

Can Peer Review Be Trusted?

The UK's 'Research Assessment Exercise' on Trial.

Tom Coupé*

ECARES

Université Libre de Bruxelles

Abstract

In the United Kingdom, the amount of public research money a university department gets is (partially) determined by the performance-rating it gets from a panel of peers in a 'Research Assessment Exercise'. In this paper, we will try to find out whether those departments that had a representative in the assessment-panel did get a higher rating than can be expected on the basis of some objective performance-measures. While OLS-regressions show some evidence of such an insider bias, we do not find that changes in panel-composition are related to changes in the ratings.

* I thank Mathias Dewatripont, Abdul Noury, Frederic Warzynski, participants of the PAI-workshop in Brussels and the Enter-Jamboree (London) for helpful comments, Rosalind Keir at HEFCE for sending me the HEFCE-circulars and the Belgian federal government for financial support (Pôles d'Attraction Interuniversitaires P4/28).

Introduction

As competition has more explicitly entered the scientific world over the last twenty years, quality-control mechanisms have become more and more important. For example, several universities now use quantitative indicators such as articles published to decide about tenure or wage-rises. Similarly, some governments fund their universities according to the universities' performance. And more strikingly: even respected scientific journals publish 'league'-tables of departments (f.e. Graves et al, AER 1982 and Dusansky and Vernon, JEP, 1998).

In these mechanisms, peer review has conquered a central place: only fellow-scientists are considered informed enough to be able to judge about the quality of scientific work. Several studies, however, have made clear that this peer review process is sometimes flawed: even scientists have difficulties to be completely objective when judging their colleagues or their competitors.

Nevertheless, the UK councils that are responsible for the funding of research use and have used such a peer-review technique to assign to each department a number between 1 and 7 which should reflect the departments' research-performance. This rating is given by a panel (each discipline has its own panel), consisting of professors (members of UK universities), scientists working in industry and representatives of some interested third parties (such as charitable organizations). In this paper, we will try to find out whether self-interest did influence the rating behavior of the panel-members¹. In other words: did being represented in the panel increase a department 's research rating?

There are several reasons to study the reliability of this specific mechanism of peer review. First, governments that have decided in favor of supporting public universities struggle with the question of how to fund these institutions in an optimal way. While supporting the 'teaching function' of universities is, in general, done by multiplying the number of students by a certain amount of money, the funding of research seems to be more awkward. Some countries make it completely dependent on the number of

¹ Self-interested is here (and later) used in the sense of interested in short-run monetary gains. Of course, even 'being honest' can enter into the utility function, in which case honest behavior would also be self-interested.

students, some create special councils to divide the money on a project-by-project basis and yet others combine such systems. The UK system of making the distribution of about half of the research budget dependent on the results of a so-called “Research Assessment Exercises (RAE)” is unique in the world and hence, “has generated considerable interest worldwide (Johnes, 1994)”.

Further, a recent study for the European Parliament² (STOA,1998), about the assessment of projects that submit for EC grants, states: ”At present the principles exclude the participation [in assessment panels] of individuals with a specific interest in programs. This has the advantage that it ensures the independence of the Panels, but it also means that they lack some of the advantages which can accrue from the inclusion of someone with an 'insider' view ”. Hence, using these UK data we can try to test whether there is a necessary trade-off between the extra information that could bring ‘interested parties’ and the risk of insider bias.

Finally, organizing such an RAE is not that cheap: the cost of the 1992 RAE was about £13.5 million (0.4% of funds to be distributed-Davies(M2/1994)). Hence, this is one more reason why it is important to know the ‘collusion-proofness’ of such a system³.

Using data for both the 1996 RAE and the 1992 RAE, we will show that even after controlling for some objective factors like the number of staff, the research income, the number of students and the number of publications, represented departments have significantly higher coefficients than departments that were not represented. However, if we control for omitted variables through a regression in first differences, we do not find that changes in ratings are related to changes in panel composition.

² <http://www.europarl.eu.int/dg4/stoa/en/publi/167406/chap3.htm>

³ Other criticisms can be found in Trow(1998) or Flemming(1991). Johnes (1994), however, notes:” Nevertheless, the UK attempts at research assessment are of considerable interest, not least because (after teething troubles in 1986) they have arguably demonstrated that performance indicators of this kind can attain a tolerable degree of legitimacy in the eyes of the assessed.”

1) The case

a) The making of the rating.

In these RAE's, professors, scientists working in industry and representatives of interested third parties (such as charitable organizations) form panels -each discipline has its own panel- and then jointly decide about a rating, ranging between 1 and 7 in 1996, which should reflect the research performance of the respective department and that is used to distribute the research budgets.

For example, for 1998-99, quality-related funding was £804 million. To split up this amount over the 69 disciplines, a volume⁴ measure for each discipline is calculated and then multiplied by a weight⁵. A discipline's part in the sum of the numbers thus calculated corresponds to its part in the £804 million. The distribution over departments within the discipline is then organized in a similar way, but now the weights of the volume are the RAE-ratings. The rating thus plays a crucial role in the amount of money a department (or the university) will receive.

The ratings are given by panels, so the first step is nominating the panel members.

Choosing panel members⁶

Panel members are chosen in the following way:

- Following consultation with the sector in November 1994 some 1,000 bodies (subject associations, learned societies, professional bodies and organizations representing users of research) were invited to nominate members of assessment panels for the 1996 RAE and to comment upon the structure and subject coverage of the panels.

⁴ The volume measure consists of

- research active academic staff - 1 x number of full-time equivalent (FTE) research active academic staff funded from general funds in departments rated 3b or above, selected for assessment in the RAE. It is up to the institution to decide which staff to enter in the RAE.
- research assistants - 0.1 x number of FTE research assistants.
- research fellows - 0.1 x number of FTE research fellows.
- postgraduate research students - 0.15 x number of weighted headcounts of postgraduate research students in their second and third years of full-time study, or third to sixth years of part-time study.
- research income from charities - 0.25/25,000 x average of last two years' income from charities. (Income from charities is divided by £25,000 (a researcher's average salary) to obtain a person equivalent).

