

## CITY UNIVERSITY, LONDON

### **Response to the joint funding bodies' review of research assessment November 2002**

#### **Executive Summary**

The opportunity to contribute to the debate on the future form of research assessment is welcomed. Key issues identified by the University include the following:

- It is recognised that the RAE has influenced the development of research within institutions. It is therefore critical, in reshaping the exercise, that the future purpose of assessment is clearly determined and made transparent to institutions, and that this informs the decision on method. Funding methods will always influence behaviour: a strategic approach to the method of assessment should therefore be adopted which introduces the appropriate drivers to effect change in accordance with agreed policy.
- Use of the terms "international" and "excellence" requires attention. It should be recognised that what the RAE seeks to do is to identify and encourage behaviours which are closely related to excellence and which are believed to lead to excellence.
- Whatever method of assessment is eventually adopted, it must take proper account of the concerns identified in the consultation document, in particular proper recognition of collaboration and partnerships, of interdisciplinary and multidisciplinary research and of impact beyond the research community including value added to professional practice, innovation and knowledge transfer.
- Assessment must be based on clearly defined and objective criteria which allow for differences between subject areas. The rules must be made as clear and as explicit as possible in advance to reduce opportunities for manipulation by institutions or by panels. Institutions should be assessed in the same way but should not be asked to take part in a game which they cannot win.
- The number of units of assessment must not be reduced to such a level that differentiation between distinct subjects is lost through averaging. Excessive merging of subject areas is likely to disadvantage middle-ranking institutions in particular, where there is greater variation in degrees of research excellence across subjects. A single panel covering many subjects will limit the number of subject representatives, increasing the risk of bias.
- A method based entirely on quantitative metrics cannot be achieved for all subject areas. This is particularly problematic for arts-based subjects but is not without problems in many of the social sciences, or in developing areas such as nursing and midwifery. It is also unlikely that a purely metric-based mechanism can be found which can assess value added to professional practice or impact beyond the research community, for example.
- A system relying entirely on historical ratings will benefit those institutions already at the top of league tables at the expense of both middle-ranking institutions and those endeavouring to build up research activity from a historically low or non-existent base. A funding model which combines (in proportions to be agreed) historical ratings, a measurement of more recent activity, and reward for innovation might be acceptable. Whatever model is adopted, the retention of an element of expert review is likely to be required.

**Response to the joint funding bodies' review of research assessment  
November 2002**

The University welcomes the opportunity to contribute to the debate on the future form of research assessment. However, it appears to us that there is one critical question missing from the range of issues raised in the consultation document which is possibly at the root of some of the concerns driving the review. The document focuses on the range of possible methods of assessment but does not seek to clarify the purpose of research assessment (other than to provide the information necessary to calculate funding levels). A number of possible aims can readily be identified, which may apply to a greater or lesser extent in different subject areas:

- to reward particular activity in research recognised as of high quality and priority (however defined – the definition might include collaborative activity, impact beyond the research community including value added to professional practice, innovation and knowledge transfer, development of researchers)
- to improve the quality of research
- to broaden the number of institutions undertaking research (assuming limited funding, this might be seen primarily as a means of informing teaching and learning)
- to concentrate research in fewer, better resourced institutions, as a means of promoting innovation and links with industry, the health sector and other relevant bodies
- to provide a convenient quality label for external audiences

No doubt others can be added to the list. It is surely vital to determine the principles underpinning research assessment before deciding on the method, given that the choice of metrics will influence the behaviour of institutions.

As a further general point, it would be helpful to know what feedback has been obtained from panel members on the 2001 RAE process to further inform the current debate. The consultation document assumes the need for research to be considered in a global context. However, there is at least some anecdotal evidence that some of the international assessors who took part in the 2001 RAE found the brief they were given very difficult to fulfil in the way they might have wished – they were given insufficient time to make their assessment and there were difficulties in matching expertise sufficiently closely. It is also arguable whether they are, by definition, in any better position to judge "international excellence" than the national assessors.

Similarly, in at least some cases the UK specialist readers were given only very limited time to reach a judgement on relatively substantial amounts of work. If a future process is to seek to validate judgements through the use of international assessors, it is critical that the methods used are robust and properly resourced, and draw on the lessons learned from 2001.

