Assessing Assessments
European experiences

Published by:

The Danish Institute for Studies in
Research and Research Policy
Finlandsgade 4
DK – 8200 Aarhus N
Denmark

Phone: +45 8942 2394
Fax: +45 8942 2399
Email: afsk@afsk.au.dk
http: //wwww.afsk.au.dk

Print: CC PRINT 92 APS, Denmark
ISBN: 87-90698-12-6
Printed: 500 copies

Assessing Assessments
European experiences

Proceedings of a conference organized by:

The Danish Institute for Studies in Research and Research Policy

in cooperation with

The European Consortium for Political Research (ECPR)

WHY ASSESS ASSESSMENTS ?

Director Karen Siune
The Danish Institute for Studies in Research and Research Policy

Since the 1980s the practice of assessments has penetrated the Political Science profession as well as many other professions. In many European countries teaching curricula and research output are being assessed every 5 years. Experiences with recent assessments in political science in Denmark, Great Britain, Germany, Austria and the Netherlands enable us to assess these assessments. The ECPR and The Danish Institute for Studies in Research and Research Policy has organised a seminar on assessments - both for those who use assessments as a policy instrument (Ministries; Universities) and those who are being assessed (Departments of Political Science; Academics).

Background

In 1995-1996 research assessment within political science took place in Denmark, the United Kingdom and in the Netherlands. Apart from these three evaluations in which members of the ECPR Executive were involved, recently, research institutes, libraries and documentation centres (Blue List Institutions) in Germany have been evaluated and so has the lnstitute for Advanced Studies in Vienna. More informal reviews have been carried out in Norway and Sweden.

The impression is there that all assessments are being conducted in a very different way. For example, in the UK the focus was on publications, the research culture and the research organisation of a department, using very detailed criteria for evaluation. In the Netherlands research programmes were evaluated and given scores for their quality, productivity, relevance and viability. And in Denmark a more qualitative approach was used with the visits to the departments as the main source of information.

Often it is the Government who demands these assessments, but the implementing institution differs from country to country. In the Netherlands it is an Association of the Universities (VSNU) which sets out the criteria and is responsible for the logistic arrangements. In the UK the Ministry of Education appoints a Research Assessment Exercise team and in Denmark the Ministry of Research (former Ministry of Research and Technology) took the initiative and established a steering group and an organisational committee.

The objective of these assessments is to evaluate the performance of a department. Depending on the financial structure, Ministries responsible for higher education can use the results when making their budget plans, granting more money to those departments with high scores, as is the case in the UK. Alternatively, individual universities or faculty boards, may use the evaluations in their yearly distribution of money between departments. For individual departments, an assessment forces the faculty to look again very closely at their curriculum or research programme. For individual members of staff, the practice of assessments may increase their productivity or change their research agenda, influencing where to publish, and what. Ultimately assessments may change the nature of the game and will decide on what academics do. For example, the British Research Assessment Exercise places a high priority on articles in prestigious journals, and a low priority on authored and edited books, and an even lower priority on convening research groups.

Finally it may affect the opportunities for career advancement, now that, for example in the United Kingdom, more and more universities compete for highly regarded scholars and their lists of publications, prior to the Census Date.

Assessments are often carried out by a group of political scientists, selected by the organising organisation. Sometimes they are retired academics or scholars from outside the country, to avoid prejudice. In a research assessment, departments of political science are asked to send the panel a selection of their best publications, together with a list of all publications written in the last five years and categorised according to scientific criteria. In the UK, where there are more than 60 departments of Political Science and International Studies with over a thousand active researchers, it is impossible to read all the cited publications. There, the need for detailed criteria and agreement on method of approach is clear. In countries with relatively few departments of political science, like in the Netherlands and Denmark, the departments are also paid a visit by the assessment panels and asked about their research performance.

The experiences with all these Research Assessment Exercises have brought some problems to the fore. A first category of problems has to do with the fact that every country is reinventing the wheel again. There exist no general guidelines for doing research assessments, no list of useful criteria, no general outline of procedure, that could be used, after minor adaptations to specific country circumstances, in different European countries. Not only do countries differ in their approach, within a country the procedure is also changing with every assessment, making a comparison over time impossible.

Other problems are more specific: how to measure and compare research quality; how much reliance and weight to give to the background statistics; how to value articles written in the mother tongue and published in journals with relatively few readers; should there be a citation index for publications written in a non-English language? or for each different language one?; how to weigh and compare policy and applied research with more theoretical work? etc.

A third category of problems has to do with the use of assessment results and the negative impact that these may have on the profession. We mentioned already the emergence of a blooming "transfer market" in the UK where, in the year before the Census Date, departments ’buy’ scholars with a high publication rate. Another danger is that assessments require change and continual change at universities may be seen as change for change’s sake and can lead to weariness and indifference among the academic staff. Research assessments may also lead to a separation of teaching and research and create two-tier faculties: one concerned with teaching and one with research. Finally, research assessments in the end may determine the research agenda of academics.

The comparative perspective

Given the importance of and interest for the practice of research assessments in Danish research policy, where contracts are prepared between The Danish Ministry of Research and universities, The Danish Institute for Studies in Research and Research Policy hosted a seminar on this topic 27th - 28th September 1998. The intention was a small working conference of about 25-30 people, where information should be exchanged about the practice of research assessments, where the pros and cons of different models for assessing research could be discussed.
The seminar was planned within a framework determined by the review process, discussing questions such as: who commissions the review (formal, Ministry or informal),

who appoints the review panels and determines the brief, who sits on the review panel, how do they do their work (visits or paperwork alone), to whom they report (public or private), what are the implications and whether there is a feed-back effect?

In this report European experiences within research assessment is described and discussed. The focus was primarily at political science but the issue whether experiences from political science assessments could be transferred to other fields of science was on the agenda as well at the seminar.

To give an understanding of the national differences in organising assessments and evaluations the practise is described for more than eight European countries including the Netherlands, Belgium, Italy, Spain, Germany and Greece, not to forget the British and the Irish experiences which has highly influences the rest of Europe. Finally the issue whether the results of assessments can be used in practise was discussed at the seminar

With this report we hope to raise greater awareness of the assessment practise in Europe.

WHO WANTS RESEARCH ASSESSMENTS AND WHAT FOR?

Monique Leijenaar
Catholic University of Nijmegen, Netherlands

Evaluations of research outcomes have always played a role in not only the control of what was accepted as valid knowledge, but also in the distribution of scarce resources (Westerheijden). Given the relative autonomy of institutions of higher education, at first research colleagues (peers control) did the evaluations. Only when governments had to cut back public expenditure and the demand for more accountability of government expenditure was growing, spending on teaching and research was being conditioned and more extensive systems of evaluation were developed.

So the short answer to the question WHY ASSESSMENTS is:

> To have a more or less objective indicator for the allocation of scarce resources, may it be research funds, additional researcher posts or in the form of extra contract research, may it be by the government, the university or the faculty
>
> To quote two administrators in the Netherlands:
> " Program review is an excellent mechanism for providing a new administration with unbiased information on the status of each department in the institution"
> " External judgements legitimise what you, as administrator, know already"

The second reason then why assessments became practice, is

- As an answer to the demand for more accountability regarding government funding.

Both objectives of research assessments can be traced for example to the British practice of RA exercises which are directly tied to governmental research funding and to the Dutch experience where publication and distribution on a wide scale of the final report is compulsory.

Given these objectives it is clear that in most countries governments or government related institutes initiate the Research Assessment Exercises (RAEs). But more goals can be distinguished. For example when you go through the many assessment reports of the Dutch Association of Universities who is in charge of the Assessment Exercises in the Netherlands you find that that are several more reasons to organise assessments. I quote again now from the Political Science Research Assessment Exercise, done in 1994:

According to this report an additional goal of the RAE was:

- Quality maintenance and improvement through feedback to the research group and the university management
- Management on the basis of quality through the provision of quality assessments to the boards of faculties and universities.
- All this should stimulate and encourage the research director and the research group in their tasks.

So the third reason main reason is to improve the quality of research. And it is this objective that is the main task of Review committees that operate in most countries. For example the tasks of the last Review Committee (1994) assessing Dutch political science were:

- To assess the quality of the research on the basis of the information furnished by the university and through local or centrally located interviews
- To advise how quality might be enhanced.

How they assess the quality of research is of course a whole other matter and the different country reports we will hear tomorrow morning will undoubtedly show us that in each country it is handled in a different way.

But I am also interested -and I hope you will address that question in your presentations-, whether review committee's present concrete suggestions for improving the quality of research. Given the task of the Dutch review committee I went through their report to check upon their recommendations and detected the following suggestions:

1. Contract research (i.e. commissioned research some other organisation than the university is paying for) should be encouraged;
2. Co-operation between research schools of political science and public administration should be intensified;
3. The original theoretical and methodological work should be more actively presented in international journals;
4. Continuity of research and the originality should receive more attentions than it is getting now.

All rather broad and vague recommendations and the question of course is whether all this did improve the quality of political science research in the Netherlands.

So RAEs may have an effect on the reallocation of research funds and on the quality of research. I like to add a few more effects.
Don Westerheijden, in an article in Higher Education, distinguishes several effects of research evaluations other than their use in university management and in quality improvement (Westerheijden). He argues that research in a university setting is a social process in which researchers depend on colleagues and administrators for input such as knowledge, facilities and money. For output like places to publish and discuss research outcomes and for feedback which consist then of peers' comparison of reputation leading to awards, additional research money. Given this dependence on other actors, Research Assessment Exercises may have the following effects:

1. The practice, at least in the Netherlands, that not individual research is evaluated but research programs has caused a much more intense co-operation among researchers.

2. A second effect comes from the pressure to publish and to publish in international (English language) journals. In all the disciplines we see a clear increase in the publication rate and publication behaviour becomes more and more strategic.

3. RA exercises, according to Westerheijden, have changed the power balance within universities in favour of the administrators to the cost of chair holders, the professors. What used to be a fragmented university dominated by the chair holders has become an administratively integrated organisation. Managerialism has taken over.

4. The outcomes of RA exercises create as he calls it a halo effect. Although the idea is not to influence individual external actors through the reports, still the reputation of

professors, leaders of research projects, is very much made or broken by the evaluation. The evaluations do influence their earning capacity for contract research, consultancy etc.

I like to add some other possible effects as well:

5. The choice for evaluating PROGRAMS and not individual research efforts structures the research agenda in such a way that in the end there will be no room anymore for independent, out of the mainstream, research. RAEs create research orthodoxy.

6. One effect of the pressure to publish preferably in ‘well respected, international (English) journals is that publications in the local language are less valued, leading to a decline of regional political science literature. A related problem, one which is important for political science, is that academics will be less inclined to become engaged in commissioned (often applied) research, since this output is often not taken into account by the Review Committee.

7. A final effect may be a further loss of the unity of teaching and research and therefore a threat to the basic distinction between universities and other types of higher education:
    - often Research Assessment Exercises are done separately from the assessment of teaching programmes;
    - take the example of the British experience where a negative judgement may lead to a complete loss of research funds for a faculty;
    - productive researchers can buy themselves out of teaching;
    - no room anymore for student related research.

Summarising I see as the main pressures for RAEs:

1. Drive for accountability and transparency in the public sector
2. Search for an objective basis for resource allocation
3. Desire to improve quality of research

And as the main of RAEs:

1. More co-operation among researchers
2. Strategic publication behaviour
3. Creeping bureaucratisation
4. Creation of Halo-Effect
5. Trend towards research orthodoxy
6. Decline of regional political science literature
7. Decline of commissioned research
8. Reduced interaction between research and teaching

RESEARCH ASSESSMENT IN BRITAIN

Hugh Berrington
University of Newcastle Upon Tyne

In Britain the research assessment exercise first introduced in the mid 1980s has been an evolving process. And it is fair to say it is very different now from what it was in the mid 80s, and it will almost certainly change in the future. However, I don't think I have the time to give a history of the assessment exercises in Britain. What I'm going to focus on is the present situation, with an occasional backward or forward glance. Now the first thing to note is that it is a nation-wide exercise.

Nation-wide Exercise

There are 4 Higher Education Funding Bodies in Britain, one for each of the component parts of the United Kingdom. They collaborate to run a single research assessment exercise. One of their most obvious tasks is to agree on a common scale and common definitions. What they did first of all was to create a scale that run from one to five, 1 being (research quality that equated to attainable levels of national excellence in none of the sub-areas of activity) - 5 being excellent (international excellence in some sub-areas… and attainable levels of national excellence in virtually all others). They have modified this. They have made it a bit more sophisticated so that 3, which was the middle ranking, is now split into two 3b and 3a, with 3a being, inconsistently by the way with the numerical ratings, better than 3b. For the outstanding submission there is now a 5*. So in effect then, a 7-point scale from 1996. And they try to offer definitions attaching to each point on the scale. It is arguable that the definitions are sometimes ambiguous and sometimes not very helpful.

Main Features

The exercise works through peer review, that is to say the assessing panels are composed almost exclusively of academics in the particular subject. In some of the scientific, engineering and medical subjects you may have one or more members whose immediate background is non-academic - people from industry and so on. The actual assessment is based on departmental submissions. Each department or each subject group in each university is invited to send in a submission, indicating the research done by its staff. In terms of publications this means that they are asked to cite up to four items. In other words, their four best items. The Politics and the International Relational panel in 1996 were eventually faced with a total of 3500 separate items.

Sub Area

This has exercised a kind of almost witch-like hypnotic influence for many years over the Funding Bodies, because they were thinking essentially in terms of scientific disciplines, and trying to apply this across the board in subjects where the concept of the sub area is less helpful. At the last time in 1996, a number of panels particularly in the arts and social sciences, dispensed with the concept of a sub-area (political thought for example or public administration) as developed by the Funding Bodies and decided to treat every individual as a sub area, which greatly simplified the whole exercise.

Quality vs. Quantity

Until 1996 the lists of four items (originally in fact two items) per member of staff was complemented by an aggregate figure showing the number of publications in different categories, the number of books, the number of articles in journals, the number of short works, the number of chapters in books and so on. These were aggregated together and were meant to be an indicator of quantity. One obvious problem is that there is no possible way of checking the veracity of the sort of figures that emerged, and anyone with a healthy scepticism would assume that they sometimes lacked something in precision, as indeed they did. The science-based subjects are still rather keen on this, but for 1996 the Funding Bodies simply decided to abolish the quantitative totals, and to rely entirely on the up to four items submitted by each individual. The quantitative totals will be coming back for those subjects that want them on the next occasion in 2001.

Dual Purpose

The next to note is the dual purpose of the exercise, which originally was to identify research excellence, and at the same time to provide a basis for funding. The two co-incited. If you got a 5 you got a certain level of funding, if you got a 2 you got a lower level of funding. Now what changed things in 1992 was the admission of the polytechnics, now called the new Universities, to university status. It became clear that if you applied the old system most of these would really be wiped out, simply because polytechnics have not been funded to do research, and a lot of their members simply had not been research productive. So what was done, was to say: we don't necessarily want a 100% percent return, but we would like departments to nominate those members of staff whom they wish to be considered, and who would be called research active. That is, a department with 20 members might decide to submit 15 names, or perhaps in case of a former polytechnic only 7 names or even 5 names. The assessment would be based on the research active staff submitted, and this had very important implications for funding.

No Account of Internal Conditions

The last point I want to make, is the fact that there is a rule, that the panels must not take any account of internal conditions in a department. In other words, you can't say: "This really is rather a good effort for this particular institution, because we know it works under very severe difficulties". We are simply not allowed to do that. We simply have to take the returned as it stands. To give a simple example, there was one former polytechnic, which submitted to us in 1996. Among the points that came out was that there was no provision in the contracts of the members of staff for research time. Clearly such a department works under a very hard handicap. But we are simply not allowed to take note of that. So when people say that department x is marvellous, then it may be that department x ought to be marvellous and it may historically have been very well funded resourced.

Grades

In the past, when you returned a 100% of staff and you were given a grade it was quite simple. But now when you submit the names of staff you have to make a calculation. For instance, you might actually do better by submitting 13 staff instead of all 18 and get a 5, where as if you submitted all 18, some of whom would be relatively weak, you might be awarded a 4. So the department then has a fine calculation to make, how many do we submit? There is a trade-off between the number submitted and the grade obtained; and of course the decision affects the volume measure on which your funding is based and the potential grade. But there are two problems. First of all you cannot predict the grade that the18 might get or the grade that the 13 might get. I think that it probably exceeds the wit of most academics to make that kind of forecast. The second problem of course is the fact that the Funding Bodies do not tell us what the formula for funding will be until after the exercise. What they did in 1996 was to make the whole gradient much steeper than it had been before, so funding became more selective. Therefore, people who had made perfectly rational calculations in 1996 on the basis of past experience found that they lost out. So what we have had is growing intensification of selectivity with more funds being channelled to the highest scoring departments.

Consequences

There are two points that I ought to mention in this connection. First of all, Monique Leijenaar raised the point last evening about those departments which do so badly that they are cut out of research funding all together. For example in 1996 those who were graded 1 or 2 were to be given no research funding at any rate until the next exercise occurs and the department has perhaps redeemed itself. However, the first point to note, if you take Politics and International Relations as an example, is that all but 6 departments after 1996 received higher than a 2. One was given a 1, which is the lowest score and five were given a 2. This might seem very hard, but bear in mind that they probably haven't lost anything at all, because they are all ex-polytechnics who did not receive research funding in the first place. So it may be very undesirable and very hard, but they haven't actually lost anything. They are not worse off than they were before, because they were so badly off to begin with!

Criteria

Let's now turn to criteria. Until 1996 the criteria were implicit. "We all know what good research is: we all know what items we should be submitting", which let to a certain restiveness. In effect, people were being asked to play a game without knowing the rules. The rules would only emerge after the game had been played and the scores had been announced. Therefore, it was decided in 1996 that each panel would develop criteria for the assessment. Unfortunately, that was done so late, that is, in the autumn of 1995 and a few months before the closing date of March 1996. No department could possibly have benefited in terms of its strategy for conducting research; it could benefit by knowing what to put in and what to leave out of the return. Let me just give a couple of examples of criteria from the Politics and the International Relations panel: "Authored books. As a general rule the panel will place great weight on research monographs taking into account the level of quality" (the emphasis on quality in its explicit sense is relatively new). Journal articles. Articles published in journals which have high editorial standards will also receive high relative weighting against other types of cited publications". However, we were not allowed to prepare a list, which says for example, that the *European Journal of Political Research* is the top and something else is down at the bottom we were not allowed to do that. The reason for that almost certainly relates to the point that Steen Sauerberg made yesterday about the challenge in the courts made by one aggrieved department in Dentistry.

Emerging Problems

I would like to conclude by talking about some of the problems that have emerged. It is fair to say that over the years the Funding Bodies have not recognised these problems. They have only been recognised in a rather vague, marginal and usually belated sense. The first problem is the common scale. A five star grade should be awarded to a department which has a majority of its sub areas producing work of international excellence, and everyone else producing work of at least national excellence. That is the formula that applies across the board. The problem is to define international excellence. It is highly subjective. A further difficulty that arises is the question whether a 5 in physics is of the same in weight as a 5 in history or a 5 in biochemistry. Funding Bodies pretend that it is of the same weight, but their argument is entirely circular. If one looks for a good example of a tautology I think you will find many examples in the conduct of the British research assessment exercises.

Secondly, I want to mention the census date. The rule is that if you are in the employment of a particular university on the census date, which last time was the 31st March 1996, then everything you have published in the period goes to that institution, even though you may have moved and done most and perhaps all of your work that is cited in your previous department. I find this prejudicial to good research planning and grossly inequitable. It also gives rise to what is called the transfer market, which is an analogy from the football field, where universities are competing against one another in order to get the stars to boost their position for the RAE. The Funding Bodies are rather complacent about it, but I thing it is scandalous.

The last point concerns problems with the items cited. If somebody cites a book, that is fine until you discover that the book was perhaps published before the beginning of the review period. You then find another book which appears to be an authored book but it turns out to be an edited book, which usually will have less weight than an authored book. Sometimes you will find a book that has been cited under one title, actually having a different title, which creates considerable problems. In 1989 the Political Studies Association played a part in drawing the attention of the Funding Bodies to these problems, and the Funding Bodies made a partial gesture by deciding to conduct an audit. They decided to select two departments for each subject and to audit the return to ensure it was accurate. The problem with that is what every criminologist will tell you, that it is the certainty of detection rather than the severity of punishment that is the crucial deterrent, and 2 out of 66 is really no deterrent at all. The Funding Bodies simply refuse to recognise it as a real problem. I can be obsessive about many things and in particular that one, but the length of time I was allotted has already doubled, so I must close now by saying that there are numbers of impending changes which perhaps can come up in discussion.

ASSESSMENT EXPERIENCES:
THE CASE OF DANISH POLITICAL SCIENCE SEEN FROM THE UK

Ken Newton
University of Essex

On request of Karen Siune Ken Newton begins his presentation by outlining his own experiences with assessment as a member of the international review panel, with a special focus on his experiences with Danish assessment.

What we learn from assessment is entirely dependent on what our experience of them is. Being in the British system in a department which is assessed is an enormous pressure, because to maintain our ratings we have to make sure that we maintain our research output. We regard this as something like ratings of restaurants. That is, there is no gossip value in improving your ratings, but if you are a 5 star department and you only get a 5 or a 4 the next time the department is rated, then it is interesting and big news. I am partly a victim but also partly a torturer. I was a member of an international review panel for Denmark a couple of years ago, and I will talk mainly about my experience as a member of that panel.

I see assessment as a process. To try to organise my fairly random thought, I will put them into this flow of events:

- Who sets up the panel of the assessment process? Is it official, private, commercial or ministry?
- Who decides what criteria that should be used?
- Who decides who is going to sit on the panel?
- How do they work?

At the end of this flow process there is the impact and feedback loop as well. Evaluation has an impact ON what it is that we all do as political scientists in our everyday life.

The Agenda and the hidden Agenda

In the beginning of this flow of events we realised that there was an agenda. The agenda was to evaluate Danish political science to give decision-makers, ministries, ministers and departments some indications whether they should continue funding and whether they should do anything about the quality of political science in Denmark, which by the way it very high. We were also aware of the possibility of hidden agendas. Was something going on that they did not tell us about? We knew that partly what was going on was an attempt by a relatively new department to legitimise itself by showing its capacity to set up this review panel in the first place. There existed bits of bureaucratic politics and we knew that we were simply pawns in that game. There are also other hidden agendas. That is, there are also assessments, which are simply used as fig leaves to justify the minister of the department doing whatever it wanted to do in the first place. They would read the report selectively and conclude that they wanted to close the whole bloody lot down, because they were no good. As proof they would then pick out the three lines that suggested that, and forget the thirty pages which suggested the reverse. We were very conscious of the fact that there were hidden agendas. We knew they were there, but we did not know exactly what they were, and it did make us a little bit cautious.

Inside and Outside Assessors

There is a difference between having inside assessors and outside assessors. It makes a difference whether they are inside the system or outside the system. It also makes a difference whether is it a national system or an institutional system. You have to have outside assessors, but it is a dangerous ploy for those who set up the assessment panel. If you have inside assessors they are likely to be accused of partisan behaviour unless you, like the British system, have pretty clear roles of engagement. In the Danish review, we were provided with a lot of background information and a huge amount of statistics. We took a look at it and discounted most of it, because we did not think that it was very helpful. We used some of the background information, but we were very dubious about this text that we were supplied with. We did not through them all away, but we used them very selectively. It did not really matter because we could visit every institution, and we did that. We thought that it was very important to do that, and moreover we found it very important to stick together as a review panel, so far as it was possible. We did not divide. As a result of that, we all got the same impression of the same institution, which we thought was a very important principle. As soon as you have different panels and different sub-panels reviewing different bits of the system then you are lost. I suspect this is what happens in Germany, where completely different people possibly with completely different criteria evaluate the blue list institutions. Therefore the result is that quite good institutions being judged by very hard criteria and quite poor institutions being judges by much more lenient criteria.

Evaluating Danish Political Science

When we evaluate, we try to understand the particular context of Danish political science. In the beginning we were worried about the language problem. We were very continuos of the fact that quite a lot of Danish political science is published in Danish. Only one of our members could read Danish, so we could not evaluate that. I cannot speak for other members of the panel but my first thought was that we were asked to judge Danish political science by international standards, and therefore we should judge the English language populations as a higher order and higher standing than Danish publications. However, I very quickly came to the view that it was completely wrong. If only because it is important that political science writes in policy areas and writes documents which are of interest and importance to bureaucrats, administrators and politicians. If you are going to do that in Denmark, you have to write in Danish. Therefore, I came very quickly to the view that the language problem was much less important than I assumed at the outset.