(See http://www.niss.ac.uk/education/hefce/pub98/98_67.html#research)

⁵ The three cost weights are:

- high cost laboratory and clinical subjects: 1.7
- intermediate cost subjects: 1.3
- others: 1.0

(See http://www.niss.ac.uk/education/hefce/pub98/98_67.html#research)

- The panel chairmen, who had been appointed by the four funding bodies in late 1994, were then invited to use these nominations to recommend panel members. In doing so, they were asked to take the following into account:
 - a. The research experience of nominees and their standing in the research community.
 - b. Their own knowledge of the pattern of research and of active researchers in the subject area.
 - c. The need for continuity and the willingness of previous panel members to serve again (although no member could serve in more than three RAEs consecutively- on average, one third of panel members and about half of panel chairmen from 1992 served again in 1996.).
 - d. The need for users and commissioners of research - in commerce, industry and the public sector - to be represented (after taking into account the participation of specialist 'assessors').
 - e. The need for panel members to have a collective knowledge of research activity throughout the UK and in a range of institutions.

- **The chairmen and members of each panel were invited to participate as individuals, not as representatives of a particular group or interest.** They accepted responsibility for providing two formal outcomes: statements of criteria for assessment and working methods; and the assessment ratings.

- The above-mentioned 'specialist assessors' could give advice but had no voting rights. Final responsibility for recommending ratings to the funding bodies rested with the main panel members alone.

The managers' report of the conduct of the exercise notes further⁷:

- Around half [of the Chairs] had served as Chairs in the previous exercise; the remainder were appointed in the light of recommendations from the outgoing Chair, and almost all of these had served as panel members in 1992. The Chairs

⁶ Taken from RAE96 2/96, see http://www.niss.ac.uk/education/hefc/rae96/c2_96.html

were then asked to make recommendations for the membership of their panels from the nominations received in response to consultations with some 1,000 outside bodies - subject associations, learned societies, and others interested in the conduct of research from a primarily subject-focused viewpoint – and having regard to the following considerations⁸:

- a. eminence as individuals**
- b. coverage of subject field**
- c. sectoral/geographical balance.**

*Criteria of assessment*⁹

To assess the departments, the panel could choose their own criteria.

- During the summer and early autumn of 1995 the panels met to agree their statements of criteria and working methods. These were published in November 1995 alongside the formal invitation to make submissions. In all cases the published criteria statements showed clearly what pieces of evidence the panel would particularly look for in the submissions; how it would interpret these; and the relative weight that it would attach to different indicators. All of the panels indicated that they would place most weight upon the listed research outputs, and described their approach to assessing the quality of these. Many indicated a general hierarchy of esteem between different media of publication.
- The submissions were entered into a computer database held at HEFCE. They were printed and sent out to panels in two stages (lists of output and prose material were issued by May 20, and the numerical data only after checking by 1 July). They were also printed out and returned to HEIs, to be checked for significant processing or factual errors, during late May and June.

⁷ http://www.niss.ac.uk/education/hefc/rae96/c1_97.html

⁸ See also circular 24/92 and 15/92 for similar statements for the 92'-exercise.

⁹ See http://www.niss.ac.uk/education/hefc/rae96/c1_97.html

- The assessment panels met again between May and November 1996. Most had three or four meetings (and those with the larger workloads typically had a two-day meeting in October).

Safety procedures

To reduce the possibilities of self-serving behavior, safety mechanisms were provided for.

- To each panel was assigned a secretary by the funding councils: “Compliance with their criteria, and with the general criteria and guidance, was ensured and monitored through the role of panel secretaries as advisers and **as guardians of procedural integrity**. “ (see Managers Report)
- And Circular 5/92 about the organization of the 1992 assessment mentions that panel members cannot vote about their own universities: “**Members will be asked to declare their interests (including visiting lectureships) on appointment, and will be asked to leave the meeting for discussion of any submission in which they may have a personal interest**”.

b) On the potential crime (collusion - insider bias- the influence of panel-members on the ratings) and the motive (money).

Collusion in organizations

The incidence of collusion in organizations has been thoroughly studied. Tirole (1986) notes the following general characteristics:

- There is manipulation of information received by the principal.
- the object of the coalition is to benefit the members of this coalition and there is reciprocity within the coalition.

As examples of manipulating information, Tirole notes minor “thefts” and perquisites that are not reported or the creation of fictitious personnel. Concerning the benefits to the coalition, he states that in many cases, these are non-monetary benefits (mutual affection, respect) and covert: “observed collusive behaviors are only the tip of the iceberg. Anticipating that their members have incentives to collude, organizations can

and do set up incentive schemes that restrict the formation and thus the effect of collusion”.

Collusion in higher education

Collusion is not a phenomenon that is restricted to the business world. Indeed, even universities are sometimes suspected of such behavior. In the US for example, the Department of Justice (DOJ) filed a case against a group of universities that organized meetings in order to synchronize financial aid to students that applied simultaneously to several of the groups' members. The universities claimed that they needed this practice to ensure that aid was need-based – rather than merit-based. The DOJ, however, saw colluding colleges that, in unfair ways, tried to raise their revenues and decrease their aid (See Hoxby(1999), Salop and White(1991), Carlton et al (1995) and Masten(1995), Netz (1998)).

