The correlation between RAE ratings and citation counts in psychology

Andy Smith and Mike Eysenck
Department of Psychology, Royal Holloway, University of London

June 2002

Summary

We counted the citations received in one year (1998) by each staff member in each of 38 university psychology departments in the United Kingdom. We then averaged these counts across individuals within each department and correlated the averages with the Research Assessment Exercise (RAE) grades awarded to the same departments in 1996 and 2001. The correlations were extremely high (up to +0.91). This suggests that whatever the merits and demerits of the RAE process and citation counting as methods of evaluating research quality, the two approaches measure broadly the same thing. Since citation counting is both more cost-effective and more transparent than the present system and gives similar results, there is a *prima facie* case for incorporating citation counts into the process, either alone or in conjunction with other measures. Some of the limitations of citation counting are discussed and some methods for minimising these are proposed. Many of the factors that dictate caution in judging individuals by their citations tend to average out when whole departments are compared.

Background and purpose

The ratings from the most recent Research Assessment Exercise (RAE) in the United Kingdom were announced in December 2001. This is the fifth attempt to assess the quality of research in British universities. Departments or other subject groupings were rated for research quality by specialist subject panels whose judgements were based on submissions made by the university departments themselves. With a government review of the future of the RAE in progress at the time of writing, it is perhaps timely to address the issue of the validity of the ratings that are produced by the RAE as it currently exists.

It could be argued that the RAE is to some extent self-validating, because it is based on research measures that are themselves valid. However, there are grounds for doubt. For example, one of the main measures consists of information about levels of research funding. There is evidence that the reliability of the assessment of grant applications is very low. Cicchetti (1991) considered inter-referee agreement on grant applications in a number of scientific disciplines. The overall correlation coefficient was only +0.32, accounting for about 10% of the variance. Another key determinant of the RAE ratings is the scientific quality of the four academic contributions (usually journal articles) listed for each person entered. These judgements, although made by a panel of experts who command a generally high level of respect, are subjective and hard to validate.

Are there better ways of assessing research quality than the present method? It is not our purpose to conduct a comprehensive review of possible alternatives, but merely to focus on one measurement that might form either (i) an alternative framework for research quality assessment or (ii) a useful additional tool for use within the existing framework.

Citation measurement

One could argue *a priori* that citation counts are likely to provide a fairly direct assessment of what is of central importance to the evaluation of research, namely, its impact on researchers around the world. It is our purpose here to argue that citation counts, when considered at the level of departments rather than individuals, provide a measure of research quality that gives very comparable results but is more objective and transparent, and is less costly to implement.

Successive RAEs have required the laborious collection by departments of a considerable amount of information (e.g., each individual's four main publications; amounts and sources of research funding; details of research infrastructure) and the production of a lengthy submission document. The evaluation of the submissions is a major time commitment for the panel members. However, information about citations per individual or per department has never formed an explicit part of any Research Assessment Exercise. There are several possible reasons for this deliberate omission. For example, it can be argued that citation measures are biased in favour of established researchers and those working in popular research areas. Many other criticisms have been directed against citation measures. Indeed, Chapman (1989) managed to identify a total of 25 alleged shortcomings of such measures, including bias against applied research, social factors, name homographs, bias against some married women, human errors, and citation without knowledge of the research in question. Chapman (1989, p. 341) concluded as follows: "With little ingenuity several of the 25 or so shortcomings...could be sub-divided and developed as independent points." There are counter-arguments to most of the points raised by Chapman (1989) and others. These will be considered later in more detail. There is also respectable evidence to validate the use of citation counts. For example, studies carried out within psychology indicate that there is a correlation of approximately +0.7 between ratings of researchers' eminence and the number of their citations (Rushton, 1984).

In this paper we report a new study aimed at estimating the average number of citations per person per annum accumulated by permanent staff in a range of UK psychology departments. On the basis of the results, we argue that it might be appropriate to incorporate citation counts into future RAEs, if indeed the RAE survives. Possible ways of countering some of the well-rehearsed shortcomings of citation counting are discussed.