We were also very concerned about what we thought was considerable inbreeding in Danish political science. We were concerned that people were born and brought up in one department and therefore seemed to spend all their lives in one department. In fact there was a rather funny episode in Aarhus. The German member of the international review panel asked me why I was so worried about inbreeding when Oxford University was so terribly inbred, and without thinking I said: “Yes, that was why I left the place”. Then we realised very quickly that inbreeding in a small country like Denmark is not terribly important, because we were enormously impressed by the amount of co-operation between the departments and institutes. There seems to be an enormously dens network of research co-operation which actually overcame the inbreeding problem.

The Impact and effect of Assessment

What impact have assessment had? In terms of the British review, which is now regular, we do have a bit of a publishing cycle. There is a last minute rush to publish and to meet the date for the research assessment exercise. We use to have a political economical cycle, but now we have a clearly review publicising cycle. Another thing is that assessment does actually improve the quantity of publish material. I am sure as a result of the research assessment exercise British political scientists are mush more productively than they use to be. They write and publish much more. I do not know if there are any clear statistic figures about that, but I am pretty sure the quantity has improved. It may be that the best thing that the government could do for the quality of political science is to abolish the research assessment exercise. You might do a favour to everybody else by publishing much less.

We spent a lot of time and a lot of money on the assessment, but did the Danish assessment have any effect? It was very enjoyable and a very useful and interesting experience but at the end of the day I do not think that we had any impact at all. The research assessment exercise pulls in £750,000 to the University of Essex in Britain, that is what our 5 star is calculated to be worth. Incidentally, I do not thing that the department of government of Essex sees a penny of the £750,000. The money goes to the university. Most of our students are undergraduate students and they are completely unaware of our research assessment rating.

Criteria Create Pressure

One important effect is the fact that research assessment and evaluation effect what we do in our daily lives. It has an impact on our lives and of the type of publication that we produce. We are under very heavy pressure to publish, what I hope you think is absolutely first class refereed journal. We are also under a pressure to get in research grounds. The Essex department has been rather successful in the research assessment exercise on the basis of rather little research money. We apparently did not need a lost of money to produce a fair quantity of high quality output. Whether we like, want or need it or not we are under pressure to go out and get research money, because that is one of the criteria.

I was involved in running a very large comparative project, which I think was a good and useful think to do. We think it was reasonably successful, but I will never do anything like that again. Whether it is useful to political science or not, it did not do my contribution to the department’s research assessment rating much good at all.

At this point Hugh Berrington interrupts to ask how Ken Newton has come to the conclusion that his effort on the comparative project had no influence on the department’s research assessment rating. Ken replies:

I cannot know, but it took me 4-6 years of my life, and in that time I could have produced 15 refereed journal articles, which would have solved my research assessment problem in the department. I will not do that kind of thing again. From my point of view and the department’s point of view it is a waste of time. From political science’s point of view it may be a sad lost. Not necessary in this particular case, but in general.

We were provided with a set of criteria in our background documents in the Danish review and they are clearly absolutely crucial. They tell the members of the review panel what they should do, and how they should do it. I most have read those criteria 15-30 times.

The 30th time I still could not understand them, so I through them away. I think that members of Danish political science can be well satisfied, that we paid very little attention to those criteria. It is said the social scientists are more intelligent than the method than the methods they use. In this case, I think we certainly were, because we did not pay any attention. We simply took the view that we may not know very much about political science, but we know good political science when we see it.

The last effect, which is of high importance, is very simple this: If you can get on an external review panel for a foreign country - by all means do it. It is a very interesting exercise, but I can tell you that we were fed and wined in that week better than I have ever been feed and wined before, and in the long run I am sure it is going to knock six month of my life.

THE ASSESSMENT SYSTEM IN IRELAND

Michael Laver
Trinity College Dublin

Within Trinity College I am the chief academic officer and as such have the job of running the assessment process for every department in Trinity College. I have only been doing it since the 12th of July and have come here today to learn how to torture people as much as anything else. Another preliminary remark is that I find it interesting to come right after the British speakers. As with many things in Ireland, the Irish are at the same time horrified and fascinated by the British, and this also applies to the attitude we take to assessments and the assessment system.

A Small University System

It is very difficult to understand any process without understanding the politics behind it, which is why it is important to understand the political science of the assessment process. We have had references to two systems at the moment, a very big system in Britain and a small system in Denmark. The Irish case is one of a small university system. There are effectively 7 universities. Two clusters are Trinity College dating from 1592 and the National University dating from the 1850s, which has 4 campuses. We can think of these two as the old universities. The Irish equivalents of new universities are Dublin City University and University of Limerick created from what were effectively polytechnics in the late 1980s and early 1990s. These facts are important because to understand any process one has to look at the coalitional structure and the power politics behind it.

There is a cartel of Irish universities linked in an association called The Conference of Heads of Irish Universities. CHIU are divided into groups one being a cartel of the presidents and an other being a cartel of people like me, who are the people responsible of the academic side of the universities. These committees meet to discuss policy and confront the government on many issues. The unity of that coalition is very important. However, the decisive structure of this game is such that essentially a cartel of Trinity College and University College Dublin can typically say no to the government, because these are two very powerful universities. If the cartel of Trinity College and University College Dublin was to be split then the government could write roughshod over the university system. This is the power structure that lies behind the assessment system.

The Irish system is funded by a state agency, which is the Higher Education Authority (HEA). Unlike the British system the Irish demographic structure favours the universities very much, and will do for the next 10 or 12 years, after which the demographics will turn against the universities in the same way as the British demographics turned against universities in Britain. There is still a large quantity of students of university age going through the system. Eventually, however, we know that the university-age population will decline. We know this because these people have already not been born. It will decline in about 15 years time and therefore the government is unwilling to build more universities, for what is really a temporarily population bulge. This means that there is at the moment a massive competition on the part of Irish undergraduates to get into universities. The universities turn away many students and export them to Britain and Northern Ireland. We thus have more students than we can deal with; we have very high admission standard and we are turning many students away. This is very important to know because we are not under any demographic pressure or threat as far as the educational system is concerned.

Voluntary self-assessment recently introduced

My predecessor in Trinity played a very large part in designing the Irish system, as one can in a small country. He was a professor of management, and he was something of a management guru to the Irish civil service. He was a very strong believer in what we call the QI/QA process - Quality Improvement/ Quality Assurance. The whole Irish assessment system is based on a fundamental distinction between these two concepts. To a large extent the research assessment exercises in Britain are what we would call a quality assurance process, which is designed to assure people on the outside that everything is okay. However, that is very secondary to the management philosophy that underlines the Irish system, which is based much more on quality improvement. This means that you take a decision-making unit that is capable of improving itself and put in place a system under which that unit will improve the quality of its output every year.

Due to the influence of my predecessor in this process, the Higher Education Authority has relatively recently been persuaded to fund a pilot QI/QA programmed in Irish universities. It is entirely voluntary and organised by the Conference of the Heads of Irish Universities. It is a self-regulation process and is enthusiastically embraced by the universities because if we do it this way, and we can be shown to do it well, then the government will hopefully keep its nose out of our affairs. This will also mean that we will be seen to be engaging in an active process of quality improvement.

The fundamental unit of analysis in any QI process is an autonomous decision-making unit. This is fundamentally based on the university department and not on a discipline. There are no disciplinary reviews and thus no general political science review in Ireland. Everything is based on a departmental review process. The fundamental unit of analysis is the university and the sub-unit within the university is the department, which is fundamental to the whole quality improvement philosophy.

The departmental reviews that we engage in are holistic in the sense that we do not separate teaching and research. These two elements are intimately related to each other. There is no separate evaluation of teaching and research but an explicit evaluation of the interaction between teaching and research. It is held to be of very high value that the teaching of a department is animated by its research. Departments are often criticised if their teaching is not animated by their research.

They spend two days in each department extensively talking to members of the department and some of the students. They also visit the library and other facilities and form an assess-ment of the department based on that. For the most part, the assessors who come are British. It is essential that they are almost always from outside Ireland because Ireland is a very small country and nobody wants to assess colleagues. To a large extent the British assessors often do make comparisons with the ratings used in the British RAE. Even though these ratings are focused upon, we tend to find that, if the relevant Irish department was informally rated as being a 4 or a 5 on the British RAE scale, the department, while totally rejecting the British assessment exercise, would quote those assessments anyway.

The quality improvement philosophy is that once the external assessment comes in, the university has to consider it. The policy priorities of the university are concerned with what to do to improve the quality of the department, which is the policy implication of the assessment. The departments have actually been very keen to be reviewed, because most departments have got additional resources out of a review, whether the review has been good or been bad.

Lessons learned to date

Thus far this has been a very consensual process in Ireland because there has been money to spend on departments being reviewed, but I am quite convinced that we are in a honeymoon period at the moment. The honeymoon arises for essential two reasons. First of all, there are resource reasons. Secondly, we are setting in place a tracking system to track each of the recommendations that are made in an assessment and report back within 18 months. At the moment we have the reports, but we have not yet began to track their implementation systematically. We do not know yet whether we have got value for the money we have spend in order to improve the department that has been reviewed. What will happen over the next three years is the implementation of a much more explicit tracking system. The key problem has thus yet to be faced since, ultimately, dealing with problem departments on an intra-university basis may well prove divisive if managed badly.

University of Dublin, Trinity College
Departmental Reviews

Section A
Procedures

1. Initiation of Departmental Review Process

- At the request of the Faculty Dean, Council and Board consider whether to instruct the Senior Lecturer to initiate the Departmental Review process.

2. Nomination of External Reviewers

- The Dean of the Faculty is asked by the Provost to nominate External Reviewers. Normally five nominations are made, from which two or three are appointed External Reviewers should not normally have had a close association with the department.
- These names are considered by a working group comprising the Provost, the Senior Lecturer, Dean of the Faculty and College Secretary. Additional names may be considered at this stage.
- The names of the proposed External Reviewers will be brought to Council for consideration and approval.

3. Conduct of Departmental Review

- Following approval by Council, the External Reviewers are invited to undertake the review.
- The review normally takes place over two days during teaching term.
- The External Reviewers visit the College, meet the Provost and are briefed by him, the Senior Lecturer, the Dean and the College Secretary.
- They carry out the Review - normally meeting members of the Department and a number of postgraduate and undergraduate students. They see facilities (including library holdings) as appropriate and meet members of relevant related departments.
- A draft report is submitted to the Senior Lecturer within three weeks of the visit.
- The Senior Lecturer consults with the Dean and any others considered appropriate on matters of factual accuracy in the draft report.
- The Senior Lecturer communicates comments to the external reviewers and requests a final report.
- The final report is made available to the Department and the Dean, both of whom have an opportunity to comment to the Senior Lecturer.

4. Council consideration of Review

- The University Council has the primary role in deciding on a response to the Report of the External Reviewers.
- The Provost summarises the Report and comments from the Department and the Dean, and presents a summary report to Council.
- Council decides on the appropriate action following consideration of the summary report.

Section B
Self-assessment

The following are sent to External Reviewers at least three weeks before their visit.

1. General Information

- Curricula vitae for all academic staff.
- A full listing of all staff, including technical, secretarial and part-time staff.
- The total FTSE of the Department, by programme, and the overall student-staff ratio for the previous five years.
- Position in College (e.g. Faculty base, relationship to other departments).
- Departmental/course handbook(s).
- Copies of relevant entries in Calendar.
- Any booklets/pamphlets/brochures of general nature (e.g. material sent to schools, prospective postgraduates, etc.).
- Website address.

2. Teaching

In relation to the teaching activities of the Department, the following is provided:

- A clear outline of the teaching programmes in which the Department is involved.
- A clear outline of the distribution of teaching (lectures and classes) by staff member.
- The syllabus for each programme/course, showing methods of assessment used, reading lists, etc.
- A statement of innovations in the teaching programme m the previous five years
- Details of student intake over the last five years to each of the programmes (including entry points).
- All assessments/reports in previous five years by external examiners or others (e.g. accreditation by learned societies/ professional bodies).
- The results of student evaluations m previous five years, where available, of programmes/courses with which the Department is involved.
- Distribution of degree results for the Department’s graduates, for the previous five years.
- Information, as available, on the progression of students immediately following graduation.

3. Research

In relation to the research activities of the Department, the following is provided:

Staff

- A full listing of all published research of Department staff in the previous five years.
- A listing of all significant work in progress by staff that is likely to lead to publication.
- Full details of research funding received by the Department in the previous five years.
- Full details of other activities of the Department that arise from the research standing of its members (e.g. membership of government-appointed commissions, officerships of learned/professional societies, editorship of academic publications).

Research Students

Details of the postgraduate research activities of the Department, including:

- numbers of research students in the previous five years;
- a table listing research students by supervisor, titles of these and/or current research topics of these students.
- publications or pending publications of these students;
- research funding (internal and/or external) for research students;
- overall supervision/support for graduate;
- rates of completion and average completion times of theses;
- brief summary of how these students have progressed on completion of these.

Postgraduate Taught courses

For each taught course, the following are provided:

- course handbook;
- student numbers over the past five years;
- sources of funding available to students;
- overall supervision/support for these students;
- completion rates;
- a brief summary of how these students have progressed on completion of course;
- external examiners' reports for the past five years.

4. Resources/Facilities

In relation to resources and facilities of the Department, the following are provided:

- The sources of funds to the Department
- Relevant information on teaching and laboratory, computing and library services, and equipment, available to the Department.

5. Organisation and Management

- Structure and management of departmental activities (e.g. department meetings, decision making, planning, departmental roles/responsibilities, course/ year co-ordinators)
- Information on the contribution that members of the Department have made in the previous five years to the administration/government of the Faculty/College (e.g. membership of Faculty/College committees, posts of responsibility at Faculty/College level) and broad resource implications of this.

6. Departmental self-assessment and perspective on future

7. Appendices

- The organisation and funding of higher education in Ireland.
- Faculty overview, including three-year planning framework (provided by the Dean).
- Information on the Library - e.g. size of collection, copyright, number of reader places, spending on books and journals by the College, by Faculty, and for the department being reviewed.

Section C
The External Reviewers' Report

1. Briefing

With effect from the academic year 1997/98, the University introduced a process whereby each Department will normally be reviewed every five to seven years. The review is based on principles of self-assessment followed by peer review. The process is initiated by a Council, at the request of the Faculty Dean, and a summary report and recommendations arising from self-assessment and peer review is presented by the Provost to the University Council. Flexibility with regard to staffing resources is determined, in the first instance, by Faculty three-year plans (all Faculties have submitted such plans for the period 1998-2001).

The Reviewers are asked to submit a report of approximately 6-12 pages. Normally a draft report is submitted within three weeks of the visit (see Section A3 above on the Conduct of Department Review). It is preferable that this should be a joint report.

The Reviewers are asked to provide the College with an assessment of the Department's standing compared to departments of international repute in this area, taking into account the size of the Department. In making their relative assessment, it is important that they indicate the departments with which comparisons are being made, whether they see the Department as having a standing in the top, middle, or lower ranges for such departments, whether the Department is of appropriate size, and whether the focus and performance in teaching of the Department is satisfactory.

In so far as is appropriate, the Reviewers are requested to comment on the issues outlined below.

2. Teaching

(a) Please outline which of these you see as the key teaching programmes for a Department such as this, giving reasons for your views.

(b) Where possible, the Department should be assessed in terms of at least two* of its main teaching programmes, in relation to other departments of international standing in this area. Give reasons for your assessment, commenting in particular on the following:
* (where a postgraduate taught programme exists in the Department, it should be one of the two programmes assessed)

- content/level of the programme;
- overall coherence/design;

- pedagogy, professional standing (e.g. accreditation by professional bodies, etc.);
- opportunities for study abroad/relevant outside experience;
- student evaluations.

(c) Comment on the distribution of teaching across staff members.

3. Research

Staff

(a) Assess the research of the Department both overall and in terms of the following:
- publications in refereed journals in last five years;
- likely publications in such journals in the next two years;
- publications in other forms - books, monographs, etc. in the last five years;
- likely similar publications in next two years.

(b) Comment on the distribution of research output across Departmental members.

(c) Comment on the research output of staff in a national/ international context.

(d) Give an assessment of the standing of the Department in terms of published output relative to that of other departments of international repute.

(e) Evaluate the Department in terms of its performance in the last five years, and likely success in the next two years, in raising external funding for research and/or teaching.

(f) Comment on other activities of the Department that arise from the research standing of its members, e.g. membership of government-appointed commissions, officership in learned/professional societies, editorship of academic publications, other achievements and standing arising from research work.

Research Students

Comment on the graduate student research activity of the Department in terms of:

(g) number of research students;
(h) research funding (internal and/or external) for research students;
(i) overall supervision/support for these students;
(j) completion times;
(k) quality of research output.

Postgraduate taught courses

Comment on the postgraduate taught activity of the Department in terms of:

(l) number of students;
(m) finding available to students;
(n) overall supervision/support for these students;
(o) completion rates;
(q) quality of programme.

Balance

Comment on the balance of published research, research supervision, and other research-related activities in the Department.

4. Resources/Facilities

(a) Give an assessment of the resources of the Department relative to those available to other departments of international standing.

(b) Assess the Department in terms of the following resources available to academic staff in the Department:

- support staff (e.g. technical, secretarial);
- teaching facilities (e.g. laboratory space, lecture/ seminar rooms, equipment);
- research facilities (e.g. laboratories, equipment, computing, library).

5. Organisation and Management

(a) Comment on the Department in terms of its overall organisation, both internal and its relationship with other relevant departments (e.g. how business is conducted; departmental roles/responsibilities, committees, decision making; planning).

(b) Comment on the contribution which the Department has made in the last five years to the overall government of the College in terms of membership of College committees and the holding of positions of responsibility.

6. Overall view and recommendations

(a) In the light of what 15 happening in other universities, comment m the Department’s self-assessment and view of the future, and in particular on its curricula and research in the last five years.

(b) Give your own views on the possible future direction of the Department.

Section D
Follow-up to Departmental Review

(a) The Department will be invited to respond within eighteen months on progress ’m addressing the recommendations arising from the review process.

(b) A report should be sent to the Provost, for discussion with the Senior Lecturer and the Faculty Dean.

(c) The Provost will subsequently report on these issues and the College’s response to the University Council.

QUALITY ASSESSMENT OF RESEARCH IN THE NETHERLANDS

Drs. Roel D. Bennink
Association of Universities in the Netherlands (VSNU)

The Association of Universities in the Netherlands (VSNU) has set up a system of quality assessment of education and research which is now in operation since 1988 for education and since 1993 for research. In this paper some aspects of this system are presented, focusing especially on the general characteristics and the information provided by the faculties.

1. General description of the system

The system of quality control consists of education programme reviews and research assessments. Both reviews are carried out by international committees of peers and are discipline oriented. This means that the programmes of one particular discipline in all Dutch Universities are reviewed by an International Review Committee every 6 years.

2. Positive and problematic aspects of the system

Evaluations of the system have shown that the positive aspects predominate. After the first round only minor adjustments have been made. Especially fruitful is the integration of the *national* reviews with the permanent local quality control, through the ’self-evaluations’ which the faculties present to the review committees. The quantitative comparison of groups or faculties has been a problematic point, because standards and definitions for quantitative data are variable and disputable for a number of reasons, and because management information systems and performance indicators are still under development.

The costs of the system are roughly DFL 27.000 per faculty for the VSNU-Committee and a total of 0,5 full time equivalent of academic staff in one year for preparing the documentation and for the site-visits. This is a total of about DFL 75.000 per faculty (about US $ 44.000). The system combines quality control with external accountability. It is sometimes that this could cause tension between honesty and strategic behaviour.

3. Different approaches in a number of disciplines

The state of the art in a discipline in terms of scientific achievement and critical capability largely determine what can be achieved in the national reviews. In some disciplines the role of the review committee is to establish a mature self-confident and analytical atmosphere by describing the strong and weak points of the discipline and by pointing out strategies for development. In disciplines with a longer experience in mutual or external assessment the reviews can have a more routine character and can build upon existing procedures. In such cases care must be taken to avoid overlap and *to integrate the existing systems of quality* control into the national approach to achieve maximum comparability.

Bibliometric analysis of the research output is especially fruitful in those disciplines that are internationally oriented and have consensus about a set of top international journals.

4. The information provided by the faculties

The faculties provide extensive qualitative and quantitative documentation for the reviews about finance, staff input, number of students, titles of publications in a number of categories over a five year period, mission statements, etc. The information is clearly defined in a number of VSNU-publications. This information is presented in Table 1. The *standardisation* of this information flow in terms of definitions, responsibilities and public use, has a positive impact on institutional management and its information infrastructure.

The standardisation of the definitions in the area of research was helped by distinguishing between the *form* and the content of publications, and by clearly separating research publications from other types of university publications (see table 2 and table 3).

TABLE 1: THE INFORMATION PROVIDED BY THE FACULTIES

(A) Research reviews

• Income (direct funding, contract research)
• Costs (personnel, all other)
• Total human resources (academic staff, support and administrative staff)
• Number of students
• Per programme: - title, subprogrammes
- programme design
- overview of results
- plans for future research
- relevance and other indications of quality
- five key publications (3 copies)
- dissertations (titles)
- publications (titles)

Quantitative data: - input of academic staff per year
- composition of the academic staff (professors, senior/junior staff)
- output (number of dissertations, scientific publications)

(B) Education reviews

Quantitative data: - number of first-year students (male/female, part-time/full-time)
- total enrolment
- success rate after 1 2,3,4,5,6 years
- number of academic staff per category (male/female, with/without Ph.D.)
- student/staff ratio

Description: - aims and approach
- programme, courses
- thesis
- number of students, success rate
- facilities, infrastructure
- alumni
- staff
- internationalisation
- internal quality control
- strengths and weaknesses

TABLE 2: Separation of tasks, output and target groups

TASKS	OUTPUT	TARGET GROUP
TEACHING	LECTURER NOTES, HANDBOOKS, PROGRAMMES, DEGREES	STUDENTS
RESEARCH	RESEARCH PUBLICATIONS, PROMOTIONS, CONGRESSES, KNOWLEDGE	RESEARCHERS
		PROFESSIONALS
		GENERAL PUBLIC
SERVICES	REPORTS, ADVICE, SERVICES	GROUPS AND INSTITUTIONS IN SOCIETY

TABLE 3: Separation of form and content

CONTENT FORM	PATENTS	POPULAR PUBLICATIONS	PROFESSIONAL PUBLICATIONS	SCIENTIFIC PUBLICATIONS	DISSERTATIONS
BOOKS					
PARTS OF BOOKS					
DISSERTATIONS					
JOURNAL ARTICLES					
EXTERNAL REPORTS					
ANNOTATIONS					
DESIGNS					
OCTROOIEN					
CONGRESS PAPERS					
OTHER FORMS					

5. Information systems, data models and Internet-access

Both in the domain of the student and education programme information, and in the domain of the research information there are several administrative information systems. In both domains the development of these systems and the exchange of data between universities is facilitated by developing data models which describe the data elements, their relations and definitions.

Most universities in Holland have central research information systems, or are in the process of implementing them. These systems create multi-purpose databases with titles of projects and publications, etc., for faculty, university, reports and quality control. Good examples are: UT-OIS, KUN-OZIS. These systems have common characteristics because they are based on the combi-format data model (VSNU, 1992). The data model identifies and defines the main data to be collected and stored in the system, and the relations between the data-elements.

Such a common data model allows you to:

i. exchange information
ii. exchange information systems
iii. reduce the number of enquiries
iv. provide combined Internet-access

6. Effects of the assessments

The results of the external reviews are published in public reports. Interviews in the universities have shown that the following effects are visible:

- the assessments create an increasing pressure to publish research results, especially in international journals with a *high impact factor;*
- the assessments provide a solid basis for managers to use the instruments of quality control at their disposal; (without the *justification* provided by the assessments many decisions would not have been made)[1];
- the assessments cause a shift of power towards the management of the faculty and the university;
- the importance of the *research policy* at university level is enhanced; (larger central stimulation funds, allocation based upon the assessments and/or strategic considerations);
- the assessments directly influence the *prestige* of the researchers; (research quality is still more prestigious than teaching);
- the assessments increase the relevance of the regular consultations between faculty and the central university level; (this increasingly leads to medium term agreements and management contracts)[2];
- the public reports have made it virtually impossible for weak and unproductive groups to continue to exist unnoticed.