Similarly, in 1990, the UK government decided to let the universities tender for students but after the receiving of the bids, it appeared that the bids were almost uniform. One of the reasons for this failure is thought to be collusive behavior (Cave et al.,1992):“ *At the same time, the Committee of Vice-Chancellors and Principals encouraged an information exchange. Universities were asked to indicate anonymously in advance the size and the level of their bids. Finally, attempts were made at the level of individual cost centers and subject groups to oppose competitive tendering. Many of these relied upon moral suasion, but in some cases the possibility of sanctions were invoked. For example a professional organization responsible for accrediting degrees wrote to university departments indicating that any discounting of the guide price might provoke an examination of the quality of teaching provided, possibly leading to a withdrawal of accreditation.*”

One could claim that the managers of universities are not real scientists and that, as scientists are more trustworthy, collusion of universities does not say anything about the behavior of individual scientists. Scientists have indeed a good image with the general public: a recent survey in Germany for example found that “*The German public considers that university professors are intelligent, hard-working and honest, do a difficult job and making a positive contribution to society (Brookman/11/06/99).*” Unfortunately, the reality is a bit less optimistic, as witnessed

by several cases of scientific fraud (see f.e. Wibble, 1998). And that they collude too, is shown by several studies that study the effectiveness of the peer review system. Blank(1991) f.e. showed that double-blind peer reviewing leads to smaller acceptance rates and more critical reviews than single-blind peer review. McDowell and Amacher (1986) found that in-house-editorship increases the number of articles written by affiliated economists. Laband (1985) shows that it increases the length of the articles written by affiliated scholars.

Collusion in the RAE

Members are barred from the discussion about their own institutions, secretaries are there to prevent overt collusion, criteria for assessment are published. One could say that utmost care has been taken to prevent misuse by panel members (in itself an indication that scientists too have to be monitored and illustrating the point of Tirole mentioned above). However, criteria remained vague (not quantitative, see for example appendix for example of economics panel) and one cannot exclude tacit or hidden (for the secretary) collusion: a silent agreement can exist to scrutinize non-panel departments in more rigorous ways. Similarly, as panels could decide about the criteria they could choose those criteria on which their universities scored high.

So there's room for manipulating the information (the ratings) supplied to the principal (the funding agency). There's also a clear, monetary motive: the RAE-rating is important because it affects the research budgets. The amount for each discipline is fixed by the councils but the division of this amount over departments is determined on basis of the RAE-rating: f.e. a department with rating 5* (i.e. category 7) receives , ceteris paribus, 4.05 times the amount of a department with rating 3b (i.e. category 3) (Ball,1997). About 50 % of the UK's research budget is divided in this way, the other half being distributed by the 5 research councils (Johnes1997) so collusion could considerably increase the budget of a department.

Is there reciprocity? The agent (the rated department) cannot influence his own score as his representative is excluded from the discussions about the departments' score. Still, it is judged by representatives of other universities whom he'll have to judge next. So there's a clear scope for exchanging favors¹⁰!

¹⁰ One could say that non-represented universities can 'collude' with represented ones. Still, the scope for the exchange of favors is in that case much more limited.

2) The Trial

Data¹¹

For each department that received a rating in the RAE, we have the data (on number of staff, number of students and so on) that were available to the panel-members. In addition, we know the composition of each panel and of these we have the university where they were employed at the time of the RAE. This allows us to create a dummy that takes the value one if the university (department) has a representative in the panel and zero otherwise. Professors that are affiliated to foreign universities, scientists that work in industry and assessors (which are mostly representatives of non-university institutions and which are only advisers) are left outside the analysis.

Empirical results

The first question is whether the ratings of departments with representatives differ from those departments without a representative in the rating-panel. Table 1 gives for each discipline the mean rating of the departments that had a representative in the panel and those that had none (RAE 1996).

Table1: mean ratings (1996 data).

Discipline	Mean rating of not repres. Depts	Mean rating of repres. depts	# not repres. depts	# repres. depts
Clin.Lab. Sc.	4.29	4.63	24	8
Com.Cl.Subj.	3.66	5.83	29	6
Hosp. Cl. Subj.	4.61	5.00	23	11
Clin. dentistry	3.83	4.89	6	9
Pre-Clin. Stud.	3.33	5.00	6	4
Anatomy	4.20	5.67	5	6
Physiology	4.40	5.60	10	5
Pharmacology	4.60	6.00	10	5
Pharmacy	4.00	4.75	12	4
Nursing	2.06	4.20	31	5
Other Medicine	3.07	3.64	57	11
Biochemistry	4.92	6.40	12	5
Psychology	3.55	5.78	66	9
Biolog. Sc.	3.78	5.64	68	14
Agriculture	3.64	5.29	14	7
Food Science	2.85	6.00	13	2
Veterinary Sc.	5.00	5.00	2	4
Chemistry	3.39	6.13	54	8
Physics	4.40	5.56	47	9
Earth Sciences	4.00	5.50	27	6
Environm. Sc.	2.44	6.25	34	4
Pure Maths	4.88	5.40	40	5
Applied Maths	3.67	5.88	57	8

¹¹ Data can be found on the HEFCE website (www.hefce.ac.uk/research/assessment/default.htm)

