/1	4/0	2/1	4/0
9	2/3	3/1	3/1	5/1
10	1/0	3/3	3/2	4/2
11	1/0	2/3	2/0	4/5
12	2/0	2/3	2/1	4/3
13	2/0	1/3	1/0	2/2
14	3/0	1/0	2/0	1/0
15	3/1	1/0	4/1	2/0
16	3/3	2/0	5/1	2/1
17	3/2	3/0	5/3	2/1
18	2/3	3/1	3/5	2/1
19	2/2	3/1	3/2	4/0
20	2/0	3/1	3/3	4/1
21	2/0	2/2	2/1	5/2
22	1/0	2/3	2/2	4/3
23	2/0	2/0	2/0	4/3
24	3/1	1/0	2/1	3/3
25	3/2	2/0	2/3	2/1

Table 4 Experimental results in terms of number of methods published and new referees.

	Treatments			
Period	3.1	3.2	3.3	3.4
1	2/5	4/5	2/8	5/8
2	4/3	4/3	6/3	8/3
3	3/3	5/3	6/3	6/3
4	4/2	4/2	5/2	4/2
5	5/1	3/1	3/2	3/2
6	4/2	3/2	3/2	3/0
7	4/1	3/1	2/1	4/1
8	3/3	3/3	3/2	5/1
9	2/3	2/1	4/2	5/3
10	3/2	3/0	4/4	6/2
11	3/1	4/1	6/5	4/3
12	3/4	4/2	5/2	6/4
13	4/3	5/3	6/0	5/2
14	3/2	5/2	6/1	3/2
15	3/1	4/2	4/2	3/1
16	2/1	3/1	4/1	3/1
17	3/1	2/2	3/2	3/1
18	4/0	3/1	3/2	3/0
19	5/2	3/1	4/3	4/1
20	4/3	4/2	5/3	4/1
21	3/2	4/3	7/5	5/2
22	4/3	4/2	3/3	6/4
23	3/2	4/1	3/2	6/3
24	2/2	3/1	3/1	5/3
25	3/2	3/1	4/2	5/2

Table 5 Summary of experimental results. Relative Q: Q as a fraction of the maximum possible in each time period averaged over all time periods (excluding periods with artificially induced method bias). Competence of referees relative to average competence.

Treatment	relative Q		Competence of referees relative to average
1. Basic experiment			
1.1 + 1.2	0.76		-
1.3 + 1.4	0.55	* (1.1+1.2)	-
1.5 + 1.6	0.92	* (1.1+1.2)	-
2. Varying competence			
2.1 + 2.2	0.67		1.4
2.3 + 2.4	0.43		1.3
3. Varying competence and non-researcher referees			
3.1 + 3.2	0.81	* (2.1+2.2)	1.2
3.3 + 3.4	0.65	* (2.3+2.4)	1.2

* (xx) : significantly different from experiments xx at 95% level.