[1] But the quality of research is not the only factor that determines the research policy; the profile of the university and the need to cover the whole range of the education programmes are also relevant.

[2] Effects in the funding at *national* level are considered as undesirable. The direct government funding is regarded as basic; money from national research councils or other funds is regarded as additional and can be increased on the basis of a positive assessment. Prestige will also help in acquiring contract research projects.

References

Frederiks, M.M.H., D.F. Westerheijden and P.J.M. Weusthof, *Effects of Quality Assessment in Dutch Higher Education,* 1994, European Journal of education, Vol. 29 (2), 181-199.

Don F. Westerheijden, E*en solide basis voor beslissingen* (A solid basis for decisions; effects of external research reviews by the VSNU, in Dutch), research report for the VSNU, May 1996.

Assessment of Research Quality - Protocol 1998, Protocol for Quality Assessment of Research, Utrecht: VSNU, 1998, ISBN 90-5588-074-4).

Gids voor de onderwijsvisitaties (Guide for the external assessments of education, in Dutch), Utrecht: VSNU, 1995, ISBN 90-5588-004-3).

Definitie-afspraken Wetenschappelijk Onderzoek (Definition-agreements about the classification of research publications), Utrecht: VSNU, 1994.

Het Combiformat, een gegevensmodel voor onderzoekinformatievoorziening (The Combiformat, a data model for research information), Utrecht, VSNU, 1992.

EVALUATION OF SOCIAL SCIENCES RTD IN BELGIUM

Luk Van Langenhove
Science Policy Office

It is always interesting to be the last speaker in a panel, because most of the things you really wanted to say have already been said. As the last speaker you have two possibilities. You can either choose to be very well prepared having made beautiful graphs and illustrations, which you would love to put on display but time is lacking, or you can choose to come unprepared, tell the audience that you will be very brief and move on to the coffee break. In this case, I have not prepared anything…

I am not a political scientist but a psychologist. At the moment my main occupation is that of a civil servant. I am the acting secretary general of the *Belgian Federal office for Scientific Technological and Cultural* A*ffairs*. As such I have come to realise two problems with assessment in the social sciences, which I very briefly will tell you about, but first of all I will make a general remark on the assessment situation in Belgium.

In Belgium there is no tradition for assessing institutions. We do not have the previously mentioned rating system. The only things we give stars too in Belgium are restaurants. The emphasis in Belgium is more on the assessment of individuals. However, there is a tendency, especially in the Flemish part of Belgium, to copy the Dutch system of self-assessment and the visitations of university departments.

An Outlining of the Main Problems

I see two main problems in social sciences research assessments on federal level. One of the problems has to do with basic research and the other with more applied research. In both cases the political and the social science issues are quite interesting.

About 11 years ago the Belgian government made an analysis of the Belgium performance in science. One of the crucial results of that analysis was that most of the research is done within the university system, but in Belgium we do not have a system of big research institutions. The government decided that they had to do something to keep Belgian science competitive and set up a scheme of funding of basic research. However, the funding was not attributed to a single department or to a group of researchers but to a number of departments across Belgium. The idea was to set up virtual centres of excellence. Belgium being Belgium, one of the constraints was that at least one of the university departments should belong to another linguistic group than the other departments. So it was also aimed to rein-force cohesion within the Belgian science system. Since then, every fourth year, a new set of about 30 to 40 virtual networks is being launched.

Each network consists of about 4-5 partners with, as I mentioned before, at least one of them belonging to the other linguistic group. They are funded to do basic research on a specific project that they have defined themselves. They are funded for 4-5 years, but a prolongation is possible which means that it is possible to be funded for a maximum of 10 years.

The very first set that started 11 years ago was set-up without any form of ex-ante assessment. This resulted in the fact that all the vice chancellors of the Universities came together, and decided how to share the money between themselves. Some quite interesting networks came out of this. What happened was that most of the universities promoted their "big stars" and together they formed groups. Everybody was happy because they now had 5 years with funding, and they did not have to do much in return. But then of course time also changed in Belgium, which meant a demand for assessment. People agreed to set up a scheme for assessing the projects, but it proved to be very difficult. First of all, every group of scientist was able to put forward a project. Their hands were totally free, which meant that the project they put forward could be in what ever they wanted it to be in. Secondly, it was basic research, which meant that the aims of the projects were, by definition, not very well defined. In 1997 we ended up with 150 proposals that we had to assess. It was impossible to organise the assessments with Belgian peer reviewers, because all Belgian peers had been working together on the project. Therefore, we had to look for foreign peers. We often used people from the Netherlands and people from France, because we needed people who spoke either French or Dutch. This proved to be very difficult to organise. But we did manage to organise it.

The next problem was the selection between the 150 projects. We ended up with a list of 40-45 projects and we only had enough money to finance 35 networks. We therefore negotiated with the vice chancellors of all the universities came out with a consensus proposal. We had selected the 35 networks we were going to fund. But, contrary to the previous sets of networks, the government now wants us to assess the networks mid-term and ex-ante.

Meanwhile, we also started to assess the impact of the previous set of networks to prepare this new assessment. We very naively thought that the result of reinforcing people to work together would be joint publications, but much to our surprise the number of joint publications did not increase. To find out what the problem was we began to interview professors. We then realised that for several of the researchers, the whole issue of collaboration is nothing but a set up. They claimed to be collaborating because they wanted the funding, but actually everybody continued to do their own research. This seems to be especially a problem in social sciences!

This faces us with a fundamental question: is it possible to ’force’ social scientists to work together on basic issues? And what does collaboration really means in the social sciences? Some social scientists claim to be collaborating when they are organising a joint conference with their colleagues... This is definitely a major problem, and one of my main reasons for coming to this seminar is to share this problem with you and listen to your feedback. Part of the problem is the fact that the classic quantitative approach does not work when assessing this kind of issues.

The second major problem I want to raise is the fact that once you have signed a contract with scientists they have a tendency to hire Ph.D. students and give the research work to them. The scientists do supervise the students’ work but they do not have any contact with the research work besides that. Even though, the scientists ensure in their submitted proposals to the government that they will come up with some results when the project is over. The results are often very interesting books and publications, but within the government, within the administration and within the ministries people want to change this procedure because it is not useful. Publications as such do not help to tackle the problems decision-makers are facing and for which they seek help from the social sciences. People who at first believed that there was a role for social science to be played in helping to decide what decisions to make are now very disappointed.

How can we solve these problems? What can we do to assess the results of social science research, and how can we organise the doing of social sciences in order for it to have some practical use? Our tactic is to combine the issue of funding research with so-called 'parallel' actions. We want to organise consensus conferences within ministries and bring lay people from the ministry together with scientists forcing them into dialogue. Some of my colleagues in social sciences claim that they do not have time for these conferences. The scientists want to be left alone with their research, which is a major problem for a civil servant like me, who's role it is to stimulate research in order to be useful to the government.

(*) Deputy Secretary-General of the Belgian Federal Office for Scientific, Technological and Cultural Affairs (OSTC). The views expressed in this paper do not engage the OSTC

R&D EVALUATION IN ITALY

Alberto Silvani
ISRDS – CNR (Institute for Studies on Scientific Research and Documentation)

Evaluation is becoming a key word for present governmental policies in Italy: its increasing role deals with modernisation in Public Administration and with a diffused request of account-ability.

Concerning R&D, the starting point of research evaluation in our country is at a low level. Evaluation is mostly confined to the selection of candidates for financing, it is basically carried out by the members of the scientific community, till now it was not a part of Government management procedures, and, when carried out, it has had little effect on decisions[3].

However in the last three years relevant progresses have been achieved both in universities and of public and private organisations. From a structural point of view, research evaluation has become an increasingly important issue in Italy as indicated in the current reorganisation process (re-ordering) of the national research system.

Not only evaluation is one of the main issues dealt with in the Report to the Parliament submitted by the Research Minister last year, but it is also included in various pieces of legislation, such as the one on the so called “brain of the research system”.

Following the main guidelines of intervention, in order to address the reform, a survey has been carried out to analyse institutions and activities related to research and innovation[4] and a parallel inquiry has shown the weak points of our research world which are essentially:

- insufficient human and financial resources;
- insufficient systemic approach;
- lack of systematic assessment;
- insufficient planning and weakness in the formulation of strategic programmes;
- inadequate evidence and diffusion of results achieved by the research system.

The first legislative act, coherent with the main guidelines above mentioned is a decree which has been adopted in June 1998. This decree establishes a new structure for the national research system following these basic elements:

- definition of a model of co-ordination and planning of the scientific and technological research referring to the role of MURST (Ministry for University and Scientific and Technological Research) and CIPE[5] (including specific actions that must take place at inter-ministerial level in order to assure harmonisation with various Ministerial policies). A general economic document, (DEPF - National Economic and Financial Planning Document), approved by Parliament, and a National Framework Programme (PNR), defined at CIPE, will be the operational tools which, according with the parallel

[3] A larger description (“A science policy view”) can be found in Silvani-Sirilli, *Research Evaluation*, vol. 5, No 1, April 1995, pp 69-78.

[4] In preparing this report, MURST has been involved in checking all existing bodies and activities (tools, incentives etc.) devoted or linked with R&D and innovation, including the revision of Directory of projects and institu-tions (“Anagrafe della ricerca”). The final report (“La Riforma del sistema ricerca Italia”) has been published in the special issue of *Università Ricerca UR*, anno VIII, No 3, 1997.

[5] Interministerial committee for economic planning, which includes all Ministries related to economic affaires.

EU Programme, allow to steer national research policies and involve public and business administrations;

- establishment of a Special Integration Fund, available to CIPE, for providing specific and strategic interventions;
- direct responsibility to the Minister for Universities and Scientific and Technological Research, who is in charge for co-ordinating CIPE activity when related to R&D, with the support of a Technical Commission and different Public Administrations involved;
- institution of two kinds of advisory bodies:
  - a Committee for research policy (CEPR), whose members are experts from scientific, social, cultural and production sectors, supporting MURST to take appropriate decisions;
  - a National Sciences and Technology Assembly, organised in National Scientific Advisory Councils, for the main scientific areas, as expression of the scientific economic and social communities;
- appointment of an Evaluation steering committee (CIVR) promoting and spreading assessment activity of scientific and technological research.

The setting-up of CIVR has finally paved the way to the creation of a central body entrusted with the promotion and co-ordination of evaluation activities in the field of research. According to the institutional decree, the new independent body shall support the quality and improve the use of research, the diffusion of methodologies and best practices in an inter-national co-operation framework. As mentioned before, the CIVR, made up of 7 members, some of them foreign, is part of a broader scheme enabling the undertaking of strategic decision by improving the advisory capacity of MURST. At the moment CIVR experts have not yet been appointed, but a specific budget item is available for the support of studies and activities on monitoring and assessment at ministerial level.

Along with structural interventions in the last years, a relevant procedural improvement has also been introduced. As first thing the university national co-financing mechanism for supporting over-local inter-university projects has been radically changed. In fact, according to existing legislation, the research financing is implemented at two levels:

a) at a first level, all universities allocate resources to their scientists on the basis of autonomous decisions and financed though the internal budget;
b) at the second level, the so-called “40% procedure” provides the universities with additional national funds for relevant research projects.

Up to 1996 the elected Disciplinary Advisory Committees, part of the general and elected body “National University Council” (CUN), were responsible for the allocation of resources. This structure adopted a sort of compromising mechanism between electors and elected to the detriment of the quality of the selection procedure and with the final result of distributing little resources to a large amount of beneficiaries.

The new procedure, which started in 1997, is based on a Committee of national guarantee (5 experts at the moment) and 2 or 3 anonymous referees for each project and on the use of telematic procedures to ensure a confidential and effective communication, has dramatically changed the situation. To implement the procedure, a list of expert evaluators for different field has been compiled. The results are quite opposite to the previous one with the financing of a smaller number of projects of larger size and are particularly relevant for the 1998[6].

The whole mechanism has been supported by the increase of available resources, from 100 billion of Lire in 1996 (the last year of the previous procedure) to roughly 200 in 1998 and with the perspective of reaching the amount of 300 within the year 2000[7].

Nevertheless the share allotted to research on the total amount of universities expenditures (about 9000 billion Lire, most of them for wages) is still very low, even including approximately 200 billion Lire from universities budget and additional resources from external contract and CNR funds.

With reference to university evaluation it is worth mentioning the implementation of the Evaluation Observatory[8] and Local Units as foreseen in 1994 Financial Law. Until now the Observatory has focused its activities mainly on the picture of the system (new universities, budgetary analyses, re-equilibrium fund) rather than on research, producing a criteria matrix for evaluation on three different areas (administration management, teaching and research) and harmonising the activities of the different local units. In the same field a joint effort developed by MURST and the National Statistical Office (ISTAT) has renewed statistics on Universities, in particular concerning Professors' activities and research budgets.

Another significant actor involved in University evaluation is the Conference of Italian University Rectors (CRUI), the autonomous association of the Vice-chancellors of all Italian Universities[9]. Since it was set up, it has expressed the necessity to guarantee greater autonomy for Universities, and has sponsored the adoption of assessment

[6] In 1998 exercise, closed at november, 4000 referees were selected among a Directory of 12000 (7000 of them foreign). All the applications have an english translation in order to allow a foreign reading. In comparison with first and pioneeristic experience (1997), a 20% reduction of applications (12% less, concerning budget requests) was registered and more than 50% of projects were positevely considered, due to a larger amount of available funds (25% more). The national contribution share was increased till 70% of the total costs, in case of inter-universities projects.

[7] More information (Reports, evaluation criteria and forms) can be found at http://www.murst.it (cofinanziamento nazionale).

[8] The Observatory (National University Evaluation Council) is an institutional body of MURST. Its functions are to evaluate the efficiency and the effectiveness of Universities' activities, to verify the development plans and to analyse the situation of the Italian University system. The Council is an indipendent body which interacts autonomously with the Universities and the Ministry. The Council relates to the Minister and to the relevant Parliamentary Commissions. The Council has a technical and administrative secretarial office, its manages its own budget and may appoint teams of experts or specialised external bodies the task of specific research and studies.

[9] The role of CRUI in introducing evaluation activities within Italian Universities has been analysed in many papers. CRUI developed methodologies, collected indicators and edited some volumes on it. Since 1996, on the basis of a specific working party, the Conference proposed a cognitive-type procedure for the evaluation of university related activities. The aim of the proposal was to begin to acquire analytical knowledge of the entities which produce research and their funding procedures and to establish general assessment criteria. Methodolo-gical criteria were defined for the evaluation of research whereby "assessing the quality of research involves identifying the characteristics which make it possible to determine the capacity to achieve or to have achieved the objectives for which a research project was conceived and performed as effectively and cheaply as pos-sible". Setting out these principles and these definitions, the CRUI proposed two distint spheres of action: the first, on a national scale in an inter-university dimention, to improve the quality of research activity; the second, at single university level, to address problems of the allocation of research funds by university administrative offices. See: Avveduto-Brandi, "In search of a systematic approach to the evaluation of university autonomous activities", paper presented at III EES Conference, Rome 29-31 Oct. 1998 and Boffo-Moscati, "Evaluation in the Italian Higher Education System", *European Journal of Education*, vol. 33, No. 3, 1998.

systems. Recent developments in CRUI activities have introduced a distinction among different scientific macro-areas, such in case of humanities in comparison with hard sciences.

To conclude this presentation of university evaluation, although not strictly related to research, the work undertaken by the Public Expenditure Technical Commission of Ministry of Treasure and Balance, must be mentioned. An econometric model was defined, in order to establish the sharing criteria for funding universities. That formula, which defines standard costs for each student using a certain number of variables, was suggested to the MURST Observatory. This algorithm is used for the allocation of the re-balancing fund (less than 10% of the total), that is to reward the most efficient and needing university.

The other main change undertaken at procedural level is that concerning industrial research. The allocation of resources for industrial research is based on a variety of instruments mostly incentives implemented through complicated procedures where the evaluation phase is replaced by the formal procedure.

The novelty of the approach consists of:

- procedural simplification and homogenisation of the different types of instruments;
- the setting up of innovative mechanism to support research mobility, the employment of young researchers, fiscal aids for collaboration between public and private bodies;
- a particular attention, also from the legislative point of view, for the South Italy by promoting the establishment of supported network and cluster.

Specific evaluation methods are seen for the three of measures.
As for the first measure, the Advisory Body, a Ministerial Technical and Scientific Committee, is the reference point for both the scientific examination phase (carried out by ad hoc expert, selected within a specific list) and administrative phase (carried out by ministerial office and bank); applications have been examined following new procedures which imply simplified acts and fixed timings.

An important novelty, in this specific case, is that the evaluation phase is no more focused, as in the past, on ex-ante view, that means on the formal and administrative aspects, but all along the project development and on the final results. These changes allow a more transparent data collection and more on-going information. The project Directory is now an important instrument representing an inventory of ideas and entities.

In the second case the instruments, recently activated, like the employment incentive, the fiscal bonus (tax credit) have successfully gained the appreciation of the research industrial world, especially among the SMEs to whom were particularly addressed. A first ex-post evaluation has been made. According to it, two measures seem to be successful: in fact, new employment of Ph.D.s and University graduates and tax credit support have achieved the goal. Concerning temporary mobility of researchers from public bodies to private firms, some operational problems have discouraged a larger participation.

In the last case the “originality” of the measure relies on the need to operate in a “weak environment” (less favoured areas). Thus trying to maintain the standard procedure used for the financing of industrial research (call for proposal, selection of projects, contract, monitoring and evaluation of results), the special need of the south of Italy required an adjustment process. The choice of the ad hoc Scientific and Technical Committee is to couple ("cluster") knowledge production of different bodies, within selected thematic areas, where desirable spin-off are expected to be. The cluster approach, by aggregating single projects in a larger scheme, required a redefinition of the contents of the projects, as well as new contractual provisions. The experts in charge of the selection mechanism were to fill in a special form, which includes specific monitoring and evaluation criteria.

The control procedure (also by advanced steps), already implemented in the case of Southern Scientific and Technological Parks, which benefit from public support, also represents, for the Italian experiences, an important novelty.

By an operational point of view, some examples must be quoted: first of all, the Nuclear Physics Institute (INFN) which promoted an external international evaluation ("Richter Report"), followed by on-going studies, requested by the Minister to different bodies for improving experiences. Another example concerns the impact of R&D Structural Funds on South, a European-National joint effort that is now running. No official activities were undertaken on Italian participation to the Framework Programme; the last deals with the third FP and was conducted by APRE, the National Agency for European Research, in 1995.

The recent new attention to the evaluation issue is confirmed by the two following examples. The first one being the development of the Treasure and Balance Ministry, that is the new Economy Ministry, which created a favourable environment for the diffusion of previous experiences like those of the Public Investment Evaluation Unit (a group of economists strongly inspired to the World Bank criteria) and of Employment Investment Fund (FIO) for the selection of targeted initiatives.

The new Department for Development and Cohesion Policy (DPS) of the Ministry, in charge of the policies co-ordination, has provided itself with a larger and internal structure (Unité di Valutalione, UVAL). The "Cabina di Regia" (Steering Unit) has been confirmed in its autonomy both on national and regional scale, and in the task of supporting the co-ordination of different policies.

The key word of DPS is: negotiation and evaluation for a new planning to enhance the involvement of different actors and to ensure effectiveness and transparency to the decision-process.

This new structure which will be involved in the defining of support intervention on South (foresight, managing, monitoring and assessment) also benefit from Structural Funds subsidiary incentives for the period 2000-2006. The evaluation issue is also part of a more general decentralisation process in public administration, which now includes items such innovation and technology transfer.

The second important element that testifies the creation of a favourable environment for the evaluation is the recent constitution of the Italian Evaluation Association (AIV), a scientific body gathering experts in the field of research evaluation. Together with the EES (European Evaluation Society), the AIV has recently organised in Rome the European Conference where the issue related to research evaluation has been discussed.

In conclusion, R&D evaluation in Italy must grow following European standard. Some improvement will be required in order to develop expertise, capacities and experiences. A relevant international co-operation could reduce costs and risks but anyway the success should be based on a large participation and involvement of Italian scientific community, not only on an official institutional approach. This is the reason why the driving force will be an enlarged and diffused presence of evaluation activities.

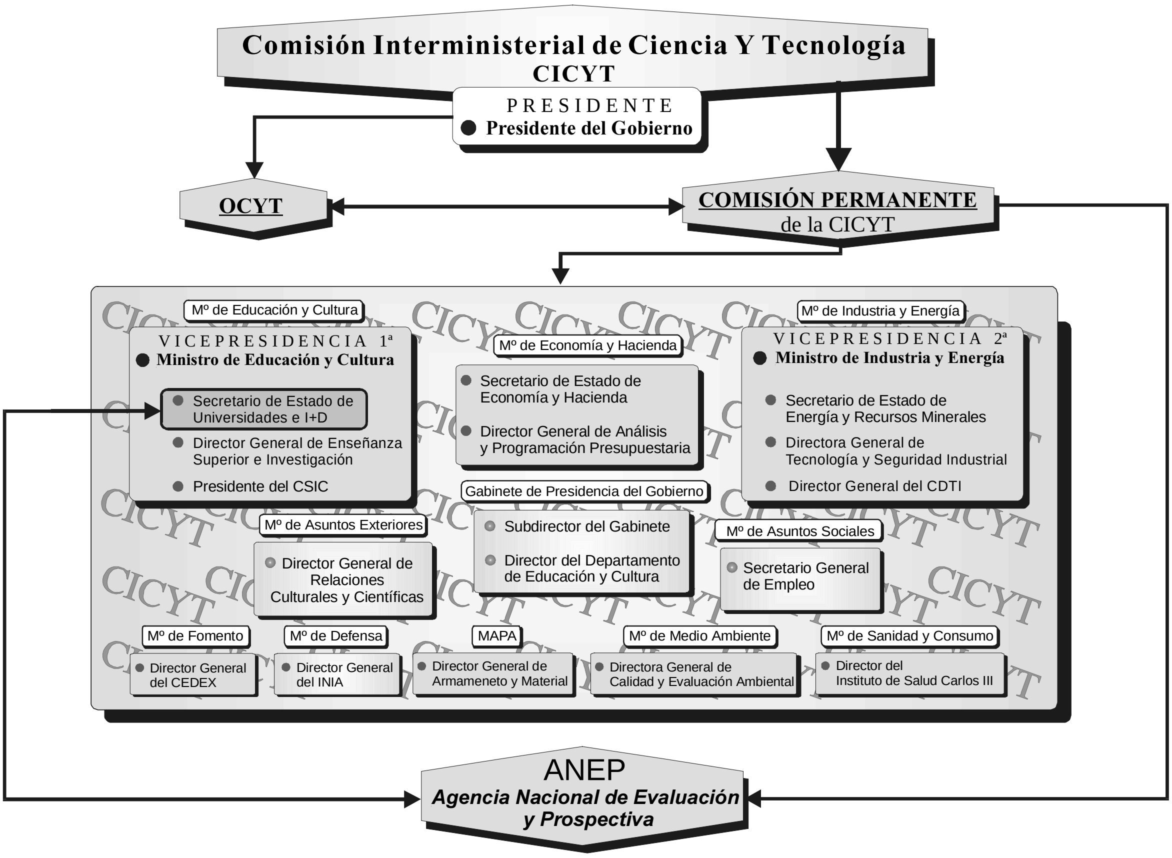
A UNIFIED FRAMEWORK FOR R&D EVALUATION AND FORESIGHT. THE ROLE OF ANEP IN SPAIN

B. Presmanes, J. Casado, and H. Guerrero
Agencia Nacional de Evaluación y Prospectiva (ANEP).

1. INTRODUCTION

The National Agency for Evaluation and Foresight (ANEP) is the official institution carrying out evaluation and foresight activities in Spain. It was created as an administrative body to give support to the CICYT and is assigned to the State's Secretary of Universities and R&D (SEUID) (Figure 1). The ANEP differs from other similar institutions in its independence from any management organism of S&T.