Discipline	Mean rating of not repres. Depts	Mean rating of repres. depts	# not repres. depts	# repres. depts
Statistics	3.96	5.60	50	5
Computer Sc.	3.68	5.80	79	10
Gen. Eng.	3.71	6.67	34	3
Chem. Eng.	3.78	6.33	18	3
Civil Eng.	3.97	5.38	35	8
Elect. Eng.	3.48	5.43	58	7
Mech. Eng.	3.76	5.00	51	6
Mineral Eng.	5.60	-	5	0
Metallurgy	4.29	5.86	31	7
Built Environm.	2.67	5.11	46	9
Town Planning	3.50	4.83	24	6
Geography	3.45	5.22	60	9
Law	3.70	5.18	53	11
Anthropology	5.00	5.40	14	5
Economics	4.56	5.44	41	9
Politics	3.98	5.78	57	9
Policy & Adm.	3.68	5.00	38	6
Social Work	3.45	4.90	22	10
Sociology	3.83	5.00	52	9
Bus. & Manag.	2.94	5.50	88	12
Accountancy	3.90	6.33	20	3
American Stud.	3.46	4.40	13	5
African Studies	4.57	5.40	7	5
Asian Studies	4.00	5.60	7	5
European Stud.	3.63	4.67	32	6
Celtic Studies	4.00	5.00	11	4
English Lang.	3.33	4.82	80	11
French	4.51	6.00	41	6
German Lang.s	4.47	4.88	38	8
Italian	4.00	5.80	17	5
East Eur. Lang.	4.73	5.57	15	7
Iberian Lang.	4.67	5.71	24	7
Linguistics	3.84	4.88	19	8
History	4.65	6.25	20	8
Archaeology	4.47	5.86	19	7
History	4.05	5.67	86	21
History of Art	3.85	5.57	40	7
Library Manag.	2.78	5.20	18	5
Philosophy	3.89	5.00	37	9
Religious Stud.	3.54	6.00	41	9
Art and Design	3.21	5.00	81	8
Communication	3.16	5.33	32	3
Performing Arts	3.00	4.60	32	10
Music	4.39	5.75	49	8
Education	3.10	5.54	91	13
Sports Subj.	2.62	4.20	29	5
Avg.	3.8	5.4	35	7

As one can see:

- Without exception, the represented departments never have a lower average than the not-represented ones. Over disciplines, the mean of the represented departments is 1.6 points higher than the mean of those that were not represented.
- On average, one out of 6 departments had a representative (some universities had 2 representatives in a panel but here we consider this as just one).

Possible explanations for this finding

Of course, it is possible that panel-members' universities are just better. However, more suspicious minds (i.e. economists) see other possible explanations:

- Panels chose as assessment-criteria those criteria on which their departments scored high.
- Panels did not apply the assessment-criteria to their departments in the same way as they did to non-represented departments.

If only the first is true, it would imply that once corrected for the assessment criteria, we would find no difference between a panel-departments' score and the score of a non-represented department. If the second was true, we would find a significant 'panel'-effect (but even then we would not be able to exclude the effect of a biased choice of assessment-criteria).

A third, and somewhat less negative explanation, could be an informational advantage of the panel-members.

- panel-members' departments are better informed about the RAE and about what kind of information should be submitted.

This last explanation could be more valid for the Teaching Assessment Exercise. This TEA's have a similar goal as the RAE's: judging the performance of a department. In a study about the panel-members of the Teaching Assessment Exercise (TAE), McDowell et al (1997) found that 48% of the members cite "*Opportunity to influence the procedures, methods and outcomes of the Assessment process*" as one of the

reasons for becoming a panel-member¹². Furthermore, “*Most Assessors (88% respondents) had experience of Assessment as a member of department or equivalent which was assessed under the HEFCE system. In comparison with the national picture, their home departments appeared to have experienced slightly more positive outcomes. Under the original Assessment method, 35% of Assessors' home departments were judged 'excellent' in comparison with the overall figure of 26%*”.

This last finding is most probably due to such an informational advantage. Indeed, the organization of the TEA and the RAE are somewhat different. The formers' panel-members f.e. are not 'elected' but 'volunteer' for the job (which makes abuse more easy) but only judge a limited number of departments (on average 3) so their influence is limited and the possibility of exchange is more difficult to achieve. And (Baty, Thesis 11/06/99):” *Universities are paying such academics [Academics employed by the Quality Assurance Agency to assess higher education teaching] up to £1000 a day to train departments to get top marks for their subject quality reviews ”. If universities are willing to pay these kinds of sums, the informational advantage must be important for the TEA. For the RAE, this effect seems to be less important as the teaching performance of departments is assessed for the first time while the RAE was organized for the fourth time in 1996¹³.*

Regressions in levels

As we have the 'objective' data on which the panel should/could have based their ratings, we regressed the rating on the number of research-staff in the department in 96, on the amount of external research income received by the department, per research-staff, in the period 92-96, on the number of research students between 92-96 divided by the number of research-staff^{14,15}. To check for the representation

¹² 'Furthering a personal interest in issues of teaching, learning and assessment' was number 1, cited by 77% of the assessors.

¹³ Though the methodology changed from one RAE to another and the former polytechnics participated for the first time in 1992.

¹⁴ Unfortunately, the 1996 RAE-database does not include a good measure of publications like the 1992 RAE. The panels received a list with bibliographic info about the publications of all selected staff but this info is not available via the RAE-database. Only a count of the number of publications submitted to the RAE is available but as the maximum of articles submitted per staff is 4, the variability is very low. When this variable was included in the regressions, we, not surprisingly, did not find a significant coefficient.

advantage, we add the above-mentioned dummy, the idea behind this being that once we control for the ‘objective’ information, the difference in ratings should disappear.

Hence,

$$R = \gamma + \alpha X + \beta D + \varepsilon$$

Where R= rating

X= explicative variables

D= dummy 1 if represented, 0 otherwise

ε = error terms

Two estimation-methods will be used: OLS and ordered probit. The first has the advantage of a clear interpretation, the second is econometrically more appropriate as our dependent variable is ordinal: the seven different ratings are ‘classes’ and hence, a rating of 6 does not necessarily mean that the department has a performance of twice the performance of 3-rated department¹⁶.

A final note: during the RAE, 2894 departments were assessed in 1996. However, for some variables, we sometimes had ‘zero’-observations. We decided to keep only those departments that had positive values on the staff, research income and students-variables.

First, we pool the disciplines into one big cross-section with 2425 observations.

Table2: pooled observations.