The study

The original objective of our study, which commenced in 1999, was to estimate the number of citations accrued in one calendar year (1998) by each person listed as a member of staff in the Association of Heads of Psychology Departments (AHPD) handbook for 1998. This handbook listed all UK psychology departments then in existence and provided staff details of each. Our source of citation data was the Institute of Scientific Information (ISI) database, which was accessed via the Web of Science website [http://wos.mimas.ac.uk]. We counted the citations of the work of each individual (whenever published) that were made in journals listed in the ISI database, separately counting the citations of papers in which the person was listed as (i) first author and (ii) second or subsequent author of the source item. Self citations (defined narrowly as a person citing a paper of which he/she was the first author) were excluded. The precise details of the search criteria and information on how to reproduce our search are contained in the Appendix. The searching was carried out by a student on placement from another university, under close supervision and according to a strict prescription. She was selected for her diligence and spot checks of the accuracy of her work were made.

The original search results were severely compromised by two recurrent problems. Firstly, the search for a given name produced citations for all persons with that name, not just the person targeted. In many cases (e.g. Eysenck MW) this was not problematic but in many others (e.g. Smith AT) it was. We felt unable to eliminate such false positives reliably. Secondly, the initials we used in a given person's search frequently differed from those used in that person's publications, resulting in some of their citations being missed.

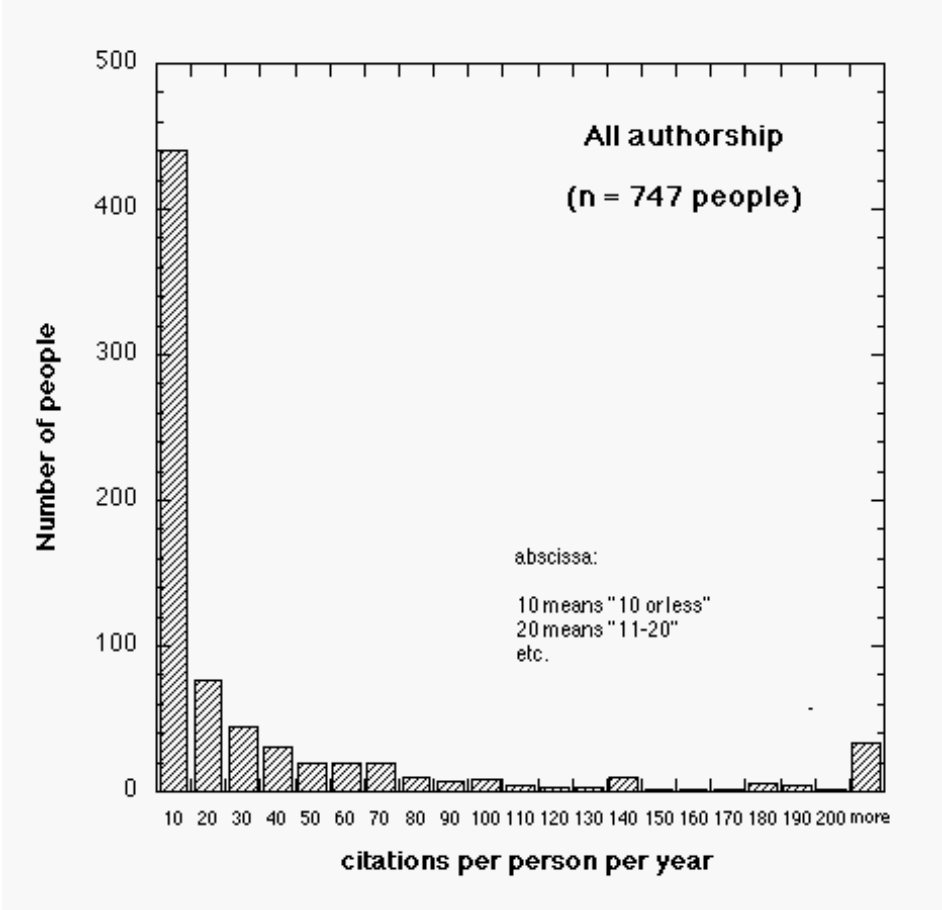
Being unwilling to commit the resources that would be needed to resolve these problems well enough to produce an acceptable level of reliability, we enlisted the assistance of departments. We contacted the head of each psychology department, providing the preliminary lists of citations for all staff in their own department. We asked them to check the lists, identifying citations of source articles that were not written by the person concerned. We also asked them to supply details of instances in which we had used incorrect names or initials. We then searched the correct names and returned the lists generated for scrutiny. Where individuals had published under more than one name, we invited them to choose the name they used most commonly and searched only that name. Clearly this is not ideal, and led to underestimation of citation levels that affects individuals differently, but it was not realistic to search all combinations of names and initials with the resources available. Even though our procedure involved discarding known, valid citations, we hoped to minimise differential effects on different departments caused by adding such citations in a non-systematic fashion.