Figure 1. Institutional Structure


Comisión Interministerial de Ciencia Y Tecnología
CICYT
PRESIDENTE
Presidente del Gobierno
OCYT
COMISIÓN PERMANENTE de la CICYT
Mº de Educación y Cultura
VICEPRESIDENCIA 1ª
Ministro de Educación y Cultura
Secretario de Estado de Universidades e I+D
Director General de Enseñanza Superior e Investigación
Presidente del CSIC
Mº de Economía y Hacienda
Secretario de Estado de Economía y Hacienda
Director General de Análisis y Programación Presupuestaria
Mº de Industria y Energía
VICEPRESIDENCIA 2ª
Ministro de Industria y Energía
Secretario de Estado de Energía y Recursos Minerales
Directora General de Tecnología y Seguridad Industrial
Director General del CDTI
Gabinete de Presidencia del Gobierno
Subdirector del Gabinete
Director del Departamento de Educación y Cultura
Mº de Asuntos Exteriores
Director General de Relaciones Culturales y Científicas
Mº de Asuntos Sociales
Secretario General de Empleo
Mº de Fomento
Director General del CEDEX
Mº de Defensa
Director General del INIA
MAPA
Director General de Armameneto y Material
Mº de Medio Ambiente
Directora General de Calidad y Evaluación Ambiental
Mº de Sanidad y Consumo
Director del Instituto de Salud Carlos III
CICYT
ANEP
Agencia Nacional de Evaluación y Prospectiva

Its functions have hardly changed since its founding, back in 1987. The successive restructures of the S&T system have reinforced its role, due to the services paid to many entities all over the country and to the firm support provided by the SEUID to the improvement of its infrastructure.

Nevertheless, its activities have evolved considerably in the last few years. Evaluation activities have experienced a remarkable increase in quantity, complexity and diversity, and this trend is on continuous growth with a constant increase in demand from new users (Figure 2).

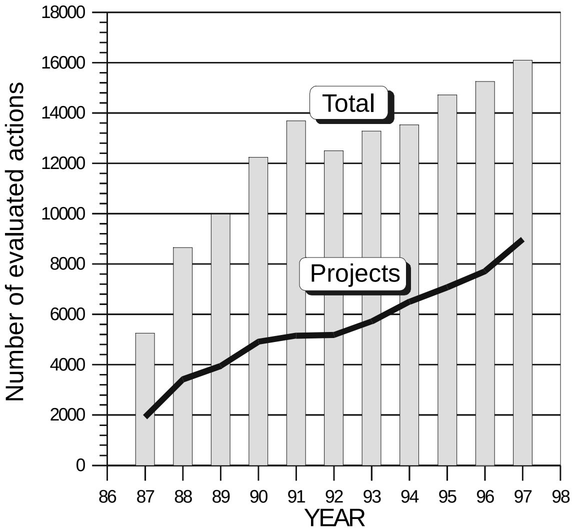


Figure 2.
Continuous growth of the evaluation activities developed in ANEP.

This strong growth of evaluation activities has hindered the development of foresight activities that have experienced two different stages. The first stage coincides with the creation of the ANEP in which it was necessary to define the programmes of the new National R&D Plan, and a second stage in which the accumulation of data and own experiences within the National R&D Plan allowed the start of specific forecast actions that advised on the design of performing lines at short and medium term. In the first stage, the ANEP, or its predecessor, the Cabinet of Studies of the Advisory Commission for S&T (CAYCIT), initiated the first foresight seminars. Their common characteristic was to reunite the most outstanding experts of the country around a specific subject, such as chemistry, physics, nutrition, history, animal science, agriculture and oceanography. The main target was to define world trends in each sector in order to guide R&D policies (Table I).

Table I. Foresight Seminars (First Phase: 1985-1988)

Year	Title	Organizer	Seminar
1985	Present Trends in Chemistry	CAICYT-CSIC	Encuentros UIMP
1986	Foresight in Earth Sciences	CAICYT-CSIC	Encuentros UIMP
1987	The Future of Food (Nourishment)	FAST-CSIC	Seminario FAST-CSIC
1987	Foresight in Animal Production	CAICYT-CSIC	Encuentros UIMP
1987	The Future of Renewable Natural Resources in Spain	FAST-CSIC	Seminario FAST-CSIC
1987	Foresight in Physics	ANEP-CSIC	Encuentros UIMP
1987	Foresight in Oceanography	CAICYT-CSIC	Encuentros UIMP
1988	Trends in History	ANEP-CSIC	Encuentros UIMP
1990	Trends in Molecular Biology	ANEP-CSIC	Encuentros UIMP

Once the programmes for the First National Plan were defined evaluation polarised almost exclusively the activities of the ANEP leaving foresight in the background. It is not until the end of 1994 when foresight is taken up again by the ANEP and this date marks the beginning of the second phase. The realisation of the R&D plans, accumulation of data and the experience gained allows them to have a field in which to work. This interest for forecasting tasks arrives to a great extent from the evolution of the system itself. After almost a decade of experience the system requires global indicators, which will allow for an assessment of the objective situation of the different scientific and technological sectors and the specific demands that afford precise orientations and directions for the planning of R&D.

The second phase proposes the following objectives:

- To gather and evaluate previous studies carried out in Spain
- To analyse the methods and practices of foresight
- To carry out pilot studies (at a sectoral and regional level)
- To promote knowledge of technological foresight and its usage as an instrument to define R&D policy.
- To establish international relations with other institutions related to foresight.

During 1995-96 intense work was carried out along these lines (Table II) which hoped to encompass both the analysis from the perspective of science (scientific and technological offer) as well as from the technological foresight (the industrial demand).

Table II. Foresight Studies (Second Phase 1994-98)

Year	Title	Editor	Method
1995	Technology evolution for advanced multimedia services	ANEP	Delphi – Panel of experts
1995	Study of the present state and foresight in mobile communication. Bibliometric analysis for the period 1989-1993	ANEP	Study – Bibliometry
1995	Advanced materials	ANEP	Study
1995	Analysis of foresight methods and its international application	ANEP	Report
1996	Technology Foresight. Perspectives for European and International Co-operation	PREST ANEP Univ. of Twente	Study
1996-98	Foresight in optics. Bibliometric analysis for the period of 1989-97	ANEP	Bibliometry
1997-99	Application of foresight techniques for the development of a sustainable tourist model in the Balearic Islands	ANEP-AIRTEL-BALEAR Goverm.	Delphi – Panel of experts

This activity did not have the expected impact on the management bodies of national R&D. Although they financed the projects carried out by the ANEP, they did not continue with their application and diffusion. It is hoped that the recent creation of the OCYT (The Office of Science and Technology) in 1998 and the growing interest that the Ministry of Industry has in foresight activities will contribute to the success of the initiatives undertaken from the ANEP. This new situation will allow the ANEP to focus its interest on foresight activities, which are much more closely linked to its evaluation function since it is in this synergy where we find the true extra value that the ANEP can contribute to foresight.

2. FORESIGHT ACTIVITIES AS SUPPORT FOR EVALUATION.

In the last few years the growth in demand for the ex-post evaluation and institutional evaluation has led to the convenience of simultaneously generating foresight activities that support evaluation. Although these types of evaluation (which were started in Spain at the beginning of the nineties) have been really scarce, the simple fact of approaching them has made us see that establishing a link between evaluation and prospective within the specific outline of the ANEP is beneficial to both activities. In this work we are presenting a common reference framework in which evaluation and foresight are only supplementary elements within the same strategy that leads the long and medium term actions of the ANEP.

Figure 3. Evaluation and foresight activities as complementary elements in a single strategy. Evaluation and foresight are projected in three scenarios: past, present and future.

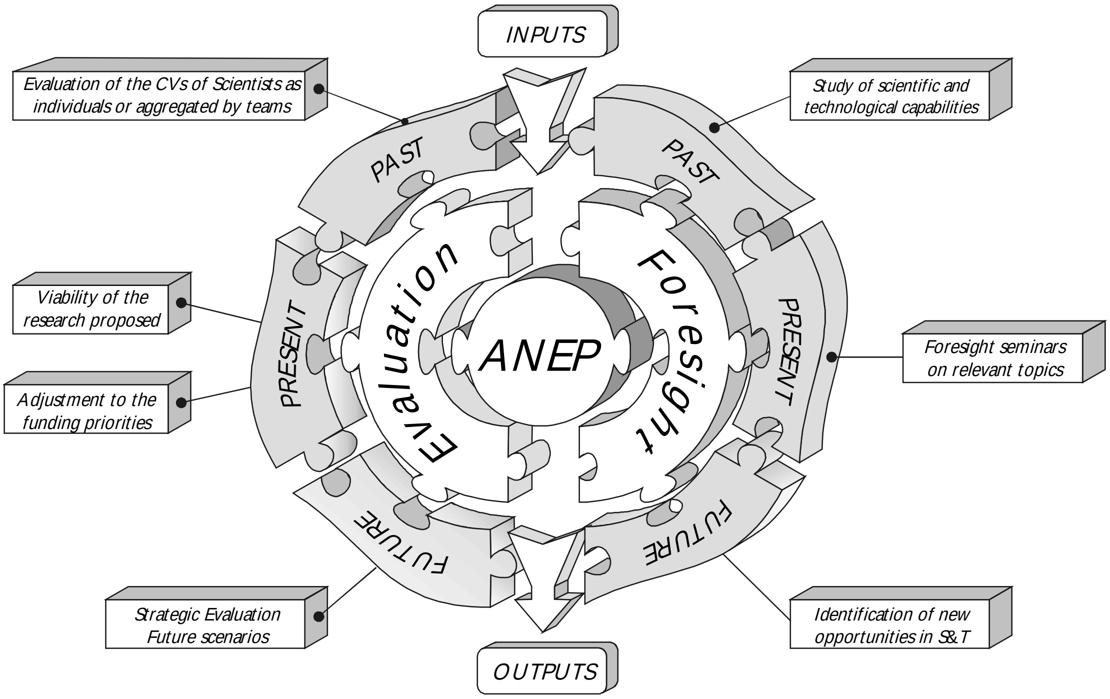

INPUTS
Evaluation of the CVs of Scientists as individuals or aggregated by teams
Study of scientific and technological capabilities
PAST
PAST
Viability of the research proposed
Adjustment to the funding priorities
PRESENT
Evaluation
ANEP
Foresight
PRESENT
Foresight seminars on relevant topics
FUTURE
FUTURE
Strategic Evaluation Future scenarios
Identification of new opportunities in S&T
OUTPUTS

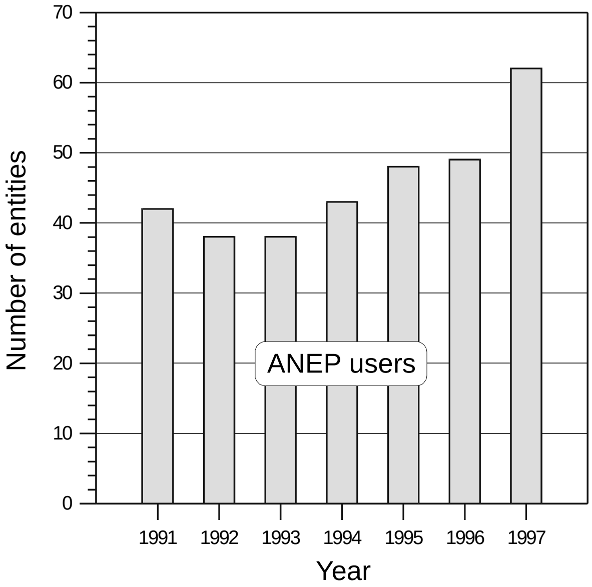
Figure 3 outlines some of the activities of the ANEP within the same *modus operandi* or strategy in which evaluation and foresight are complementary activities. In this sketch, and as a puzzle, we represent the different evaluation and foresight functions that the ANEP has been developing with the aim of obtaining the necessary information for a high quality evaluation of the research activity and its institutions or programmes. As we will be able to see, there are three scenarios in which we can project the foresight and evaluation activities: the past, the present and the future. The synergy between ANEP's evaluation and foresight appears at the level of each of these scenarios and the individual results of evaluation and foresight improve if they are considered together.

Now we will develop each of the characteristic elements of this diagram.

3. INPUTS AND OUTPUTS: WHO DEMANDS THE SERVICES OF THE ANEP?

The ANEP is a singular case in Europe as it concentrates in a single institution the evaluating activity of a very broad number of institutions whilst it maintains its independent status of political power. This has been respected throughout time and thanks to this it has a tremendous amount of support in the scientific community of the country. With the aim of fostering the evaluating culture the ANEP has given support to many other institutions that voluntarily have requested its services. For example, in 1997 62 institutions from both the public and private sectors have made use of the services of the ANEP (Figure 4). It is worth highlighting though that the majority of the regional autonomous government are users of ANEP's services.

Figure 4. Entities demanding ANEP services.



Ayuntamiento de Lérida, Centro de Desarrollo Tecnológico e Industrial CDTI (MINER), Centro de Investigación y Documentación Educativa CIDE (MEC), Clínica Puerta de Hierro, Comunidad de Madrid, Consejo de Universidades (MEC), Consejo Superior de Deportes (MEC), Consejo Superior de Investigaciones Científicas CSIC (MEC), Diputación General de Aragón, Dirección General de Enseñanza Superior (MEC), Dirección General de Relaciones Culturales y Científicas (MAE), Fondo de Investigación Sanitaria FIS (MSC), Fundación Caja de Arquitectos, Fundación Caja Madrid, Fundación Canaria de Transplantes, Fundación de Investigación Cardiovascular, Fundación La Caixa, Fundación La Marató TV·3, Fundación Marcelino Botín, Fundación Rich, Fundación Universitaria de Las Palmas, Fundación Valenciana de Investigaciones Biomédicas, Generalidad de Cataluña, Generalidad de Valencia, Gobierno Canario, Gobierno de Navarra, Gobierno Vasco, Hospital de Cabueñes, Instituto Danone, Instituto de la Mujer (MTAS), Instituto de Migraciones y Asuntos Sociales IMSERSO (MTAS), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria INIA (MAPA), Junta de Castilla y León, Junta de Castilla-La Mancha, Junta de Extremadura, Junta de Galicia, Principado de Asturias, Región de Murcia, Secretaria de Estado de Universidades Investigación y Desarrollo (MEC), Secretaría General del Plan Nacional de I+D (CICYT), Secretaria General Técnica (MEC), Universidad Autónoma de Barcelona, Universidad Autónoma de Madrid, Universidad Complutense de Madrid, Universidad de Extremadura, Universidad de Granada, Universidad de Huelva, Universidad de La Rioja, Universidad de las Palmas, Universidad de Lérida, Universidad de Málaga, Universidad de Murcia, Universidad de Oviedo, Universidad de Salamanca, Universidad de Santiago de Compostela, Universidad de Sevilla, Universidad de Valencia, Universidad de Zaragoza, Universidad del País Vasco, Universidad Jaime I, de Castellón, Universidad Nacional de Educación a Distancia, Universidad Pública de Navarra.

The reports elaborated by the ANEP focus on scientific and technological aspects. This is the reason why experts of different scientific and technological fields participate in evaluation and prospective activities. The most common actions are the assessment of the viability of research proposals and the opportunity, from a scientific point of view, of the acquisition of resources. It is also usual to assess results of research already carried out.

The evaluation provides information that allows:

- To elaborate recommendations for the implementation of results
- To advise managers about potential interest of emerging lines
- To identify outdated research lines
- To guide researchers for a better use of resources
- To find out the potential interest of international collaboration
- To favour contacts between researchers and enterprises.

All these kinds of advice require nowadays more than an individual action, case by case, of experts. Foresight activities, therefore, could be understood as a way to sort, in an aggregated manner, a number of indicators and favour a collective thinking around strategic themes that will facilitate decision making. The connection between evaluation and foresight functions and the diversity of users that converge at the ANEP is its real added value.

4. A unified framework for R&D Evaluation and Foresight.

4.1 - The PAST: The search for INDICATORS.

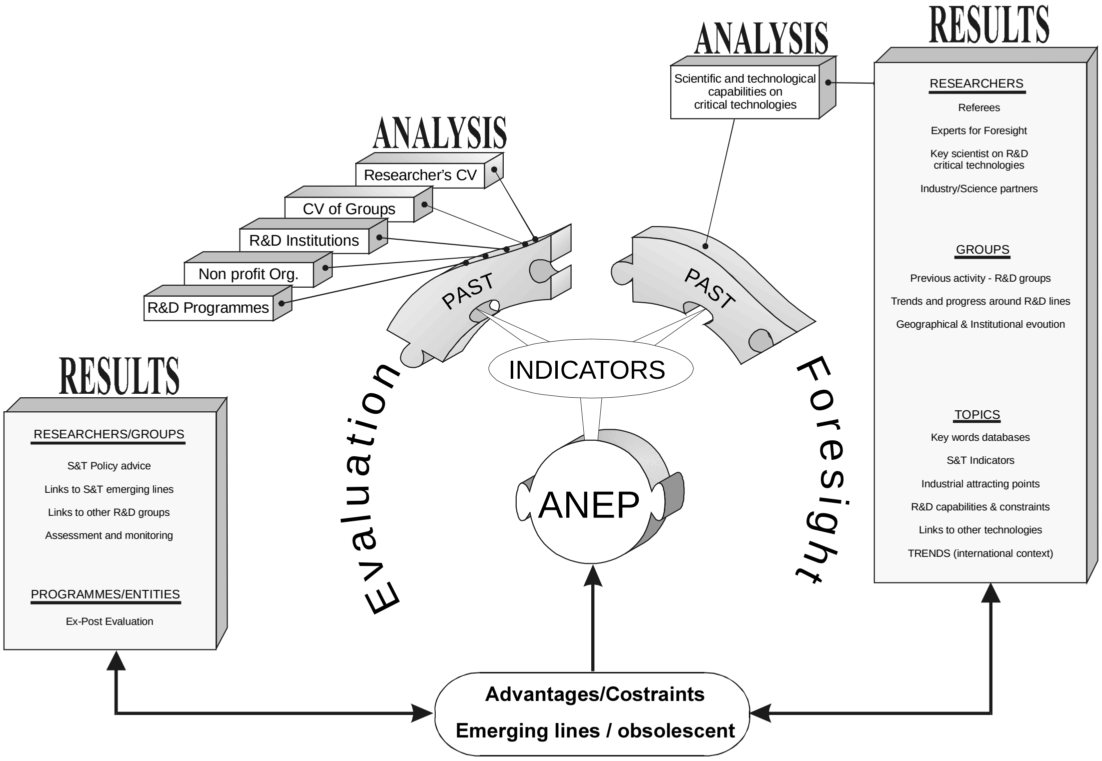
Evaluation and foresight together considered within the framework of the past leads to what could be termed the “indicators phase” (Figure 5). The results of both activities increase the collection of R&D indicators related to quality, activity, performance, etc.

Evaluation projected to the past would focus in the past capability of parties. Any initiative of R&D policy planning or implementation demands an assessment of the research capability of individual researchers, teams, institutions, etc. For example, CVs in the case of isolated researchers, annual reports in the case of R&D institutions, etc. This information comes from the same parties and will be provided together with the project under evaluation.

Past evaluation results relevant to the present scenario by generating a number of recommendations regarding the R&D of parties. For the latter, this process could be of great value since they are being audited by highly qualified third parties. When projecting foresight to the past, we must make sure to use valuation criteria with updated scientific and technological information, as well as the current environs. In such way foresight becomes a diagnosis tool allowing the assessment of available and lacking resources as opposed to strategic lines for the future of an institution, a region, a country, etc. Even when obtaining the expected results, they do not always fulfil expectations. On the other hand, some could provide unforeseen opportunities.

This foresight activity produces indicators useful to managers and to information sources close to needs set up by potential R&D users, especially the entrepreneurial sector. It also allows the creation or actualisation of a data bank of experts in emerging strategic lines as a support of evaluation activities (precise identification of evaluators) and even of foresight itself (i.e. Delphi consultations).

Figure 5. Results obtained considering both Evaluation and Foresight during the PAST.

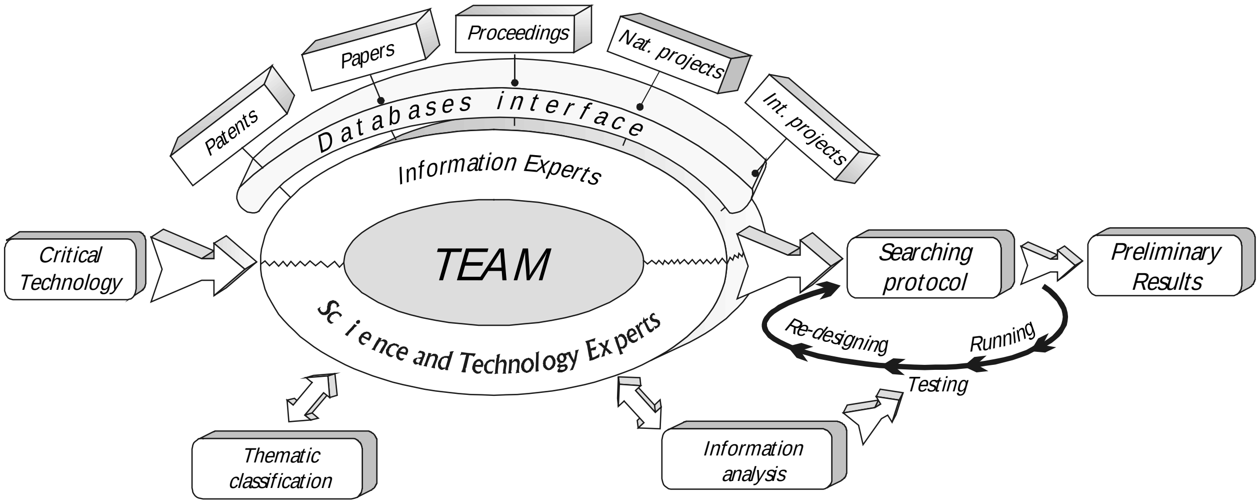

ANALYSIS
Researcher's CV
CV of Groups
R&D Institutions
Non profit Org.
R&D Programmes
ANALYSIS
Scientific and technological capabilities on critical technologies
RESULTS
RESEARCHERS
Referees
Experts for Foresight
Key scientist on R&D critical technologies
Industry/Science partners
GROUPS
Previous activity - R&D groups
Trends and progress around R&D lines
Geographical & Institutional evoution
TOPICS
Key words databases
S&T Indicators
Industrial attracting points
R&D capabilities & constraints
Links to other technologies
TRENDS (international context)
PAST
PAST
INDICATORS
Evaluation
Foresight
ANEP
RESULTS
RESEARCHERS/GROUPS
S&T Policy advice
Links to S&T emerging lines
Links to other R&D groups
Assessment and monitoring
PROGRAMMES/ENTITIES
Ex-Post Evaluation
Advantages/Costraints
Emerging lines / obsolescent

When studying in a global and aggregated way the evolution of broad and representative segments of the Spanish scientific and technological panorama, we obtain a retrospective vision that allows to plan a successful consolidation of lines. In this way, foresight is similar to ex-post evaluation, in which there is a retrospective approach to what has been achieved according to objectives set at the start. In this case the reflection raised should be considered even deeper.

An example of a study on the scientific and technological capacity of critical technologies is being carried out by the ANEP. It is on OPTICS in Spain. Figure 6 shows a sketch of how it is being carried out. The first step is to identify a critical technology (Optics in this case), then a team of specialists in both the subject and the use of information sources, is set up. After several interactions a classification by subjects is made and based on this a compilation of the information relevant to the study (patents, workshops, articles, etc.).

A searching strategy is developed and it will be verified several times until it shows trustworthy preliminary results. Based on this a study is published and the cited researchers inquired. Out of this interaction (the current status of the optics study) a second study is made to complement the first one, in which the state of the art in Spain will be described over a meaningful period (11 years) hat allowing to collect enough information to confront other analysis. An example of the information obtained is shown in figure 5.

Figure 6. Schematic view of the process followed to identify capabilities in critical technologies.