A related point is the definition of "international" in this context – how is international excellence to be defined as differing from national excellence? Is it a question of standard, which will clearly be differently defined across subject areas, one of impact, or simply one of geography? Is research which has significant implications for a region of the UK but is not relevant more widely, and may therefore be of no interest to a global audience, considered to be of international excellence? Is work done at international level, but which might be uninteresting or not at the cutting edge (for example comparisons of systems between countries), automatically to be rated as 5 or 5\*? Were the approaches taken by the 2001 panels in this regard as consistent as could have been expected, given subject differences, and has this issue been studied in the light of the 2001 experience? There is at least anecdotal evidence that some panels took "international excellence" to mean work published in international (American, in one case) journals without further analysis. The definition is clearly subjective and the differentiation between international and national may be blurred, depending on the cultural viewpoint.

The use of the term "excellence" itself also requires attention. The assessment exercise cannot in reality be about identifying excellence, rather it is about identifying activities which are considered to be closely related to excellence. Contribution to knowledge can only be properly identified retrospectively and over a long timeframe - this cannot be achieved through the form of assessment under consideration here. The funding councils should recognise that what they are seeking to do is to identify and encourage behaviour which will lead to excellence and that, for example, peer review of journals or the award of grants are taken to be closely associated with this. It would be better to avoid the use of "excellence" in the rating definitions and to refer instead to behaviours or proxies, and to attempt to find reasonable proxies. It would be an interesting exercise to look at work from five or

ten years ago which is now acknowledged as groundbreaking, selecting something from each of the principal subject domains, and apply the RAE outcome measures to the departments from which the work originated, as they were then, to establish whether they would have been rated as excellent. This would test the strength of the proxy measures in use.

Whatever method is eventually adopted, it must take proper account of the concerns identified in the introductory section of the consultation document. In particular, the current system does not provide sufficient recognition of collaborations and partnerships and, indeed, can have the effect of penalising institutions involved in collaborative activity if they are not the "lead" institution holding the grant or providing the payroll facility for shared staff. The position of interdisciplinary or multidisciplinary research is also a key concern. We welcome the greater emphasis placed on the contribution made by institutions to the supply and development of researchers and the consideration of the need for targeted development funds to support new, or newer, subject areas. We note that HEFCE and the Department of Health have already made a commitment to a capacity and capability development fund for nursing and the allied health professions, for example, although information on the magnitude of the fund has not yet been forthcoming.

### **Expert review**

A passing reference is made in this section of the document to the possibility of combining assessment of teaching and research. This is not picked up again and is therefore perhaps not intended for serious consideration but we would note that it would be deeply problematic to produce a rating system which took proper account of the different issues surrounding teaching and research activities. It would also serve to aggravate the difficulties relating to the assessment of interdisciplinary activity. The link between teaching and research, however, needs to be vigorously defended. If the effect of future assessment exercises is to increase selectivity in the allocation of research funding, the funding councils should address this issue by other means.

The teaching of research methods might usefully form one element of assessment (noting that institutions contribute to the supply and development of researchers in this way via activities which fall outside the current scope of research assessment – City University, for example, has a very successful Masters programme in research methods and analysis).

Prospective assessment is difficult to carry out entirely objectively given that circumstances can change substantially and at short notice, for instance through the departure of key members of staff, or a reduction in levels of research grant income. Greater objectivity can be achieved on the basis of assessment of activity already undertaken or clearly in hand. A limited amount of weight can reasonably be given to prospective assessment based on a combination of previous record of achievement and credibility of research plans. Here a funding model might be applied in which most of the funding (perhaps 80%) is awarded on the basis of rewarding recent past performance, with the remainder awarded on the basis of anticipated future performance.

The system applied in 2001 already incorporates most of the readily identifiable objective data - research student numbers and degree awards, grant income, staffing information and actual outputs. To support the supply and development of researchers, data on research student publications and progression, and objective data relating to the progression and development of newer research staff could usefully be added. The principle of the dominance of quality criteria, however uncertain, needs to be retained.

A return to the inclusion of a publication count is strongly opposed as this distorted activity and was not necessarily meaningful (an assumption of quality on the basis of quantity). If impact beyond the research community is to be treated seriously, panels need to take proper account of it in judging outputs. There is a tendency in some subject areas to treat refereed journals as superior to books or other forms of reporting – implicit in this is an assumption that communication with academic audiences is of greater importance than wider impact.