A total of 38 psychology departments checked the lists we supplied and returned amended versions. The 38 departments covered the range of RAE grades fairly evenly. The inaccuracy of the initial lists caused some consternation in some departments, but eventually data that were reliable (within the limitations of our criteria) were obtained. From the corrected lists, we obtained a citation count for each individual and then combined these to produce the mean number of citations per person in each of the 38 departments.

The final stage of the study involved correlating the estimated citation figures for 1998 with the Research Assessment Exercise ratings for 1996 and 2001, to test the level of agreement between research impact and the judgements of research quality made by the RAE psychology subject panel. This gives only an approximate indication of the true correlation, because there were substantial differences in the people entered in the two exercises, for two reasons. Firstly, we searched *all* members of each department (as listed by AHPD), whereas many departments entered only a subset of these individuals in their RAE submissions. Secondly, our chosen year (1998) fell between the two RAE submission dates, creating a mismatch with both arising from staff changes, which may well have been numerous in some cases.

Results

The total number of AHPD-listed staff in the 38 departments was 747. Fig. 1 shows the distribution of citations per annum per person for these 747 people. Data are shown separately for (i) first-author citations and (ii) all citations (including first-author citations). More than half the names searched yielded less than 10 citations per annum. Impressively, however, about a quarter of British psychologists received 30 or more citations of their work during 1998. At the top end of the distribution, 27 people (3.6%) received more than 200 citations in that year.


All authorship
(n = 747 people)
abscissa:
10 means "10 or less"
20 means "11-20"
etc.
Number of people
citations per person per year
0
100
200
300
400
500
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 more