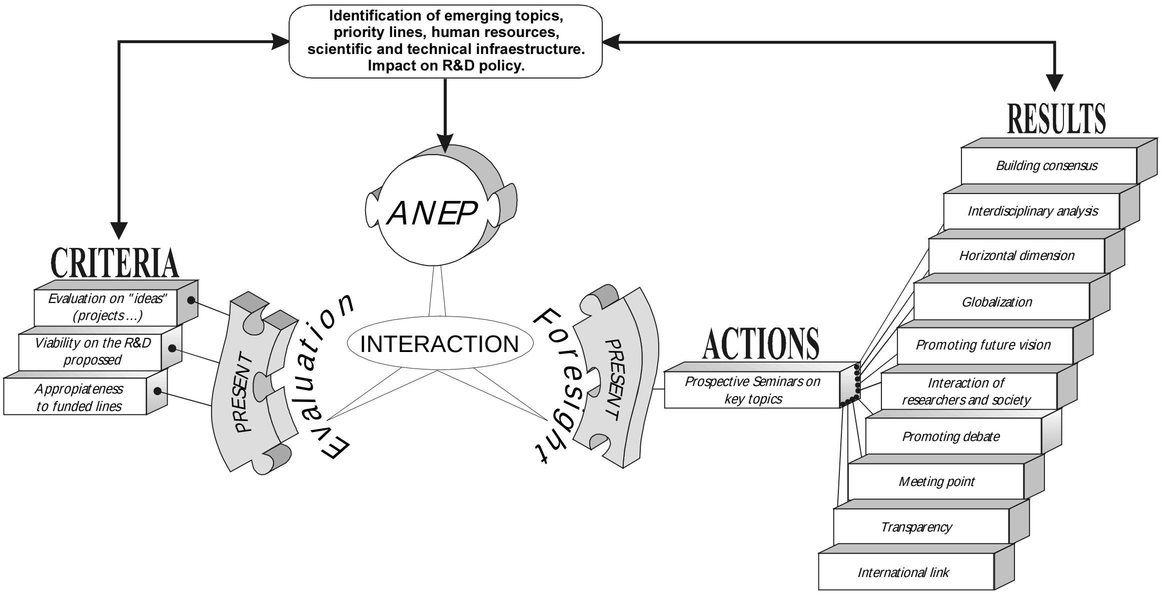
4.2 - The PRESENT: Producing INTERACTIONS.

ANEP’s evaluation activity is mainly focused in the present, since it has thousands of projects every year that have to be assessed for its financing. What has to be assessed is the “concept” that wants to be developed and which is translated into projects that include additional measures such as equipment, scholars, foreign support researchers, etc.

It is important that the idea is framed under a global context, reflected on the priority lines of guided programmes. When referring to basic research its importance should mainly be qualitatively valued.

ANEP seminars are the projection of foresight activity into the present (figures 7 and 8). In this sense, they are a very useful tool to propitiate debate on different subjects of strategic interest for a specific sector, region or country. ANEP’s mission is to identify those subjects and put in contact researchers of different origins and users -mainly from the entrepreneurial sector-, to promote a strategic and joint vision of the sector’s needs, identify the research lines to be promoted based on the interest of the region or sector - such as Mediterranean- with the purpose of influencing the design of R&D policies, facilitate the specific identification of strong and weak points, identify the lines of priority interest for international collaboration, identify available resources of the sector to favour its valorisation, strengthen the European position and improve transfer policies both in Spain and third countries. Besides, seminars are also a reference point for diffusion of foresight methodologies.

Figure 7. Evaluation and Foresight: The PRESENT.


Identification of emerging topics, priority lines, human resources, scientific and technical infraestructure. Impact on R&D policy.
ANEP
INTERACTION
CRITERIA
Evaluation on "ideas" (projects...)
Viability on the R&D propossed
Appropiateness to funded lines
PRESENT
Evaluation
Foresight
PRESENT
ACTIONS
Prospective Seminars on key topics
RESULTS
Building consensus
Interdisciplinary analysis
Horizontal dimension
Globalization
Promoting future vision
Interaction of researchers and society
Promoting debate
Meeting point
Transparency
International link

How do these seminars work? A nucleus formed by the ANEP and equivalent bodies of Spain and other countries establish a number of horizontal activities which are the core point around which the basic tools for evaluation and foresight activities are generated.

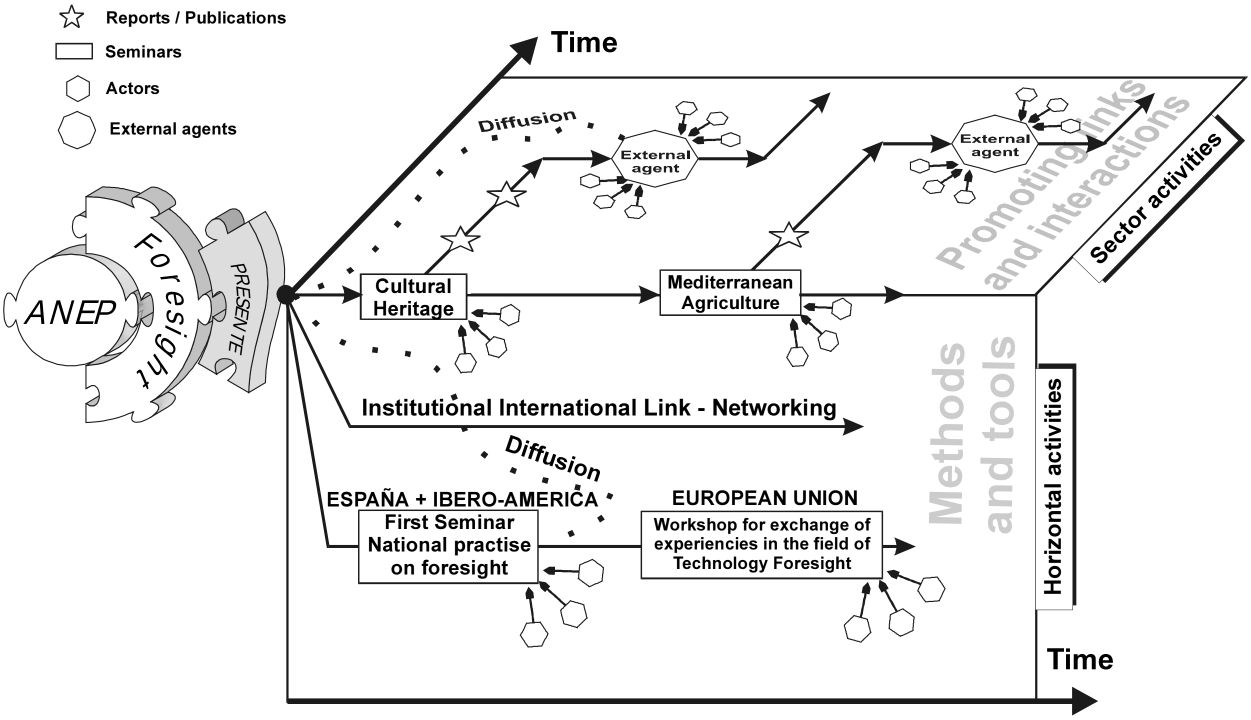
At the ANEP seminars current policies are analysed, innovative cases in collaboration with the entrepreneurial sector are presented and debates are organised on tentative future strategies. After the seminar, a written report with the most relevant information is elaborated, including:

- Analysis of national and European policies, trends of the sector.
- Main technological demands and future challenges (non-strictly scientific aspects that may condition the sector).
- Researchers and most important companies involved in R&D.

This first encounter aims to generate a regular foresight activity organised according to the participants' demands. This activity does not necessarily have to be carried out by the ANEP. A group or institution with enough drive to keep this activity alive by developing a catalogue of experts, following a pre-established model, organising a consultation work on investigation lines and related technological demands and maintain the organisation of seminars as a communication and diffusion tool could very well take the baton.

Figure 8 shows the foresight activities developed by the ANEP at present time. There are two main types of activities. One called "methodologies and tools" and the other "promoting links and interactions".

Figure 8. Dynamics on PRESENT foresight related activities.



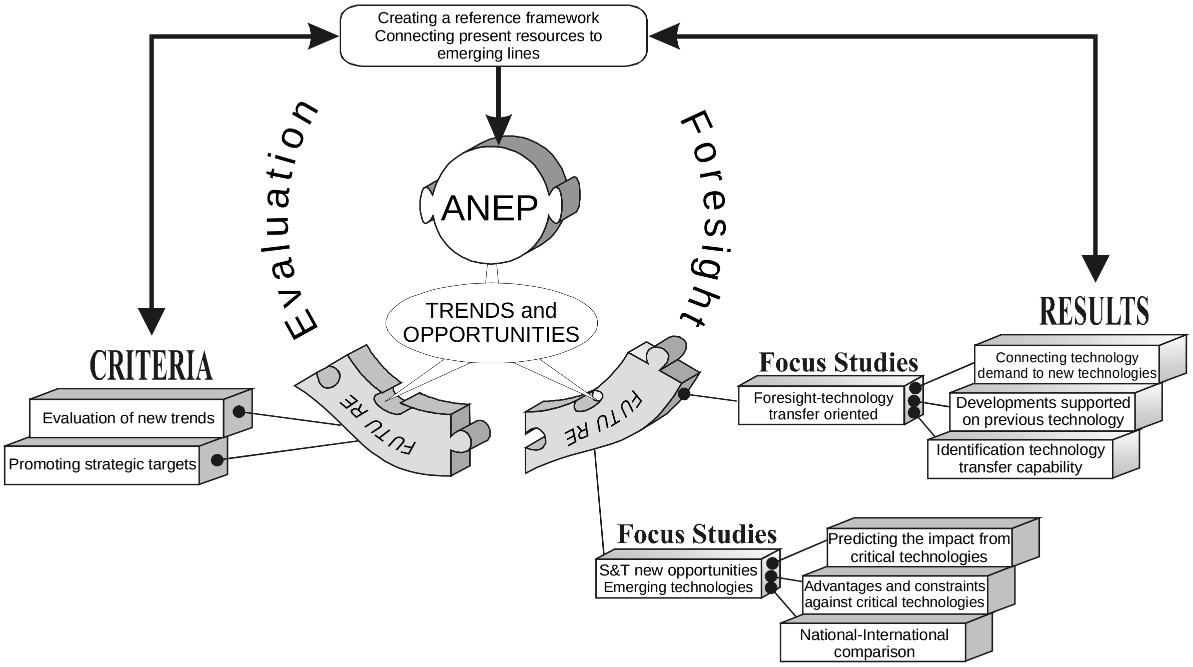
The first type of activities is characterised by organising and participating in foresight seminars covering a wide spectrum of agents of our cultural and geographical spheres (South America and the EU). The objective would be to share experiences and generate a specific methodology that would have a direct application in other foresight activities.

The second type usually consists of seminars in which the ANEP plays a neutral role, acting as a forum for the various interested parties on a specific subject. In a way, the ANEP acts as a catalyst, enabling the interaction between parties and allowing a horizontal processing of information and facilitating its flow. Once the seminar is concluded and its findings published, the neutral role of the ANEP gains again interest by transferring its initial prominence to another party that will act as a nucleus and will organise a network of parties around itself to continue with those activities.

4.3 - The FUTURE: identification of TRENDS and OPPORTUNITIES

It is normal to associate foresight with future and it could be thought that the incidence of foresight in evaluation activities is rather poor. Notwithstanding, to evaluate a project looking to future possibilities should be seriously considered. At this point, evaluation and foresight naturally converge.

Figure 9. Evaluation and Foresight: the FUTURE.



Creating a reference framework
Connecting present resources to emerging lines
Evaluation
ANEP
Foresight
TRENDS and OPPORTUNITIES
FUTU RE
FUTU RE
CRITERIA
Evaluation of new trends
Promoting strategic targets
RESULTS
Focus Studies
Foresight-technology transfer oriented
Connecting technology demand to new technologies
Developments supported on previous technology
Identification technology transfer capability
Focus Studies
S&T new opportunities
Emerging technologies
Predicting the impact from critical technologies
Advantages and constraints against critical technologies
National-International comparison

To evaluate looking to the future implies considering criteria about new trends, the need to strengthen specific strategic interests and assess the possibility to fit the evaluated action into future scenarios. This kind of evaluation could be called “strategic evaluation”. For it to be successful and play a decisive role in the proper implantation of R&D lines, evaluators should have information and general criteria about priority aspects for the future. This information will arise from the different activities carried out by foresight projected into the future.

There are two types of foresight in a future scenario that have started to be taken into account by the ANEP. The first one is technology foresight transfer oriented for the assimilation and performance of technologies in sectors where it is not well established. As a result of the foresight seminars (for example “cultural interest”) and thanks to the contacts created among different specialists from multiple areas, there have been ideas on the use of technology that could be very beneficial for certain sectors. Structuring the present by looking to the future (with an adequate panel of experts) would open new ways in which to make available technology that could be of great strategic importance to the recipient sectors.

A possible ramification of the foresight activity would be to study the possible adaptation of our technology in less developed third countries.

The second foresight activity carried out by the ANEP in a future scenario is the identification of new scientific and technological opportunities, which are a consequence of the different lines that have lately been conceived world-wide. As an example, following is an outline of a study on nano-technology.

4.3.1. Example of identifying new opportunities and trends: NANOTECHNOLOGY

Since the mid Eighties all areas of S&T have incorporated the term nano-technology, associated to new behaviours as a result of the study of processes and means at a nanometric scale.

The nano-technology is beginning to be considered one of the essential elements in which the next scientific and technological revolution will lean on. It could be seen as a train that has only recently started to move and countries like Spain need to make sure they catch it!

From the scientific policy point of view, the nano-technology is one of the key topics to take into consideration. Performance lines should be drawn at national and trans national level to allow not only its follow up but also to actively participate in its generation and gradual transfer to industry. It is essential to be aware of the evolution of nano-technology to be able to make a prospective analysis of those trends that nano-technology reveals and establish possible future scenarios. It is also necessary to identify its current resources (laboratories, researchers, etc.) at a national and international level. Once aware of the state-of-the-art of our research teams, it will be time to outline an adequate policy to our National R&D Project. The objectives within the wide spectrum of nano-technology should be set up after a detailed analysis of nano-technology

Evaluation of initiatives and research projects on nano-technology will be guaranteed if a previous action plan exists providing information about the parties, their backgrounds and capabilities. Evaluation has always played an important role from the appropriate distribution of resources and is a keystone that should rely not only on quantity aspects of research and researchers that present proposals but also on the scientific and technological opportunity of its project.

Evaluation has traditionally been sustained on past and present capabilities of evaluated groups rather than on the future incidence within a framework of common interests.

By applying foresight methods to evaluation, the project will also be evaluated according to its opportunity on previously defined scenarios. This is one of the great values of foresight, to provide a framework for the future in which evaluated projects could be adapted in this way and only at experimental level. In the case of nano-technology it would be possible to try a combination of evaluation and foresight.

5. CONCLUSIONS

After more than 10 years of intense work in the evaluation field, ANEP has successfully spread an “evaluation culture” in Spain. This culture is a further sign of the interest shown by scientific policy makers in our country to develop a top quality research. This has been reflected at an international level, where the improvement of Spain’s position is already a fact.

However, these achievements demand an additional effort in analysing the way forward, the reasons to take one way or another and to really know our possibilities towards the future. Evaluation duties should always bear in mind this strategic vision and make good use of indicators and criteria that facilitate such work.

Foresight, both as analysis of possible futures made from different perspectives and languages and as a mobilising process integrating various visions, facilitating communication between parties, promoting consensus and conciliating opposite interests, can be seen as complementary and enriching element of evaluation.

For all these reasons, ANEP has decided to focus in foresight as added value for evaluation. Our activities are and will be based on the support provided by the scientific and technological community that collaborates with the ANEP on a regular basis as well as on those institutions that demand our services.

Foresight should not be considered sole domain of one body or entity, but as a mobilising process that leads to a constructive and open attitude to the future. To build a common future in Europe and within our own individual countries it is important to facilitate the inclusion rather than the exclusion of parties, to profit from gained experience and to favour a dialog that will lead to consensus and credibility.

ACKNOWLEDGEMENTS

Authors acknowledge Mrs. Laura Ramos for her valuable help in the manuscript correction.

CAN RESEARCH ASSESSMENT CRITERIA BE USED BY POLITICAL SCIENTISTS? CAN THEY BY USED BY PERSONS IN OTHER FIELDS THAN SOCIAL SCIENCES?

Stefan Kuhlmann
Institute Systems and Innovation Research, Germany

In principle, I fully agree in many of the things that have been said in the preceding presentations. To a very high degree I share Mogens Petersen's sceptic view in terms of university resource evaluation practises. As Dagmar Simon and Martina Räbbecke I see a need for more self-evaluations procedures in the research institutions, and I also understand Hanne Foss Hansen's warnings against the problems of too mechanistically applied indicators in evaluations. Nevertheless, I want to use this opportunity to present a more strategic view on evaluation policy making and the role of evaluation within that. My attempt is to put evaluation or assessment in the context of related policy making. In order to do that, I will briefly illustrate the complexity of my systems of references.

The German Research System

One has to have the German research system in mind in order to understand my presentation. Figure 2 (p. 58) is an attempt to describe the German research system. The vertical axis illustrates the type of research carried out in various institutions of the research system, and the horizontal axis illustrates the differentiation between the fully publicly funded research and the fully privately funded research. The different areas illustrated in the graph characte-rise the related institutions in the system. The size of the different areas illustrates the amount of money spend in the different institutions annually.

In the future we will see a growing development of europeanization and the emergence of a European innovation policy of research. One of the reasons for showing you this figure of the German system is to ensure you that the diversity of the future European research system is at least as complicated as the German system. We will face problems of complexity at the same size as we have faced in this system.

Figure 2 (p. 58) illustrates the flow of public money. More theoretically or conceptually one could also illustrate the system in this way: One could design it as an arena with various stake-holders in it, which belongs to various parts of the society of the research system. One have the scientist actors in the narrow sense, the industrial actors, the variety of policy actors in that field, and depending on what area of research technology we are looking at, other society actors are also trying to participate and influence the content and the directions of policies. All the stakeholders are expressing their expectations of what research innovation should do. The fact that there is no dominant political player in this game is very characte-ristic for the German system. This applies for all levels both the national and the regional and federal state level. Of course there are some relatively strong political bodies, but they do not dominate the whole system in a top-down way. *The National Ministry for Research and Technology* is within this system a prominent player, but not a dominant player, which I will return to later on in my presentation.

I would like to relate these general considerations to the theme of this seminar. Listening to the previous presentations, I got the impression that we probably should differentiate our subject in at least three dimensions of assessment. What I have heard today has manly been concentrated on the dimension of the assessment of the individual performance of researchers, departments or research groups. One can also keep in mind the arena model. The conditions under which the research is made are of highly importance. Most of the things that are done are not only depending on the quality of what people are doing, but also on the frame condition under which they are working. These frame conditions have developed over a decade. For instance, if an institute receives 100% institutional funding it is a very important frame condition, which influences the way this institute can do its work. Other institutes may only receive 15% institutional funding and other 15% of their funding by programmes, which they have to compete for. This of course means that they follow other incentives, which have consequences for the way in which they perform their research. So we can differentiate these innovation policies and try to assess them. All this can be embed-ded in a general strategic discussion of what is going to or should happen in the research system as such. One could talk about the appropriateness of certain funded research fields for given institutions in relation to their cost and their performance.

As you know there is a variety of evaluation methods which have been developed during the last three or four decades. I do not wish to go into a detailed description of these methods, but simply say that we have qualitative and quantitative methods according to type, data and kind of analysis that in principal can be used on all levels - from the work level to the strategic level.

Research Evaluation Practices in Germany

Figure 3 (p. 59) illustrates the history of the evaluation practices in Germany. In Germany we have a historical development concerning the application of assessment tools and evalua-tion tools. It concerns the assessment of the quality of single projects and researchers by peer reviews and recently also by bibliometrics. During the 1970s, 1980s and 1990s the assessment was supplemented (not substituted) by an attempt to assess the social econo-mic impact of programs. By the way, it is worth noticing that 40% of the nearly 10 billion Marks spend by the German Research Ministry run through programs and not through institutional funding. More recently in the 1990s one find attempts to assess the perfor-mance of institutions, which we have heard about previously in the seminar.

In this quite complex innovation system large parts have never been evaluated so far. However, things are changing. Universities on an institutional basis have only been evalua-ted in a few cases. There are some very new attempts to assess this kind of institutions in a systematic way (figure 4, p. 59). The dominating kind of assessment is the peer review within the DFG funding, which by the way has to be mentioned in the context of universities since most of this research is carried out within the universities. However, keep in mind that assessment is not institutionally but only on a project basis. The most impact analysis we find in the area of federal technology and innovation programmes developed in the 1980s and 1990s.

How and in which way can we use evaluation methods as a tool for improving the strategic debate on the further development of research systems in general and on selected areas of research? If it is true that there is no dominant actor in this field, the evaluation cannot simply be applied in a top-down manner. Political decisions have to be made by various actors. Evaluations should be used more as a moderating or mediation tool trying to put advanced information on research performance into the arena as a basis for contested decision making. It is important to know that evaluation, at least as I am trying to illustrate it, will never happen without a certain piece of authority in it.
I have a concrete example within the university system. It concerns clinical research at university hospitals. Clinical research at university hospitals in Germany has repeatedly

been judged as a research without progress compared to other highly industrialised countries e.g. United Kingdom, the Scandinavian countries, the Netherlands and especially compared to the United States. In relation to the size of the sector, this sector consumes 40% of the university budget in Germany. There are of course exceptions, but the general tendency is quite a low level of research activity and quality, which especially is true when it comes to clinical research. Design counsellors have repeatedly told this fact during the last twenty years, but nothing has happened. The reason why nothing has happened is mainly because the universities in general and the hospitals within them enjoy a high degree of autonomy, and at the same time within this autonomy the hospitals are highly scattered around single chairs.

The other problem is of course that the universities and especially the hospitals have to serve various purposes at the same time. They have to serve the education system, they have to serve the health system and they more or less also try to serve the scientists. Only a few hospitals collaborate with industry, but they are all indirectly related to the political system, which provide them money. Therefore, a couple of years ago, the federal research ministry in Germany decided to launch a program which aimed at the creation of new clinical research centres in the context of the university hospitals. This was due to the fact that they realised that they would never achieve a change in the existing structure concerning the existing institutes and the existing chairs without setting a new strategy. The new strategy was to tell the university hospitals that they would receive a considerable amount of money for many years if they manage to create a new research infrastructure bringing together various chairs and various institutes in an interdisciplinary context. If they succeeded the university hospitals would receive 50 million marks distributed over the following eight years. This idea was organised as a competition. 20 universities wanted to be a part of the new project, but only eight universities succeeded. The eight centres have now been existing for 2½ years.

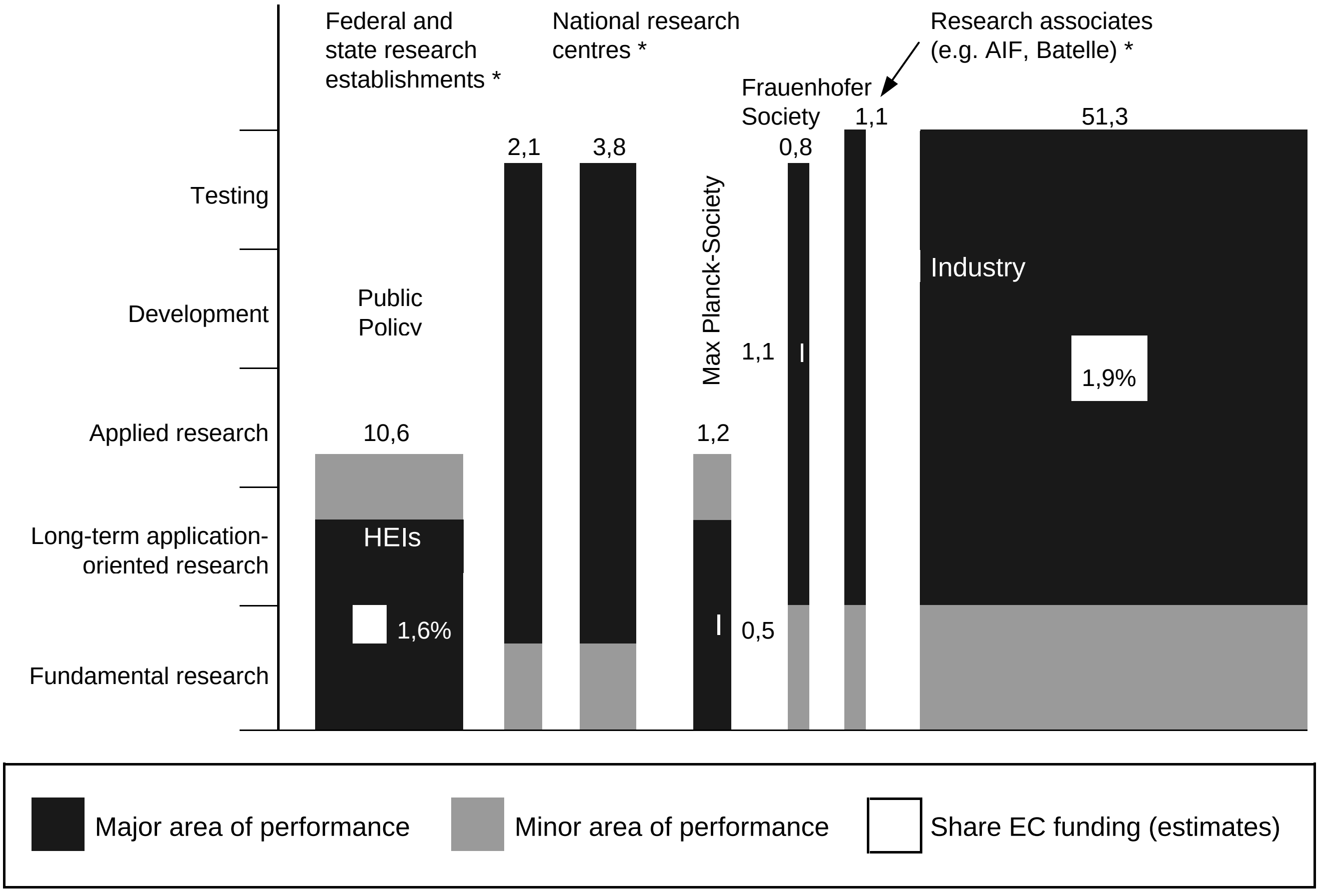
The arena of health research is at least as complex as the one I mentioned in general terms earlier on. The problem is that from the very beginning it seems to be clear that there was a high danger for these new interdisciplinary research systems to fail. Therefore it was decided to accompany the whole approach with two kinds of evaluations. One of the evaluation methods was to attach a scientific board to each of the centres. The scientific board consisted of peers mainly from foreign universities. These peers should evaluate the quality of the research projects done within the centres. At the same time a structural monitoring evaluation project was created, which tried to look critically at the structural frame conditions in order to help the centres understand their own situation better. However, we are still facing the problem that the people who are running the centres are confronted with contradictory interest. Therefore, we cannot follow the normal way of doing impact evaluations. We cannot only concentrate on certain kind of indicators, but we have to develop these evaluations as multi-prospective evaluations, taking into account all the various perspectives.