	OLS 1996	Ordered Probit 1996	OLS 1992	Ordered Probit 1992
Constant	3.1 (0.05)	1.1 (0.05)	2 (0.08)	0.33 (0.12)
Staff	0.02 (0.002)	0.014 (0.002)	0.02 (0.001)	0.02 (0.002)
Insider-Dummy	1.13 (0.07)	0.8 (0.05)	0.73 (0.06)	0.8 (0.07)
Resinc/staff	$0.17 \cdot 10^{-5}$ ($0.27 \cdot 10^{-6}$)	$0.14 \cdot 10^{-5}$ ($0.24 \cdot 10^{-6}$)	$0.1 \cdot 10^{-5}$ ($0.25 \cdot 10^{-6}$)	$0.12 \cdot 10^{-5}$ ($0.32 \cdot 10^{-6}$)
Stud/Staff	0.1 (0.009)	0.08 (0.008)	0.08 (0.007)	0.08 (0.01)
Stud/Staff	-----	-----	0.15 (0.03)	0.15 (0.05)

Standard errors between brackets. For the ordered probit estimates, we used the Eicker-White standard errors.

¹⁵ Such a ‘hedonic’ regression is the traditional form used to reveal the importance of possible determinants of research ratings f.e. Ehrenberg and Hurst (1998) for the US, Taylor and Izadi(1996) and Taylor (1995) for the 1992 UK RAE and Johnes et al. (1993) for the ’89 UK RAE.

¹⁶ Note that for the distribution of the funds, the HEFCE rescales the ratings into a cardinal scale: 5*(7),5 (6), 4(5), 3a(4), 3b(3), 2(2) 1(1) become 4.05, 3.375, 2.5, 1.5, 1, 0 and 0 (Ball,1997).

As one can see, both specifications clearly indicate positive effects of the number of staff, the number of students (per staff) and the research income (per staff), and most important for us of the insider-dummy. The interpretation of the OLS-coefficients is straightforward: for each 10 staff extra, a department gets 0.2 research-points extra, for each 100000£ external research money per staff a department gets 0.2 points extra and for each student per staff a department gets 0.1 point extra¹⁷. Being represented in the panel further adds one point.

We next estimate equation (1) for the 92 RAE. For the 1992-RAE, data about the number of publications are available for each department which should allow us to explain more of the variation in ratings as we now include both input- and output variables¹⁸. The inclusion of the publications, a simple unweighted count of the number of ‘outputs’ (articles in journals, books, editorships etc) results in a positive and significant coefficient for the latter variable but does not really affect the other conclusions. We still find a substantial positive effect of being represented in the panel.

For the interpretation of the insider-dummy in the ordered probit specification, we compare the probabilities of falling in each category for those represented and those not represented, taking the other variables at their sample means.

Table 3: probabilities of falling in a particular class for represented and unrepresented departments (1996-data).

Category	represented	Not represented
1	0.004	0.035
2	0.035	0.14
3	0.07	0.16
4	0.15	0.23
5	0.33	0.28
6	0.28	0.12
7	0.12	0.02

¹⁷ A bit surprising at first sight is the fact that the student-teacher ratio has a positive sign. Note, however, that the students here are research students (masters or doctoral) who (with a bit of goodwill) can be seen as an aid rather than a burden to the professors

¹⁸ note that the dependent variable now varies between 1 and 5 and not between 1 and 7 like before, so strict comparisons are not permitted.

The above table clearly indicates that being represented increases the chances of falling into a higher category¹⁹.

Because one could argue that disciplines are too different to be pooled together, we next re-estimated the above equation for 38 disciplines²⁰. Table 4 gives for each variable, the number of significantly positive and significantly negative coefficients and the mean of those significant coefficients (significant: $t > 1.75$).

Table 4: number of positive and significant coefficients.

	OLS				Ordered Probit			
	#pos	Mean	#pos	Mean	#pos	Mean	#pos	Mean
	1996	1996	1992	1992	1996	1996	1992	1992
Res. Inc/staff	27	$1 \cdot 10^{-6}$	21	$1 \cdot 10^{-6}$	26	$2 \cdot 10^{-6}$	23	$2 \cdot 10^{-6}$
Staff	30	0.06	29	0.04	29	0.094	29	0.1
ResStud/staff	17(1)	0.21	20	0.2	21(1)	0.28	23	0.2
Insider dmy	21	0.95	9	0.76	23	1.3	11	1.4
Public/staff	---		8(2)	0.55			10(4)	1.03

Between brackets are the numbers of negative significant coefficients. The mean is the average of the positive and significant coefficients.

The mean adjusted R-squared for the 1996 regressions is 0.54, implying that (only) 58% of the variation in ratings can be explained by ‘objective’ factors. Of these objective factors, the size of a department seems to be the most important variable, playing a significant role in 79% of the disciplines. For each 10 researchers, the performance rating increases a bit more than half a point²¹. External money earned by the department is also considered as an indicator of quality, as for each 100000 £ (per staff) that a department wins 0.14 rating-points. Now, we come to the variable of central interest to us: the panel-membership dummy. For 21 disciplines, the coefficient on the dummy is significantly positive, and the effect for these 21, is on average about 0.95 points. So for these 21 disciplines, being in the panel increased the rating with 0.95 after controlling for the objective quality-factors. Sixteen other disciplines had positive but insignificant coefficients while 1 had insignificant

¹⁹ Note that if we take the levels of research income and students rather than relative to the staff variable, only the insider-dummy and the number of students are significantly positive.

²⁰ We selected those disciplines with more than 25 departments and having departments in 6 or 7 categories.

²¹ Note that this is only valid on average for those disciplines where the variable was significant! This remark remains valid for the other variables that follow! Results for the individual disciplines can be found in the appendix.

negative ones. In other words, for more than half (55%) the disciplines, panel-members' departments are classified about one 'class' higher than expected using objective data. The ordered probit estimates largely confirm these results.