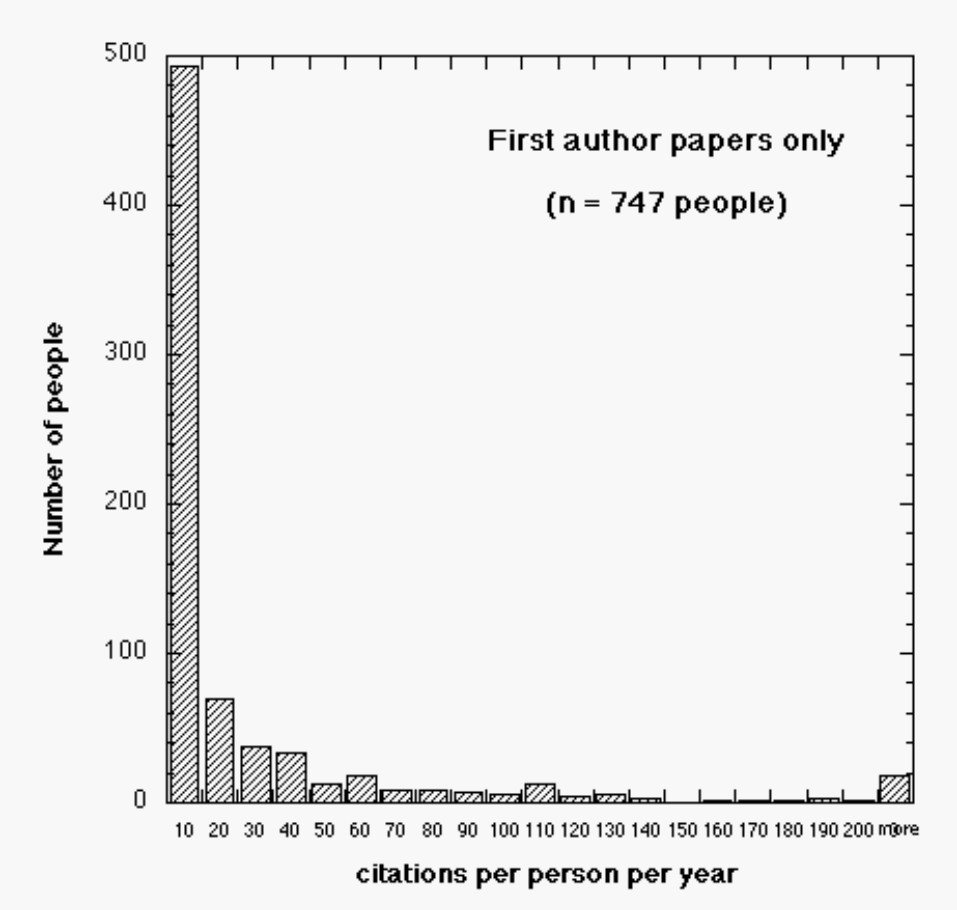

First author papers only
(n = 747 people)
Number of people
citations per person per year
0
100
200
300
400
500
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 more

Figure 1

Table 1 shows the average citations per person per annum, based on one searched name per person and excluding self citations, in each of the two authorship categories, for each of the psychology departments involved. All heads of department concerned were invited, in a letter, to comment on whether they felt departments should be identified by name. One department requested that its identity be withheld, but subsequently rescinded on seeing the flavour of the results. All other HoDs either confirmed that they were happy for their departments to be identified or (more commonly) did not respond.

	1996 RAE	2001 RAE	1st author mean citations	All author mean citations
Birmingham	4	5*	58.05	86.07
Brunel	3a	4	26.92	31.54
Cambridge	5*	5*	99.21	180.53
Central Lancs	3b	3a	9.25	8.54
City	4	4	14.5	20.15
Dundee	4	4	17.94	21.63
East London	2	3a	4.27	5.02
Edinburgh	3a	5	12.5	20.64
Exeter	4	5	25.59	30.35
Glasgow	4	5*	16.71	25.42
Goldsmiths	4	4	19.19	28.81
Greenwich	2	3b	8.00	12.00
London Guildhall	2	3b	5.0	6.04
Keele	3a	4	6.21	6.43
Kent at Canterbury	3a	4	12.57	14.71
Lancaster	4	5	9.11	13.47
Leeds	4	5	36.25	59.79
Lincoln	2	2	1.24	14.53
Luton	2	2	1.17	1.35
Manchester Metro.	3b	3a	7.17	7.74
UC Northampton (Nene)	1	n/a	0.28	0.89
Northumbria at N/cle	2	4	3.93	9.4
Nottingham	4	5	35.25	46.08
Nottingham Trent	2	3a	6.6	8.47
Oxford	5*	5*	64.85	164.6
Portsmouth	3b	3a	8.0	9.72
Reading	5	5*	38.5	59
Royal Holloway	5	5	46.56	65.13
Sheffield	5	5	49.69	75.54
Sheffield Hallam	n/a	3b	11.71	15.29
Southampton	3a	5	38.81	64.14
St Andrews	5*	5*	54.95	99.47
Staffordshire	2	3a	6	8.24
Stirling	4	5	39.19	64
Sussex (Exp Psychol)	(BioSci) 5	(BioSci) 5	46.27	86.6
Thames Valley	1	1	1	1.78
Warwick	4	5	20.59	35.24
York	5*	5*	48.76	88.71

Table 1

Fig. 2 shows the relationship between citation counts and RAE grades. The RAE ratings plotted have been converted to a 7-point scale in which 7 indicates a 5* grade, 6 indicates a 5, 5 indicates a 4, 4 indicates 3a, 3 indicates 3b, while 2 and 1 are unchanged. This was done

in order to facilitate correlation with citation counts. Each point represents one department. Separate plots show data for the two RAE years and for first- and all-author citations. In all cases there is a strong positive correlation. The function describing the relationship is compressive, rather than linear (this is only meaningful, of course, if the RAE scale is regarded as an interval data scale, which, strictly speaking, it isn't). The relationship is slightly more compressive for 2001 than 1996, reflecting a shift to higher average grades with no change in overall range.