In regard to the question on level of assessment, it is not clear whether this refers to published ratings or to assessments which inform the eventual rating. Published assessment of individuals would be invidious and probably unmanageable; research institutes do not exist in all institutions. Assessment at the level of HE institutions would lose sight of areas of strength, in middle-ranking institutions in particular. It is therefore appropriate in our view to continue with assessment at the level of groups and/or departments, maintaining the current flexibility as to the relationship between submitted groups and organisational units, although we would note also that the current system encourages game-playing as institutions attempt to maximise the results

against a somewhat uncertain set of rules. It may or may not be necessary to reach the overall assessment by way of assessment of individuals – clearly this depends on the method used.

There is no clear alternative to organising assessment around subjects or thematic areas. The number of units of assessment should not be reduced to such a level that differentiation between distinct subjects is lost through averaging. Administrative convenience should certainly not over-ride academic issues and the burden placed on individual subject specialists who might find themselves the sole representative of their area on a broadly defined subject panel must not be underestimated. The inevitable limitation on the size of a single panel covering many subjects also risks the introduction of bias. Excessive merging of subject areas would be most likely to disadvantage the middle-ranking institutions where there is greater variation in degrees of research excellence across departments. A single rating for an institution would be meaningless.

### **Algorithm**

It is not made clear who would be surveyed to arrive at a measure of reputation based on surveys but it is assumed that this would be other academics in cognate areas. It might be an interesting study to try this out in a sample set of areas to see how the results relate to the 2001 RAE outcomes. Since it would be tantamount to voting, however, it is unlikely to produce robust and sufficiently credible results.

Any algorithm must allow for subject differences. For example, citations are not appropriate to all areas, there is considerable variation in research student applications across areas (ie it is easier to recruit good students in some subjects than in others), there is greater availability of external research income for some subjects, and the importance given to conference presentations and other forms of output (eg books versus journals) varies considerably. It would not be possible to achieve an acceptable outcome in an assessment based entirely on metrics across all areas of activity. A set of metrics which would provide an acceptable result sufficiently close to the likely outcome of a peer review or other system might be achievable in some subject areas, most obviously in the science-based subjects where agreement can perhaps be reached on a ranking of journals and conference contributions (although the treatment of books might remain an issue). However, research outputs must be a key component of any form of assessment and judgements of quality of research cannot be reached purely on the basis of the location or nature of the output for some subjects without more detailed assessment of content. This is particularly problematic for the arts-based subjects but is not without problems in many of the social sciences, or in developing areas such as nursing and midwifery. Equally, if the exercise aims to reward factors such as value added to professional practice or impact beyond the research community, it may not be possible to find a purely metric-based mechanism which provides an adequate basis for assessment of these aspects.

While arguing for subject differentiation in some respects, there is nonetheless a need to retain consistency of approach between panels as far as possible in the application of rating definitions, in the interests of fairness across subject areas. The lack of consistency between subject panels in applying criteria for excellence in research can create difficulties in the development of a coherent and equitable research strategy within a multi-disciplinary school, or indeed across an institution.

A metric-based method would need to differentiate between more established and newer researchers (who, at least in some subject areas, may not yet be publishing in the best journals, for example), in order to avoid penalising institutions for including the latter group. The impact of the inclusion of all staff, should this be the model adopted, would itself need careful consideration. There are many reasons why staff may not be research active – there may have been a clear management decision to differentiate between teaching, administration and research time amongst staff; institutions may contain bodies of staff who are not expected to be research active (eg in nursing schools or those whose focus is the delivery of professional skills courses); there may be cases of management failure where staff are not performing at the level expected of them; there may be particular circumstances which have impacted on an individual's research performance during the period under review. Equally, the effect of the RAE in some institutions has been to distort staff recruitment such that research activity is promoted over the requirements of teaching, and to increase the use of part-time visiting staff to support teaching. (This last point suggests a need for the funding councils to consider the impact on teaching of the RAE and its associated funding model alongside the more specific review of research assessment itself.)