A bit surprisingly, adding the publication-variable adds only 2% to the mean adjusted R^2 for the 1992 regressions. In addition, only in one out of four cases the number of publications is significant. For these an extra publication (per staff and over a 4/6-year period!) increases the rating with half a point²². Signs and coefficients of the other variables are fairly similar to those found for the 1996 RAE's but our dummy variable indicates that now only 25% of the disciplines could have been 'insider'-biased²³.

The above indicates that at least some panels might have chosen 'objectively' the criteria but applied them less strictly to their own departments (though combination of criteria-choice and non-strict application cannot be excluded either). However, before accusing the panel-members we should check for yet other explanations. Indeed, it is possible that some variable influences both the rating of the university and the choice of the panel members.

$$D = \gamma + \alpha * \text{individual characteristics} + \beta * \text{departmental characteristics} + \varepsilon$$

Insofar that these departmental characteristics are the 'objective' factors we included in the ratings-regression, this will not pose a problem because it will be captured by the α -coefficient of the ratings-regression²⁴. More problematic would be the situation where there are some 'subjective' factors that influence both decisions, causing an 'omitted variable' problem²⁵. In this case, there would be no 'unfair' effect of representation: panel-membership would just be an indicator for 'subjective'-elements of quality.

²² Note that it is an unweighted count of journal articles, books, editorships etc!

²³ The fact that there are only 5 classes might be an explanation for this as it makes 'mis'-classification more obvious.

²⁴ F.e. if good departments are bigger, then their sheer size also increases their chances of having a panel-member (*ceteris paribus*). As we include the number of staff in the regressions this effect should not be captured by the dummy but by the coefficient of the staff-variable.

²⁵ Omitted variable rather than endogeneity as the real rating is not known in advance it cannot have an impact on the panel-membership-choice. It's rather a third variable 'subjective department quality' that is correlated with individual research performance (and hence panel-membership-choice) and also the rating.

While we cannot exclude that the positive correlation between panel-membership and research rating is caused by a subjective but quality-related omitted variable, we do know that other reasons (that are not linked to the departmental quality) such as regional balance were also taken into account, that even in non-top departments one can find excellent scholars and finally that the number of members of a panel is limited.

An example of a 'subjective' quality-related variable could be the 'name' of the institution, f.e. if from two 'objectively' similar scholars, one has a position at Oxbridge and the other at Rummidge, the former could still be considered as better and hence more likely to be asked to become a panel-member, simply because of the reputation of the university. Quite naturally, the same effect could play on the departmental level, two clone-departments could be judged differently because of the name of the university. If such an element played, the dummy would capture such a 'brand-name' university-effect.

How to solve this omitted variable-problem?

A first possibility is pooling different disciplines and including a university dummy: It is likely that the 'subjective' factors are university-related so including a university dummy would at least take some of the 'omitted variable'-influence away. Therefore we include university dummies in the pooled specification. Together with disciplinary-dummies this halves the OLS-coefficient: while before we found, using the 1996 data, that being represented was worth about 1 research-point, we now find a significantly positive value of 0.5²⁶.

The availability of the number of publications in the 1992 data-set allows for another robustness-test: if the departments that are represented are simply better and if this is captured by the dummy-variable, then a regression like (1) but taking the number of

²⁶ Another reason is that HEFCE notes: 'Within the 1996 Exercise, considerable stress was laid on the need for a common approach. In particular this meant all panels interpreting the rating scale in the same way, and adopting criteria which were different only where subject-related considerations clearly required it. This should mean that the 1996 ratings are comparable between subjects. However, these judgements have been made in different subjects by different panels against different criteria and it would therefore be unsafe to assume that what a particular rating indicates about a department is the same across all subjects.' (http://www.niss.ac.uk/education/hefc/rae96/c1_96.html#part1)

publications as the dependent variable should also show that the represented departments are better. Indeed, if after controlling for staff, research income and students we find a positive and significant effect of the insider-dummy, we would be forced to conclude that this insider dummy reflects unobserved quality rather than an insider effect as panel-membership is unlikely to influence the number of publications per se.

Table 5 gives the results of this test for two specifications, one with the number of publications as dependent variable, the other with the number of publications per staff-member as dependent variable.

Table 5: number of publications as dependent variable (1992 data).

	OLS-pubs/staff	OLS pubs
Constant	2.1 (0.02)	-4.2 (0.5)
Staff	$0.4 \cdot 10^{-3}$ ($0.8 \cdot 10^{-3}$)	2.5 (0.02)
Insider-Dummy	0.01 (0.03)	-0.86 (-0.8)
Resinc/staff	$0.5 \cdot 10^{-6}$ ($0.1 \cdot 10^{-6}$)	$0.8 \cdot 10^{-5}$ ($0.3 \cdot 10^{-5}$)
Stud/Staff	0.04 (0.004)	0.28 (0.09)

Standard errors between brackets. For the ordered probit estimates, we used the Eicker-White standard errors.

The coefficient of the insider-dummy is not significant which goes against the idea that the insider-dummy captures unobserved quality-differences. Including university and disciplinary dummies even leads to finding a negative coefficient.

A more convincing strategy is combining data of both the 1992 and the 1996 exercise. As the panel-composition differs over the exercises, looking at the ratings of departments that only had once a representative could solve our problem: if departments that had a representative in 1992 but not in 1996 lose (or win less), while those that had in 1996 a representative but not in 1992 win (or lose less), alternative explanations would lose a lot of their power! Indeed, the assumption that we then make is that changes in the panel-composition are unrelated to the changes in the omitted subjective quality variable, which we suspect to be quite stable over time. That this assumption is reasonable can be illustrated by the findings of Dusansky and Vernon (1998) for US Rankings: *"We have compared four different rankings*

methodologies and eight series. cursory statistical analysis suggests that there are discrepancies between the subjective and the objective assessments of the quality of economics departments in the United States. Established programs appear able to maintain their reputations in the face of declines in their publication productivity, while more aggressive upstart programs must be patient in realizing the full returns from their substantial investments in professorial capital.”