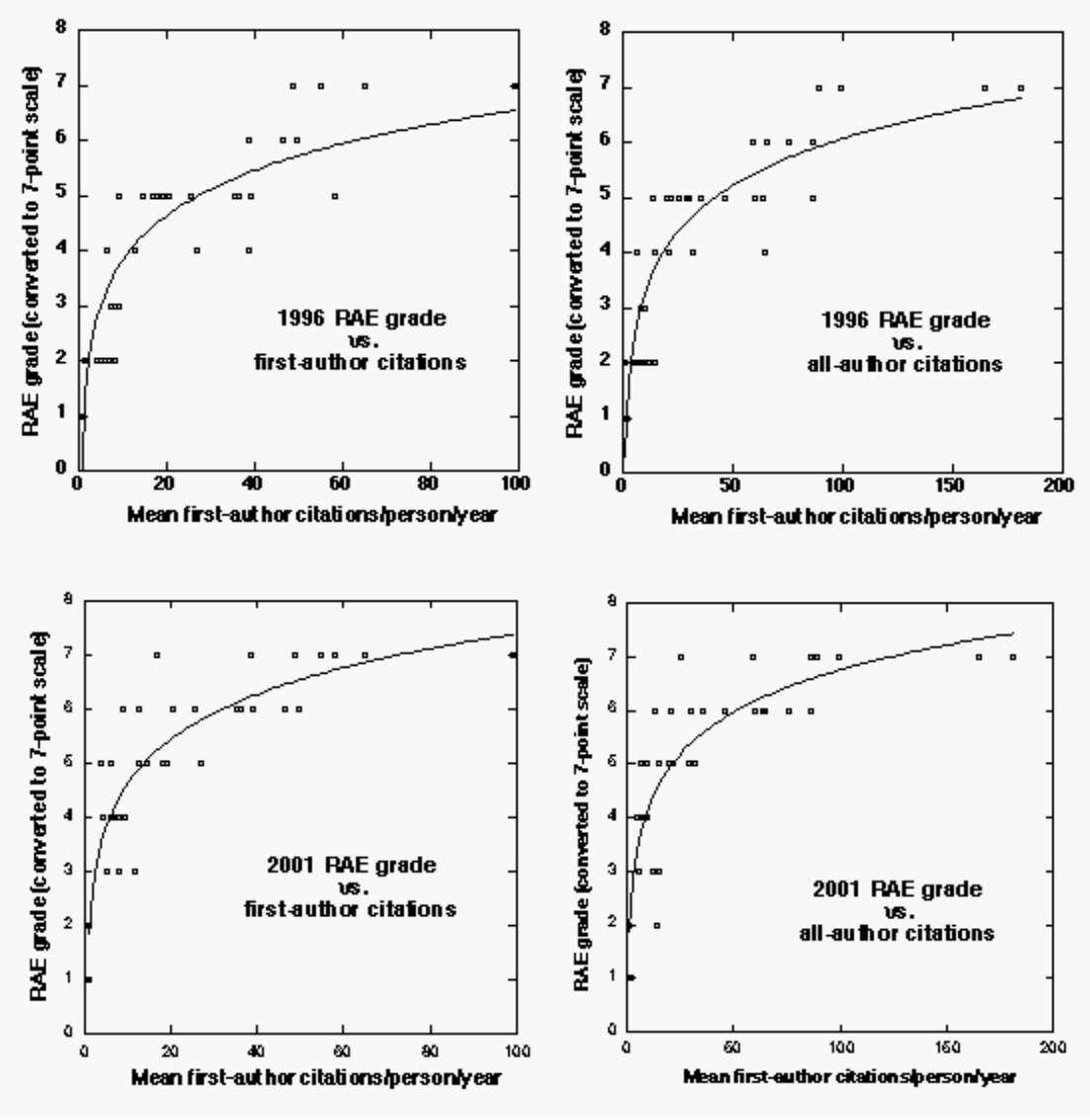


Figure 2

The correlation between the RAE ratings in 1996 and mean departmental citations was +0.91 when based on first-author citations only and +0.90 based on all-author citations (Spearman's rho). The corresponding figures for 2001 were +0.86 and +0.85 respectively. The difference between the correlations for the two RAE years is small and statistically non-significant. However, the fact that citation counts are historical, in the sense of referring to

research reported some time earlier, means that it is reasonable to expect them to predict previous RAE ratings better than future ratings.