In the beginning there was a high reluctance towards the two evaluation method and their policy, but now after 2-3 years the different actors on all levels of research and science policy are asking for these kind of analysis and evaluation methods.

Figure 1: Outlook

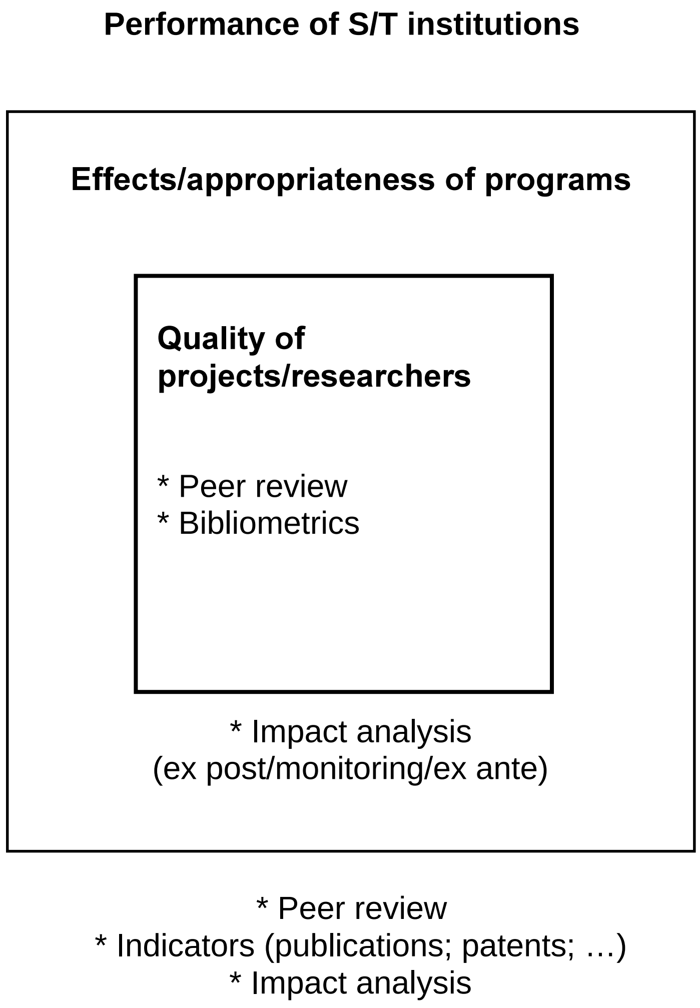
- Evaluation as Communication/Moderation Process
- Combination of Evaluation Practices and Foresight Procedures /Technology, Social Development, Policy Options)
- Use of Advanced Procedures as Basis for Strategic Debating and Shaping of S/T Policies
- ”Advanced Science & Technology Policy Planning:
  Towards the Integration of Policy Evaluation, Technology Foresight and Technology Assessment” (ASTPP 1996-1998, funded by EU)

Figure 2: Science & Technology System in Germany
(expenditure in billions of DM, 1991, old Länder)



- No information supplied on EC funding

Figure 3: S/T Evaluation Practice - Historical development in Germany



Performance of S/T institutions

Effects/appropriateness of programs

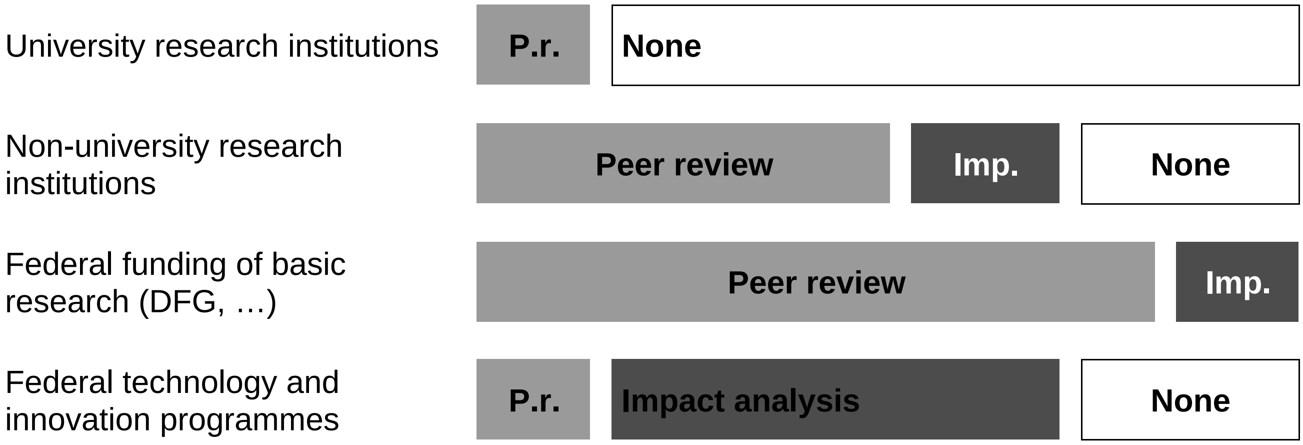
Quality of projects/researchers

* Peer review
* Bibliometrics

* Impact analysis
(ex post/monitoring/ex ante)

* Peer review
* Indicators (publications; patents; …)
* Impact analysis

Figure 4: Research evaluation practices in Germany



University research institutions: P.r. | None

Non-university research institutions: Peer review | Imp. | None

Federal funding of basic research (DFG, …): Peer review | Imp.

Federal technology and innovation programmes: P.r. | Impact analysis | None

REFERENCES

1. Casado J., Presmanes B., Guerrero H., La prospectiva como apoyo a las tareas de evaluación: Experiencia de la Agencia Nacional de Evaluación y Prospectiva (ANEP). *European Workshop for the exchange of experiences in the field of technology foresight*. Oviedo (Spain), 26 & 27.11.1998.

2. Presmanes B., Guerrini A., (1998) Procedure e metodi per la valutazione della ricerca in Spagna, in *La Valutazione della Ricerca, Roma: Università Ricerca*, Numero 3 , pp. 39-41.

3. Ben R. Martin (1996), Technolgy foresight : capturing the benefits from science-related technologies. *Research Evaluation* V.6 nº2., UK.

4. Presmanes B. (1998) Chapter XXI. Las previsiones tecnológicas en la Agencia Nacional de Evaluación y Prospectiva (ANEP). J.F. Tezanos and R. Sánchez Morales (eds.) in *Tecnología y Sociedad en el nuevo siglo*, Madrid: SISTEMA, pp. 635-667.

5. Guerrero H., Casado J., Presmanes B., Sorozábal T., Vázquez L. (1998) Study of the Scientific and technological capabilities on critical technologies. An example: the Optics in Spain from 1987 to 1996. *The R&D Management Conference. Technology Strategy and Strategic Alliances. Proceedings*. Madrid: COTEC.

6. Guerrero H., Antón m., Sorozábal T., Millet C. (1998) *Estudio de la capacidad científico técnica española en el campo de la Óptica a través de las publicaciones científicas en el decenio 1987/96*, Madrid: ANEP-BBV. (Freely accessible by **www.seui.mec.es/Inves_Cientifica_Tec/anep.html)**.

7. Blanco M.T., Presmanes B. (1998) Primer informe sobre: Tendencias en la conservación del patrimonio cultural: Demandas tecnológicas y científicas. Madrid: ANEP. (Freely accessible by **www.seui.mec.es/Inves_Cientifica_Tec/anep.html)**

8. Romero L., Ramos Saralegui L. (1998) Políticas y tendencias en la investigación agrícola en el ámbito del Mediterráneo. B. Presmanes and A. Guerrini (Coord) Madrid: ANEP-CNR.

SELF-EVALUATION AS A CONTROLLING INSTRUMENT

Dagmar Simon and Martina Räbbecke
Wissenschaftszentrum Berlin für Sozialforschung, Germany

In this presentation, we go into the contextual factors of science policy and research policy underlying the Federal Republic of Germany’s present discussion about different evaluation procedures. We will first outline the discussion as it stands in the universities and in publicly funded non-university research institutions. We will then explain our self-evaluation approach and present our current project, which deals with instruments and procedures of self-evaluation.

International economic competition in recent years has had direct impact on the various systems of science in nearly all western industrialised countries. Because it is assumed that knowledge-based technologies will play a major part in staying ahead and securing future markets, such phrases as ”international competitiveness” and ”market orientation and a focus on innovation” have become prominent additions to the vocabulary of science policy discussions. Germany is no exception. The federal government has set research policy priorities in the areas referred to as ”future technologies,” and the productivity of scientific institutions is being critically examined. In view of country’s the dwindling resources for research and technology, publicly funded scientific institutions are being expected to publicly demonstrate high-quality, efficient, and profitable performance. These expectations are the background of today’s ongoing discussions about objectives, instruments, and procedures of evaluations and of the planning and monitoring entailed.

In the higher education institutions in the Federal Republic of Germany, debates about objectives and instruments of evaluations began around the mid-1980s, much later than in other countries. Given the obvious scarcities in the universities, evaluations centred on teaching and learning, not on research. These efforts, however, encountered considerable resistance within the universities. A fundamental change did not occur until the early 1990s, when German reunification strained university budgets. At the same time, the ethic of equal treatment increased pressure on West German institutions of higher education to offer evidence of their accomplishments just as required of their counterparts in the new federal states, all of which were undergoing detailed evaluation.

The influential German Rector's Conference initiated the first pilot projects on quality control in 1991. This step was prompted by three considerations in particular. First, documenting productivity can be important for gaining and retaining public acceptance and hence for receiving government funding. Second, in a bid for increased autonomy, the universities were willing to accept evaluations and accountability if they were largely freed from government regulation. Third, some universities had meanwhile had positive experience with evaluation procedures and had recognized them as a way to improve teaching and learning.

The different procedures for ensuring quality are meeting with increasing acceptance, a response that may well have something to do with the fact that research issues are not usually involved and that funding has been only marginally affected by the results of evaluations. In other words, the impact of evaluations is limited by the strong position that professors have and by comparatively weak academic decision-making bodies. It is true that differentiation and emphasis on setting priorities have increased both within and across universities, but it has been recognized that this contentious processes has to be combined with a professionalization of university management and with new decision-making and control structures. Seeking future models for decision-making is therefore one of the central topics in discussions of university policy.

A significant amount of the research in the Federal Republic of Germany is conducted in publicly funded institutions outside the university system and industry. It is them we now give attention. They are supported by both the federal and state governments. There are four of them in all, each pursuing largely its own type of research. The Max Planck Society is dedicated to pure basic research. The Fraunhofer Society conducts applied research. The Helmholtz Society has several large-scale research institutions. The Wissenschaftsgemeinschaft Gottfried Wilhelm Leibnitz (WGL) is an exception, which is described in detail below. By the way it also includes the Social Science Research Centre Berlin (*Wissen-schaftszentrum Berlin für Sozialforschung*) - the WZB -. Though all these institutions are publicly funded for the most part, concerning research none of them is directly dependent of federal or state ministries. Each is protected by the academic freedom of knowledge guaranteed by the country's constitution, the Basic Law of the Federal Republic Germany.

The non-university research institutions are presently confronted with evaluations in two respects. Firstly, the federal and state governments call upon the research institutions to undergo external evaluation. These evaluations, so-called system-evaluations, are designed to assess basic structural and organizational principles rather than scientific achievement. Secondly, the federal and state governments expect each research institution to develop its own internal procedures for quality control. Unlike the evaluation procedures of the universities, those of each non-university research institution are to be used in all of its institutes. As internal controlling instruments, these procedures are intended to have direct impacts on research planning and the use of resources.

Most of the non-university research institutions have been slow to comply with the government's calls for evaluation, for they fear that evaluations will inevitably restrict their scope of action and increase pressure to pursue the objectives of the federal government's research policies. Externally conducted systems evaluation of these autonomous institutions is unprecedented, and calls for uniform evaluation procedures have thus far been parried on the grounds that each institution has its own quality-control procedures. Nevertheless, there are indications that research institutions such as the Max Planck Society have not only bowed to political and financial pressure. In recent years, for example, voices within the Max Planck Society itself have expressed concern about whether the principles underlying its internal structures are commensurate with the demands of modern interdisciplinary science.

The fear of losing autonomy is not the only reason that these institutions have been reticent about undergoing evaluation. Thus far, external evaluations in the non-university research sector have become known primarily as processes for deciding on the continuation or closure of research institutes. The institutes of the WGL in particular have been confronted with evaluations of that kind. Unlike the other non-university institutions we have mentioned, the WGL is quite heterogeneous. Along with institutes dedicated to pure research, it also has libraries and museums. It includes institutes in the fields of natural science and engineering as well as a great number of institutes in the humanities and social sciences. Its work ranges from basic research to applied research and development. The one common feature is that all these parts are funded by the both the federal and the respective state governments.

Since the mid-1970s, the federal and state governments have agreed to regularly assess whether the non-university research institutions deserve funding. The key criteria are ”supra-regional significance” and ”relevance to the science policy of the nation as a whole.” These evaluations were adopted by the German Science Council, an influential advisory body in the area of research policy. Half of its members are high-ranking representatives of scientific organizations, and the body enjoys great esteem in the scientific community.

In 1990 and 1991 the German Science Council conducted its first systematic, comprehensive evaluation of several research organizations and their institutes, including those of the Academy of Sciences of the German Democratic Republic (GDR). Working quickly and using a highly controversial procedure, the Council used the results of these evaluations as the basis for deciding which research institutes were to be terminated and which research groups were to be integrated into the West German research system. Many of the evaluated institutes were given a chance to survive as part of today’s WGL. It encompasses 80 organizations, half of them stemming from the former GDR’s Academy of Sciences.

This systematic evaluation by the German Science Council has continued with a slightly modified procedure since 1995. The stated objective is to examine the quality of all institutes and, by a process of elimination, to create the financial conditions to include other, more efficient institutes in public funding. The striking thing about this second round of evaluations is that the federal government’s previously mentioned orientation to the market and innovation has not become the primary measure, despite the funding criteria set forth in official research policy. Presumably, the composition of the German Science Council and its groups of evaluators on site have done much to lend the evaluation procedures a markedly academic bias. However, this emphasis, too, is a cause for considerable concern to the institutes undergoing evaluation, given their heterogeneous research tasks.

What are the merits of self-evaluation procedures in research institutions? What are the intended objectives? In the first part of this presentation, we concentrated on the contextual factors of science policy, in other words, on what it is that motivates actors in science and research policy to establish procedures for ensuring quality research and teaching. We shall now leave this terrain and turn attention to the interests and motives of the major actors in the research institutes. In doing so, we would like to change the perspective by casting an eye on procedures for monitoring and assessing what these actors have accomplished.

Reduced resources, increased external pressure, intensified calls for legitimation, and demands for quality ”products” are confronting research institutions with new challenges to their management competence and their ability to monitor themselves. However, many of the institutes of the WGL, which are the main interest in this presentation, have an organizational structure that cannot cope very efficiently with the demands of modern research management, especially when resources are increasingly scarce and competitiveness is sought. The long-dominant opinion that the excellence of the researchers ensures quality research and hence the survival or even further development of research institutes no longer bears up. As shown by a close look at the everyday research world, that assumption has led to a squandering of resources, especially of human resources and time, and that waste can no longer be legitimated.

The changed conditions drive home the point that internal procedures are needed to check and foster the productivity of non-university research, for research outside the university system is under particular pressure to prove its legitimacy. We call these procedures processes of self-observation that permit critical consideration of the research institution's productivity. Their purpose is to ascertain the institution's current performance, analyze its strengths and weaknesses, and develop perspectives on future development. The institutionalization of such procedures is based more on an understanding of learning processes than on routine control.

Evaluating performance and ensuring the research quality can succeed credibly only if the instruments and procedures of self-monitoring take the institution's organizational structure, tasks, and modus operandi into consideration and thereby capture what sets the institution's apart. In essence, our research project also has this goal. It is an attempt to modify and develop existing procedures and instruments so that they meet the needs of the WGL's diversely structured institutes and remain in keeping with both the research orientation of the respective institutes and the intentions of self-evaluation procedures.

The self-observation procedure conducted in the WZB in 1994 is the point of departure for this project. It was necessary to develop instruments that take into consideration the general standard criteria of research evaluation as well as the special criteria of the WZB's program of problem-centred basic research and its specific contexts. In an interactive process involving the president of the WZB, the directors of the research areas, the fellows, and the organization's democratic decision-making bodies, the proposed instruments were discussed, critiqued, and developed. This approach underscored our conviction that effective evaluation rests on basic shared understanding among the actors.

The set of qualitative instruments in this repertoire is focused on the research units of the WZB. Respondents were asked to describe the objectives and projects of their research program, the thematic linkages with other research units and external co-operating partners, and the ways in which the research results were communicated. Assessments of strengths and weaknesses, the structural conditions bearing on the research, and future orientations were also inquired into. This information provided the background for an appropriate interpretation of the quantitative data from the descriptors, which we have categorized as input (basic funding, positions, features of the scientific personnel), throughput information (project features, co-operation with others), and output information (publications, third-parting funding). The significance these factors have as indicators depends "on the special characteristics of scientific practice, which differ substantially from one discipline and research field to the next" (Neidhardt, 1995). The intention behind this procedure is to lay the groundwork for statements about institutional productivity, the conditions governing it, and the possibilities for continued development and to initiate a process of discussion about implementation. At no time has there been any attempt to gather data or descriptions pertaining to the performance of individual fellows.

The kind of institutionalized system of self-observation that has been tested on WZB research practice also entails the danger of routinization, which is not conducive to the intended critical reflection on the status quo. An instrument of this kind must therefore be continually examined and developed to ensure that it remains able to help a research institute and its environment perceive problems and formulate goals precisely. Even then, this conception of self-monitoring procedures cannot eschew a specific internal view or perception of problems, nor can it avoid blocking out processes of social development or scientific reorientation that could become important for the institution's research concept. For that reason this procedure cannot and should not replace external evaluations. On the contrary, it is suitable for supplementing and preparing external peer reviews, that is, for conceiving of research evaluation as an interactive process.

In our investigation, five case studies on the WGL provide the foundation for an attempt to develop procedures of self-observation and self-monitoring and to develop discerning criteria for evaluating research. In order to come anywhere close to representing the diversity of the institutes within the WGL, the selected organizations are quite heterogeneous. They represent a variety of disciplines and are either pursuing primarily basic research or are engaged mainly in applied research and/or consulting.

Initial results of our study show:

- the significance that the structure and organization of research has for ensuring the quality of the work, especially the success of internal self-evaluation processes;
- the significance of different research concepts: basic research, problem-centred basic research or applied research;
- very different understandings of the objectives and tasks of research evaluations; and the issue of having a strictly academic bias to the German Science Council's evaluations of the institutes, which are committed - both by themselves and by others - to ensuring the social relevance of their research.

METHODOLOGICAL PROBLEMS / WARNINGS
DISCUSSION OF THE VALIDITY OF THE RECOMMENDED CRITERIA

Hanne Foss Hansen
University of Copenhagen, Denmark

Introduction

My background for speaking about this is partly research carried out within the sociology of science and science policy studies, among other things two projects on research evaluation, partly practice as member of evaluation committees concerning two research programmes, one in Denmark, one in Norway, a research centre in Norway, and as member of a committee designing and supervising department evaluations at the Copenhagen Business School. In addition my own department, the Political Science Department at University of Copenhagen, has been evaluated, in fact by some of you, and at last I have as a member of the Danish Social Science Research Council read and commented on a number of different evaluations and policy documents concerning evaluation in the last six years. I am not a member anymore so recommendations are strictly personal.

Under the heading “Methodological problems and warnings. Discussion of the validity of the recommended criteria” at least two approaches to the discussion can be chosen.

One approach is narrow and technical discussing different criteria for assessment, e.g. the validity of counting publications, counting citations, using journal impact factors and so on in order to develop measures for quality in the scientific community.

Another approach is broader discussing both the organization and the context of evaluation as well as the more technical methodological problems. If we aim at assessing assessments not only on a theoretical level but also in relation to experiences and utilisation, we have to choose the second approach - to discuss evaluation design in relation to both content and context. And that’s what I will be doing in the following. I will be discussing evaluation in a historical, methodological perspective, characterising four different models or standards used for or proposed for evaluation. The four models are:

- the classical peer review model;
- the modified peer review model;
- the informed peer review model; and
- the performance indicator model.

Each model or standard will be characterised on the dimensions: process (what’s going on?), organization (how is the work organised?), output (what are the results of the process?), logic (which is the more principal philosophy the standard is based on) and warnings (which are the weak sides of the standard?).

Having characterised the four models and the problems and warnings related to them, I will conclude by drawing some recommendations for evaluation practice.

First of all, **the classical peer review standard**.

Classical peer review

Process: reading basic documents (narrow focus: a limited number of persons, research proposals or manuscripts)

Organization: several individual assessments or committee work (overlapping competence)

Output: Yes/no/work harder recommendation/decision – clear agenda

Logic: professional, meritocracy

Warnings:

- time-consuming
- disagreements reflecting the complexity of the concept of scientific quality and especially in the social sciences and the humanities cognitive particularism
- conservatism?
- the extent of institutional particularism very much debated
- characterised by the luck of the reviewer draw

You are all familiar with this I am sure. The process central in this standard is reading, reading basic research documents. The focus of this standard is narrow, on a limited number of persons applying for appointments, on a limited number of research proposals applying for grants or on a manuscript seeking publication in a journal or seeking an award or a degree of some kind. The organization of the assessment may be either several individual assessments asked for by an editor or a research council or committee work implying a discussion of concepts and levels of quality. To some extent assessors in this model have overlapping competence making possible a kind of interpersonal social control.

The agenda for the assessment is clear, the output of the process is a recommendation or a decision implying either yes, no or work harder. The logic of the standard is the professional meritocracy.

What are then the problems and warnings?

First of all, these kinds of assessments are time consuming.

Secondly, from time to time, some would probably say often, there are disagreements, reflecting the complexity of the concept of scientific quality, and especially in the social sciences and the humanities reflecting cognitive particularism due to competing paradigms.

Thirdly, the logic of meritocracy may lead to conservatism and Mathew effects in recommendations and decisions, thus there may be a risk that classical peer review limit the development of new paradigms and research fields.

Add to this that the extent of institutional particularism is very much debated. Do classical peer review assessments reflect the existence of old boys network? Do they discriminate according to gender, colour, size of departments or whatever? Different studies show different results. In fact there is a rather huge bulk of literature on biases in classical peer review. What seems to be for certain is the existence of the luck of the reviewer draw. Some of you probably know the classical study of National Science Foundation (Cole, Cole and Simon, 1981:855) showing that 25% of research proposals moved from the category worthy of support to refusal or vice versa, when proposals were re-refereed. From the point of view of the evaluated classical peer review is to some extent a matter of chance, as very many other things in life.