Should the rating outcome be influenced by any or all of these circumstances? Should an algorithm produce the same rating for a research unit which contains only a number of high performing staff and a unit which contains the same number of high performing staff along with a number of lower performing staff (who might, for example,

be excellent teachers)? Is it the achievement of high quality research which is important in itself, or the average level of activity across a unit? It may be possible to use student:staff ratios as a normalising factor provided that subject differences are taken into account.

As previously suggested, it would be useful to reward institutions for supporting the development of researchers by including publication by research students prior to the award of the degree in any set of metrics.

Funding methods will always influence behaviour. If specific elements of funding are attached to particular metrics, this will cause institutions to focus greater attention on these metrics – there might be some perceived benefit to this if there is clear value obtained from increasing activity but this would need to be approached with great care and in the light of explicitly stated objectives. There might be merit in varying the balance of reward and incentive funding between subject areas in recognition of the different stages of development of some areas or a need to support new subjects.

### **Self-assessment**

Attempts were made to include an element of self-assessment in submissions in 2001. Feedback from panels on the value and credibility of this (albeit limited) component would be helpful.

It would not be acceptable to rely entirely on self-assessment and it is doubtful that much weighting could be given to this as a component of a rating system unless it was based largely on clearly defined objective criteria (in which case it could, for example, form part of a process of application for additional funds). Detailed review of self-assessment, particularly if it was done at individual level, would be burdensome on the assessors and any attempts at validation would be very time-consuming. The inclusion of individual self-assessment in a system of review would also be administratively burdensome. As to whether assessment should be prospective, retrospective or a combination of both, the same issues apply as for peer review. Criteria would also have to vary between subjects for the reasons already given.

### **Historical ratings**

How would an "institutional" rating be derived? Many institutions have a good deal of variation in the level of performance across subjects which would be lost in any averaging process and performance could well be improving in some subjects while declining in others, making it difficult to arrive at a robust rating. What would be meant by "the value of [an institution's] research infrastructure" outside the engineering and the sciences? Total research income? Money spent on research development and training? How would "failing" institutions or "value for money" be defined? A set of explicit and objective measures would clearly need to be established taking into account the likely results of the impact on behaviour these would produce.

Research strength is capable of changing more rapidly than is assumed here. A system relying on historical ratings will benefit those institutions at the top of league tables at the expense of both middle-ranking institutions and those endeavouring to build up research activity from a historically low or non-existent base. This is not in the interest of the furtherance of research potential and would result in the elimination of opportunity.

A funding model which combined historical ratings, a set of metrics measuring more recent activity (subject to the issues set out under the algorithm heading), and innovation in some proportion to be agreed might be acceptable. This would provide some degree of funding stability, would reward recent improvements or punish failure to improve and would look to the future. The use of historical ratings would clearly have to be revised in any further round of assessment to take account of changes. Additional rewards or penalties might be considered for above average improvements or reductions in performance.

## Cross-cutting themes

The high-level question of the purpose of research assessment was addressed at the beginning of this response. The use of ratings once derived is difficult to police or to control – witness the use of QAA subject review scores. Even overseas institutions make use of RAE ratings – for example, a new college in the Arab Emirates is seeking partnership only with groups rated 5 or 5\*. Research councils already make use of RAE ratings in reaching award decisions: we would argue that all sources of funding should be treated as equal within the assessment exercise itself if new development is to be encouraged. The case of the research councils supports this if we are to avoid a "chicken and egg" situation where departments are prevented from obtaining research council funding if they have not achieved a high RAE rating, which in turn impacts on their future RAE rating.

Five yearly intervals between assessments are probably the most acceptable – this allows some stability in planning on the basis of funding expectations, while much longer would be detrimental to areas which are improving. The present longer duration for the output period in arts subjects needs to be retained. It is important to know the basis of assessment methods well in advance.

A rolling basis of assessment where a different group of subjects was assessed each year would create additional administrative and resource burdens in managing the process (again, the QAA subject review process provides a point of comparison) and would make consistency of approach more problematic. It would rule out proper assessment of interdisciplinary research in many cases – even if broad subject groupings were introduced, this would still be likely to exclude some forms of interdisciplinary activity. It would also either lock in a particular system approach until a full cycle of assessment had been completed, or risk inequities caused by mid-cycle changes to the system. Ratings in some subject areas would also become out of date sooner than others, creating possible disadvantage.