The above correlations are remarkably high, especially when account is taken of the fact that there is nothing like a perfect overlap between the individuals included in our analysis of citation counts and those included in the two RAEs. Some inaccuracy is inevitable because of the limited effort we were able and willing to devote to ensuring accuracy. The effect of any inaccuracy of counting, omission of individuals, use of incorrect initials, etc. is to add noise to the data. This noise could be substantial. Noise added to one of two correlated datasets will reduce the correlation between them. Thus the correlations between citation counts and RAE ratings that we obtain represent a *lower bound.* More accurate measurement might make the correlation higher, but would be most unlikely to make it lower. Since it is already very high by most standards, any inaccuracies of procedure or implementation in our study cannot be used to challenge our assertion that citation counts and RAE ratings are well correlated.

Despite very high correlations between citations and RAE grades, there are a few individual departments that depart from the trend. These appear as 'outliers' in Fig. 2. Herein lies a danger. Although using citations in place of professional judgements would give a similar overall result, a few departments might receive very different ratings. It is therefore important to consider the reasons for the differences. We do not have the detailed knowledge of the reasons for specific RAE panel decisions that would be necessary to judge this issue properly, and those who do are not at liberty to discuss such details freely. But a few illustrations may be helpful. When studying the comparison with 1996 RAE data (top panels in Fig. 2), the biggest outliers are York (above the curve) and Southampton (below the curve). York received a 5* while its citation rate is more in line with departments that obtained a 5, although it is not far behind St Andrews, which also received a 5*. This does not suggest a major travesty of justice. Southampton received a 3a, whereas comparison of its citation count with others suggests that it should have received at least a 4, or even a 5, in 1996 (as indeed it did in 2001). If one were to assert that citation counting is a sound method of evaluation then this department might be considered to have been short-changed in 1996. Alternatively, it could be argued that the use of citations would have advantaged it unfairly. We cannot judge which viewpoint is correct. Looking at the comparison with 2001 RAE ratings (lower panels), there are no significant outliers below the curve. The biggest outlier above the curve is Glasgow, which received a 5* while its citation count is more in line with other departments receiving 4 or 5. However, a significant fact about Glasgow is that its 2001 RAE grade was based on a lower proportion of its staff than most other 5* departments, being a "5* C". Thus, the discrepancy may arise because the difference between the staff we searched and the staff submitted to the RAE is extreme. Clearly there are two completely separable issues here: whether citation counting is a good measure and whether it is appropriate to allow departments to submit only their best researchers to the RAE. If we were to search the names submitted to the RAE, we might well get a much higher average citation count and the discrepancy might vanish. Against this argument, Birmingham (one of two other 5* C psychology departments, the third being Newcastle, which did not participate) has a citation count very much in line with the average for 5* departments. Thus, there may be some other factor affecting Glasgow.

Conclusions from the study

There are several conclusions that follow from our study.

1) Citation counting gives very similar results to the existing Research Assessment Exercise, at least in the psychology subject area. It would therefore seem that there is a *prima facie* case for simply replacing the RAE with departmental citation counts. The case is pragmatic: it would be considerably more cost-efficient than the current, time-consuming RAE and would give essentially the same results. It could also be argued that citation counts are intrinsically less subjective than RAE ratings, and provide a clearer goal for individuals and departments. However, there are also counter-arguments; these will be discussed in the next section. A less radical approach would be to incorporate citation counting into a slimmed-down RAE in which the final judgement remains in the hands of a subject panel. This would permit incorporation of other measures that address points missed by citation counting.

2) The limitations of the RAE (e.g., subjectivity) are not the same as those of departmental citation counts (e.g., biases in favour of certain types of research). Accordingly, the extremely high correlations between RAE ratings and citation counts provide substantial cross-validation of both measures.

3) Following on from point (2), it is very hard to argue that RAE ratings provide a seriously erroneous assessment of research quality within departments. If RAE ratings were much affected by subjective biases, then it would hardly be expected that they would correlate so highly with citation counts.

4) Similarly, it is very hard to argue that departmental citation counts are badly flawed. If many citations were spurious, then it would not be expected that such counts would correlate very highly with RAE ratings. Only if the existing RAE is regarded as flawed can citation counting be so regarded.

Incorporation of citation counts into the RAE: practicalities

In this section we discuss the limitations of citation counting as a means of evaluating research quality. It is important to distinguish between (i) problems that afflict our limited study, that could perhaps be solved with the use of greater resources, and (ii) problems that would persist, no matter how thorough the counting and no matter what transformations were applied to the counts before using them. Our discussion of the various limitations focuses on possible solutions to the problems that would arise if citation counting were to be incorporated into the RAE.