Assessing the standard of peer review my personal conclusion is, that it is not at all a perfect standard but nevertheless it is an indispensable element in the reward system in the scientific community. What is important is continuously to reflect upon whether procedures can be developed as to function better and more fairly.

Secondly **the modified peer review standard**.

Modified peer review

Process: reading overall profiles, descriptions, statistics, research plans and self-evaluation reports: site visiting (broad focus, one or several departments, a programme, a discipline or a research field)

Organization: committees made up on the principle of complementarity

Output: qualitative assessments – unclear agenda

Logic: professional, meritocracy but audience consisting of both professionals and political-administrative actors

Warnings:

- time-consuming
- cognitive particularism
- institutional particularism?
- luck of the reviewed draw – conflicts in the selection of peers
- tendency towards soft conclusions
- tendency towards drift against organisational focus
- mainstreaming

For the last 15 years, in some European countries probably even longer, the modified peer review standard has been widespread. In Denmark the standard was introduced in the beginning of the 1980's by the science policy advisory body, at that time called Council for Research Policy and Planning. Inspiration partly came from Sweden from the practice developed by the Swedish Natural Science Research Council.

The modified peer review standard has a much broader focus than the classical peer review standard. It has focus on one or several departments, a research programme, a discipline or a research field. The process includes the reading of overall profiles, descriptions and statistics, research plans, maybe self-evaluation reports and it includes site visiting. Some-times it also includes collecting a number of important publications, but are they read by panel members? It is my impression that panel members involved in modified peer review processes typically do not have working conditions giving room for reading basic research products systematically. In fact the only example I have heard about where panel members seem seriously to have been reading basic research products is the discipline evaluations in Sweden initiated by the HSFR - the humanities and social science research council. These evaluations seem to build on Swedish traditions of thoroughness but then an evaluation of a discipline also takes about 3 years.

The work is done by a committee made up in order to cover the area in focus. Competence is no longer overlapping but complementary, often there are even uncovered gaps. To mention an example which some of you are familiar with, the important field of public administration, was not well covered, if at all covered, in the evaluation of Danish political science.

Concerning output the agenda related to the use of the modified peer review standard is most often unclear. The outputs are statements and opinions qualitative in their character.

The logic of the standard is still the professional meritocracy but the audience differs from the classical peer review standard, the audience includes both professionals and actors in the political-administrative system.

What are then the problems and warnings of this standard?

First of all the modified peer review standard has common problems with the classical peer review standard, but due to the principle of complementarity in the composition of committees, and thereby the loss of both professional competence and interpersonal social control in committee work, problems grow even stronger in the modified peer review model.

Using the standard is still time consuming. Cognitive particularism is a serious risk. I have seen several examples of this in my time in the research council. Also institutional particularism is a risk. The luck of the reviewer draw certainly is dominant, which is also illustrated by the discussions and fights most often going on in selecting and recruiting peers to committees.

But there are other problems too. Due to the political context of the process, and the mixed audience, assessments often tend to become formulated in soft language, and assessments tend to focus on organisational questions (e.g. the size of departments, networking with other departments international and national as well as with users, the organization of Ph.D. programmes and so on) rather than on scientific profile and quality. A focus on organisational questions of course may be relevant but the problem is which expertise do peers have concerning organisational questions? If organisational assessments are requested call for or at least include organisational experts.

Further there seems to be a risk that modified peer review lead to mainstreaming. There are to my knowledge no studies of this, but there seems to be a risk that modified peer review evaluation lead to standardisation of research profiles, publication profiles and organisational features across departments and across countries. If this is the case, the question whether it is a desirable development ought to be discussed.
Assessing the standard of modified peer review my personal conclusion is that using this standard most often results in wise opinions and suggestions to be discussed and

considered, but also that conclusions are no more than wise opinions. There is no reason to believe that the conclusions of committees are objective or to use a grand word is the truth. As part of reflexive processes in the scientific community modified peer review undoubtedly is fruitful. But there seems to be a tendency that some actors in the audience ascribe to much authority to the conclusions.

Thirdly **the informed peer review standard.**

Informed peer review

Process: more systematised, quantified data as input; otherwise a process very much like modified peer review

Organization: committee; complementarity

Output: quantified assessment: rating

Logic: translation of professional assessment to bureaucratic system

Warnings:

- the concept of scientific quality is multidimensional: is it reasonable to add up assessment of scientific quality in one mark?
- mechanical coupling between rating and resource allocation strongly influences behaviour; productivity rather than quality is increased

The standard informed peer review might be understood as a reaction to the tendency of the modified peer review standard to either drift away from assessments of scientific quality or drift towards very soft conclusions. Informed peer review compared to modified peer review implies two things, that peers are informed and that peers are informing. First of all assessments are based on more systematised often quantified data. The idea is that peers should be better informed in order to do their work. Secondly peers are asked to formulate their results in a more clear, informing language, and asked either to give marks on several dimensions as in the Dutch system or rate departments as in the UK system.

In the following I briefly characterise the UK system, because this to me is a very clear example of an informed peer review standard in use. The Dutch system may be characterised more as an example of a mixture of the modified and informed peer review standards.

The process of the informed peer review is to some extent alike the process of the modified peer review. Only the data is more systematised, maybe it includes bibliometric analysis, probably secretarial back up is stronger and probably site visits are not used.

Also the organization of the work is much alike. A committee based on the principle of complementarity.
The big difference between the modified standard and the informed standard concerns the output. The informed standard results in quantified assessments, in rating, and the agenda is clear, the rating is input to resource allocation.

The logic is translation, to translate the assessment of scientific quality into a bureaucratic system. This kind of system is I think the fulfilment of the bureaucratic dream. Bureaucracy gains an easily managed system to be used for resource allocation but at the same time bureaucracy stays totally free from responsibility.

What are then the problems and warnings?

Others in this room have much more experience with this standard than I have. I will briefly characterise the lesson to be learned from the UK in this way:

1) The concept of quality is multidimensional. It is not at all obvious how to add up dimensions as scientific quality, productivity, scientific relevance, societal relevance and so on in one mark.

2) The mechanical coupling between marks and resource allocation strongly influences behaviour; productivity rather than scientific quality is increased.

The performance indicator model

Performance indicators

Process: collecting data, calculating

Organization: “calculator”

Output: quantified assessment: ranking

Logic: logic of auditing

Warnings:
- validity of indicators is not strong
- the easily used indicator, JIF, is senseless
- use of indicators require both knowledge about the validity of indicators and knowledge of the research field in question

Finally, the performance indicator model or standard is the radical model build on the notion that it is possible directly, without peer review, to measure scientific quality.

Following this model the central process is collecting data and calculating. The organisation is simple, call for a “calculator”, a person with bibliometric, scientometric or may bee economic skills. Outcome is quantified assessments, often ranking; and the logic may be characterised as logic of auditing.

In my opinion there are strong warnings related to this model. Most important, the validity of indicators can be questioned.

To sum up the most important elements in the discussion:

- Indicators focussing on input such as the ability to attract external grants intend to measure quality and relevance but also to a considerable extent are influenced by

differences in financial structures, networks and differences in persistence in application behaviour;
- Indicators focussing on output such as publications in general or articles published in refereed journals in particular intend to measure work performance and scientific quality but to a great extent are reflecting also productivity, publications market structures and networks.
- Indicators focussing on effect, the most important one being citation analysis, intend to measure quality but rather reflect impact, visibility and networks. There is a huge literature on citations, and warnings related to the use of citation analysis. I will not go into this in detail, but only stress that there are problems related to the databases used for citation analysis, problems related to how scientists choose references and problems related to how the development of a research field (its size, dynamics and network structures) influences citation behaviour.

Add to this the even more problematic indicator, called the journal impact factor, measuring the average number of citations given to articles published in a specific journal. As clearly documented by the Norwegian medical professor Per Seglen, this indicator only has meaning for evaluating journals, whereas it has no meaning at all used mechanically for evaluation of research. The only reasonable way to use journal impact factors in evaluation of research is to do analysis comparing what is called reel impact, that is the counting of achieved citations, with expected impact, measured by journal impact factor analysis. To do these kind of analysis however imply considerable bibliometric competence.

Summing up the **warnings** of the performance indicator model is:

- The validity of indicators is not strong
- The easily used indicator, journal impact factor, is senseless
- Use of indicators require both knowledge about the validity of indicators and knowledge of the research field in question.

Assessing the standard, I would say, that it is a good idea to have (and in Denmark further develop) statistics not only on the allocation of resources between fields and institutions and so on but also on performance, on publication patterns, citation patterns and so on. These kinds of statistics are fruitful in order to monitor the overall development in a country. But it is not a good idea to use this kind of statistics mechanically for evaluation purposes. Performance indicators simply are not objective, valid and strong enough measures of scientific quality.

Conclusions and recommendations

Now it is time to sum up and conclude. I will present two conclusions and three recommendations:

Conclusion 1 is illustrated by this figure, the researcher scale.

Figure 1: Performance indicators:
Differences between the intentions of measuring and the results

Category: Characteristics:	Input	Output	Effect
Content	ability to attract external grants	publication (types)	citations
Intended to measure	quality, relevance	work performance, quality	quality
In addition measuring	external conditions, networks, persistence	productivity, publication markets, networks	impact, visibility, networks

The figure illustrates the serious limits in the possibilities to measure scientific quality by quantitative measures in the form of performance indicators. Validity of indicators are low. Indicators have to be used very cautiously. Use of indicators require bibliometric, scientometric skills as well as professional knowledge about publication profiles, journal markets and so on. Also this figure raises a more fundamental question: Is it at all possible to measure the value and quality of creativity? And in what time horizon? We had a Nobel Prize taker in Denmark in chemistry last year - Jens Christian Skou. How was his work evaluated and measured real time 40 or 50 years ago? According to his own version of his life story other scientists at that time found his work very peculiar. How would he have managed working in the present age of evaluation?

Conclusion 2 is illustrated by this figure, the animals. Here you see an evaluator, and several research groups or departments. The evaluator says: “To secure a just selection you will all have the same examination: You have to climb the tree!” This figure illustrates the huge problem of comparison. There are very narrow limits for comparison in research evaluation. The advice in the literature always is only to compare like with like. But what in fact is like with like? To further illustrate the problem of comparison I would like to tell you a story from the real life. In a large University Hospital in Denmark around five years ago, it was heavily discussed how to allocate the sparse basic resources for research between departments. Some professors suggested that a department should gain resources corresponding the share of quality journal articles published. A quality journal article was defined as an article published in a journal with a impact factor above a given value, if I remember correct above 2,5 or 3 or something like that. There was just one problem. Citation behaviour and publication markets differ very much across fields, even within medical science. Within anaesthesiology for example only one out of the 10 most recognised journals in the figures available at that time had a journal impact factor above the suggested level. If the system had been approved the result would had been an abandoning of all research in this field. A thought-provoking example of the limits of comparison.

Finally I wish to put forward three recommendations, and now I am probably talking mostly to the Danish part of the audience.

Recommendation 1: We need better statistics and more general monitoring of the development the Danish research system. We are far behind the countries we usually compare with in this area: the Netherlands, Sweden, Norway, just to mention a few. Here is an important task for The Danish Institute for Studies in Research and Research Policy in the bibliometric part of the field preferably in collaboration with experts in bibliometric and scientometric.

Recommendation 2: I am not arguing that evaluation is impossible and ought to be abandoned. My position is: in Denmark we need to discuss and develop a new evaluation policy. After abolishing the Science Policy Research Council some years ago and recently transferring the institutional responsibility for the universities from the Ministry of Education to the Ministry of Research, old notions concerning research evaluation seem to have faded away. Maybe new ones are emerging on the horizon? To my knowledge, but I have not been analysing this in any depth, by the way somebody ought to do that, the experiences with institutional evaluation at the sectorial research institutes, that is research institutes outside the university system, are rather good. One way to go also in the university sector is to place the responsibility of evaluation locally with the individual universities and faculties, following the advice of such a prominent observer as Martin Trow from the University of California, Berkeley.

Recommendation 3: We need in Denmark to do more follow-up and analysis of the classical peer review standard both in the system of the research councils and in relation to the appointment system at the universities.

To our guest from abroad I apologise for this narcissistic closing. But I simply had to make use of this chance to try to influence the local agenda.

Selected literature for further reading:

Cole, Stephen: Jonathan R. Cole & Gary A. Simon: Chance and Consensus in Peer Review in Science, 214, 1981.

Foss Hansen, Hanne & Birte Holst Jørgensen: Styring af forskning: Kan forskningsindikatorer anvendes? København: Samfundslitteratur, 1995.

Seglen, Per O.: The Skewness of Science in Journal of the American Society for Information Science, 43 (9), 1992 (628-638).

Seglen, Per O.: Causal Relationship between Article Citedness and Journal Impact in Journal of the American Society for Information Science, 45 (1), 1994 (1-11).

Trow, Martin: Academic Reviews and the Culture of Excellence. Stockholm: Kansler-ämbetets skriftserie, 1, 1994.

CAN A COMMON SET OF CRITERIA BE ESTABLISHED ?

Mogens Pedersen
Odense University, Denmark

A Discussion of Research Assessments, which takes into consideration, the following two questions:

Can a common set of criteria be established?
Can we avoid inventing the wheel every time a new assessment is made?

I was somewhat surprised, when I noticed these questions in the programme. My first reaction was to find the questions rather silly. I answered them immediately with a NO and with a YES.

NO, we cannot establish a common set of criteria for all branches of learning. It is just impos-sible. YES, we should of course build upon past experiences in this field, as we do it in other fields of science.

After a little while I did, however, start to speculate and to waver. One has to admit that mea-surements and indicators based on various criteria already exist for assessment purposes, either one likes it or not. They are being used in many scientific disciplines and to some extent also used in a number of research assessment exercises within the social sciences, including, alas, my own field, political science.

My immediate reaction to question number two was, as already mentioned, an affirmative one. We do not have to start all over each time an assessment is made. We can build on past experiences and cumulated wisdom. We may e.g. just copy, totally or partially, either the English or the Dutch model. At least in these two countries a lot of experiences have been made with a variety of quantitative and qualitative indicators. Several rounds of assessments have been completed, and the procedures have become routinized. The question to be discussed really is, whether we should try to copy these institutionalized systems or not, or, how much we should eventually copy. In the gloomy light of the not too happy experiences with the most recent Danish assessment of economics and political science the two questions also ring some bells.

The idea behind the two questions seems to be the assumption that it is possible to base research evaluations upon one and only one set of criteria, which are considered legitimate and good by everyone and everywhere, and which consequently will be accepted, as soon as they are invented or discovered, and will be used thereafter forever and forever.

But as is the case with all such general normative rules, you should always expect to find people, who do not want to adhere and obey. People, who will say that it is possible on defensible value premises to raise arguments against and fight any kind of common criteria of research evaluation. Given my practical and theoretical experiences I tend myself to take this negative view of the “common criteria business”, even if I have to admit that my own practice does not always conform to my principled objections.

Looking closer at the two questions I did, however, notice a semantically problem, which I will have to address, before we are able to proceed.

Who are *We*?

In my mind the most interesting problem with these questions is the ambiguous use of the word *we*. Can criteria be established - by *whom*? Can *we* avoid inventing the wheel every time a new assessment is made? Who are *we?* Are *we* those who decide on evaluations? Are *we* those, who carry out the evaluations? Or are *we* the subjects, or, as I will call them henceforth, the *victims* of the evaluation? The problem is that these *we's* can be very different people in very different situations. The situation and the problems tend to look different, when observed from the various points of observation. The interests of the groups of *we's* can be very different indeed. They may even be contradictory.
Let us look briefly at some of the actors in the assessment processes.

There are some people, whom we might describe as the A*ssessment Bureaucrats*. They differ from country to country, dependent upon the organizational setup of the political and administrative agencies that try to control national research matters. You may even wish to make a distinction between the political masters, the ministers, and their administrative helpers. These people do not put a lot of emphasis on openness about their agenda. They are, after all, the masters. They call the tune. It is often very hard for other participants to see, what really is going on in the research assessment, exactly because the real agenda is at the same time, complex, multi-dimensional, and not at all public. There could be a lot of hidden agendas each time there is an assessment. This is not only the case in Denmark, where it is quite obvious to everybody, but in other countries as well, e.g. in The Netherlands, where I once had the chance to participate in an assessment exercise. Those, who participated as external evaluators in the Dutch 1996-evaluation of political science, were never sure of what kind of hidden agendas there were behind everything that happened. This was the case, even if the Dutch assessment methodology otherwise should be praised for its openness. Seen from the perspective of the victim a lack of openness about the intentions of the *Assessment Bureaucrats* is of course an even more serious problem, than it is for the evaluators. But *Victims* and *Bureaucrats* never meet during the assessments. That is part of the game!

There also is another group of actors in assessments. These people I shall call the *Assessment Experts* or the *Assessment Executors*. They may or may not be experts, but they are defined as those people, who conduct the assessment. They tend to develop a peculiar relationship with the assessment bureaucrats. I have never been able completely to under-stand this relationship, but it is beyond doubt that it is very much conditioned by the way in which the experts are selected, how well they are paid, and to what extent they are international or local experts people. This is an important relationship, which we have to discuss. Let me here only mention that international experts are often preferred to national experts, and that is so for very good reasons, especially in small countries. But experts coming from outside the country may be impaired by the lack of knowledge of the contextual problems and the various national political agendas surrounding the assessment exercise. Haphazard personal knowledge is another weakness of the international experts, a weakness I shall return to below.

There is yet another group, which we will call *Assessment Victims*. The victims of assessments can be individuals - scholars. But it is often also universities, departments or programmes. Any organizational unit within the research system may become a victim of an assessment.

Once again we have an interesting relationship. One would expect that a social relationship would develop between *Executors* and *Victims*. As the *Stockholm Syndrome* about the emotional bonds between the terrorist and his victims have taught us strange things may happen if and when the two groups meet each other.

As far as I have understood, the English system is based on the idea that the experts should be situated far away from the victims. They should not meet in person. The experts thus always have clean hands. At least in physical terms. The victims do not know, how the assessment executors are thinking, because to some extent their criteria are hidden as well.

In Denmark we have seen an illustration of the exact opposite situation, which could be observed during the recent research assessment of political science. The nice international scholars, who had been hired as experts, were travelling first-class around in Denmark, met with the people of the departments and had gastronomically satisfying dinners and other social occasions all the time. Since the Danish assessment bureaucrats also were kind and polite people, or maybe just innocent people, they had decided to send international assess-ors with whom many of the scholars in the departments were familiar. Even if efforts were made to avoid legal incapacity or other formal problems, the combination of a small national profession and an - also - small international circle of experts - led to interesting situations. The assessors at least knew all the heads of the departments very well from ECPR circles. They would also have formed relations further down the hierarchical ladders. If the assessors and the victims did not know each other previous to the assessors' visit they might, during the visit, have a chance to cultivate a friendly relationship, given the way the meetings were planned and actually conducted.

In contrast, a rather strange and somewhat boring social situation developed during the Dutch meetings, where the evaluators were locked up in a provincial hotel in the middle of nowhere and were fed milk for breakfast, lunch and dinner for several days, days, which were spent in solitude, apart from the very formal meetings with the *Victims*. Even if social bonds may have existed between some of the experts and some of the Dutch scholars, the very assessment process tended to minimize the possible dysfunctional effects of such bonds. The assessment team had to listen to 33 program directors and their accompanying deans etc. Each program was given half an hour for its presentation, and the experts got very tired. After a few days they were not able to remember many faces.

Even though this relationship between *Expert* and *Victim* looks different in Denmark, England and Holland, it is very interesting to study it as an example of a social relationship with an unstable power balance. I am, however, not aware of any such studies.

Evaluation equals comparison

To evaluate means to compare. This is a very basic thing to say about evaluations. Each time you evaluate e.g. Trinity College, Dublin, you do it against some kind of explicit or implicit standard. Normally you would in this case do it against the standards of Oxford, Cambridge, Harvard, Paris etc. There will always be comparisons made. If we want to discuss the methodology of evaluations, we should accordingly discuss evaluations in terms of what a good comparison is and what good comparative method is. One should first and foremost try to establish what can be compared and what cannot be compared at all.

A comparison always addresses three sets of questions: (1) WHO are to be compared? (2) With regard to WHAT are they compared? (3) HOW are they compared? I shall address briefly some of the comparative problems identified here.

Who are to be compared?

You can compare many different units in these assessment exercises from individual scholars, over projects, programmes, departments, faculties, universities, and national academic disciplines, to global comparisons. I always thought that the only meaningful comparison was the one performed at the level of the individual scholar. After all, he or she is the one, who thinks, writes, and publishes. Basically, all evaluations that I know of are based on individual data. These data are then aggregated to some level, and then they are being compared. But down below there is the individual scholar with his work. This is also a fact, when the individual scholar participates in projects and programmes. It is my sincere conviction that a good set of criteria for evaluation or comparison always should be one that is fair to the individual scholar. Conversely, any set of measurements that is unfair to the individual scholar should be scrapped immediately. The research production of the individual scholar should be "measured" and taken stock of as precisely as at all possible. The scholar always should be compared on equal terms to comparable scholars. This is, however, as we all know easier said than done.

The next level among my units is the university department. This is the most important organizational unit, because this is where most of us will spend most of our time. The department normally is where we teach and where we do research. When comparing departments you should therefore also do it in a fair way. You should not compare a polytechnic department with an Oxford department. A highly specialized research unit with only modest postgraduate training responsibilities should never be compared with an all-purpose department which take in hundreds of students at many levels and with many different backgrounds. From the very beginning the research conditions in these two units are uneven, and therefore it is unfair to compare without "controlling" for the variations.

In recent times some smart assessment bureaucrats have developed the idea of assessing projects or programs instead of departments. On the surface this change or transformation of units for comparison sounds tempting and a possible way to establish fairness. One should, however, be aware of the fact that projects and programs are often quite artificial units, and that they can be tampered with in the context of an assessment.

Thus, when I participated in the Dutch assessment, I very soon started to wonder about what the definition of a program is. In the final days of the evaluation I realized what it could be, or, at least, how it is practiced by some.

It can be practiced in such a way that the assessment exercise itself "creates" artificial units. Many tricks can be made. Suppose, for the sake of argument, that in your university you have a brilliant star. He has published hundreds of publications over a wide range. Then you may create a programme that in terms of manpower covers this particular area, but where the research output comes from the desk of this scholar mainly. But you may also allocate the star scholar to several weak programs at the same time. In that way his light is spread over several programmes, and in the most bizarre situations his publications may end up being "counted" several times, unless the assessment experts call the bluff. It is of course unfair to compare these artificial programs to a "naive" department or programme. In the Netherlands assessment the evaluators also discovered that some of the departments had ‘forgotten’ some of their employed scholars, and thereby also had decided not to send in for assessment some of the weaker results.

This is an illustration of the problems and unfairness you may create by using artificial units as programmes and projects. As I mentioned earlier, the only natural unit of evaluation is the department.

It is also in principle possible to compare universities, but that could be even more dangerous. A university could eventually be closed down if it comes out badly in the assess-ment report, or some of its research resources could be taken away. Therefore, on this level fairness is also of high importance in the comparison.

We all know, and I am quite sure that deans in particular will be familiar with, the problems involved in comparing academic disciplines. Let me remind you of the recent evaluation of Danish political science. It happened at the same time as the parallel evaluation of Danish economics took place, and as a result of this coincidence quite a few comparisons were made between the two research disciplines and their output. Especially the newspapers would tend to make these comparisons. When the results of the assessment exercise was presented by the Ministry of Research to the Danish public, it happened in a public meeting, in which the two assessments were presented and discussed jointly. But very little considera-tion was given to the fact that these disciplines always have had different publishing traditions, and that their historical trajectories are different as well. Comparisons, which do not take into consideration these and other differences, are bound to result in unfair conclusions.

On what dimensions do we compare?

Comparisons in connection with research assessments normally cover at least six different, albeit inter-linked, dimensions:

* Quantity,
* Productivity,
* Relevance (social/political),
* Relevance (academic),
* Quality, and
* Viability

A few remarks will be made about each of these dimensions.