In addition, this strategy makes it possible to distinguish collusion-effects from the influence of ‘informational advantages’ in as far as being once a panel-member is enough to know the tricks of the game²⁷.

Unfortunately, merging the 1992 database with the 1996 database complicates things a bit. First, the rating scale in 1996 was between 1 and 7, while in 1992 between 1 and 5. About this comparability between the 1992 and the 1996 ratings, the HEFCE notes²⁸: *“The rating scale used in 1996, reproduced at Annex C, maps directly onto the scale used in 1992. The definition of the scale points is the same as in 1992 except that point 5 on the 1992 scale has been extended to create a new 5* rating, and point 3 has been subdivided into points 3a and 3b. Nonetheless, comparisons between the ratings awarded in 1992 and 1996 should be made with some caution. In some cases the definition of the subject units of assessment has changed; and an HEI may have chosen, as a result of staff changes or for other reasons, to submit different groups of researchers in the two exercises within the same UOA.”* We decided to run regressions for both a 1-5 and a 1-7 1996 rating scale.

Second, the definition of the variable research income changed from ‘*the value of expenses from external research income in year X*’ to ‘*the value of the grant received in year X*’. This forces us to exclude the research-income in our regressions.

Third, some departments have only been rated once (or data are missing in 92 or 96) which reduces our sample even further, to 155 universities.

²⁷ But remember note 12!

²⁸ http://www.niss.ac.uk/education/hefc/rae96/c1_96.html#part1

So now we estimate

$$\Delta R = \gamma + \alpha \Delta X + \beta \Delta D + \varepsilon \quad (2)$$

The dummy variable now equals 1 if the university had a panel-member in 1996 but not in 1992 (218 times), -1 when the university had a panel-member in 1992 but not in 1996 (137 times), 0 when either the university had a representative in both years or in neither of the two years (1714 times). As it is likely that the influence of subjective factors is quite time-invariant, a positive and significant β coefficient cannot be explained but by an ‘insider’-bias.

The other independent variables are the change in the number of research-staff and the change in research students/research-staff ratio.

We look at the pooled sample in three different ways: we take first the difference between the 1996 rating (varying between 1 and 7) and the 1992 rating (varying between 1 and 5), second we looked at the difference between the rescaled 1996 (mapping of the 1-7 scale into 1-5 scale) and the 1992 scale, third we estimated an ordered probit where the latter difference is rescaled into increased rating, same rating and decreased rating.

Table 6: first differences.

	OLS				Ordered Probit
	1-7 vs1-5	1-7 vs1-5	1-5 vs1-5	1-5 vs1-5	
Ct	1.6 (0.02)		0.6 (0.02)		
Staff	-0.007 (0.002)	-0.007 (-0.003)	-0.01 (-0.002)	-0.01 (-.002)	-0.01 (-0.004)
Insider	0.1 (0.03)	0.2 (0.05)	0.07 (0.03)	0.11 (0.04)	0.2 (0.06)
Stud	-0.009 (0.005)	-0.03 (-0.008)	-0.007 (0.005)	-0.02 (0.007)	-0.03 (0.01)
Dumm	NO	YES	NO	YES	YES

Standard errors between brackets. For the ordered probit estimates, we used the Eicker-White standard errors. Dummies are university and disciplinary dummies.

The regression in first differences gives quite mixed results. At one side, we find positive and significant effects of changes in panel-composition on changes in ratings. But the estimated coefficient is very small. At the other side, we find significant negative effects of both the number of staff as the number of students, a result that is

counterintuitive. In the hope to get a clearer picture, we next looked at the individual disciplines. However, much more than an almost general absence of significant coefficients cannot be reported which seems to indicate that changes in rating are unrelated to changes in the panel composition or to changes of any of the ‘objective variables’. This confirms the above statement of Dusansky and Vernon (1998) that subjective ratings do not seem to follow instantaneously objective changes.

Conclusions: The verdict of the jury

When looking at the raw data of table 1, one could be inclined to think that the self-serving scientists have been at work: departments that were represented in the rating-panel have higher ratings than non-represented departments. Our econometric analysis, however, showed that it’s quite unlikely that this difference is caused by such an insider bias. For several disciplines objective factors like the number of staff or the external research income received could explain this difference, while omitted subjective variables are the probable cause for most other disciplines.

Does this mean that scientists are not self-interested? Not necessarily, we just can say that there are not any indications that they colluded. This might be due to the guardian-secretaries or maybe the scientists just could not come to a collusion-agreement. Anyhow, alternatives that make collusion even more unlikely do exist: in the US, the National Research Council lets at least 100 faculty members rate the disciplines, thus making the law of large numbers do the job (see Ehrenberg and Hurst, 1998).

References

Ball, D (1997), “Quality Measurement as a Basis for Resource Allocation: Research Assessment Exercises in United Kingdom Universities”, *R&D Management*, vol. 27, nr. 3, p.281-289.

Baty, P.(11/06/1999), “Experts Cash In on Grades Rush”, *The Times Higher Education Supplement*.

Blank, R. (1991), “The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review”, *American Economic Review*, vol. 81, Nr. 5, p.1041-1067.

Brookman, J. (11/06/1999), “German Professors Get Thumbs-Up”, *The Times Higher Education Supplement*.

Carlton, D., Bamberger G. and Epstein R. (1995), “Antitrust and Higher Education: Was There a Conspiracy To Restrict Financial Aid?”, *Rand Journal of Economics*, vol. 26, nr. 1, p. 131-147.

Dusansky, R and Vernon, C. (1998), “Rankings of US Economics Departments”, *Journal of Economic Perspectives*, vol. 12, nr. 1, p.157-170.

Ehrenberg, R. and Hurst, P. (1998), “The 1995 Ratings of Doctoral Programs: a Hedonic Model”, *Economics of Education Review*, vol. 17, nr. 2, p. 137-148.