What are the objections to citation counting and can they be overcome?

A number of problems and objections have been raised over a number of years. These can be categorised as follows.

<u>1) Practical problems that could be solved by reference to the Department concerned</u>

Mistaken identity. Where two individuals have identical names, it is impossible to distinguish their citations without sifting through the cited works and checking the identity of the author of each. Failure to do this can lead to large errors. We avoided this problem by asking Departments to do the sifting. About 65% did not do so and our survey is based on the remaining 35%. Sifting could be accomplished more efficiently and comprehensively by

requiring each department to supply a full list of the publications of each member of RAE-entered staff, in a specified format, as part of the RAE submission. Source items could then be checked against this list. The checking could be automated (allowing for a specified degree of mismatch between the publication as listed by the author and the source item in the database).

Name changes. Individuals who change the name they use for publication, or include different initials on different occasions, may be credited with an artificially low count if only one variant is searched. This problem could be solved in the same way as mistaken identity problems. If departments were to supply full lists of publications, it would be easy to search citations of all publications.

2) Problems that could be solved by adjustment of citation counts

(i) *Variations in age profile of staff.*
Citations are partly a function of age (or career years) and so a citation count may disadvantage a Department with a low average staff age. It would be possible to compensate for this by establishing the relationship between career age and citation rate in a large sample of individual researchers and then applying a correction to each Department, based on the mean and variance of its age profile.

(ii) *Variations across disciplines.*
Citation rates may vary across disciplines and across sub-disciplines. There are at least three separate aspects to this problem:

(a) *Size.* It has sometimes been claimed that citation counting disadvantages individuals working in small or new disciplines or sub-disciplines, or in inter-disciplinary fields. In principle, this is not expected. It should be the case that citation rates per paper are independent of the size of the discipline or sub-discipline. This is because a large discipline yields a large number of citations but these are distributed among a correspondingly large number of authors. However, extremely high citation counts are likely only in a large sub-discipline, so the variance of citation rates may be greater in large disciplines. In this respect the criticism is valid, since those operating in a small field and positioned in the high tail of the distribution for that field will be adversely affected. But, given the shape of the distribution (see Fig. 1), rather few individuals are to be found in the tail of the distribution. The problem will thus tend to average out across groups of individuals, so it should affect whole departments much less than it affects individuals. If this is not, in fact, the case then discrepancies could be ameliorated by normalising the counts such that the counts for all disciplines (or sub-disciplines) have the same mean and variance.

(b) *Conventions*. Citation rates may vary across disciplines according to convention. If either the average number of published papers per individual or the average number of citations made in each paper varies significantly across disciplines, this will compromise comparisons across disciplines. Again, normative data could be obtained and used to calculate an appropriate adjustment. To prevent such compensation leading to cynical changes in practice, dynamically changing correction factors would be needed.

(c) *Funding opportunities.* In some cases, high-quality science may be impeded by a generalised and persistent lack of funding opportunities. For example, some British social psychologists consider that their funding opportunities are worse than those in other areas of

psychology. In other disciplines, the excellence of work requiring expensive equipment or exceptional manpower may be reduced by inadequate funding. To the extent that such factors relate to whole disciplines or sub-disciplines, an adjustment to citation counts based on normative data would address the problem.

(iii) *Living off reputations*.
Some academics are not currently research-active but have been so in the past. These individuals may accrue a high level of citations based on their past work. This is legitimate in terms of assessing the individual's contribution, but is misleading as regards assessing the current research strength of the person's department. This problem might be solved by weighting each citation by the recency of the cited work, again based on normative data.

3) Problems that cannot readily be solved

(i) *Inaccurate citations.* Citations that are inaccurate, particularly if the name of the author is mis-spelled, may be missed. This problem would be very hard to solve. It may well affect some individuals more than others (e.g. individuals with an unusual spelling of a common name). The differential effect on whole departments will usually be minor, but it could be serious in some instances. The average error introduced by this problem could be estimated.

(ii) *Citations of poor work*. Authors may be cited for the "wrong" reasons, e.g. because they say something contentious that is discussed irrespective of its merit, or because they say something incorrect, on an important topic, that is corrected by others. This problem is unavoidable, but probably accounts for a rather small proportion of citations (e.g. Chubin & Hackett, 1990). Again, the extent of the problem might be estimated empirically, although defining "wrong" might be problematic.