Quantity and Productivity

It seems very obvious that quantity is among the items that we compare. It looks simple, for quantity is by definition measurable. As we all know real life is a bit more complicated than that. There is not such thing as pure quantity. Quantity is always qualified by some kind of implicit quality measure. Even naive experts will notice the difference between a mimeo-graphed report about a municipal implementation problem and a theoretical article in *American Political Science Review*.
Most important, quantity is normally measured against resources, so what we get are mostly measures of productivity, which on the surface looks fair enough, and which of course is also exactly what the *assessment bureaucrats* and *their* masters want. At the end of the day those who order assessments want to have something, which is measurable and fairly simple, and productivity measures can be understood by anyone

outside *Academia.* The masters could not care any less about quality, academic relevance and that kind of things. They primarily want to know, whether we are productive scholars or not. At least in a country like Denmark a disproportionate interest is given to the so-called "Zero-Researcher", the poor guy who has not published anything recently - or not so recently. He has wasted the tax payers money because he has not published. In contrast, everyone sees the scholar who publishes lots of junk as OK, maybe apart from the closest colleagues and competitors.

For these and other reasons productivity is the central, if not the one and only, indicator in most of the evaluations. Let me about this measure say, that it is as good as its components: if the numerator or the denominator contains rotten data, then the measure of productivity is also bound to be misleading. If you put garbage in, you will get garbage out of your productivity measurements.

Social/Political Relevance and Academic Relevance

The Dutch research assessment procedures, which by the way have been negotiated with the universities, are based on the idea that you are able to evaluate relevance. You are expected to pass judgments, according to which programme x is rated very relevant, while programme y is rated less relevant. In our assessment group in the Netherlands political science evaluation we were very much in doubt of the fairness of such a measure and the comparisons involved. Most of us did in fact doubt the validity as well as the reliability of any such measurement.

Relevance is a relational concept. It is always "Relevance with regard to Something". In actual research assessments it is decidedly a political construction. Social and political relevance is what our Masters want. Research, which is in demand by some major groups of citizens or by some segments of political elites, is by definition relevant. Studies of the welfare state, and how to improve it, are bound to end up at the peak of relevance, irrespective of the quality of the project. There are on the other hand many kinds of social science research, which are not seen as relevant, simply because it is irritating, deconstructive, difficult to see in a means-end-perspective. Hence unwanted. If it is critical to the values of the "ruling ideology" or the "political formula" in the terminology of Mosca, it may even be dangerous. It would be stopped immediately, if some of the political masters had their say. In my opinion the concept of social relevance should never be used for any assessment purpose.

It is important to make a distinction between social/political relevance and academic relevance, and this distinction is also sometimes made in assessment practice. But is academic relevance really much safer and better than the just mentioned social relevance concept? What is academic relevance? In my humble opinion it is just another way of determining, whether the research under assessment is mainstream-research or not. If it is the kind of research, which fairly easy would go into dominant journals and fairly easy would be published in respected book series etc. That is in daily parlance, what we call academic relevance. I do not think that there is any need for such a concept, because the interesting aspects of it could as well be discussed in terms of quality.

Quality

Quality is something that cannot be counted or weighed on a scale. Many number fetishists have tried hard, but at the end of the day one is still reminded about what Aristotle said in his *Nicomachean Ethics*: “It is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject matter admits.” For quality measurement many “surrogate indicators” have been suggested, primarily some which utilize citation countings, visibility measures etc. The validity problems in this area have been analyzed thoroughly, and no “educated man” can recommend use of any of these indicators. For quality assessment one simply has to rely on *peer reviews*.

Assessment experts, who serve as peers, are mostly, but not always, very good and respected scholars. But assessment bureaucrats sometimes ask too much of these poor people, more than they can be expected to deliver, even if they are well trained in their trade. In the assessment of Dutch political science, a group of five experts were asked to evaluate 4000 published items, spread all over the academic disciplines of political science, public administration, and communication science. Only limited time was available, and some of the members of the team were in addition not too familiar with the Dutch language, which had been used to some considerable extent by the Dutch scholars under scrutiny. My point here is that even peer reviews have their obvious limits, when we speak about large-scale assessments on the level above the individual department. But since the political masters behind the assessment bureaucrats demand measurements, they do of course get it even under these impossible conditions. The peers in the large-scale assessment exercise just tend to use some kind of implicit quantitative criteria in order to support their expertise and their common sense. To provide an illustration: We, and I mean all of us in such situations, tend to evaluate an article on the basis of the kind of journal the article appears in. We do so, even if we know that this surrogate measure is ridden with validity problems. If Cambridge has published a book or Oxford University Press we tend to think that it is a better book, than if it has been published by a local publisher in Brabrand or Esbjerg. The countless number of implicit assumptions present in these “short-cut”-operations should be discussed openly.

Viability

In the Dutch assessment protocol the review committees are asked to assess the viability of research programmes. The long-term viability of a programme is assessed “in view of what has been achieved so far, in view of the plans and ideas for the future, and in view of foreseeable developments in personnel and facilities.” Not an easy task for experts, who are flown into the country at short notice. Not even easy for the Dutch experts themselves. The idea behind this measurement of viability is that a team of peers by looking at the publications, some plans, and even some loose ideas, can determine if a program is something that could - and eventually should - go on for a number of years until the next assessment, or if it is something that will die shortly - or should be terminated immediately. This kind of prognostic activity, mixed with possible, more or less explicit, recommendations, is dubious and dangerous as well. In my opinion reviewers should never engage in such activities, simply because they are not equipped, and even less legitimized, do so. The peer reviewers are not familiar with the background and the contextual position, not to speak about the local or national political agenda, of the program.

At one point during the evaluation of Dutch political science in 1996-96 I had to give a numerical score to a program in these terms of academic viability. When the Dutch chairman of the committee asked me to do it, I at first rejected the request. Hard pressed I ended up giving in, cowardly. I told the chairman of the committee that I would give the program in question a top grade on viability, a “fiver”. Why? Not because of the inherent quality of the program, even if it was very good, but because I happened to know, that the leader of the program - and department - was holder of the chair as professor at Leiden University. I also knew, that this chair had been established several hundred years before. I was therefore quite sure, that even the most malevolent Dutch minister under the direst conditions would probably not close down activities dear to this chair. And he would certainly not dare to close down the - arguably - oldest chair in political science in Europe. Therefore I was prepared - under pressure - to give the program the highest possible score on this invalid measure. This incidence illustrates my criticism of the concept of viability.

Conclusion

My message in this introduction can be summarized as follows:
We are not able to produce a satisfactory system of measurements, indices, criteria etc, which will be fair to all disciplines and to all units at all times. A few reasonably simple indicators of productivity can be identified, but even with these, we should be aware of validity problems. Interpretations of such measurement data ought always to respect the differences between disciplines and between national research systems. At best indicator data send out signals that may lead to further activities by those who are responsible for the research units. Given this data situation research assessments will have to rely mainly on peer reviews, which actually are bound to be impressionistic and highly debatable, especially when the research assessment takes place at a level higher than the individual department.

After Mogens Pedersen’s presentation the members of the audience made comments and raised questions:

How about victims, who pay the taxes which support research?
Why do researchers always regard themselves as victims when they are being evaluated, when the evaluations exist in order to support good researchers against bad researchers?

Of course we should all of us feel responsible for the use of government resources and other resources. We ought to do a lot of self-control and self-evaluation inside the university. Such evaluations actually also take place in most good universities. I am aware of several universities, which as part of their ordinary activities maintain databases on inputs and outputs in order to be able to provide e.g. simple productivity measures to the outside world. I have even seen a few faculties, including my own, where internal research assessments of a kind are conducted continually or periodically. Such internal evaluations should, however, be done with some caution and finesse. They can be very unpleasant for the individual scholar or for a department in internal crisis. Effects of some kind are highly visible, and sometimes even dramatic. In contrast the other assessments - the big institutionalized ones - often are without much effect. What were in fact the effects of the recent Danish assessment of political science and economics? Could it pass a simple test of productivity, quality, and relevance? What is the viability of the idea?

I believe very much in self-evaluation. However, to some considerable extent the internal assessments will have to solve the same measurement problems as the other kinds of assessment.

Do you make a combined measure of the input of resources and the output of research? Can such measurements become an operational tool?

Yes, it is possible to devise many measures of productivity, but, I have to repeat, they are all of them liable to criticism.

There are a number of input indicators that with some caution can be measured within the university and across disciplines and universities. Let me just mention the following:

* Annual government appropriations;
* External grants from research councils and from contract research;
* Actual “consumption” of resources (in money terms);
* Actual “consumption” in terms of manpower to research as well as technical assistance;
* Etc.

In the same way we may also point at a number of output indicators. Caution is here even more important. I mention here just the following, which I have seen used:

* Publications (total number; according to language; according to types of media; according to degree of pre-publication quality control);
* Citations and Awards;
* Evaluation activities (editorships etc.);
* Number of Ph.D.-students or Ph.D.-degrees;
* Licenses, patents etc.;
* Incoming/outgoing visitors;
* Etc.

Any output indicator can be used as numerator and any input indicator can be used as denominator. The result is a tempting indicator of productivity.

Instead of repeating what I have already said about the validity problems, I will instead mention a major problem, which I see in connection with all this fashionable measurement business. This is a problem, which I have not yet addressed at all, but which has a lot to do with my introductory remarks about the individual scholar as the true source of all research output.

When you move “down” to the individual level of measurement, the level of the scholar, you cannot make all these distinctions, which the measurement bureaucrats want us and even ask us to make. As a scholar, I cannot divide myself up into 40% researcher, 50% teacher and 10% administrator. My activities throughout the year, yes, even during my normal working day, are a mixture of everything. At the level of the individual scholar the ongoing production is an example of what economists call *combined production*. Stated more precisely, it is impossible with a sufficient degree of precision to break down resource consumption on “research”, “teaching” and “other purposes” at the lower (individual) levels of the university. Any attempt to do so will in the short run, but especially so in the longer run, violate basic principles of the autonomous university and free research.

When measures of productivity are created, based on input and output measures, which themselves violate these principles, the results of measurements can be very unfair to individual scholars. Economists will tell us that behavioral mechanisms related to “moral hazards” will influence future behavior among scholars. Research assessments can never be neutral. Their effects can be divided into the intended and the not-intended effects, and into functional and dysfunctional effects. These problems need to be discussed more and they need to be communicated to the *Assessment Bureaucrats* and their political *Masters*.

Comment from the audience:
I have been evaluated 4 times, twice in the US and twice in Denmark, but I have also myself been evaluating. Therefore I recognize the roles that you have defined. Two of the roles I feel very uneasy about and one I feel very good about. I feel very good about being evaluated because the process of preparing to be evaluated is a very constructive one. I feel very uneasy about the role of the assessment bureaucrats and the assessment executors. When I have been abroad evaluating I feel very bad about the role which is pressed upon one, because you know within your heart that you are doing something that is not honest. I agree with Mogens Pedersen that the universities should try to develop their own type of benchmarking on their own sort of criteria. I find that a department is the closest you get to a natural unit, but it requires the departments to develop mechanisms and to develop themselves, instead of waiting for politicians to come and stop them, when something has gone wrong, or when something has gone too far. This is really a problem in Danish research.

I tend to agree. Let me just add that even if the department is a natural unit, and most of us will agree on that, we know that departments will not reform themselves without some help or “help” from the outside. I have at least never myself seen one single example of a department, which entirely took it into its own hands. Some input from outside was always instrumental.

CAN ASSESSMENT REPORTS BE USED AT FACULTY LEVEL ?

Dean Niels Christian Sidenius
University of Aarhus, Denmark

At the Faculty of Social Sciences we have had three research evaluations during the last three years: Economics, Political Science and Psychology.

One will not go into them in detail; allow me only to say that the reports were very favourable, using different words, though. One will rather put forward some pragmatic reflections on research evaluations as an instrument of quality assessment and improvement.

Research evaluations are necessary - in one form or another - in conjunction with evaluation of education, which cannot stand alone. But what use can be made of external research assessments as an isolated phenomenon?

To answer that question one have to specify - in very general terms - the focus of a Faculty Board and a Dean of a Social Sciences Faculty:

Set up rules, procedures and frameworks productive for the pursuit of high standards of quality within research and education Allocate resources to the Departments Ensure proper relations between research and education

This means that, generally, the choice of research topics are left to the Departments, groups of researchers or individual researchers. And research evaluations do most directly aim at this level, the Department.

When the Faculty receives a research assessment it is put on the agenda of a Board meeting where it is debated. The Head of the Department is asked to give an evaluation of the assessment stating the points on which the Department agrees or disagrees with the report and its recommendations. As one said, our research evaluations have been very favourable and the inducement to further action has been limited.

This would have been somewhat different, though, if the conclusions of the reports had been negative. In that case, the Faculty Board would ask the Department to reflect upon its possibilities of a required improvement of standards and return later to the Board with a plan for implementation.

But what else would the Faculty be able to do in case of a negative research assessment? Actually not very much.

It might, of course, change the allocation of resources. But to be a serious incentive the taking away of resources would have to be of a certain magnitude. And in that case you would have to decide on the future existence of the Department, including its educational obligations. 1 do not see that route of action as a very feasible one. - But it might, of course, be quite different with centres with only research obligations.

The second line of a more dramatic action would be to fire those researchers who did not live up to the standard of being an academic at a university and in this way did not produce sufficiently for the benefit of the Department. Such cases are usually very complex, and one is quite certain that external research evaluations would prove to be a very poor basis for such an exercise. Fortunately, most research evaluations are not geared for that purpose.

To sum up, I would say that external research assessments are only required at those Departments that by themselves are more or less unable to define international standards of quality and to improve inadequate quality - whereas external research assessments do not make much sense at Departments with a high quality performance.

The necessary agenda for these last-mentioned Departments is on a permanent basis to reflect upon questions such as: organisational changes, infra-structural improvements, ensuring of a productive academic environment, mobilisation and recruiting of top talented students, etc. - But these questions are not very convincingly analysed and reflected upon in external research assessment reports.

CAN ASSESSMENT REPORTS BE USED BY UNIVERSITY ADMINISTRATORS?

Head of Department Peter Nannestad
University of Aarhus, Denmark

A. The functions of research assessments and of the Head of Department

1. What use are research assessments intended for?
   - The background of the current assessment-boom in NPM and principal-agent theories
   - Research assessments as monitoring devices for principals

2. Is the Head of Department a principal or an agent?

B. When the Head of Department acts as a principal

1. Will research assessments tell something about the department's research, the Head of Department does not know?
   - The focus of research assessments: the institutional level
   - The source of information for research assessments

2. Why are research institutions so hard to steer and change - lack of monitoring information or disincentives?
   - The distributional role of institutions and the cost of changes
   - Dispersed benefits and concentrated costs

3. Can research assessments change the cost-benefit calculus?
   - Changing the incentive structure of key actors
   - Strategies of blame avoidance etc.

4. Prediction
   - In most cases, research assessments will not be instrumental in bringing about major re-orientations or changes at the departmental level

C. When the Head of Department acts as agent

1. Information asymmetry between principal and agent
   - Is research assessment the solution?
   - Why research assessments may be advantageous to the agent

2. Can research assessments be used in negotiations with the principal?

D. Conclusion

1. Research assessments, as we have seen them, are of limited use to a Head of Department

2. From the point of view of a Head of Department, the process may be more valuable than the result

CAN ASSESSMENT REPORTS BE USED BY UNIVERSITY ADMINISTRATORS?

Head of section Kari Lindbekk
Academic and Student Affairs Department, University of Oslo, Norway

Can assessment reports be used by University Administrators?
My answer to this question is yes. Assessment reports can be and will be used by University Administrators. But will they also be useful ?

Personally I have been an University Administrator for only six weeks. I came to my present position as Head of Section at the Section for Contracts and Research Administration after 18 years as an Administrator in Research Councils. First 13 years at the Norwegian Fisheries Research Council and then five years at The Research Council of Norway. My personal experience with evaluations dates from that period. The evaluation challenges that I shall have to face in the time to come is already at my desk where I have found that I must get started four evaluations in the weeks to come.

One indispensable condition for the usefulness of assessment reports is, in my experience, that somebody intends to use them, that there is an identified purpose. This leads me to the first point that I want to make, namely the importance of competent and committed users. I have been fortunate enough to work with highly qualified users, first in the Norwegian fishing industry and later as co-ordinator of the Evaluation of Research in Lithuania by The Research Council of Norway. In both cases I found that qualified users who wanted to make use of the resulting recommendations, transformed our work from what could have been boring routine to meaningful efforts.

Another point that one want to make is the importance of competent administrators. Again I want to go back to the evaluation of Lithuanian research where information had been compiled from all the Academies and Institutes, where additional information was added when asked for, and where the Norwegian evaluation Committees were met on time and by the right people at every one of the 97 site visits. In this case the admirable preparations and following-up by my Lithuanian colleagues was a necessary condition for the carrying out of the actual evaluation.

My most important point is the indispensability of competent evaluators. We may evaluate research organisation and research institutions, but the main issue is scientific quality and relevance. The legitimacy of every evaluation depends on sound and qualified judgement from respected peers.

The purpose of most evaluations is to legitimise change or modifications. When I look at the evaluations that I as an administrator must get started at the University of Oslo, one see that they will almost certainly also provoke discussion. I hope that we will be able to carry out these evaluations in such a ’ way that the discussions will be constructive. This leads me to the last point that I want make, namely the importance of empathy.
Evaluation of research is an important matter both for society and for the interested parties and should be carried out with due respect for the task at hand and with readiness and ability to enter into discussion with the parties involved.

CAN ASSESSMENT REPORTS BE USED BY POLITICIANS?

Christine Antorini
Member of the Danish Parliament 'Folketinget'

This spring I became a new member of the Parliament, and we have not yet discussed evaluation neither in the research committee nor in a public discussion. Therefore, it has been very interesting for me to be here and listen to how much there is going on in this field in Denmark and in other countries.

As a member of the research committee I was puzzled by the drawing Mogens Pedersen presented us with during his presentation. He seemed to have left out the politicians. It made my wonder if the politicians in this context function as the link between the minister and all the autonomous universities because the universities know that they can create some pressure on us and indirectly create some pressure on the minister.

I will make a short political and personal comment on the discussion on assessment. I find that evaluation is very important, also on the political level, because that is one of the only ways we can gain information of what is going on within the system which we are making rules for. I also find it very important that we have a high quality of education and research, both in Denmark and abroad. Even though you all have mentioned the problems of measuring and defining good quality no one of you draws the conclusion that we should stop assessing, because we all agree that assessment is necessary. As a politician I find assessment necessary, but I am not qualified to evaluate and to define good quality. It is your job to define the criteria of assessment and present the politicians with models of your conclusions.

All through Mogens Pedersen's presentation he mentioned the universities and the researchers as the victims of the systems, which is an attitude I would categorise as typical Danish. Danes are very anti-authoritarian, and we are not very comfortable with the fact that other people and institutions are evaluating us. We would prefer self-evaluation. Therefore, we have to come up with a combination of self-evaluation and evaluations from external institutions.

In conclusion I will mention three things that I find important:

1. On the political level we are lacking clear model of how to improve the quality of research.
   There should be clear rules for evaluation, and it is extremely important that there are no hidden agendas.

2. You have to define the criteria. I fully agree with those of you who have mentioned that the criteria cannot only be defined by numbers.

3. We need a clear conclusion of how to change the system and we need to act on the recommendations on how to make the system better. After the evaluation is done we need to follow up on the evaluated institution, and we need to make sure that the money we spend on evaluation actually pays off in form of better quality.

CAN ASSESSMENT REPORTS BE USED BY POLITICIANS?

Nick Constantopoulos
GSRT Evaluation Section, Greece

The problems and the arguments which have come up in this discussion have overwhelmed me. I do not find it necessary to contribute with a lot more, but I will tell you a few snippets of bits and pieces of activities going on in Greece. I am comforted by the idea that Stefan Kuhlmann explained that university research has been exclusively assessed in Germany, only 10-15% on a peer review basis. Since we do not want to overdo the Germans, we have done even less than that.

The university assessment level in Greece has a few characteristics, and what I have collected is a few opinions by academics themselves. First of all the assessment in Greece has a non-obligatory character. It is carried out when it is carried out, and it has a high degree of freedom both in designing the criteria and setting up the targets of why we are doing it. It is of course an ongoing process within the university departments and within universities themselves. From the administrators' point of view there is a disparity in forms and aims of the assessment, they are not always the same, taking into considera-tions the fact that we have polytechnic schools, technical universities, general universities and agricultural universities. Some of the experiences expressed in the past by academics in Greece are for example that we should carry out the assessment under the supervision of the Prime minister. Nevertheless, other professors have thought that the interaction between university research and industry, which comes very close when it comes to the technical universities, has not been productive for a numbers of reasons. They do feel that as far as the technology production is concerned within technical universities there is a need for a quantum leap in finding the solution of the problems in the industry. As long as you do not know what the industry wants and what kind of interface you can establish with them, it is very difficult to assess yourself and to be assessed.

The national research committee is concerned with the relationship between internal researchers and external researchers to carry out contracted research. They also measure the relationship of the research carried out by the funds of the university itself and the research contracted through European Union funds. Overall they believe that the evaluation of the university researchers should be done according to criteria established by the rectors.

Universities have been active in assessing themselves like the agricultural university in Athens. Nevertheless, in 1992 there was an activity that has not been mentioned today. The Council of Ministers of Education of the European Union addressed a memo to universities and rectors' associations, and they informed them about a few things. In the memo the Council mentioned some basic principals in the assessment process. One of the few values that were expressed was a demand for a dual level of assessment. National authorities should continue to do their own evaluation and assessment exercises, in addition the Council should feel free to do the assessment of the national systems.

Another aspect, which has importance for Greece, is that the measurement of performance could be done on the distribution of public investment funds, as it has been explained several times today. The public investment budget is the most important for Greece as far as non-operational budgets are concerned. This is the budget given to all construction works, research programs and special projects. In 1992 the law on education allowed for the creation of a special evaluation unit for university level assessment. It has not been materialised yet, but it is on its way and we are stepping in the right direction in preparing something substantive and holistic.

The access of the assessment depends on different things. When the government request the assessments it is evident that they see it in a global way, and they have to take all performance indicators into consideration. There are of course assessments carried out by the Universities, by University Departments and Research Centres as a routine procedure. They have a different type of acceptance and rate of utilisation. I suppose that there are assessments that nobody asks for, but they still have some value. Another way of looking at the performance of the assessment is the level of the content quality. The research and technology development and demonstration assessments have their own value as long as they can be translated into guidelines.

Finally, there is the question of the participation of the user of the assessment experience into the planning and carrying-out process. Sometimes the user is the authority that delegates it, which is very clear-cut, some other times it could be a third party. Trade unions could be the user when it concerns employment of specialised staff, and it could also be other social groups. The more we involve the user in the planning of the assessment experiences and exercise the higher level of quality in the assessment.

CAN ASSESSMENT REPORTS BE USED BY POLITICIANS?

Director Ove Poulsen
The Danish Ministry of Research and Information Technology

Since the discussion we have had today already has been very good and thorough, I will only make some very brief remarks.

The question about roles is a very important one. Who is supposed to do what in the society? What is the role of a central administration and policy makers as myself? It seems very clear to me that it is not our job to run institutions. We represent the owners – e.g. the society. This means that we should try to develop tools to maintain our ownership in a clear and definite way, but we should never redefine the classical play that Universities do have to develop themselves, and to develop mechanisms and new incentives so they can change themselves. When it goes wrong our job is to give advise to the political level. In this warning system we do not need many institutional evaluations.

We have used the evaluation tool very reluctantly because we do not find it to be a very good tool in understanding the dynamical behavior of an institutional system. Hanne Foss Hansen pointed out an extremely important issue which is the availability of high quality indicators; indicators which not so much look at the performance of individual institutions as trying to define our knowledge system. The universities and the research institutions represent a system which produce knowledge, but knowledge in itself is not of any interest if it is not moved. This has to do with the question of how we transmit knowledge from one person to another, and from one institution to another. How do we move knowledge internally? How do we move knowledge externally? We ought to be concerned about those questions, because it is in that process that we generate added values of our knowledge system. It is in that way we can formulate new innovation policies.

Society invest a lot of money in research, not to keep the individual scientist happy but because it is good for the society. Research produces students and knowledge, and this knowledge is to be used in society. Those countries that have a well-defined knowledge system with innovation potential attached to it are the wealthy countries. That is why science and technology indicators in a broad sense are much more important to policy makers like myself compared to strict institutional evaluation. Institutional evaluation is of interest to institutions. Therefore, institutional evaluation could never be our job. The political level always tries to get close and tries to control, but basically, when working inside the administration you realize that the political system only have the obligation to step in when things go wrong, or to create a framework on a general level for better institutional development. It is a clear message to the institutions that they have to make the thinking themselves. This is the general philosophy of the way we try to develop this thinking in the coming years.