The use of the term "excellence", or indeed "quality", and the problems associated with defining "international standards", were addressed in the first part of this response. However defined, measures of or proxies for excellence in research must be established at the subject level and should seek to avoid subjectivity. There are clearly different aspects of research activity which demand recognition, but it would be highly complicated to achieve a robust method of assessment which addressed this. It is not considered that these issues were captured by the 2001 RAE in any coherent or explicit fashion.

There are a number of issues which must be taken into account in determining the distribution of funds between subject areas: these include an allowance for any legitimate differentiation in units of resource; the need to ensure that departments receive a level of resourcing appropriate to the rating achieved (and avoiding a repeat of the 2001 funding debacle); and the need for institutions to know what level of rating will or will not be funded, to avoid wasting time on pointless submissions at the lower end of the scale. The controversy which followed the 2001 RAE over the different distribution of ratings within units of assessment also needs to be avoided in future. Is there any objective evidence that the overall quality of research in one subject area is better than another? Should a system which ranked all submissions and normalised the distribution of ratings within a subject area be introduced? Any game will be played by those involved (at whatever level) to maximise their own benefit - the only way to deal with this is to make the rules as clear and as explicit as possible in advance to reduce opportunities for manipulation.

A metric based on external funding in the subject which differentiates between sources is unlikely to be equitable – even at the level of the research councils, there are differences in the total sums of money available, and there are considerable variations in other sources of funding between subjects. There is an added difficulty of defining interdisciplinary or non-subject specific "pots".

Each institution should be assessed in the same way, although a capacity building fund might be introduced to assist development where needed (assuming this had been determined as part of the purpose of the exercise). If a separate pot of funds were to be introduced for this purpose, the assessment exercise might usefully provide the mechanism for determining its distribution. However, institutions should not be asked to take part in a game which they cannot win.

In regard to assessment at the subject level, the proposal to define broad subject areas to determine similar assessment methods should be explored further to establish the extent to which it is feasible, but the system must have regard to the need to maintain appropriate subject differentiation.

We would dispute the assertion that institutions currently have a large degree of control over by whom submissions are assessed. As previously indicated (see the section on algorithm), the degree of discretion over what submissions are made and who is included must be linked very clearly to the mechanism for determining the outcome. While a system which would minimise the ability of institutions to vary or manipulate their submissions is supported, this must avoid inappropriate penalties.

The question which links the research assessment process to equality of treatment for all groups of staff in higher education is inappropriately framed. The research assessment process does not encompass all groups of staff in higher education but even restricting the question to academic and research staff is inappropriate. If the funding councils wish to assess and reward diversity in the research community, which might well be perceived to be an entirely legitimate aim, a different set of measures would be needed. The research assessment exercise is not the means by which this should be undertaken. There are a number of very complex matters relating to the nature and culture of institutions which underpin diversity issues and the RAE is not an appropriate vehicle for engineering social change. The only data currently collected by the exercise which have any bearing on this relate to age (which itself is problematic – length of experience as a researcher would be a more appropriate measure) and gender. These data were not presented to the panels. It was understood that some analysis was to be undertaken as a separate exercise; however, these data have yet to be made available to the sector. If there is any evidence that suggests that age or gender bias has occurred during the exercise, this should be published, but the exercise itself cannot be used to address such problems. If discrimination is believed to have occurred in relation to these or any of the other characteristics listed, this can only be within institutions, in determining what research is supported or who is to be submitted, or at the panel level, where subtler forms of bias could occur in the decision-making process which might be impossible to detect. Efforts might be made to ensure that assessors are as representative as possible, but the credibility of assessors in terms of their experience in the subject area must remain the principal criterion for selection. At the institutional level, it is clearly a matter for internal systems to address. The encouragement given to institutions within the 2001 process to include newer researchers notwithstanding the absence of four publications was a positive feature of the exercise which it is hoped was appropriately reflected in the outcomes.

The features of an assessment process listed are all important (although some are clearly closely linked eg not burdensome and minimally expensive). We would identify rigorous, fair and transparent as the most important – resistant to games-playing is also important but should be derived from the other three features.

### **Anything else?**

We would recommend strongly that studies are carried out which experiment with the outcomes of different metrics, applied both separately and in combination to the 2001 RAE data and including sensitivity analyses, to see what differences emerge before any decision is taken as to the future form of research assessment.