4) Other alleged problems and our responses

"If citation counts were to form part of the RAE, this fact in itself might shape researchers' behaviour in undesirable ways."
It is true that the business of whom authors choose to cite (and not to cite) might become more personal or political if there were implications for their own finance and prestige. Some people in the UK would cite their friends and not their rivals. But (a) this happens already and (b) citations are an international measure, so the scope for UK researchers to distort them is limited. Nonetheless, there may indeed be some degree of undesirable behaviour change. But on the other hand, the main changes may be that researchers will tend to make greater efforts to publish in high-impact-factor journals (which is fine) and to publicise their work at conferences and elsewhere in the hope of securing citations (which would seem to be positively desirable). Overall, then, the cost/benefit equation of any shaping of behaviour would perhaps be neutral.

"The database is incomplete and the software for consulting it is not foolproof"
This seems to be a widespread suspicion, perhaps sustained by the fact that it is hard to verify whether or not it is true. Clearly, proper checks would be needed if citation counting were to be incorporated in an assessment exercise.

In conclusion, many of the limitations of citation counting could be overcome by a combination of (a) use of source data provided by academics themselves, (b) intelligent interrogation of the database, perhaps involving the development of special-purpose software run under a specific contract with ISI, (c) appropriate weighting and adjustment of the resulting counts. These measures would not completely eliminate all problems and it might remain dangerous to judge the quality of an individual's work purely on his/her (adjusted) citation count. But any residual problems affecting individuals should be small enough to average out across departments to an extent that would avoid major misclassification of departments.

References

Chapman, A.J. (1989). Assessing research: citation-count shortcomings. The Psychologist, 8, 336-344.

Chubin, D. E. & Hackett, E. J. (1990) *Peerless Science: Peer Review and U.S. Science Policy* Albany, NY: State University of New York Press 165-190 (chapter 6: Augmenting peer review: The place of research evaluation).

Cicchetti, D.V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences, 14, 119-186.

Rushton, J.P. (1984). Evaluating research eminence in psychology: The construct validity of citation counts. Bulletin of The British Psychological Society, 37, 33-36.

APPENDIX

How to check your citation count, using our method.

1) Go to the Web of Science website http://wos.mimas.ac.uk/ Click on Login: WoS. You will only be allowed in if your institution subscribes. There's a link to a list of licensed institutions on the home page. If necessary, enter your Athens user name and password.
2) Click 'full search'
3) Tick the boxes for all three databases (SCI, SSCI, AHCI)
4) Click 'limit search to years selected below'
5) Tick the boxes for 1997, 1998 and 1999. This will ensure that all 1998 entries are found, even if they were entered in the database in the year before or after.
6) Click 'cited ref search'
7) In the 'cited author' box, enter your name in the form shown in the example, using the name and initials you normally use in your publications. Leave the other two boxes empty. Then click the 'lookup' button. A list of articles that were written by you (or someone with the same name and initials) and were cited at least once during 1997-1999 will appear.

For each item on the list, first check that the paper was indeed written by you and not someone else. If it was, place a check in the check box beside the item and click the 'search' button. This will bring up a list of articles that cited your paper during 1997-1999. They are listed 10 at a time and there may be more than one page. Go through these and count the total number that meet the following criteria:

(i) the year of publication was 1998 (not 1997 or 1999)
(ii) the first author is not you (i.e. it is not a self-citation)

Use the 'back' button to return to the list of your articles. Repeat the above counting process for each cited paper. You could speed things up by checking all the buttons and getting one big list of articles that cited you, but counting these may give you an underestimate of your citation count. This is because articles that cited more than one of your papers will be counted as a single citation, whereas we counted each cited paper as one citation.

When you have finished, add up the total eligible citations for all your papers. Add up citations of your first-authored papers separately from the others. (If you are the first author, your name is capitalised in the list of your cited articles. If you are not, it is in lower case and preceded by several dots).

If you have published with different names or initials on different occasions, do not repeat the search with each variant, but instead do it once only, using the most common variant.

The totals you have counted should be close to the number we counted. It may not be identical because of changes in the database in the two years since we did the counting (or possibly in the software that interrogates it).