Graves, P., Marchand, J. and Thompson, R. (1982), “Economic Departmental Rankings: Research Incentives, constraints , and Efficiency”, *American Economic Review*, vol.72, nr. 5, p. 1131-1141.

Hoxby, C. (1999), “Benevolent Colluders? The Effects of Antitrust Action on College Financial Aid and Tuition”, *Harvard University Working Paper*.

Johnes, G. (1997), "The funding of Higher Education in the United Kingdom", in Hare, P. (ed), "Structure and Financing of Higher Education in Russia, Ukraine and the EU".

Johnes, J., Taylor, J. and Francis, B. (1993), "The Research Performance of UK Universities: a Statistical Analysis of the Results of the 1989 Research Selectivity Exercise", *Journal of the Royal Statistical Society A*, vol. 156, part 2, p. 271-286.

Laband, D. (1985), 'Publishing Favoritism: A Critique of Departmental Rankings on Quantitative Publishing Performance', *Southern Economic Journal*, Vol. 52, p.510-515.

Masten, S. (1995), "Old School Ties: Financial Aid Coordination and The Governance of Higher Education", *Journal of Economic Behavior and Organization*, vol. 28, p. 23-47.

McDowell, J. and Amacher, R. (1986), "Economic Value of an In-House Editorship", *Public Choice*, vol. 45, p. 101-112.

McDowell, L., Colling, C., Sambell, K., Dove, P. and Dobbins, M., (1997) "Improving the Quality of Education: The Impacts of Subject Specialist Assessor Experience", HEFCE M15/97. (http://www.niss.ac.uk/education/hefce/pub97/m15_97.html)

Netz, J. (1998), "Non-Profits and Price-Fixing: the Case of the Ivy League", *Purdue University Working Paper*.

Salop, S. and White, L. (1991), "Policy Watch: Antitrust Goes to College", *Journal of Economic Perspectives*, vol. 5, nr. 3, p. 193-202.

Taylor, J. (1995), "A Statistical Analysis of the 1992 Research Assessment Exercise", *Journal of the Royal Statistical Society A*, vol. 158, part 2, p. 241-261.

Taylor, J. and Izadi, H. (1996), "The 1992 Research Assessment Exercise: Outcome, Outputs and Inputs in Economics and Econometrics", *Bulletin of Economic Research*, vol. 48, nr. 1, p.1-26.

Tirole, J. (1986), "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations", *Journal of Law, Economics and Organization*, vol. 2, nr. 2, p.181-214.

Wibble, J. (1998), "The Economics of Science", Routledge, London.

APPENDIX

Economics and Econometrics Assessment Panel (1996)

Professor David Hendry (Chairman), University of Oxford
Professor Philip Arestis, University of East London
Professor John Beath, University of St Andrews
Professor Anne Booth, School of Oriental and African Studies
Professor Ken George, University of Wales, Swansea
Professor C A E Goodhart, London School of Economics and Political Science
Professor D Greenaway, University of Nottingham
Professor J M Malcomson, University of Southampton
Professor Peter J Sloane, University of Aberdeen

Assessors

Economic and Social Research Council, Mr Adrian Alsop
Government Economic Service, Mr Norman Glass

Secretary

Ms Celia Hunt – HEFCW

The Rating Scale

- 5* Research quality that equates to attainable levels of international excellence in a majority of sub-areas of activity and attainable levels of national excellence in all others.
- 5 Research quality that equates to attainable levels of international excellence in some sub-areas of activity and to attainable levels of national excellence in virtually all others.
- 4 Research quality that equates to attainable levels of national excellence in virtually all sub-areas of activity, possibly showing some evidence of international excellence, or to international level in some and at least national level in a majority.
- 3a Research quality that equates to attainable levels of national excellence in a substantial majority of the sub-areas of activity, or to international level in some and to national level in others together comprising a majority.
- 3b Research quality that equates to attainable levels of national excellence in the majority of sub-areas of activity.
- 2 Research quality that equates to attainable levels of national excellence in up to half the sub-areas of activity.
- 1 Research quality that equates to attainable levels of national excellence in none, or virtually none, of the sub-areas of activity.

Abbreviations

HEFCE: Higher Education Funding Council for England

OLS: Ordinary Least Squares.

RAE: Research Assessment Exercise

TAE: Teaching Assessment Exercise

Table 5: results for 1992 data.

	OLS	OLS with dummies	Ordered Probit
Constant	2 (0.08)		0.33 (0.12)
Staff	0.02 (0.001)	0.02 (0.001)	0.02 (0.002)
Insider-Dummy	0.73 (0.06)	0.46 (0.05)	0.8 (0.07)
Resinc/staff	$0.1 \cdot 10^{-5}$ ($0.25 \cdot 10^{-6}$)	$0.3 \cdot 10^{-5}$ ($0.3 \cdot 10^{-6}$)	$0.12 \cdot 10^{-5}$ ($0.32 \cdot 10^{-6}$)
Stud/Staff	0.08 (0.007)	0.05 (0.008)	0.08 (0.01)
Pubs/staff	0.15 (0.03)	0.18 (0.03)	0.15 (0.05)
Univ. & Disc. Dummies	NO	YES	NO

Standard errors between brackets. For the ordered probit estimates, we used the Eicker-White standard errors.

Table 7: number of positive and significant coefficients.

	OLS			Ordered Probit		
	#pos	#neg	Mean pos	#pos	#neg	Mean pos
Res. Income/ staff	21	-----	$0.1 \cdot 10^{-5}$	23	-----	$0.2 \cdot 10^{-5}$
Staff	29	-----	0.04	29	-----	0,1
Res. Students/ staff	20	-----	0.2	23	-----	0.2
Insider dummy	9	-----	0.76	11	-----	1.4
Publications/staff	8	2	0.55	10	4	1.03